Chapter 2
Exploring and Discovering Data

In this chapter, we discuss different approaches for exploring data. Data exploration is probably the single most important step in any data analysis. While the availability of huge amounts of data often tempts the user to jump directly into sophisticated models and methods, one of the main messages of this book is that it is of extreme importance to first understand one's data and to thoroughly explore it for patterns and anomalies.

So, why do we perform data exploration? The answer is very simple: to better understand our data and get intimately familiar with it. We simply should not base business decisions on complex methods and models unless we are certain that these methods capture the essence of our data. For instance, much of this book will talk about *linear* models. But what if the reality is not quite linear? What if our business processes are subject to "diminishing returns"? How could we detect such "non-linearities"? We could have a "hunch" that our process requires a somewhat different model, but sometimes (especially when dealing with new business processes) we simply don't know. But, as it turns out, our data typically knows much more about our processes than we do, and data exploration will tease out all of its knowledge. Moreover, data exploration is useful not only to detect trends and patterns but is equally important for uncovering anomalies and outliers. Not every single one of our customers behaves in the same way. In fact, there are typically a few customers who behave in ways very different from the bulk of our customers. It is important that we can identify such customers and deal with them in the appropriate way. Data exploration will help us pick out such atypical customers and their behavior.

In this chapter, we present an array of data exploration methods and tools. In Section 2.1, we start with *basic* data summaries and visualizations. We use the word "basic" since they should be part of everyone's toolset and should be consulted every single time we explore new data. These basic tools include summary statistics (such as the mean, median, or mode), frequency tables, and histograms and boxplots for exploring the distribution of variables, as well as scatterplots, correlation tables, and cross-tabulations for exploring pairwise relationships among variables.

W. Jank, *Business Analytics for Managers*, Use R!, DOI 10.1007/978-1-4614-0406-4_2,

Many of these basic tools discussed in Section 2.1 can be found in spreadsheets (such as Excel) and are not necessarily a special or distinguishing feature of specialized data mining software. In subsequent sections, though, we also discuss "more advanced" (or more powerful) tools for data exploration. Many of these advanced tools cannot be found in spreadsheets, and they illustrate the power of more advanced data mining solutions. To that end, we will discuss scatterplot matrices and trellis graphs (Section 2.2), time series graphs (Section 2.3), spatial graphs (Section 2.4), density and spine plots for categorical responses (Section 2.5), or a combination of several different types of graphs and data-aggregation techniques for panel data (Section 2.6).

We also want to emphasize that, in contrast to many standard textbooks on statistics, we do not explicitly separate *numerical* data summaries (such as the mean or the standard deviation) from *graphical* displays (e.g., a histogram) since we believe that both numerical and visual data exploration should be used simultaneously, as one informs the other and their joint and simultaneous application leads to better insight about patterns and anomalies in the data.

2.1 Basic Data Summaries and Visualizations: House Price Data

We start out by discussing some of the most basic tools for exploring data. We use the word "basic" because these approaches constitute the minimum toolset that each analyst should possess. They also can often be found in spreadsheets and are therefore in widespread use. Either way, mastering these tools is an absolute must!

Data: Table 2.1 shows a sample of house prices (and associated house characteristics) for a major US metropolitan area. In particular, it shows a house's ID, its selling price (in US$), its size (in square feet), the number of bedrooms and bathrooms, the

Table 2.1 The house price data. See also file `HousePrices.csv`.

ID	Price	SqFt	#Beds	#Baths	Offers	Brick	Nbhd
1	114300	1790	2	2	2	No	East
2	114200	2030	4	2	3	No	East
3	114800	1740	3	2	1	No	East
4	94700	1980	3	2	3	No	East
5	119800	2130	3	3	3	No	East
6	114600	1780	3	2	2	No	North
7	151600	1830	3	3	3	Yes	West
8	150700	2160	4	2	2	No	West
9	119200	2110	4	2	3	No	East

number of offers it has received while being on the market, whether or not it has brick walls,[1] and the neighborhood where it is located.[2]

Goal: One of the main goals of this analysis is to determine what drives the price of a house. For instance, it is reasonable to assume that larger houses (i.e., those with larger square footage) will fetch a higher price. But *how much more* does the price increase for each additional square foot? Also, does the siding material (i.e., brick vs nonbrick) have a significant impact on price? Or, does it matter in which neighborhood the house is located? Answers to these questions could help a potential buyer decide how much to bid for a house. It could also help the seller (or his realtor) price the house properly.

We accomplish this goal in several steps. First, we investigate the *distribution* of individual variables. For instance, we investigate *summary statistics* such as the average (or median) price to obtain a general sense of a typical home's value. We also compute the standard deviation of price to understand how much house prices are fluctuating around that typical value; high fluctuations could be indicative of a market in which it is hard to compare the value of one home with that of another home (which may be an advantage for the seller). We compute the *histogram* of price in order to gauge the shape of the price distribution, which could help us determine whether there exist unusual homes (with unusually high or low values). After investigating the distribution of all variables *individually*, we look at *pairwise relationships*. Pairwise relationships let us understand whether, for example, the price of a house increases with its square footage, or whether an additional bedroom has a stronger impact on price than an additional bathroom. Pairwise relationships are explored using *correlation measures* or *scatterplots*. We advocate the use of both correlations and scatterplots simultaneously since each conveys different pieces of the (big) picture: while scatterplots allow us to determine whether there exists any (practically relevant) relationship and the *form* of that relationship, correlation measures allow us to quantify (and hence compare) the strength of this correlation. We start out by discussing summary statistics for the house price data in more detail.

Summary Statistics: Table 2.2 shows summary statistics for the house price data. In particular, we compute the minimum (Min) and the maximum (Max), the first and third quartiles (1st Qu and 3rd Qu), the median and the mean (or average), and the standard deviation (StDev).

Looking at the first column of Table 2.2, we can learn that the average (or mean) house price equals $130,427. We can also see that house prices are slightly *skewed* since the mean price is a bit larger than its median value ($125,950). The most and least expensive houses sold for $211,200 (Max) and $69,100 (Min), respectively. The first quartile ($111,325) implies that 25% of all homes have sold for *less* than $111,325; similarly, the third quartile implies that 25% of homes have sold for *more* than $148,250, so there is considerable variability in house prices. In fact,

[1] Many homes in the United States have vinyl or other types of siding.

[2] Neighborhoods in this data are characterized as East, North, or West.

Table 2.2 Summary statistics for the house price data.

	Price	SqFt	#Beds	#Baths	Offers
Min	69100	1450	2.00	2.00	1.00
1st Qu	111325	1880	3.00	2.00	2.00
Median	125950	2000	3.00	2.00	3.00
Mean	130427	2001	3.02	2.45	2.58
StDev	26869	211	0.73	0.51	1.07
3rd Qu	148250	2140	3.00	3.00	3.00
Max	211200	2590	5.00	4.00	6.00

Table 2.3 Frequency table for *Brick* and *Neighborhood*.

Variable	Categories		
Brick	No	Yes	
	86	42	
Neighborhood	East	North	West
	45	44	39

the standard deviation ($26,869) measures the precise amount of this variability. One way to interpret the standard deviation is as follows: if house prices were perfectly symmetrically distributed around their mean, then a standard deviation of $26,869 implies that 95% of all house prices fall within $\$130{,}427 \pm 2 \times \$26{,}869$, (i.e., between $76,689 and $184,165), a considerable range. The general formula for this relationship is $Mean \pm 2 \times StDev$. Of course, before applying this formula, we have to check first whether the distribution is in fact symmetric around the mean. We can do this using, for example, a histogram of price (see below).

We can also learn from Table 2.2 that the typical house has three bedrooms and between two and three bathrooms. (Notice that while the median number of baths equals 2, its mean is 2.45, which suggests that there are a few "outliers" with a surprisingly large number of bathrooms; in fact, the largest number of bathrooms (Max) in our data equals 4.) The typical house also has a size of 2,000 square feet, and it appears that the variability in home size (standard deviation = 211 SqFt) is not very high. And finally, we learn that most homes get between two and three offers; however, there also exist some rather unusual homes that have received as many as six offers.

Frequency Tables: Note that while there are a total of seven different data columns available ("compare" Table 2.1), Table 2.2 shows summary statistics for only five of them. The reason lies in the differences in data types: while the first five columns are all *numerical* (i.e., measured on an interval scale), the last two columns are *categorical* (e.g., "Brick" assumes the values "Yes" or "No" but no numbers). We cannot compute summary statistics (such as the mean or the standard deviation) for nonnumeric data. Instead, we explore categorical data using *frequency tables* that compare the frequencies between individual categories. For instance, Table 2.3 shows that most houses (i.e., over 67%) are built from nonbrick material.

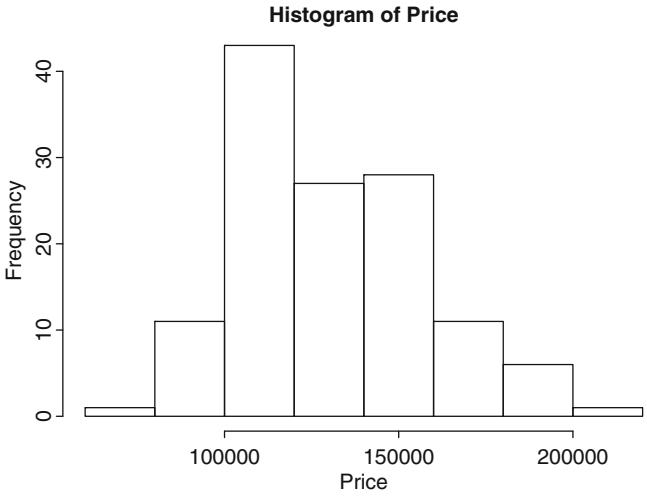

Fig. 2.1 Histogram of price.

Histograms: While summary statistics are a great way to summarize important aspects about a distribution in a single number, they are limited because they only capture a single aspect of that distribution. Most of the time, data are too complex to be summarized by a single number only. For instance, while the typical house in our data sells for $130,427, the price distribution could be skewed (i.e., there could be some houses that sell for much more), it could be multimodal (i.e., there could be not just one "typical" house but two or even three typical houses), there could be outliers (i.e., some houses that sell for an exorbitantly larger amount), or there could be other anomalies that cannot be detected using only a single number. To that end, we want to visualize the entire data distribution. This can be done using a histogram.

Figure 2.1 shows a histogram of price. We can see that the distribution appears rather symmetric around its mean, although there appears to be an unusual "bump" between $100,000 and $120,000. This suggests that while the "typical" house sells for $130,427, there is a rather large proportion that sell for significantly less.

Figure 2.2 shows histograms for the other numerical variables from Table 2.1. We can see that while the distribution of a home's size (i.e., square footage) is very symmetric, the distributions of the remaining three variables are skewed. For instance, while the average number of bathrooms is 2.45, there are some (but few) houses with as many as four bathrooms. Similarly, while a house typically receives 2.58 offers, some receive as many as six offers. We also want to point out that in the context of *discrete variables*, the average may not always be a meaningful way of summarizing data. For instance, note that the variable "number of bathrooms" assumes only discrete values (i.e., a house can have either 2 or 3 bathrooms but not 2.5). Thus, concluding that "the average number of bathrooms equals 2.58" does not make much sense. We can interpret this as the average house having between

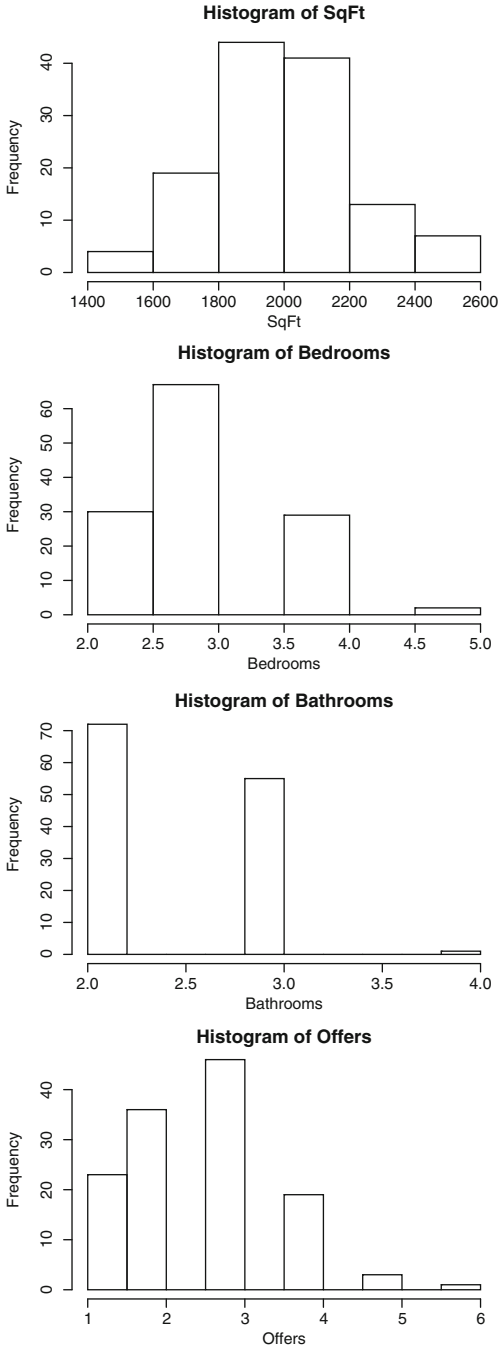

Fig. 2.2 Histogram of other numerical variables.

Fig. 2.3 Boxplot of price.

two and three bathrooms. Alternatively, the *median* tells us that a typical house has two bathrooms. The median is not affected by the difference between discrete and continuous data, and thus provides meaningful answers in both cases.

Boxplots: An alternative way of visualizing the entire distribution of a single variable is via *boxplots*. A boxplot graphs the *quartiles* of a distribution. That is, it draws a "box" between the first and third quartiles and marks the median by a vertical line inside that box. Furthermore, it draws "whiskers" between the outside of the boxes and 1.5 times the *interquartile range*; the interquartile range is the distance between the first and third quartiles and can hence be used as a measure of variability in the data. Data points beyond the whiskers are considered *outliers* and are marked by circles.

Figure 2.3 shows the boxplot for price. It conveys information similar to the histogram in Figure 2.1. However, we can now see more clearly that the price distribution is slightly right-skewed. (Note the longer whisker to the right side of the box, and the larger area inside the box to the right of the median.) A right-skewed price distribution suggests that some sellers manage to fetch a significantly higher price for their home than the rest; from a seller's point of view, it would be important to understand what these successful sellers do in order to get such a price premium. We can also identify one potential outlier on the boxplot; this outlier marks a house with a price that is above and beyond the rest. In that sense, the boxplot conveys information similar to the histogram, but it presents this information in a more detailed fashion.

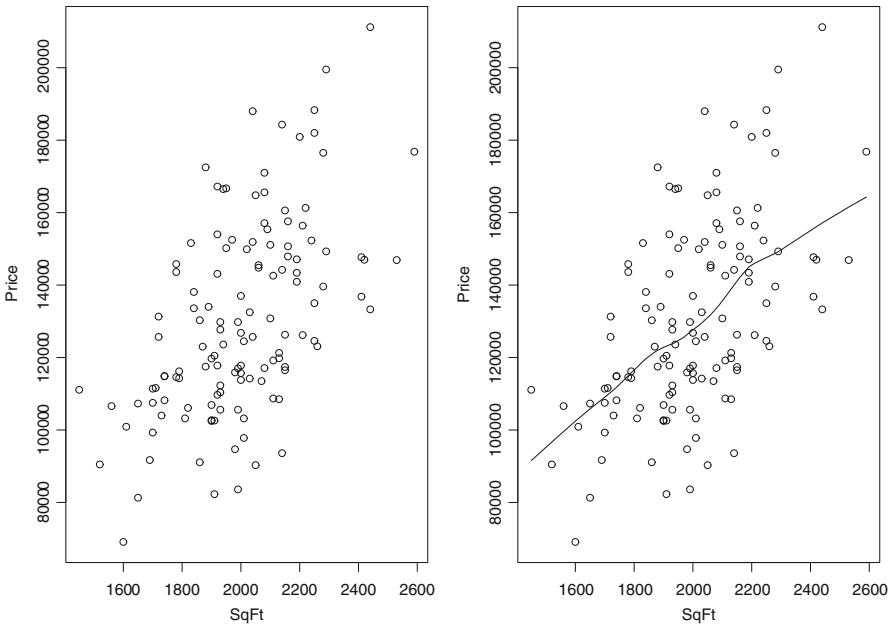

Fig. 2.4 Scatterplot between price and square footage. The left panel shows a traditional scatterplot displaying only the individual data points; the right panel shows an enhanced version with an overlaid smoothing line.

Scatterplots: After exploring each variable individually (using histograms, boxplots, and numerical summaries), we now want to investigate *pairwise relationships* between variables. The most common (and also most powerful) way of exploring pairwise relationships is via scatterplots. Scatterplots graph pairs of two variables' values on an *X*- and *Y*-coordinate system.

Figure 2.4 (left panel) shows a scatterplot between price and square footage. We can see that, unsurprisingly, there exists a positive relationship between the two (the larger the square footage, the larger the price). We can also see that this relationship appears almost *linear*; that is, it appears as if for every increase in square footage the price increases by the same (constant) amount. This observation will become important later on when our goal will be to *model* the relationship between price and square footage. The (almost) linear relationship between price and square footage becomes even more apparent in the right panel of Figure 2.4, which shows the same scatter of data points but with a smooth trend line overlaid.

While a scatterplot can be used to identify general trends, we can also use it to scrutinize individual data points. For instance, Figure 2.4 shows that while most houses have the same positive relationship between square footage and price, there are a few houses (at the top right corner of the graph) that appear to "fall off" that trend. Deviations from a general trend might be indicative of segments, pockets or geolocations that behave differently from the rest. Such segments or pockets are

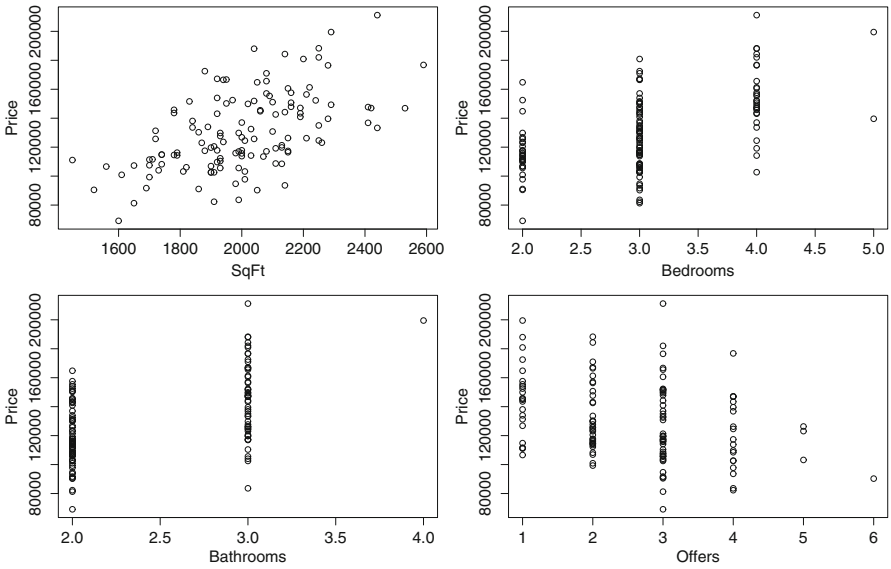

Fig. 2.5 Scatterplot between price and all four numerical variables.

important to identify, as we usually have to treat them with different marketing strategies (e.g., target different buyers, use different advertising or promotion strategies, etc.).

Figure 2.5 shows scatterplots between price and all four numerical variables. We can make several observations. First, while there exists a positive relationship between price and the number of bathrooms (and similarly for the number of bedrooms), the relationship between price and the number of offers appears negative. This last observation is curious since one may expect more offers to result in a higher level of competition, which, as one could argue, should result in a higher price. We also see that the scatterplots pertaining to the number of bathrooms and bedrooms are of rather limited use. In fact, since both variables assume only three and four different values, respectively, the information gleaned from the scatterplots is restricted. This illustrates that the use of scatterplots in connection with discrete variables should be done rather carefully.

Correlation Measures: While scatterplots provide a *graphical* way of investigating the relationship between pairs of variables, we can augment this graphical approach with a numerical assessment using pairwise *correlations*. In fact, while scatterplots are a great way of "seeing" relationships, the eye can sometimes betray us. Moreover, two people looking at the same graph may see two different patterns. It is thus often desirable to augment the visual (sometimes subjective) impressions gleaned from a scatterplot with objective numerical measures. Correlations provide such an objective measure.

Table 2.4 Correlations between all numerical variable.

	Price	SqFt	#Beds	#Baths	Offers
Price	1.00	0.55	0.53	0.52	−0.31
SqFt	0.55	1.00	0.48	0.52	0.34
#Beds	0.53	0.48	1.00	0.41	0.11
#Baths	0.52	0.52	0.41	1.00	0.14
Offers	−0.31	0.34	0.11	0.14	1.00

A correlation (also referred to as *Pearson's correlation*) measures the strength and direction of the linear relationship between two variables. A large positive value implies a strong positive relationship. It is important to remember that correlations only capture *linear* relationships between two variables; that is, for two variables that have a nonlinear relationship (e.g., curvilinear, circular, etc.), the correlation may lead to wrong conclusions.

Table 2.4 shows the table of correlations between all five numerical variables for the house price data. We point out again that since both "Brick" and "Neighborhood" are categorical, we cannot compute their correlation with price (at least not directly). We can learn from Table 2.4 that price has the strongest positive association with square footage (0.55) and that its correlation with the number of bedrooms and bathrooms – while still positive – is weaker (0.53 and 0.52, respectively). This illustrates one of the advantages of correlation measures over scatterplots: while scatterplots also allow us to conclude that price has a positive relationship with all three variables, we could not readily see which variable had the *strongest* association with price. We again observe the negative relationship between the number of offers and price, but we can now also see that this relationship is not very strong (correlation = 0.33), so while the negative relationship is rather surprising, it may actually not matter (at least not for all practical purposes).

Table 2.4 shows additional important information. For instance, we can see that there is a rather strong correlation between square footage and the number of bedrooms and bathrooms, respectively. This is not too surprising since one needs a larger home in order to fit a larger number of rooms. However, this finding also suggests that some of the information contained in square footage is already captured by the number of bedrooms and bathrooms. This observation will become important later on (we will refer to it as "multicollinearity") when we try to find good models for price.

Cross-tabulations: We have pointed out repeatedly that computing numerical summaries or correlations for categorical data is not possible (at least not directly). One alternative to computing the correlation between two categorical variables is to inspect their *cross-tabulation*. Table 2.5 shows the cross-tabulation between brick and neighborhood. It appears that there is some relationship between the two variables, as the percentage of brick homes in the North is significantly smaller compared with the East (or the West). In fact, there exist alternative correlation measures for categorical data. These measures are referred to as *Kendall's Tau* or *Spearman's Rho*. For our data, Kendall's correlation between brick and neighborhood equals −0.03 (and similarly for Spearman's correlation).

Table 2.5 Cross-tabulation for *Brick* and *Neighborhood*.

	Neighborhood		
Brick	East	North	West
No	26	37	23
Yes	19	7	16

Lessons Learned:

- There exist several fundamentally different data types: numerical data vs. categorical data and continuous data vs. discrete data. Numerical data is recorded in the form of numbers and can be "measured"; categorical data is recorded in the form of classes or categories and it typically cannot be measured. Continuous data is numerical, which can be recorded on a "continuous scale" (i.e., with as many decimal places as desired); discrete data, on the other hand, only assumes a set of fixed (typically integer) data values. Depending on the data type, we need to apply different tools for data analysis and exploration. In particular, most tools for exploring numerical data do not work (at least not directly) for categorical data. In addition, certain summary statistics (e.g., the mean) may be more meaningful for continuous data and may require more careful interpretation when dealing with discrete data. The exploration of categorical data often needs special attention.
- There exist many different tools for exploring the distribution of a single variable. Among them are summary statistics (e.g. the mean, median, mode, standard deviation, minimum, and maximum), tables, or graphs (e.g., histograms and boxplots). All of these tools should be used jointly and simultaneously, as they complement one another. In fact, while graphs (such as a histogram) provide a visual impression of a distribution, they do not allow easy quantification (and hence make comparisons of two distributions challenging). Summary statistics explore distributions quantitatively and hence can be compared readily across two (or more) variables.
- There also exist many different tools for exploring relationships between pairs of variables. Among them are correlation measures, cross-tabulations, and scatterplots. As with tools for single variables, tools for exploring pairwise relationships complement one another and should be used simultaneously. While scatterplots provide a visual assessment of the relationship between two variables, correlation measures can quantify this relationship (and subsequently be used for comparison).

2.2 Data Transformations and Trellis Graphs: Direct Marketing Data

In this section, we discuss a few more advanced and powerful ideas for exploring data. First, we introduce the concept of *scatterplot matrices*, which can unearth relationships between many different variables in one single graph. In fact, the version of scatterplot matrices that we use here is one of the most powerful available, as it combines scatterplots, correlation measures, and histograms in one single view. We also discuss data *transformation* as a means to obtain more consistent (and typically also more linear) relationship patterns. Then we discuss *trellis graphs*. Trellis graphs are powerful because they allow conditional views of the data. Trellis graphs are one of the most useful tools for unearthing new and unsuspected relationships in subsegments (or "pockets") of the data. It is often exactly these pockets that are of greatest value to the marketer or the investor, as they may offer opportunities that are otherwise impossible to detect.

Data: Table 2.6 shows data from a direct marketer. The direct marketer sells her products (e.g., clothing, books, or sports gear) only via direct mail; that is, she sends catalogs with product descriptions to her customers, and the customers order directly from the catalogs (via phone, Internet, or mail). The direct marketer is interested in mining her customers in order to better customize the marketing process. She is particularly interested in understanding what factors drive some customers to spend more money than others. To that end, she has assembled a database of customer records. These records include a customer's age (coded as young, middle, and old), gender (female/male), whether the customer owns or rents a home, whether the customer is single or married, the location of the customer relative to the nearest brick-and-mortar store that sells similar products (coded as far or close), the customer's salary (in US$), and how many children the customer has (between 0 and 3). The marketer also records the customer's past purchasing history (coded as low, medium, or high, or NA if the customer has not purchased anything in the past), the number of catalogs she has sent to that customer, and the amount of money the customer has spent (in US$).

Goal: One of the main goals of the marketer is to understand why some customers spend more than others. She is particularly interested in understanding the relationship between the number of catalogs and the amount of money spent since every catalog costs a fixed amount of money to produce and ship. Moreover, as the relationship with a customer matters, she is also interested in investigating whether customers with a high purchasing history in the past indeed also spend more money in the future. And lastly, as the marketer suspects that her product and service offerings may appeal more to some demographics than others, she is particularly interested in detecting "pockets" of customers that are most profitable (which she may ultimately decide to target with coupons and promotions).

We again accomplish these goals using only exploratory tools (graphs and data summaries). Some of the tools we use here were introduced in Section 2.1, but here

Table 2.6 The direct marketing data. See also file `DirectMarketing.csv`.

Age	Gender	Home	Married	Loc	Sal	Chld	Hist	Ctlgs	Spent
Old	Female	Own	Single	Far	47500	0	High	6	755
Middle	Male	Rent	Single	Close	63600	0	High	6	1318
Young	Female	Rent	Single	Close	13500	0	Low	18	296
Middle	Male	Own	Married	Close	85600	1	High	18	2436
Middle	Female	Own	Single	Close	68400	0	High	12	1304
Young	Male	Own	Married	Close	30400	0	Low	6	495
Middle	Female	Rent	Single	Close	48100	0	Med	12	782
Middle	Male	Own	Single	Close	68400	0	High	18	1155

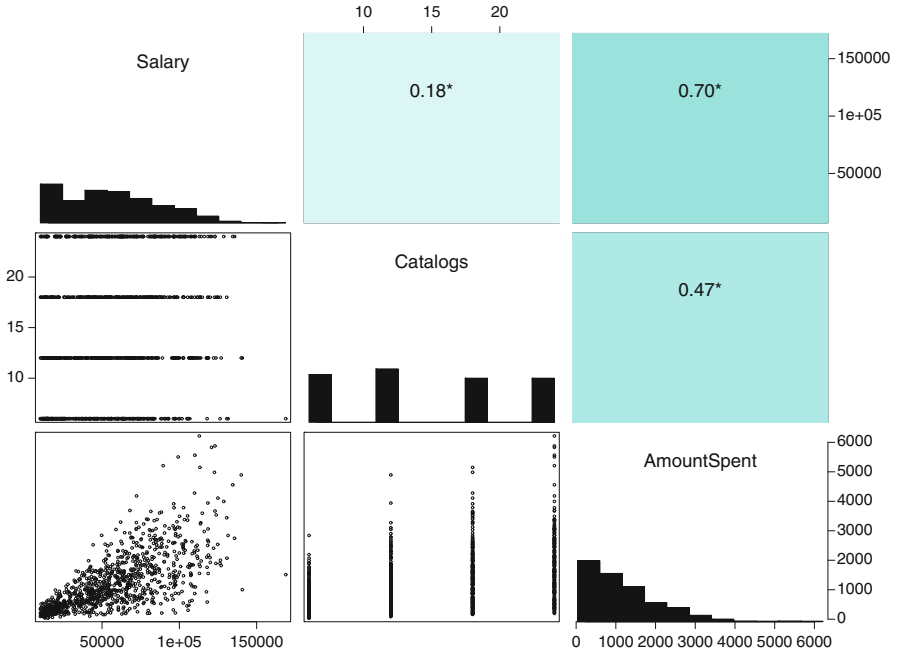

Fig. 2.6 Scatterplot matrix for salary, number of catalogs, and amount of money spent.

we use them in a slightly more advanced fashion. In addition, we also introduce new tools and concepts that are especially useful in the context of mining large databases. These include *scatterplot matrices*, *data transformations*, and *trellis graphs*.

Scatterplot Matrices: Figure 2.6 shows a scatterplot matrix for the variables salary, number of catalogs, and amount of money spent. In particular, it shows three different types of visualizations in one graph. Along the diagonal axis, it shows histograms for each of the three variables; below the diagonal, we see scatterplots between each of the three variable pairs; and above the diagonal we see the corresponding correlation values for each pair. Note that the correlation values are accompanied by different colorings, where darker colors indicate stronger correlations.

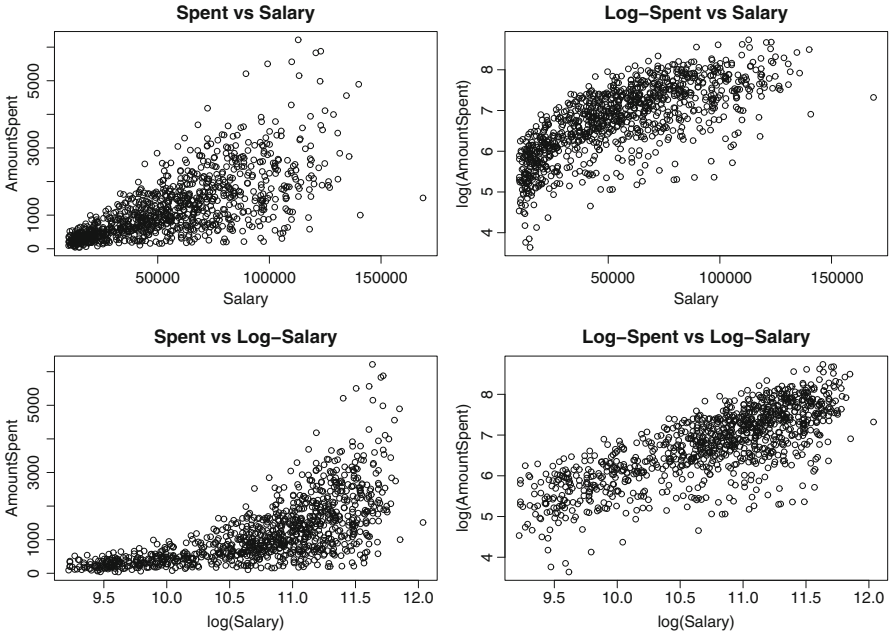

Fig. 2.7 Data transformations between salary and amount of money spent.

We can learn from Figure 2.6 that salary has the strongest correlation with the amount a customer spends. This is not too surprising because customers with little income are unlikely to spend a lot of money. But we can also learn that both salary and the amount spent are very right-skewed (notice the shape of the histograms) and, as a result, the relationship between the two is not at all consistent. In fact, if we look at the corresponding scatterplot (bottom left corner), then we notice that the points are "funneling out." In other words, while there is only a little variance at the lower levels of salary, the variance of amount spent increases with increasing levels of salary. An increasing variance is a problem because it implies that we cannot *predict* the spending behavior of the high-salary customers very accurately and, as a result, cannot target our potentially most profitable customers very well.

Data Transformations: Problems with skewed distributions in histograms or funnel effects in scatterplots can often be overcome (or at least smoothed out) by applying a suitable transformation to the data. Note that the scatterplot between salary and amount spent suggests that as salary and amount spent increase, the variation between the two also increases. We can eliminate this effect by transforming the data in a way that reels in the very large data values while leaving the smaller values unchanged. The logarithmic (or "log") transformation has this property. Figure 2.7 shows the changing relationship between salary and amount spent as we apply the log-transform to either salary, amount spent, or both. We can see that applying the log-transform to both salary and amount spent results in a

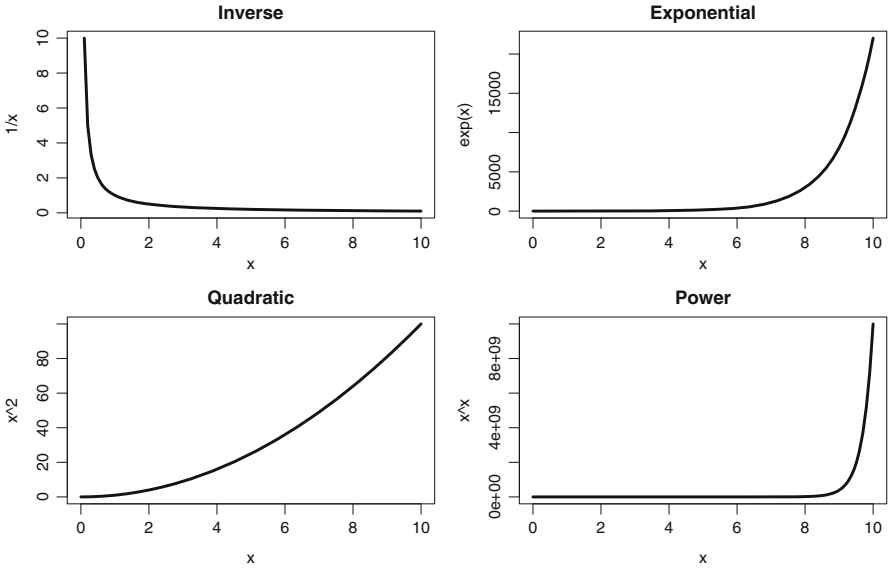

Fig. 2.8 Typical data transformation functions.

pattern that resembles a straight line. Moreover, the funnel effect has vanished; that is, the variation between the two variables is now the same at all levels. Thus, if we base our targeting efforts on the relationship between log-salary and log-spent, then we can target the high spenders with the same accuracy as the low spenders.

We have seen in the previous paragraph that a logarithmic transformation can ease data problems and in particular make relationships between variables more consistent. The logarithmic transformation is not the only transformation that can achieve that goal. There are many more transformations (such as exponential, inverse, quadratic, or the power transform) that can lead to similar results in different applications. Figure 2.8 illustrates some of these transformations.

Trellis Graphs: Our analysis thus far has revealed that there is a (linear) relationship between (log-) salary and (log-) amount spent; in other words, our most profitable customers will be the ones with the highest incomes. But does this relationship apply equally to all our customer segments? For instance, could it be that the rate at which customers spend their earnings varies between old and young customers? Figure 2.9 shows one answer to that question. It shows a *trellis graph*, which displays the relationship between two variables (log-salary and log-spent in this case) *conditioned* on one or more other variables (age and marital status in this case).

Figure 2.9 shows that the relationship between salary and amount spent varies greatly across different customer segments. While there is a strong linear relationship for old customers, there is almost no relationship for young married customers. In other words, while we can predict very accurately how much an old customer will

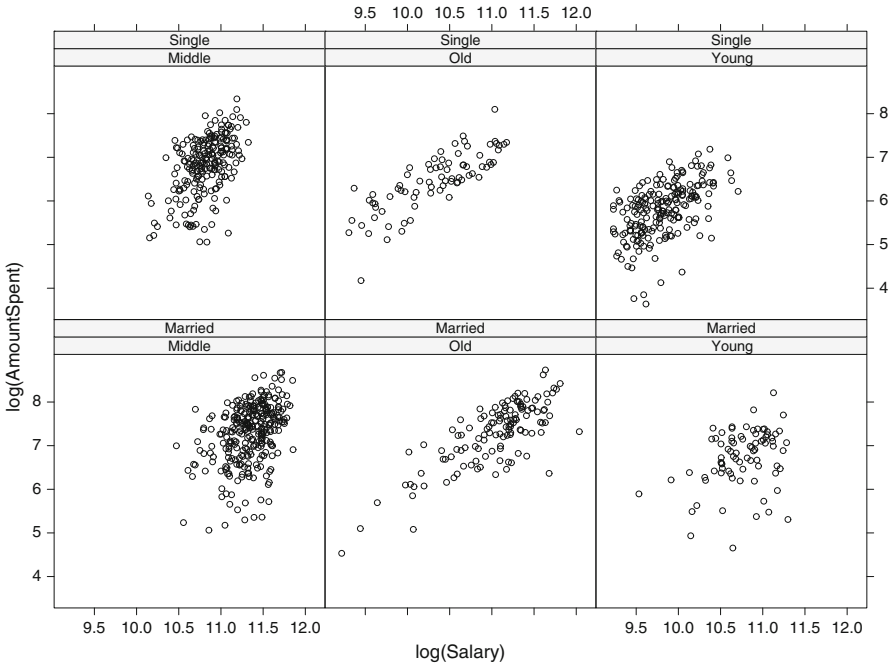

Fig. 2.9 Trellis graph for the direct marketing data. The relationship between salary and amount spent is conditioned on age and marital status.

spend, we cannot do the same for the young and married customers; we take this as an indication that it would be better to avoid this customer segment altogether. We can also see that compared with the previous two segments, the relationship for single middle-aged customers is much stronger (i.e., there is a much steeper trend, hence the rate of spending is much faster). Thus trellis graphs allow for a more granular inspection of the data and for the discovery of new segment-specific relationships. This is further illustrated in Figure 2.10, which shows another trellis graph, this time conditioned on a customer's spending history and location.

Lessons Learned:

- Scatterplot matrices allow us to visualize the relationships between many different pairs of variables on one single graph; they also allow us to incorporate additional information such as correlation values or the distribution of individual variables. Scatterplot matrices are a great tool for giving an overview of the most important data features in one single snapshot.
- Data transformations can be used to render more consistent relationships between variables. In fact, data transformations can be used to get rid of "funnel effects" or skew in variables. Data transformation includes

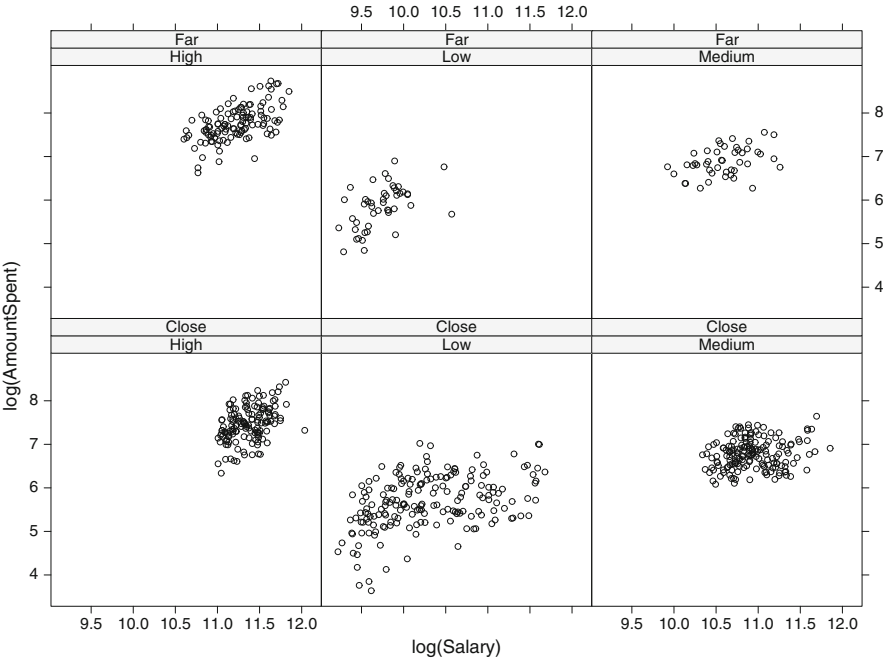

Fig. 2.10 Another trellis graph for the direct marketing data. Now the relationship between salary and amount spent is conditioned on purchasing history and location.

many different functional operators such as the logarithm or the quadratic function. The main goal of data transformation is to render the relationship more linear (i.e., to transform the data pattern so that it more-closely resembles a straight line).

- Trellis graphs allow us to investigate segment-specific relationships and detect pockets where relationships change. Unearthing this change of relationship could lead to a different managerial action: it could either lead to more specific, custom-made marketing or advertising, or it could lead to abandoning entire market segments altogether. In other words, trellis graphs allow us to detect pockets of opportunity and, equally, segments where no competitive edge exists.

2.3 Time Series Graphs: Soft Drink Sales Data

In this section, we discuss time series graphs. Time series graphs are different from the other visualizations discussed in this chapter, as they capture dynamic information that changes over time. While time series graphs are, at least in principle, a very simple concept, we discuss good and bad examples of them.

Table 2.7 Soft drink sales data. See also file `Softdrink.csv`.

Quarter	Sales	t	Q
Q1-86	1734.83	1	1
Q2-86	2244.96	2	2
Q3-86	2533.8	3	3
Q4-86	2154.96	4	4
Q1-87	1547.82	5	1
Q2-87	2104.41	6	2
Q3-87	2014.36	7	3
Q4-87	1991.75	8	4

Data: Table 2.7 shows a different kind of data. It shows sales (recorded in millions of dollars) of a major soft drink company. What makes this data different is that we only have two pieces of information available: information on the quarter (e.g., first quarter of 1986, Q1-86) and sales in that quarter. Note that Table 2.7 also has a quarter count t (which ranges from 1 to 56 since there are a total of 56 quarters in this data) and a quarter indicator Q, which denotes the quarter of the year (1 corresponds to the first quarter, 2 corresponds to the second quarter, etc.), but this is merely a recoding of the quarter information in the first column.

Goal: Our goal is to understand company sales, if and why they vary from quarter to quarter, and the rate at which they grow (or decay). Ultimately, a manager will want to use this information to *forecast* future sales for planning purposes. Specific items that we may want to identify are a *trend* (i.e., whether the data grows in a systematic pattern) and *seasonality* (i.e., whether that data fluctuates systematically; e.g., higher sales in the summer months and lower sales in winter).

Time Series Plot: While the data above appears to be rather simple (after all, it contains only two different pieces of information, time and sales), only a very careful analysis will reveal all the knowledge hidden in it. Figure 2.11 shows two different graphs of that data. The left panel shows a simple (scatter-) plot of sales versus time (quarter in this case). We can see that there appears to be a positive trend (sales grow over time), but we can also see that there appears to be a lot of noise around that trend. In fact, while sales appear to trend upward, individual data points scatter quite heavily around that trend. This would suggest that sales are quite variable from quarter to quarter, making sales forecasting quite burdensome and unreliable.

The right panel reveals the reason for this "noise." The colored boxes represent the type of quarter, and we can see that sales are generally higher in spring and summer (blue and green boxes) compared with fall and winter (light blue and red boxes). We can thus conclude that sales exhibit not only a positive trend but also a strong seasonal pattern. In other words, once we control for *both* trend and seasonality, the data aren't all that variable after all and there is good reason to believe that we can forecast sales quite accurately into the future.

We can make an additional observation: the dashed grey line shows a *smooth trend* through the data, and we can see that while sales are generally growing,

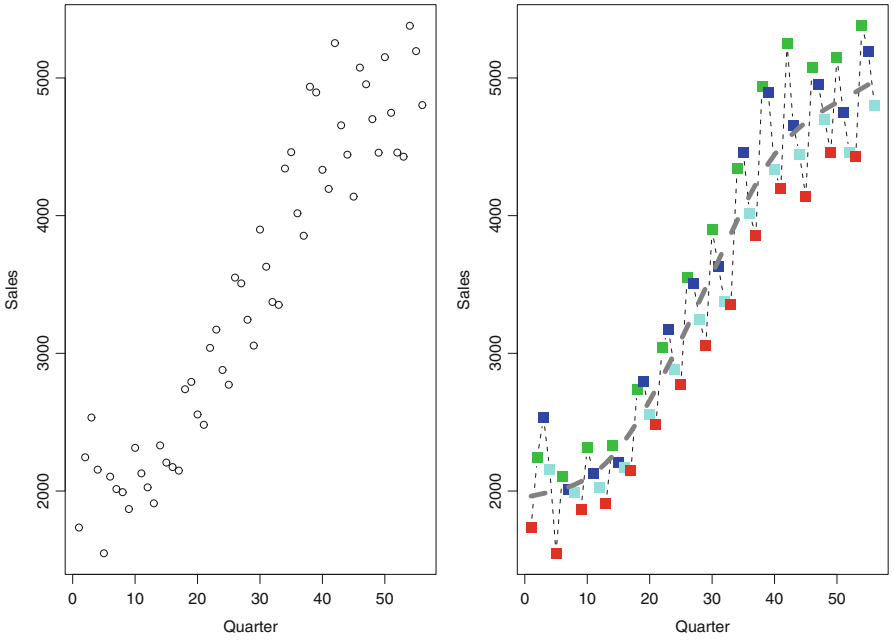

Fig. 2.11 Time series graphs for the soft drink sales data. The left panel shows a very simple graph of the data. In the right panel, colored boxes represent different quarters; the grey dotted line shows a smooth trend through the data.

the *rate of growth* is decreasing toward more recent quarters. In other words, sales increases are diminishing. Note that none of this information was directly obvious from the much simpler scatterplot in the left panel of Figure 2.11.

Lessons Learned:

- We can use time series graphs for visualizing trends and seasonality in data that is recorded over time. Time series data often appears dangerously simple when in fact it could hide a trove of valuable knowledge. This knowledge can be unearthed only by using the right graphical tools (such as color-coding different quarters differently or overlaying a smooth and flexible trend line). Time series data often shows a trend; that trend can be linear (i.e., growing at the same rate over time) or nonlinear. Nonlinear trends can occur in the form of increasing or decreasing growth rates and could capture real effects such as diminishing returns or an explosion of word-of-mouth referrals. Time series data can also show seasonality. Seasonality can occur on a quarter-to-quarter basis (e.g., summer vs. winter) or on a monthly basis. Seasonality can be detected more easily when it arrives in very regular patterns (e.g., winter sales are always lower

compared with spring sales). But seasonality can be less regular (and hence a little harder to detect and capture). For instance, sales could bottom out in January in one year but hit lows in February in the next year. While in both years sales are lowest in the winter months (January or February), it is hard to pinpoint the exact timing of the low on a year-to-year basis. Seasonality can become even more complex (and hence harder to detect) when it occurs only every few years. For instance, while the economy may grow in some years, it may experience a downturn in other years. Such "ups and downs" in long-term economic data are often referred to as cyclical (or business cycle) rather than seasonal variations. From a modeling point of view, we would need many years worth of data in order to properly account for such business cycles.

2.4 Spatial Graphs: Online Purchase Preferences Data

We next discuss spatial graphs. By spatial graphs we typically mean maps, and we use them to visualize geographical dependencies. Spatial graphs are becoming increasingly important with the increasing availability of spatial information. Take for instance the very recent development of Google Latitude,[3] which allows Google users to share their geographical locations. This is only one example, but it suggests that geographical information on customers, products, or services will explode in the upcoming years. Spatial information is extremely valuable because it allows us to geotarget consumers. Local searches and searches on maps are only two recent applications that rely heavily on geotargeting. Most spreadsheet-based software packages (such as Excel) have no way of exploring geographical data. This limitation does not allow managers to access and learn from one of the most important pieces of business information.

Data: Table 2.8 shows data on geographical differences in product preferences across the United States. The table shows sales data for books that were offered both in print format and as downloadable PDF files (i.e., in electronic format). The table also shows the price differences between print and PDF versions: PrPRINT denotes the price of the print version (in US$); PrPDF denotes the corresponding price of the PDF file. The electronic format was typically priced lower than the print format, and RelPrPDF records the relative price difference between the two formats. The table also records whether a customer purchased the PDF version (PurPDF) or

[3] See www.google.com/latitude.

Table 2.8 Geographical preference data. See also file `SpatialPreferences.csv`.

Long	Lat	PrPRINT	PrPDF	RelPrPDF	PurPDF	PurPRINT
−74.058	42.83326	34.95	17.48	50%	1	0
−163.11894	60.31473	39.95	29.96	75%	0	1
−163.11894	60.31473	39.95	29.96	75%	0	1
−86.1164	32.37004	28.00	7.00	25%	1	0
−111.82436	33.32599	24.95	18.71	75%	0	1
−111.82436	33.32599	18.00	13.5	75%	0	1
−118.29866	33.78659	49.95	0.00	0%	1	0
−118.29866	33.78659	57.95	14.49	25%	1	0

the (higher-priced) print version (PurPRINT).[4] Moreover, Long and Lat denote the longitude and latitude of the customer's location (i.e., it denotes the geographical area of the purchase).

Goal: One of the goals of the analysis is to determine whether there are geographical differences in product preferences. For instance, we may want to ask whether customers on the East Coast are more likely to purchase a book in the electronic format. Moreover, we would like to understand how product preferences vary as a function of the price difference between the print and PDF formats. Understanding customers' geographical preferences and price sensitivities allows retailers to better market their product, geotarget their customers, and offer the right coupons and promotions in the right locations.

Spatial Graphs: Figure 2.12 shows a map of the United States. On this map, we record the location of each transaction; a black circle represents a print purchase and a red circle represents a PDF purchase. The size of the circle corresponds to the price of the PDF for that relative to that of print. In other words, very large circles indicate that the PDF version was priced (almost) as high as the corresponding printed book; small circles indicate that the PDF version was available at a steep discount relative to the print version.

We can see that the preference between PDF and print varies significantly throughout the united states. While in some areas (e.g., in the South) print was the predominant format (unless the PDF was offered at a steep discount), in other areas (e.g., the West Coast or the Northeast) customers preferred the PDF format, even at a higher price. This insight can help marketing managers determine the right price for their product, geotarget their customers, and offer spatially varying coupons and promotions.

[4]We only show the transactions that resulted in either a print or a PDF purchase; of course, some transactions resulted in no purchase, but we do not show such data here.

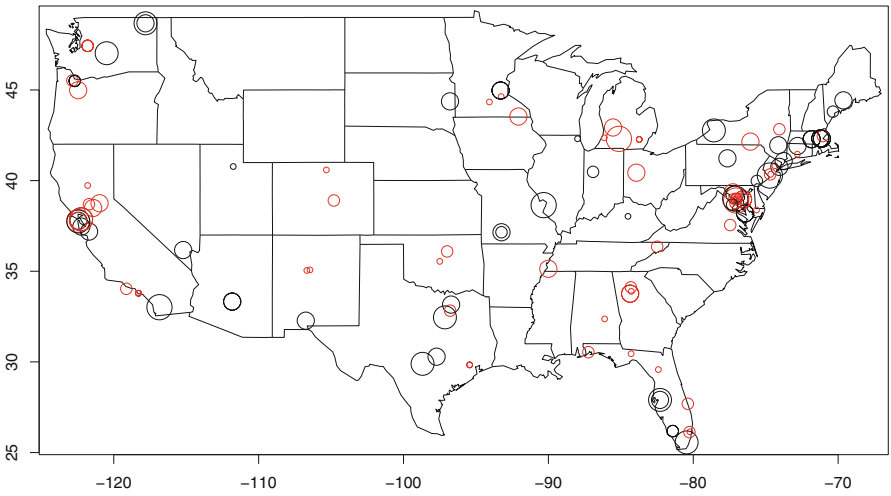

Fig. 2.12 Map of geographical preferences for the print vs. PDF format. Black circles indicate print purchases; red circles indicate PDF purchases. The size of the circle represents the price of PDF relative to print.

Lessons Learned:

- Spatial graphs, in particular maps, can be used for exploring geotagged data; that is, data with geographical information attached. Spatial graphs can be used for geotargeting and identifying geographical pockets of varying consumer demand.

2.5 Graphs for Categorical Responses: Consumer-to-Consumer Loan Data

In this section, we take a spatial look at data where the outcome of interest is categorical. While in principle similar to the data types discussed in earlier sections, categorical outcomes pose a challenge because standard scatterplots or correlation measures often are not meaningful. On the other hand, categorical outcomes are becoming more and more prevalent, especially in marketing, where managers are often interested in understanding the choices that consumers make – and choice data is inherently categorical.

Data: Table 2.9 shows credit data for a consumer-to-consumer (C2C) lending market. In this market, consumers ("borrowers") can post loan listings, and other

Table 2.9 Consumer-to-consumer lending data. See also file `LoanData.csv`.

Status	Credit Grade	Amount	Age	Borrower Rate	Debt-Income Ratio
Current	C	5000	4	0.150	0.040
Default	HR	1900	6	0.265	0.020
Current	HR	1000	3	0.150	0.020
Late	HR	1000	5	0.290	0.020
Current	AA	2550	8	0.079	0.033
Late	NC	1500	2	0.260	0.030
Current	HR	3001	6	0.288	0.020
Current	E	2000	6	0.250	0.020

consumers ("lenders") can invest in those loans by bidding on borrowers' loan rates. The data shows the status of the loan (current, late, and default), the borrower's credit grade (AA is the highest grade, followed by A, B, ..., E; HC stands for "high risk" and denotes the lowest grade; NC stands for "no credit rating"). The data also has information on the amount borrowed (in US$), the age of the loan (in months), the borrower rate (i.e., the interest rate the borrower pays the lender), and the debt-to-income ratio of the borrower.

Goal: The goal is to distinguish good loans from bad. In other words, we want to investigate how a lender can determine which loans will result in timely payments ("Current") and which will result in late payments or even in defaults. Note that the prediction problem is slightly different from all the other examples we have studied before: while previously the goal was to predict the outcome of a numerical variable (e.g., house price, amount spent, or quarterly sales), now we need to predict a categorical variable, "Status." Status assumes the values current, late, or default and is thus not measured on a numerical scale. The problem with predicting categorical variables is that traditional models (which assume numerical variables) do not apply. This is also important for the exploration task since we need to choose our data visualizations carefully, as otherwise we will not get the right answers to our questions.

When visualizing data with categorical outcomes, one typically visualizes the distribution of input variables at all levels of the outcome variable. For instance, in the case of the loan data, we may want to investigate if the distribution of the loan amount differs between loans that are current and those that are late. In fact, if we detected a systematic difference, then this would indicate that the size of the loan amount is a good indicator of future loan performance. Similarly, we may also want to investigate whether the distribution of credit grades differs systematically across different loan statuses because if we found a systematic difference, then the conclusion would again be similar to that above, namely that credit grade is a good predictor of loan performance. Thus, while in both cases we want to investigate the *distribution* of an input (or predictor) variable at all levels of the outcome (or response) variable, the exact way we accomplish this depends on type of input

variable. In the following, we discuss two examples, one in which the input variable is numerical and another where the input variable is categorical. To that end, we will use *density plots* and *spine plots*.

Density Plots: A density plot is similar to a histogram. In fact, the only difference between a histogram and a density plot is that while the former selects "buckets" of a certain length and then plots the frequency in each bucket, density plots can be thought of as histograms with arbitrarily small buckets. Thus, their advantage is that they represent the data distribution in the most granular form.

Figure 2.13 shows a density plot for the loan data. In fact, we see density plots for each of the four numerical variables: amount borrowed, age of loan, borrower rate, and debt-to-income ratio. Moreover, for each variable, the density is broken up by the status of the loan: the black lines correspond to densities of current loans; the green lines correspond to late loans; and the red lines correspond to loans in default. We can see that while the distribution of loan amount (top left panel) is almost identical across all three loan statuses, it is very different for the age of the loan (top right panel). In fact, the graph suggests that many current loans are young (i.e., only a few months of age), while most defaulted loans are old (i.e., five or more months old). While this result is not completely surprising (a consumer typically defaults after a certain period of time and not immediately after taking out the loan), it does suggest a way to distinguish between good and bad loans. Figure 2.13 suggests additional ways in which loans can be distinguished. The bottom left panel

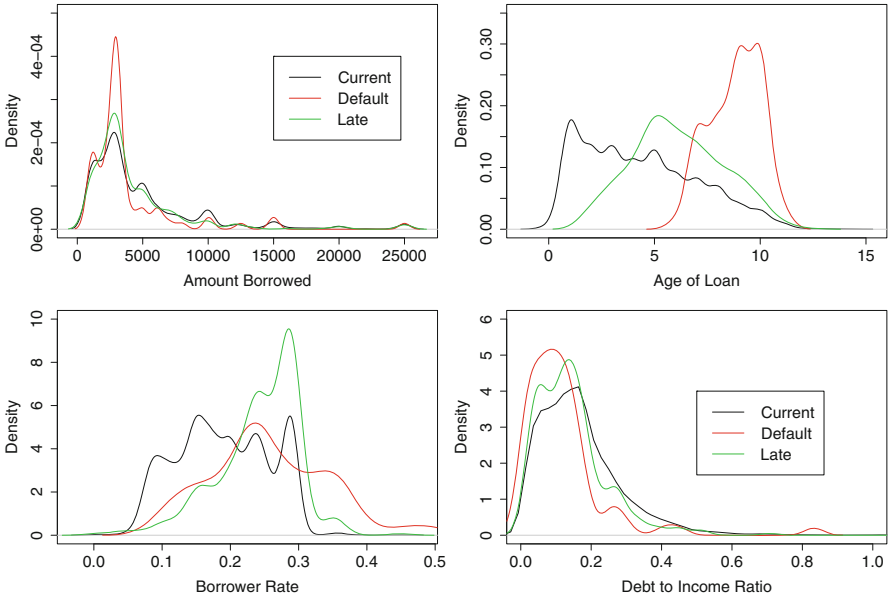

Fig. 2.13 Distribution of amount borrowed, age of loan, borrower rate, and debt-to-income ratio, broken up by different loan outcomes (current, late, or default).

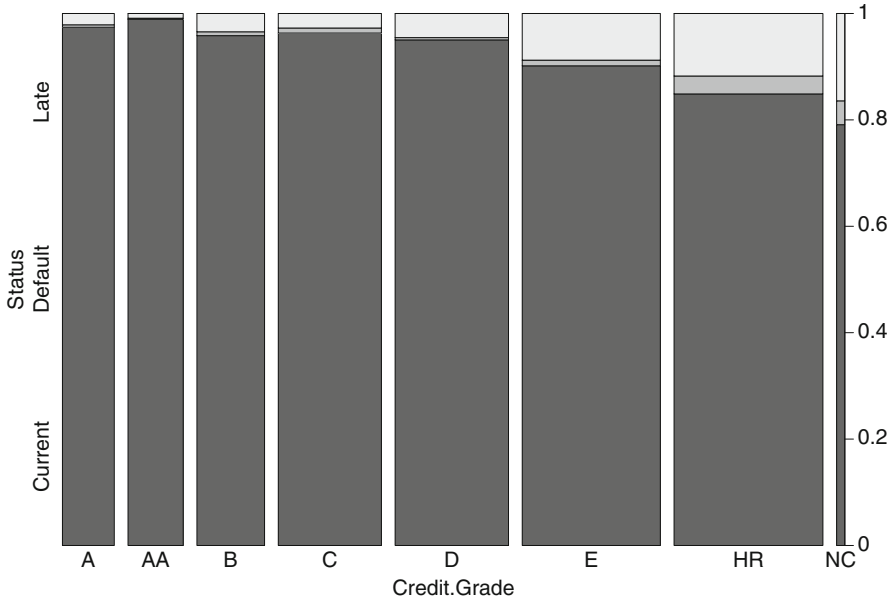

Fig. 2.14 Credit grade by loan outcome.

(borrower rate) suggests that late or defaulting loans have higher loan rates; the bottom right panel also suggest that a low debt-to-income ratio does not necessarily signal a good loan.

Spine Plots: Spine plots are a way of visualizing cross-tabulations (e.g., relationships between two categorical variables). Figure 2.14 shows a spine plot between the loan status and the credit grade. Black bars correspond to current loans, dark grey bars correspond to loans in default, and light grey bars correspond to loans that are late. The width of the bar corresponds to the number of loans with a particular credit grade. (For instance, the "A" bar is thinner than the "C" bar, suggesting that there are many more loans graded C than A.)

We can learn that, unsurprisingly, as the grade deteriorates, the number of late and defaulted loans increases. In particular, HR (high-risk) loans have the greatest number of loans in default or that are late. It is interesting to note, though, that while there are only a small number of nongraded loans (NC), their default and late-payment rates are is even higher than for high-risk loans. Thus, credit grade is a very strong predictor of loan status.

Lessons Learned:

- Density plots and spine plots are very powerful tools for investigating data where the response is categorical. The main idea of these plots is to split up one of the input variables (e.g., age of the loan) by the different levels

of the target variable. For instance, we could plot histograms for different levels of loan status or, as in the case of density plots, create histograms of extremely fine granularity.
- Categorical responses occur frequently in business data, especially in marketing, where we study customers' choices and preferences. While traditional scatterplots are not very useful for exploring data with categorical responses, density plots and spine plots can unveil new, previously unknown knowledge.

2.6 Graphs for Panel Data: Customer Loyalty Data

We finish this chapter by giving a glimpse into yet another challenging form of business data: panel data. Panel data occurs when we follow a set of customers over time and record their behavior and preferences. Thus, panel data share similarities with time series data (for each panel member, we have a time series of observations); on the other hand, panel data also has cross-sectional features, as the panel contains a sample of different customers. Understanding all the information that panel data carries is not easy, and we want to explain the challenges in the following section.

Data: Table 2.10 shows purchasing data for nine randomly selected customers of an Internet DVD rental business for a period of 12 months (January–December). Each of the 12 columns refers to the amount of money a customer spends in a given month. For instance, customer 2 spends $114.33 in the month of February and continues to spend rather frequently until the end of September (after which he makes no further purchases). On the other hand, customer 1 made a single purchase ($25.74) that occurred in February, and she did not spend any more money in subsequent months. We also see that some customers spend nothing at all during the entire year (e.g., customers 6, 7, and 9).

Table 2.10 Customer loyalty data. See also file `CustomerLoyalty.csv`.

ID	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0	25.74	0	0	0	0	0	0	0	0	0	0
2	0	114.33	108.56	51.28	0	52.28	70.07	40.1	47.96	0	0	0
3	21.54	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	8.79	42.1	0	0	10.77	63.25	27.93	0	63.45	83.94	30.98
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	24.48	13.97	0	48.37	0	52.27	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0

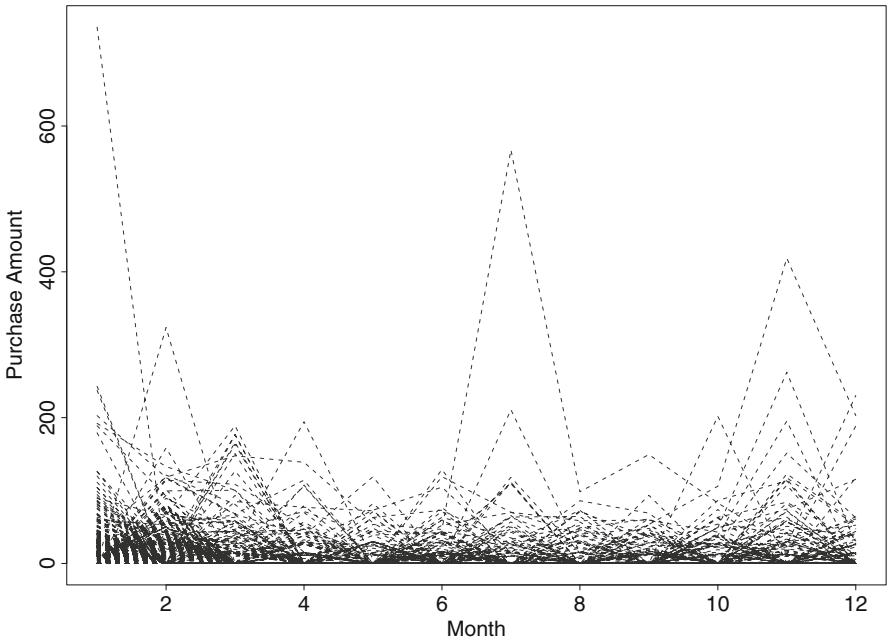

Fig. 2.15 Time series plots for all customers in the panel.

Goal: The goal of the analysis is to understand purchasing patterns. Why do some customers spend more money than others? And when do they spend their money (early in the year or towards the end of the year)? And why do some customers make only one-time purchases while others spend frequently? Can we segment customers by their purchasing patterns? Answers to some of these questions could help our business target individual customers with tailored coupons and promotions. For instance, customers who only spend at the beginning of the year could be enticed to spend additional money in later months via coupons that are valid only for the holidays. On the other hand, if we knew that a customer only made one-time purchases and if we knew the timing of that purchase, then we could maximize that single purchase with the right advertising and promotions offered at the right time.

Overlaid Time Series Plots: We introduced time series plots in Section 2.3. We have also argued above that panel data are essentially a bunch of individual time series – one series for each customer. So, why not plot all those time series (one per customer) into one single graph, you may ask? Well, the answer is that this kind of approach often leads to information overload and clutter; that is, our attempt at exploring all the available data at once leads to more information than the graph can carry and, as a result, we don't learn much at all!

Take as an example Figure 2.15, which shows the purchase pattern for all customers, across all months, in one graph. We refer to this graph as an *overlaid*

time series plot since we essentially took many individual time series and overlaid them all together on one page. Note that we attempted to make the graph as clean as possible by choosing dashed lines, which results in as little clutter as possible.

All in all, we cannot learn much from Figure 2.15. We can learn that there exist some (in fact, rather few) customers who make remarkably large purchases at select instances throughout the year. (Notice the high spikes at the beginning and at the end of the year, and also the unusually high spike at month 7.) However, while these few customers bring exceptional value to our business at select occasions, they are not representative of the *typical* customer. The typical customer is "hidden" in the line clutter at the bottom of the graph.

The main problem with Figure 2.15 is that it tries to accomplish too much: it tries to represent both the temporal information (i.e., the purchasing pattern of each customer over time) and the cross-sectional information (i.e., the variation across customers). While preserving as much information about the data as possible is often a very valuable objective, this is an example where data aggregation will lead to better insight. What we mean by that is that we should first try to aggregate the data (either by its temporal or cross-sectional component) and only then graph it. In the following, we discuss several ways of accomplishing this aggregation task. It is important to note that the actual graphs that we use are standard and have been introduced in earlier sections (e.g., histograms); however, we apply these graphs in an innovative way to take advantage of the special structure of panel data.

Aggregating the Cross-sectional Dimension: Panel data have two main dimensions: temporal information and cross-sectional information. If we want to explore trends over time, then we should aggregate over the cross-sectional dimension and keep the temporal dimension intact. Aggregation can be done in a variety of ways. For instance, we could – for each month of the year – compute the average purchase amount; that is, we could compute numerical summary statistics for each month of the year. Alternatively, we could visualize the purchase distribution in each month using month-by-month histograms. This is shown in Figure 2.16.

Figure 2.16 shows that purchasing patterns differ from month to month. While January features a large number of high-value purchases (i.e., purchases with amounts up to $20 or $40), purchase amounts decline in subsequent months. In fact, January, February, July, and November appear to be the months in which a customer spends the most money during a single visit.

While the amount a customer spends matters, it equally matters whether a customer spends anything at all. In fact, Figure 2.16 does not quite tell us what proportion of our customers made any purchase at all. To that end, we can employ a similar rationale as above and compute month-by-month pie charts (see Figure 2.17). Note that each pie chart compares the proportion of customers who did not make any purchase (denoted by "0" and colored in white) with those who made a purchase (denoted by "1" and colored black).

We can learn that January, February, March, and maybe June are the months in which most customers make a purchase. In fact, as pointed out above, January and

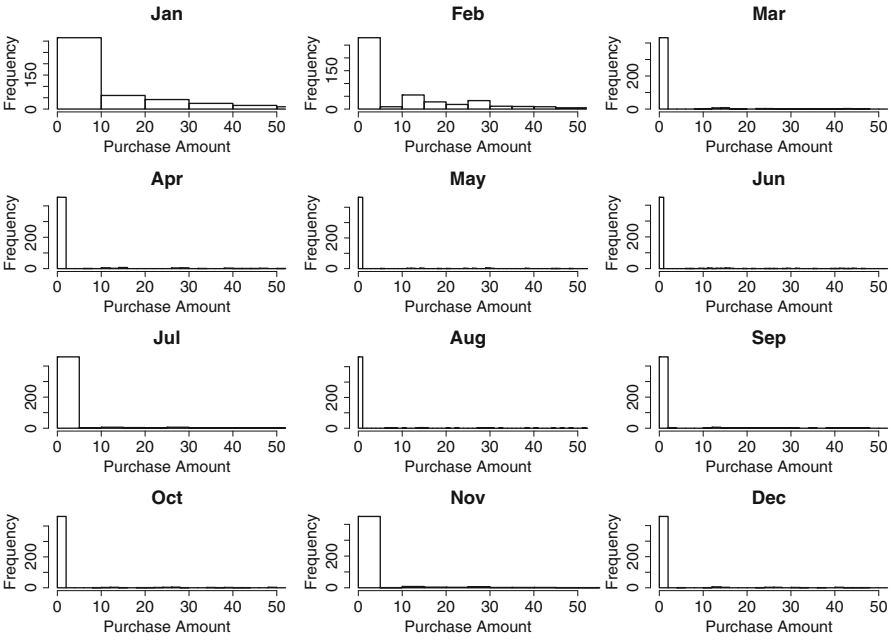

Fig. 2.16 Month-by-month histograms of customer purchases.

February are also the months in which a customer spends the most money during a single visit. Hence, these two months appear to be our most profitable month. The month of March is interesting because it is marked by many visits but relatively low spending per visit. On the other hand, while the month of July does not see a high number of visits, the amount spent per visit is rather high.

All in all, the aggregation of our panel data has led to new insight regarding the timing and amount of purchase decisions. While some months see more frequent customer visits (but are marked by lower purchase amounts), other months see higher purchase amounts (but less frequently). This insight could lead our marketing department to devise seasonally varying advertising and promotion strategies that during some periods aim at increasing the amount a customer spends ("budget focus") and during other periods aim at increasing a customer's purchase frequency ("frequency focus").

Aggregating the Temporal Dimension: Instead of aggregating over the cross-sectional information, we could also aggregate over the temporal information (and hence keep the cross-sectional information intact). In our situation, the cross-sectional information corresponds to the variation from one customer to another. Figure 2.18 shows customer-specific histograms (for the first 25 customers in our data). Each histogram shows the distribution of purchases made by this customer over the period of one year. In other words, while Figure 2.18 pre-

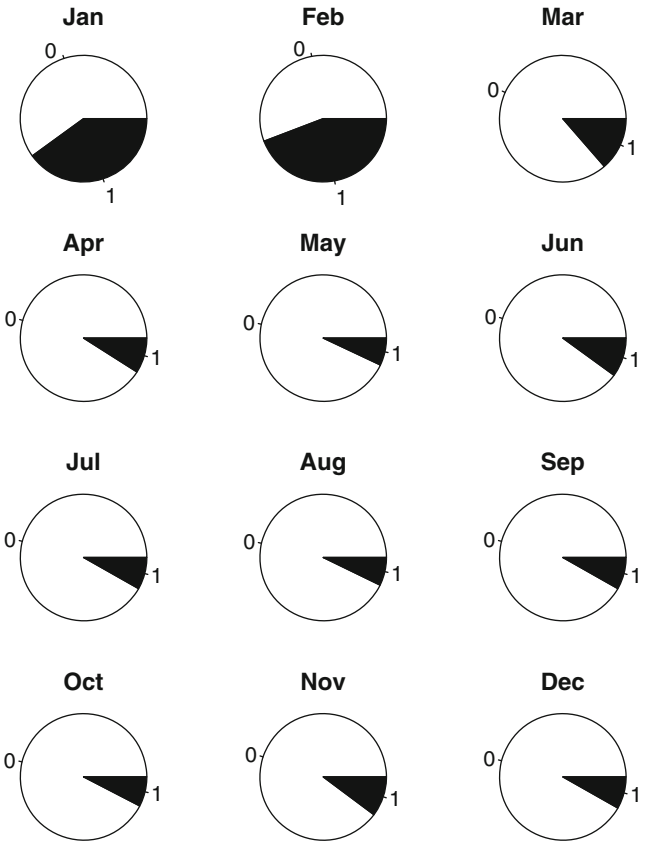

Fig. 2.17 Month-by-month pie charts of customer purchases.

serves the customer-to-customer differences, it loses the temporal information: we are no longer able to determine whether customer 1 made the purchase of $25 in January or in July. This is what we mean by aggregating over temporal information.

The usefulness of Figure 2.18 is limited and depends on our objective. If our goal is to develop customer-specific spending patterns, then Figure 2.18 tells us that, for example, customer 5 has a very different pattern compared with customers 6 and 13. However, recall that Figure 2.18 shows only a snapshot of the first 25 customers – if our panel contains several thousand (or even million) customers, then this approach would be quite cumbersome. Moreover, since we plot different histograms for different customers, we don't quite learn what is *common* across customers. In other words, panel data are challenging and one has to think very carefully about how best to extract the kind of knowledge that supports one's business goals.

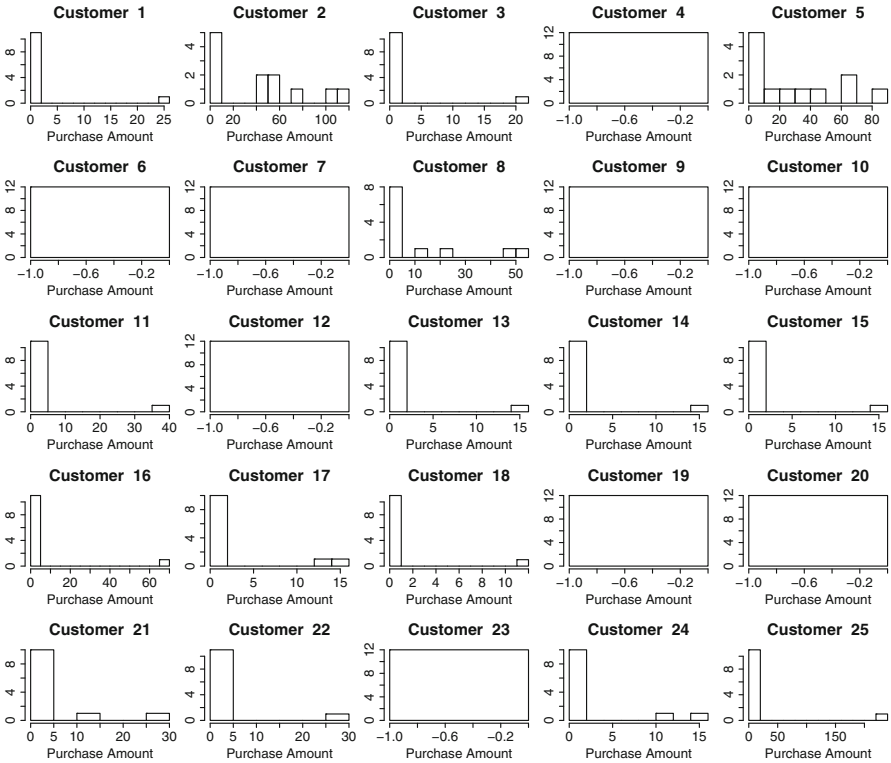

Fig. 2.18 Distribution of purchases for each customer.

Lessons Learned:

- Data aggregation is often useful prior to exploring panel data; data can be aggregated with respect to either temporal dimension or cross-sectional dimension.
- Sequences of histograms, pie charts, boxplots, or any other form of numerical or graphical summary can be useful for exploring aggregated panel data.

http://www.springer.com/978-1-4614-0405-7

Business Analytics for Managers

Jank, W.

2011, XI, 189 p. 100 illus., 63 illus. in color., Softcover
ISBN: 978-1-4614-0405-7