6: Explaining the City

Learning Objectives

By the end of this chapter students will understand the following:

- Models can be descriptive, predictive, or explanatory.
- Exploratory data analysis can help uncover meaningful patterns in data, which can in turn help guide model development.
- Regression is a flexible tool for helping to understand complex relationships within cities.
- Urban data are spatial data and these can be statistically problematic when used in a model, but there are techniques to explicitly account for spatial patterns.

Understanding the Interconnected City

The complexity of cities fosters opportunities for interaction that emerge between a diverse set of actors including households, businesses, and governments, whose interests may or may not align. Further, what we observe at any particular moment is the outcome of a myriad of processes, those past and present. These and many other tensions make explaining cities difficult.

Consider something simple. Many cities have bike-sharing systems (Figure 6.1) that also allow the automated tracking of bicycle ridership. The data collected by such a system over a five-year period may show that ridership has increased. This increase could have been *caused* by many factors including government intervention: for example, an increase in the number of bike lanes or traffic-calming measures aimed at making roads safer for cyclists; or changes in demographics, for example, an influx of younger residents or increased desire to live in denser parts of the city; or broader externalities, for example, a rise in global oil prices, making driving more expensive, an unusual number of days with pleasant temperatures, or a decline in the relative price of bike sharing compared to other public transit. The list could go on and on. In this chapter we will consider a variety of strategies to explain what is happening within cities through developing and testing hypotheses. We will primarily be considering the interactions of variables, and how variables interact in space.

Figure 6.1 Changes in bike commuting could be caused by a variety of factors

Source: Pixabay (CC0 Public Domain)

In previous chapters we mentioned that a key goal for urban analytics is the conversion of *data* into *information*. Modern data sources from sensors to social media, along with traditional data from surveys and censuses, have resulted in a proliferation of attributes available for study. In their raw form, such data are however not particularly useful, and its preparation for use is often the most time-consuming part of the urban analytics process (Chapter 3). After initially exploring data using graphical methods (Chapter 4), it is common to progress through a series of other research goals. In this chapter we formalize these as *descriptive*, *predictive*, or *explanatory* statistics, and present various modeling frameworks, in particular exploratory data analysis and regression models.

Models and Data

Models and data are two key pillars of quantitative research. A *model* is a mathematical representation of some real-world phenomenon. By their very nature, models are approximations or simplifications of the processes they hope to represent. This is exemplified in the aphorism attributed to George Cox: *all models are wrong, but some are useful*. Actual phenomena are far too complex to capture exactly in an equation; this is especially true for urban phenomena, which are subject to all the intricacies discussed throughout this book. Nonetheless, models can help us sort out the relative importance of contributors to complex phenomena. For example, taking the bike share example, do demographics or the presence of bike lanes explain the change in usage?

Another more detailed example: urban researchers have repurposed Newton's law of universal gravitation, the so-called "gravity model," to understand the economic, social, and migratory attraction between two locations. Newton's law explains the gravitational attraction between two objects. The mathematical formulation is:

$$F_{i,j} = G\left(m_i m_j / d_{i,j}^2\right)$$

which says the force ($F_{i,j}$) between two objects relates to their masses (m_i and m_j), the squared distance between them ($d_{i,j}$), and the gravitational constant (G). In an urban context, m_i could be the amount of goods produced in region i and m_j the amount demanded in region j, and $F_{i,j}$ the amount of goods flowing between the two regions. This general model, with empirically calculated values for G and the exponent on d, has been effective at explaining various types of aggregate flows between places.

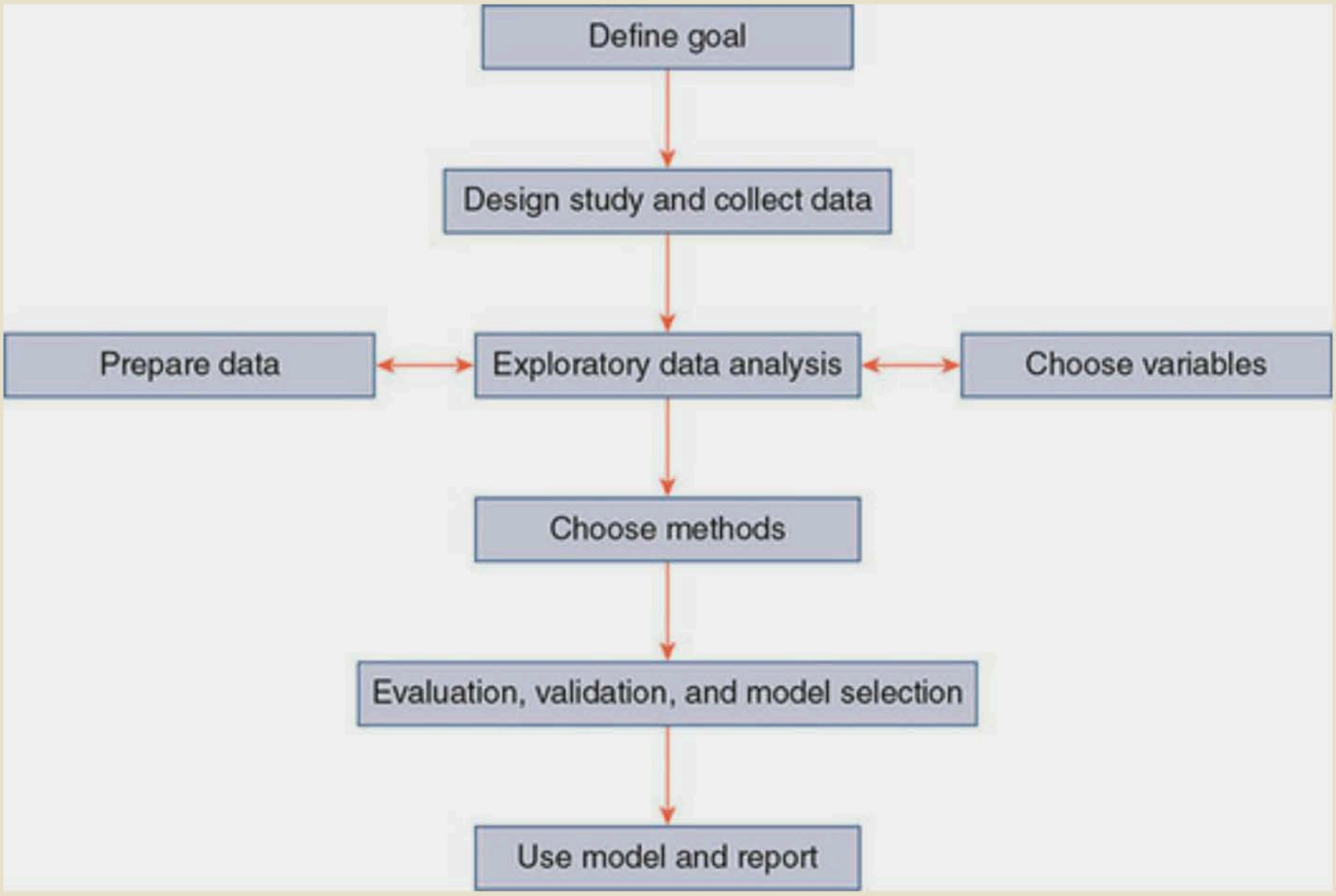
Data has been discussed earlier in the book (Chapter 2). In the context of modeling we conceptualize data as a set of *observations*, with each observation containing *attributes*. An observation is the organizing unit for the data put into a model. In almost all cases, models provide more useful and reliable results when more observations are included in the calculations. When studying where to deploy police (Figure 6.2), the organizing unit might be calls for service; in contrast, a study of effective police management might use the police precinct as the organizing unit since this aligns with the organization of leadership, staffing, and equipment. In turn, each observation has one or more attributes. An attribute is a characteristic of the observation. For example, calls for service might include the location of the event, the nature of the crime/problem, the address and phone number of the complainant, the number of officers dispatched, the date and time of the call and response. Attributes of observations can be qualitative, such as the type of crime (e.g., burglary, car theft, etc.) or whether the precinct uses body cameras, or quantitative, such as the value of goods stolen in a theft or the number of officers in the precinct.

Figure 6.2 The nature of a crime, policing, and location are all examples of attributes that might be collected by the police

Source: Pixabay (CC0 Public Domain)

Some common stages of quantitative modeling are presented in Figure 6.3, which are based on Shmueli (2010). Most of the steps precede the actual modeling. Effective research is based on having solid research goals and study design, and a clear understanding of the available data. With this grounding, the specific model can be chosen, tested, and reported.

Figure 6.3 Modeling flowchart



Source: Authors' own, but based on Shmueli (2010)

Modeling Goals

We consider three research goals that underlie most modeling efforts: *description*, *prediction*, and *explanation*. While there is considerable overlap between the three, and conflation of the terms within the research community, they have distinct characteristics and as a result distinct data and modeling requirements. Descriptive analytics involves *describing* the data, that is, summarizing large amounts of data into meaningful statistics and computing basic relationships between variables. Predictive analytics is used to *predict* the values of attributes that we do not have. Explanatory analytics aims to *explain* the relationships between variables, that is, to test causal relationships. Related terminology to these three are *exploratory data analysis* (EDA), *forecasting*, and *prescriptive analytics* respectively.

Descriptive analytics is by far the most prevalent of the three. It encompasses both means and correlations, among many other methods, with the goal of finding patterns from a wealth of data available to researchers. These methods tend to be straightforward to implement and are not overly burdened with statistical assumptions on the structure of attributes explored within the data. Related research questions might be: "What was the average home sales price last year?"; "Are the number of residential swimming pools increasing at a faster rate than new homes?"; and "Has the average drive time from fire stations to homes increased or decreased over the past ten years?". Another strength of descriptive analytics is that the methods and results tend to be easy to explain to a broad audience.

Predictive analytics answers questions such as: "What is the market value of homes for which we do not have a recent sales price?"; "What is the estimated demand for water to fill residential swimming pools next year?"; and "Which of the 10 candidate locations for a new fire station will have the lowest response time?". In all these cases the goal is to fill a data gap – to use existing observations to gain insight into things that do not (yet) exist. Also of note is that these questions do not ask "why": they simply challenge the analyst to make the best possible prediction using the available information. Of the three goals, prediction has benefited most from recent increases in data availability and computational power. These approaches thrive on large datasets that capture not only the main trends, but also the rare situations that can lead to large misses in predications. Machine learning and

data mining are increasingly popular tools for predictive analytics. In a predictive model the proof is in the pudding – how often is the model actually right?

Explanatory analytics tests hypotheses about urban phenomena. Questions might include: "Does proximity to the city center increase or decrease a home's sales price?"; "Does the presence of a swimming pool have a significant effect on residential water consumption?"; and "Does firefighter experience have an effect on response times?". The gold standard in explanatory analytics is to identify *causal* relationships. A causal relationship implies "if this, then that"; that is, with an appropriately defined causal model not only can one explain phenomena, but also one can identify the factors that can change outcomes. When an explanatory mathematical model achieves a causal explanation, there is often a relationship to some underlying social, economic, or other theory. This way of thinking can be flipped: if a theory is true and can be converted into mathematical form, one can test it with a model. Assessing the quality of an explanatory model can be difficult. The goal of the model is to show how variations in attributes *cause* variations in outcomes, so quality is typically based on how well the model fits observations. However, there might be many models that fit a dataset; choosing the "best" explanation among a set of alternative explanations is difficult (and is the subject of many hundreds of textbooks!).

Descriptive Analysis

The modern world of large datasets is both a blessing and curse. A large volume of data are only useful if we can make sense of them, which implies using techniques to (1) summarize and visualize the data in search of meaningful patterns, and (2) identify problematic data that might impede or obscure that search. Although these goals are distinct, there is overlap in the methods used to achieve them. While these methods can be the end goal of analysis, they also contribute to the modeling steps (Figure 6.3), by helping us design the study and choose variables.

Data irregularities can hinder both explanatory and exploratory modeling activities. It might be reasonable to say that Big Data often also means bad data – that is, in large datasets there are often records that are malformed, incorrectly recorded, missing, or just broken. Developing procedures to *clean* a dataset and identify *missing data* and *erroneous* attributes is a critical early step in an analysis. Missing data are not hard to find, it is simply an empty cell in a data table. It is important to maintain a distinction between values that are "missing" or "zero". A missing record is simply the absence of information, due to a broken sensor, human error, or some other cause. A "zero" is implicitly a substantive measurement. In the case of a zero, a sensor could be working properly and whatever it measures is simply absent. For example, if observations are neighborhoods and the attribute of interest is the number of property crimes, a missing value implies "we do not know the number of crimes" and a zero implies "no property crime occurred". Replacing a missing value with zero is thus bad practice.

Similarly, extraordinary data values, called *outliers*, can be real or erroneous. An outlier is a value that is extreme relative to the other values in the series. It can sometimes be substantive: Mark Zuckerberg's wealth will be extreme when put in a table with nearly any group of people; an outlier can also be the result of measurement error: a count of 10 global tweets in a day is clearly erroneous considering that the actual counts are hundreds of millions per day. Differentiating these two types of outliers usually requires some knowledge of the data generating process. Table 6.1 considers summer high temperatures for 11 days in Las Vegas, Nevada. The actual data has an average of 102.3 degrees Fahrenheit (column 1), but the presence of an outlier (July 3) reduces the average to 96.2 (column 2). In general, outliers should only be removed when they are the result of measurement error since "unexpected" findings are not necessarily wrong. The median can be an effective measure of central tendency when outliers are present; the median of columns (column 1 and column 2) is not affected by the presence of an outlier in this example. Missing values are often recorded as "NA" in a data table; some software raise an error in this case (column 3), while others silently ignore the missing value, resulting in 10 values being used in the computations (column 4). Often missing values are replaced with a zero (column 5), but in this example that is clearly an incorrect strategy as the lowest temperature ever recorded in Las Vegas was 8 degrees Fahrenheit in the winter of 1963, and this is data on summer high temperatures!

Table 6.1 Outliers and missing data can cause problems when creating models

	(1)	(2)	(3)	(4)	(5)
	Actual	**Outlier**	**Missing (error)**	**Missing (ignored)**	**Missing (zeroed)**
July 1	97	97	97	97	97
July 2	97	97	97	97	97
July 3	103	**40**	**NA**	**NA**	**0**
July 4	105	105	105	105	105
July 5	105	105	105	105	105
July 6	103	103	103	103	103
July 7	104	104	104	104	104
July 8	106	106	106	106	106
July 9	104	104	104	104	104
July 10	101	101	101	101	101
July 11	100	100	100	100	100
Average	102.3	96.5	NA	102.2	92.9
Median	103	103	NA	103.5	103

Note: Temperatures in degrees Fahrenheit

Source: Authors' own

Note: Temperatures in degrees Fahrenheit
Source: Authors' own

A key tenet of EDA is the identification of patterns via data visualization. This embrace of graphics as part of the modeling process is grounded in the astonishing capacity of the human brain to identify patterns via visual input. Our ancestors who could reliably distinguish the lion from the grass (Figure 6.4) were more likely to survive and thus pass on their pattern-identifying genes to the next generation. EDA leverages these skills to fill the gaps that social science theory and computer algorithms cannot. Through EDA, visualization can be used both to better understand the data we have and to identify potential research questions. However, the human capacity to see patterns in visual representations of data cuts both ways. People are notoriously bad at interpreting a random pattern as such. When presented with a graph of random numbers many people will find (spurious) patterns, this is especially true of maps. Chapter 4 introduced urban data visualization in a geographic context; here, however, we will consider a few additional techniques.

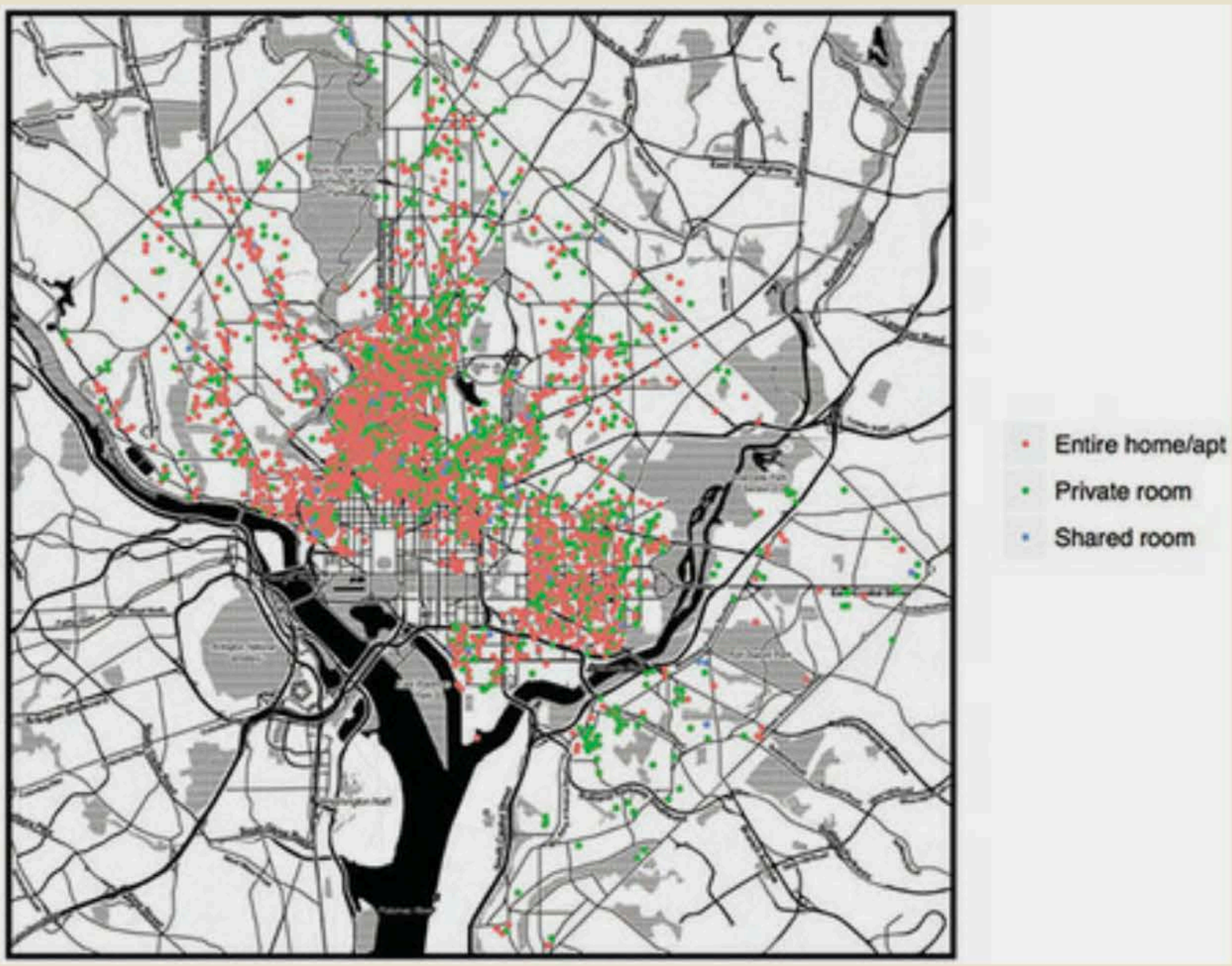
Figure 6.4 EDA leverages the human brain's remarkable capacity to see through noisy data (the grass) to identify substantive information (the lion)

Source: Pixabay (CC0 Public Domain)

Two common univariate data distribution visualizations are the *histogram* and *boxplot*. A histogram is a specialized bar chart that shows the shape of the distribution. The width of each bar represents the size of the class interval, with the intervals being adjacent, non-overlapping, and typically of equal width; the height of the bar is related to the number of observations falling in the corresponding interval. A boxplot is a visual representation of five key points in a distribution: the minimum, 25th percentile, median, 75th percentile, and maximum. The "box" highlights the interquartile range (IQR), that is, the area between the 25th and 75th percentiles, with a line through the box marking the median. The data outside the box can be presented in a number of ways to highlight how far away extreme values are from the bulk of the data, in particular outliers.

We will use Airbnb data for Washington, DC compiled on October 3, 2015 by Inside Airbnb (http://insideairbnb.com) to illustrate these visitations (Figure 6.5). Airbnb is a marketplace that connects people with short-term accommodations to rent with those looking for a place. Figure 6.6 presents histograms of price data for rentals costing up to $500 per night. The figure highlights the effect of different class interval or bin size choices. The representation of the distribution becomes smoother as the bar width goes from $5 in Figure 6.6A to $50 in Figure 6.6D. We can observe in Figures 6.6A and 6.6B that "hosts" tend to price their rentals around benchmark values such as $100, $200, $300, etc. This feature of the distribution nearly disappears with a bin width of $20 (Figure 6.6C) and is entirely gone with a width of $50. Choosing an interval size to maximize the interpretive value of a histogram can be a matter of trial and error.

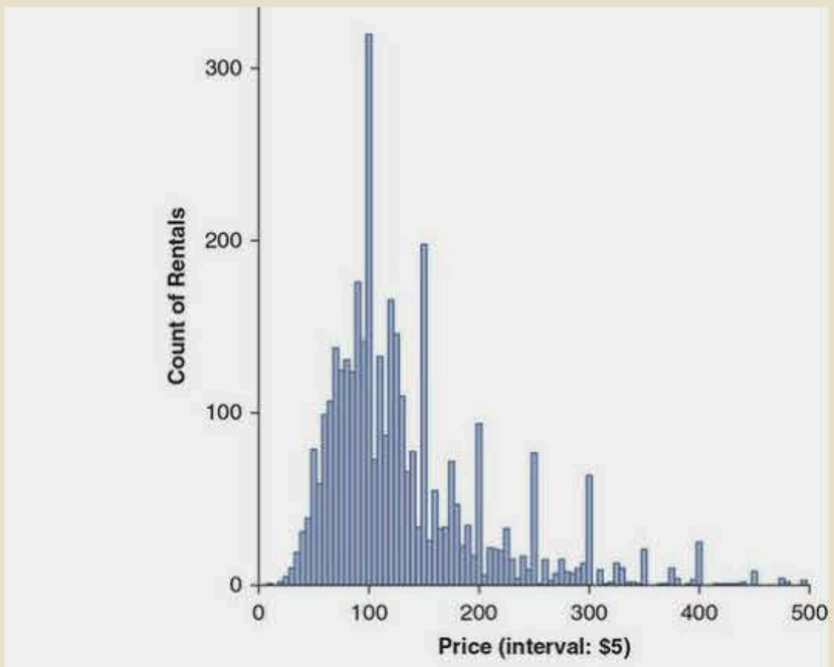
Figure 6.5 Airbnb locations in Washington, DC



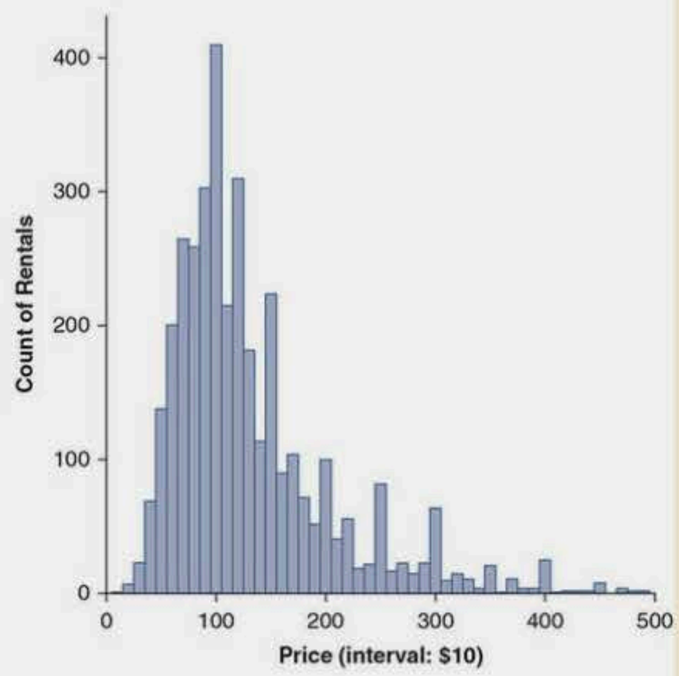
Source: Authors' own

A boxplot presents similar information to the histogram but in a more compact form, and is therefore conducive to comparisons across different subsets of the data. Figure 6.7 splits the Airbnb rentals into three types: entire house or apartment, private room in an otherwise occupied home, and a shared room. The figure shows at a glance a number of features of these distributions: the narrow boxes for private rooms and shared rooms show that most of these rentals occupy a relatively narrow price range when compared to entire homes; median price declines as privacy decreases; approximately 75 percent of entire homes are priced higher than 75 percent of private rooms (indicated by their boxes not overlapping). The vertical lines in the plot, commonly called "whiskers," stretch from the box edge to the furthest point that is within 1.5 times the IQR. Values beyond the whiskers are considered outliers and are represented by points.

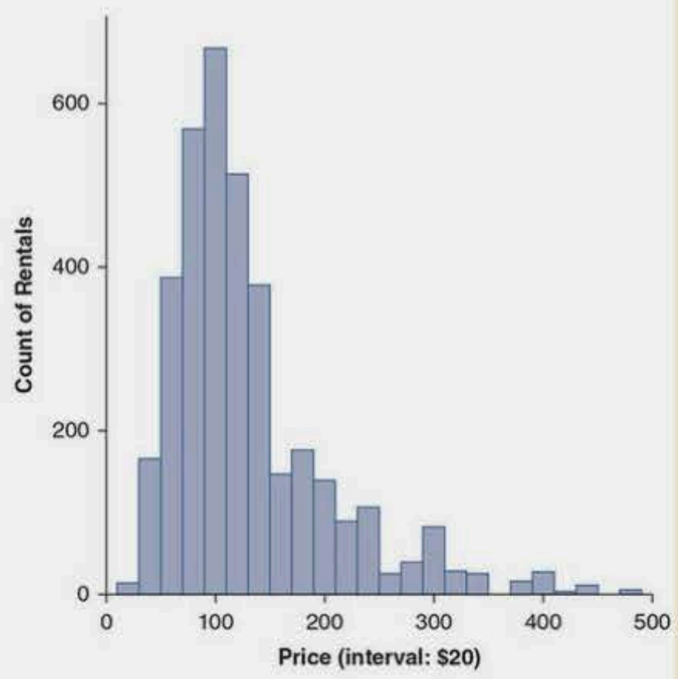
Figure 6.6 Histograms of Airbnb prices in Washington, DC



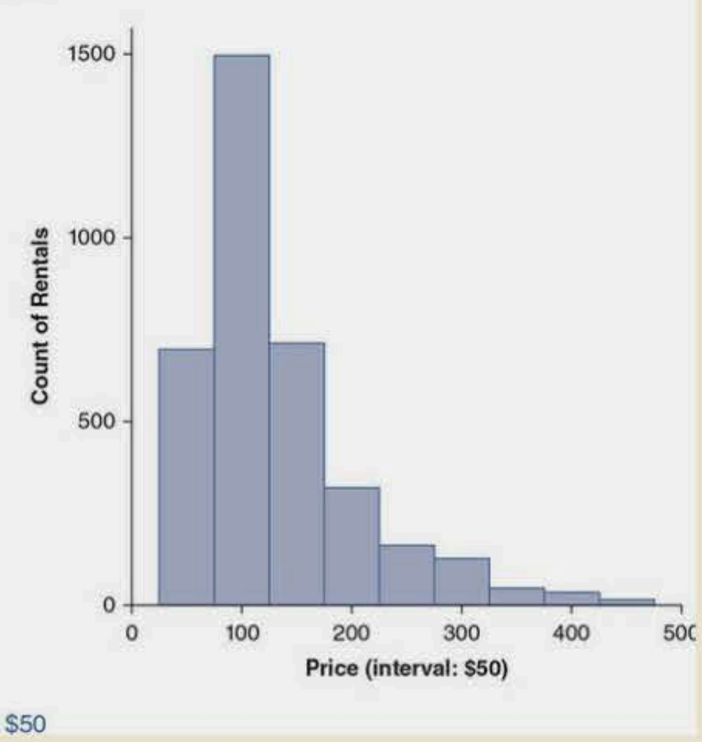
A Interval $5



B Interval $10



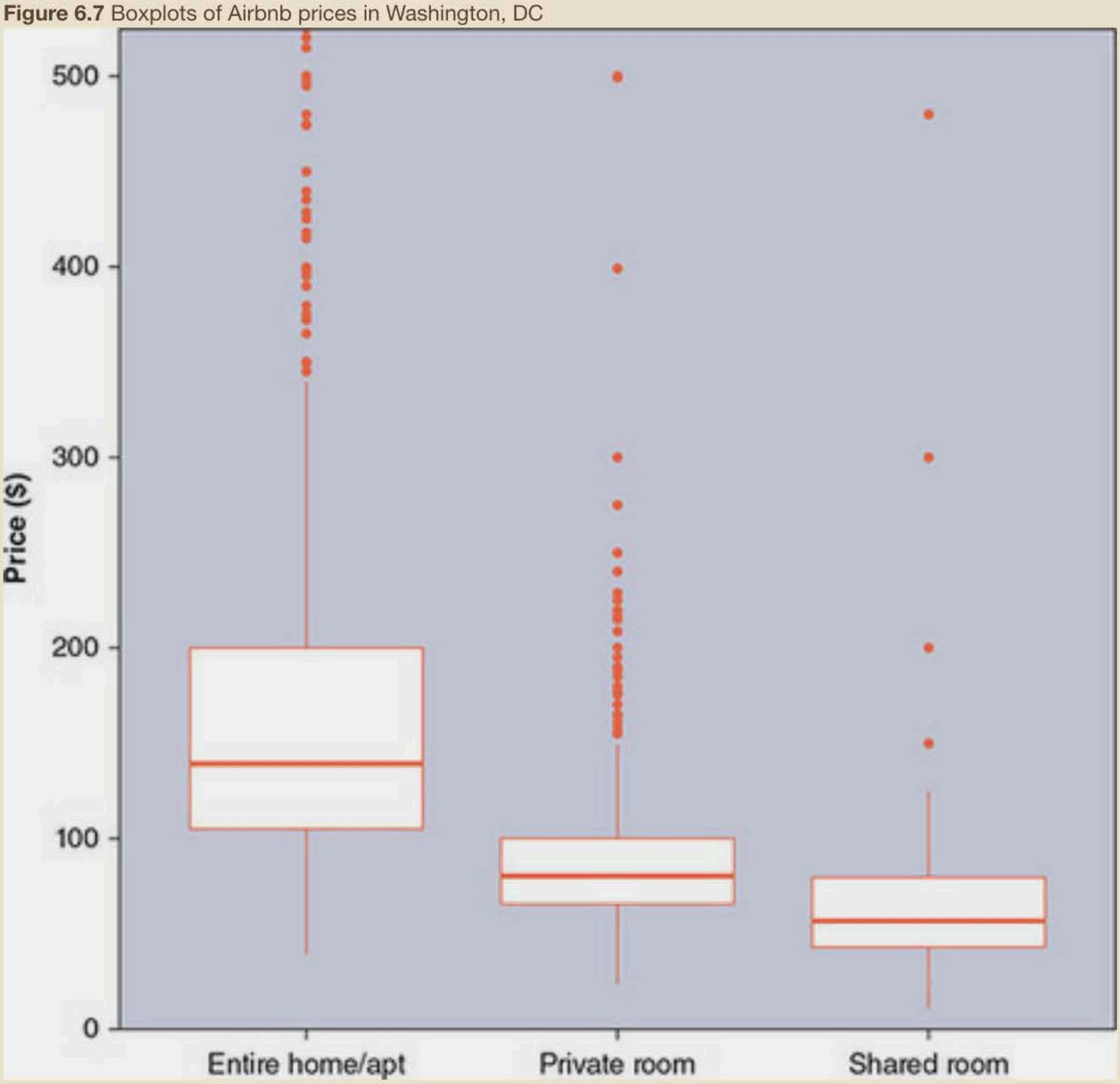
C Interval $20



D Interval $50

Source: Authors' own

Figure 6.7 Boxplots of Airbnb prices in Washington, DC



Source: Authors' own

The observant reader might have noticed that the histograms and boxplots were limited to rentals up to $500, and wondered about additional outliers. There are 69 rentals (out of 3,654 total) priced above $500, with the most expensive priced at $2,822. Initial plots of all the data revealed this handful of extreme values, but also confined the vast majority of the data to a small portion of the plot. Restricting the range of the *X*-axis in the histogram and *Y*-axis in the boxplot to $0 to $500 allowed the patterns of the remaining data to be visible. Note, though, that the outliers were not removed from the computation of the distributions. The choice to truncate the plots at $500 is somewhat arbitrary, but again is based on the goal of maximizing the visual information conveyed by the plot.

There are many other EDA techniques available for interrogating your data. These two visualizations, and the summary statistics that they contain, are examples of those most commonly used.

Spatial and Non-spatial Patterns

In urban analytics maps are one of the most powerful forms of description (Chapter 4), and simply showing spatial patterns in data can be enormously useful. However, quantifying those patterns is more difficult. Firstly, we consider how we can explain patterns in aspatial data, then we will extend this approach to deal explicitly with geography.

Correlation: Patterns in Data

The most widely known way to "explain" a pattern in a dataset is correlation. Stated simply, correlation measures the extent to which attributes are related. For example, we are used to hearing things like, "there is a high correlation between education and income". In this section we will explain how to measure correlation. There are three steps: (1) measure the average deviation from the mean (standard deviation); (2) compute the *z*-score, which tells how far away a given observation is from the mean; and (3) take the average of the product of the *z*-scores.

We already discussed means (also called "averages") earlier in this chapter. Given a list of integers *x* from 1 to 5, the mean would be (1 + 2 + 3 + 4 + 5)/5 = 3; we denote the mean of *x* with $\overline{x}$. Once we know how to compute the mean we can measure how far away each observation is from the mean. If we use the letter *i* to refer to an individual observation we could measure how far the *i*th observation is from the mean $x_i - \overline{x}$. The average of all of the deviations from the mean will be zero, by definition. However, we might want a measure that tells us, on average, how far away a data point is from the mean. To do this we compute a statistic called the standard deviation (s.d.), which is calculated by taking the square of the deviations from the mean, averaging these squared deviations, and taking the square root. In mathematical notation:

$$\text{s.d.} = \sqrt{\left(x_i - \overline{x}\right)^2 / n}$$

Once we have the s.d. we can compute how far x_i is from the mean; we do this in terms of standard deviations. So if x_i is zero standard deviations away from the mean it equals the average. If it is one standard deviation away from the mean it equals $x_i = \overline{x} + \text{s.d.}$; if it is two standard deviations below the mean, $x_i = \overline{x} - \left(2 \times \text{s.d.}\right)$. When we express a number in terms of how many standard deviations it is from the mean we call it a *z*-score. To calculate the *z*-score we divide the deviation from the mean $x_i - \overline{x}$ by the s.d.: that is, $z_i = \left(x_i - \overline{x}\right) / \text{s.d.}$ Although complex at first sight, these are built up from components that are likely quite familiar.

So if we have a neighborhood with a crime rate that has a *z*-score of 2 we immediately know it is far above the average; a neighborhood with a crime rate *z*-score of –0.5 is one with a crime rate slightly below the average. Here is why *z*-scores are useful. Let us imagine we have a list of *n* neighborhoods; if in places where we saw a higher than average crime rate (i.e., observations with a positive *z*-score) we saw a lower than average number of police on patrol (i.e., a negative *z*-score) we could say that there was a negative correlation between crime and police patrols. The data indicates that more police patrols are associated with less crime, or that more crime is associated with less police (remember that correlation does not equal causation!). Just because we see fewer police on patrol in places with more crime we cannot say the lack of police *causes* the crime.

We can use the *z*-score to compute the correlation. The correlation statistic *r* is just the average of the product of the *z*-scores. If the average of the product of the *z*-scores is positive it would mean the places that are above average on variable 2 tend to be above the average on variable 1. At the risk of being too pedantic, remember that a positive number times a positive number equals a positive number; the same is true of the product of negative numbers. For example, high crime and high policing or low crime and low policing would equal a positive correlation. Conversely, if low policing is associated with high crime the correlation would be negative. Written formally:

$$r = \left(z_1 * z_2\right) / n$$

When you compute the correlation coefficient you get a single number that describes the entire city. You might have reason to think that the relationship between variables is not the same everywhere in the city. For example, maybe suburban areas have a different pattern than urban areas.

Spatial Proximity

The spatial proximity that drives many of the most important relationships within cities can be problematic for statistical models, and is a widely researched issue within geography and other spatial disciplines (Profile 6.1). A key assumption in most statistics is that the observations are independent from one another. This means that the magnitude of a variable at one location should not be correlated with the magnitude at nearby locations, a phenomenon called *spatial autocorrelation*.

Spatial autocorrelation is quite common in an urban context. When a person maintains their house well, it has a positive impact on their home's value; in addition, it has a spillover effect by positively impacting their neighbor's home value since it improves the overall aesthetic of the block. Similarly, the values of all the homes in the catchment area of a good elementary school benefit from the school's success.

The types of spatial autocorrelation can be grouped into *substantive* and *nuisance*. Substantive relationships are those deriving from the interdependence of actors and activities in nearby locations. These are relationships for which there is some social or economic theory that might explain the spatial relationship; for example, competition between retail locations, constraints on municipal resources, desire to live near similar people (i.e., segregation), etc. Nuisance spatial autocorrelation is grounded in data problems stemming from a mismatch between the actual spatial footprint of the phenomenon being studied and the available data. Urban data available for enumeration areas defined by the government will not likely match the pattern of, say, gentrification within the city. Similarly, air pollution data are derived from monitoring stations scattered throughout the city; such raw data needs to be spatially interpolated to provide a predicted pollution level at other locations. This mismatch can induce non-substantive spatial patterns in the actual data being used for modeling. In either case, the issue is that there is less unique information than appears on the surface since autocorrelated observations share information. When the values of two nearby homes are both positively impacted by the same high-quality school, their values are at least partially based on some shared information. When a good school's catchment area is split among multiple enumeration areas, those enumeration areas are all benefiting from the same school's success.

Testing for spatial autocorrelation and incorporating it into a model requires a formal specification of the spatial structure. This is done through a *spatial weights matrix* and *spatial lag*. The spatial weights matrix is an n ™ n square matrix, designated as *W*, where each element, w_{ij}, describes the strength of the spatial relationship between each of the *n* spatial observations. The weights matrix embodies the analyst's knowledge and assumptions about the study area and data within the model; it is defined exogenously from the data. Typically, most pairs of observations are assumed to have a relationship of zero, indicating that they are too far apart to have any meaningful interaction. We will consider three types of weights matrices: *contiguity*, *distance* and *nearest neighbors*.

A contiguity weights matrix is useful when the observations are polygons. It defines two polygons as *neighbors* if they touch. This concept can be further refined as *rook* or *queen* neighbors, a naming convention based on chess pieces. Two polygons can be considered queen neighbors if they share either an edge or a single point; the queen can move in all directions on the chessboard. In contrast, rook neighbors are only those polygons that share an edge; the rook can only move up, down, left, and right. The contiguity matrix is binary: a one is entered in the weights matrix if two polygons are neighbors, and a zero otherwise. A distance weights matrix considers the distance between observations when defining the spatial structure. It works well with point data since precise distances can be captured. The weights matrix captures the strength of the relationships, so two points that are close in space should have a larger weight than those further apart, typically going to zero at some cutoff point. A straightforward way to implement this relationship is by using the inverse distance between the two points, but there are other approaches. The nearest neighbors approach defines the *k* closest points to an observation as its neighbors, where *k* is an integer defined by the analyst. The nearest neighbor matrix is also binary. When the density of points varies greatly around the map, the nearest neighbor approach inherently adapts to that variability by always designating a fixed number of observations as neighbors. While these are the most common spatial weights matrices, the key goal is that the weights provide a reasonable approximation of the spatial structure relative to the study area and the research question.

The spatial weights matrix formalizes the neighborhood around each observation. A spatial lag is created by multiplying the weights matrix by a variable measured for each observation. A common convention is to row-standardize the weights matrix by summing the values in each row and then dividing the values in that row by the sum. The result is that each row sums to one, and when the row-standardized matrix is multiplied by a variable the resulting spatial lag gives the average value of that variable in the neighborhood around each observation.

Profile 6.1: Professor Luc Anselin

Figure 6.8 Luc Anselin, Stein–Freiler Distinguished Service Professor of Sociology and College Director, Center for Spatial Data Science Senior Fellow, NORC, University of Chicago

Source: Luc Anselin

Briefly describe your research interests

I am interested in methods to analyze data where the spatial aspect (location, distance, interaction) is central. This ranges from techniques for geovisualization and cluster detection to formal spatial econometric methods and models. I am a firm believer in turning these methods into open-source and accessible software, such as GeoDa. Substantively, most of my recent work has been dealing with public health issues, neighborhood transition and urban housing markets.

What do you see as the relationship between "traditional" analysis methods and those used in data science?

I see data science and spatial data science as a repackaging of many existing insights, but with a more effective integration of ideas from both statistics and computer science. In many ways, this is what I have been doing in my own research for many years, so I wholeheartedly embraced the data science label. I never liked how the traditional statistical (and spatial statistical) texts avoided the messiness of dealing with real data and issues of computational efficiency. The data cleaning and manipulation by itself reportedly takes up 80–90 percent of the effort in a typical empirical analysis. Moving towards making those aspects more efficient and automatic is an important part of data science that was largely ignored in the traditional analysis methods.

Can data science methods benefit from a more sophisticated consideration of geography?

I believe that data science and especially the aspects inspired by machine learning and data mining can benefit a lot from a more explicit consideration of spatial characteristics and spatial effects. In part this of course reflects my own personal bias, since for most of my career I have worked towards demonstrating how incorporating spatial effects in standard statistical techniques affects the properties of the methods as well as the results. I see a similar potential in a more explicit treatment of spatial effects in the evolving data science methods, rather than treating locational coordinates as just an additional feature. This is where I see a potentially important contribution for a true spatial data science, rather than just data science applied to spatial data, similar to the evolution of ESDA (Exploratory Spatial Data Analysis) out of EDA for geographic data.

How do you see new forms of data affecting our understanding of cities?

I like to use the term “new data” rather than Big Data to characterize the emerging streams of information about many characteristics of our cities, such as provided by real-time sensors, open data portals, volunteered information, social media, etc. These data tend to be geo-coded and time stamped, which allows a much greater flexibility in the treatment of scale, both geographical as well as temporal. Working up from micro-geographic data allows for new insights into the dynamics over space and time, such as provided by detailed commuter flow data, the measurement of air quality, crime, house prices, and rental rates. However, these new data also represent interesting challenges pertaining to their representativeness, issues of scale, and other aspects of data quality, and lead to new tensions between the need for public information and privacy. While I see much of the new information currently employed to improve urban operations, there is also great potential to gain a better quantitative understanding of the pulse of the city and the space–time dynamics that operate simultaneously at multiple scales.

Explanation

Regression is a statistical framework that allows us to predict how the mean of a *dependent variable* changes as an *independent variable* changes. This simple premise spawns many different specific models based on the characteristics of the data and the goals of the analysis. We will consider a few of these models that are particularly relevant to urban analytics.

Ordinary least squares (OLS) is one of the most accessible methods for creating regression models. Regression models are useful when you have reason to believe that the mean of one variable is dependent on some other variable or variables. For example, we could compute the average weight of everyone in California. However, if we know that a person’s weight is probably dependent on their height, a person who was 6′ 5″ and weighed 110 lb (1.96 m and 50 kg) would be much more surprising than someone who was 5′ tall (1.52 m) and weighed the same amount. In regression language we believe that the average weight of a person is dependent on their height. But how dependent? How strong is the relationship between height and weight? OLS regression provides a means to answer this and similar questions. Regression models require the use of some mathematical notation. We have the thing we want to explain as y; X is the data that we will use to help us understand y; and β describes the strength of the relationship between y and X. We have many observations so we use the subscript i to refer to a single observation. If $i = 5$ and y referred to neighborhood crime levels, y_5 would refer to the fifth neighborhood observation in our table. The letter i is used as an index for our observations.

Finally, we have more than one potential explanatory variable x, so we use a numeric subscript to indicate a specific *independent* or *explanatory* variable, such that x_{i1} might mean the number of police patrolling a neighborhood and x_{i2} might be the percentage of the neighborhood’s population that is under 18.

So how do we use this notation? Practically, we start with some independent variables, x_1 and x_2 in this case, that we have reason to believe have meaningful relationships to the dependent variable, y. Multiple observations ($i = 1,2,3,\ldots,n$) are needed for each of these variables, where n is the total number of observations in the dataset. The dependent variable might be the percentage of workers who commute by bike, with the independent variables being the total miles of bike lanes and number of days with measurable precipitation, and the n observations being all cities in the United States with more than 100,000 residents.

These data are used to estimate the parameters in the model, β_0, β_1, and β_2.The parameters in OLS are global in the sense that they do not change with each observation (notice that they do not have a subscript i). The magnitude of a parameter, in absolute value, describes the strength of the relationship between the independent and dependent variables, and the sign, either positive or negative, describes the direction of the relationship. In this example, one might expect that more bike lines will correspond to more bike commuters (positive relationship), while more inclement days will correspond to fewer bike commuters (negative relationship). This is all put together in Equation (1):

(1)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \tag{1}$$

The last term in the equation, ε_i, represents the variation in the dependent variable that is not explained by the independent variables. Our simple model of bike commuting with just two independent variables likely misses many factors that contribute to residents’ decisions to use this mode of transport to get to work, the implication being that ε_i will likely be large in this example.

While the math needed to compute the parameters is beyond this high-level overview of regression concepts, Equation (1) can be interpreted as any other equation. Before running the regression, we know y, x_1, and x_2; we use regression to estimate β_0, β_1, and β_2 based on the known data; finally, once we have the parameters we can plug everything back into Equation (1) and solve for the remaining term ε_i.

In effect, ε_i is the difference between the known value of y_i and the modeled value for y_i, which is $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. Due to the math underlying regression, in the average value of ε being zero, but particular values of ε_i can vary widely across the different observations. If ε_i is small, then the model does a good job of estimating the dependent variable for observation i.

Health insurance coverage is a politically contentious issue in many countries as not all people have the same coverage rate; we can use regression to help understand these patterns. In the United States health insurance is typically provided by a person's employer during working age years and then by the federal government after retirement. Employer health programs often also allow employees to add their spouses and dependent children to their plan for an additional cost. People who do not fit into any of the previous categories are left to purchase health insurance directly from insurance companies, seek out government assistance, or go without insurance. Using data from the ACS, we can investigate the characteristics of places that tend to have higher insurance coverage.

Table 6.2 presents OLS regression results for Census Tracts in Phoenix, Arizona. Percent insured is the dependent variable and the independent variables are percent unemployed (pct_unemp), percent 65 and older (pct_65over), percent married (pct_mar), percent white (pct_white) and percent Hispanic (pct_hisp). The latter two variables are included to account for racial disparities often present in the United States. The results indicate that, as expected, places with higher unemployment have lower rates of insured residents (indicated by the negative sign on the parameter) and places with more married people have more insured people. These results are *statistically significant* (indicated by the stars on the coefficients), meaning that the reported relationship holds in most cases. In contrast, having more people 65 and older is only marginally significant (p-value = 0.08), meaning that the relationship often holds, but not enough to consider it a strong relationship between the dependent and independent variables. The percent Hispanic in a neighborhood shows a strong negative relationship to the insured rate in Phoenix. Hispanic residents represent 30 percent of the population in the region. Having a larger white population has no relationship to the insured rate. It is important to remember that the pairwise relationships presented in the table reflect a situation where we are controlling for the variation in the other variables in the model. Therefore, while largely white neighborhoods may have lower unemployment rates, the regression framework essentially isolates the impact of white from the other variables in the model. The R^2 statistic in the table indicates that the variables selected explain 72.82 percent of the variation in the insured rate across Phoenix neighborhoods.

Table 6.2 Ordinary least squares (OLS) results for health insurance coverage rate in Phoenix

	Estimate		*p*-value
(Intercept)	0.82627	***	< 2.2E-16
pct_unemp	−0.29705	***	1.25E-14
pct_mar	0.21831	***	< 2.2E-16
pct_65over	0.02385	·	0.0801
pct_white	0.01144		0.5752
pct_hisp	−0.23333	***	< 2.2E-16

Significance codes: 0.001: ***; 0.01: **; 0.05: *; 0.1: ·

Source: Authors' own. Data from Minnesota Population Center: National Historical Geographic Information System (www.nhgis.org)

Significance codes: 0.001: ***; 0.01: **; 0.05: *; 0.1: ·
Source: Authors' own. Data from Minnesota Population Center: National Historical Geographic Information System (www.nhgis.org)

Spatial Regression

Just as correlations might vary spatially, space might have an impact on regression coefficients. Spatial regression is a technique for accounting for these relationships. When using regression on observations located in space, the potential for spatial autocorrelation is always present. When it is, it can cause imprecision and bias in the parameter estimates. The imprecision results in overconfidence about the model results, that is, the statistical significance of the parameters is overstated. This can result in a statistically unimportant variable being deemed important. A biased parameter estimate is actually incorrect. To correct for these problems, spatial autocorrelation can be incorporated into a regression model through the dependent variable, y, or through the error term ε. The former is called a *spatial lag model* and the latter a *spatial error model*.

The spatial lag model is appropriate in the case of substantive spatial autocorrelation. The spatial autocorrelation is introduced as a spatial lag on the dependent variable, Wy. The full regression model is as follows:

(2)

$$\gamma_i = \rho \sum_j w_{ij} \gamma_j + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad (2)$$

where the spatial lag is added as another regressor in the regression model introduced earlier in Equation (1). The parameter ρ is estimated in the model and gives the strength of the spatial autocorrelation.

In the spatial error model, the spatial autocorrelation is operationalized through a spatial lag on the error term, *Wε*. In effect, the error term from Equation (1) is modeled with the equation:

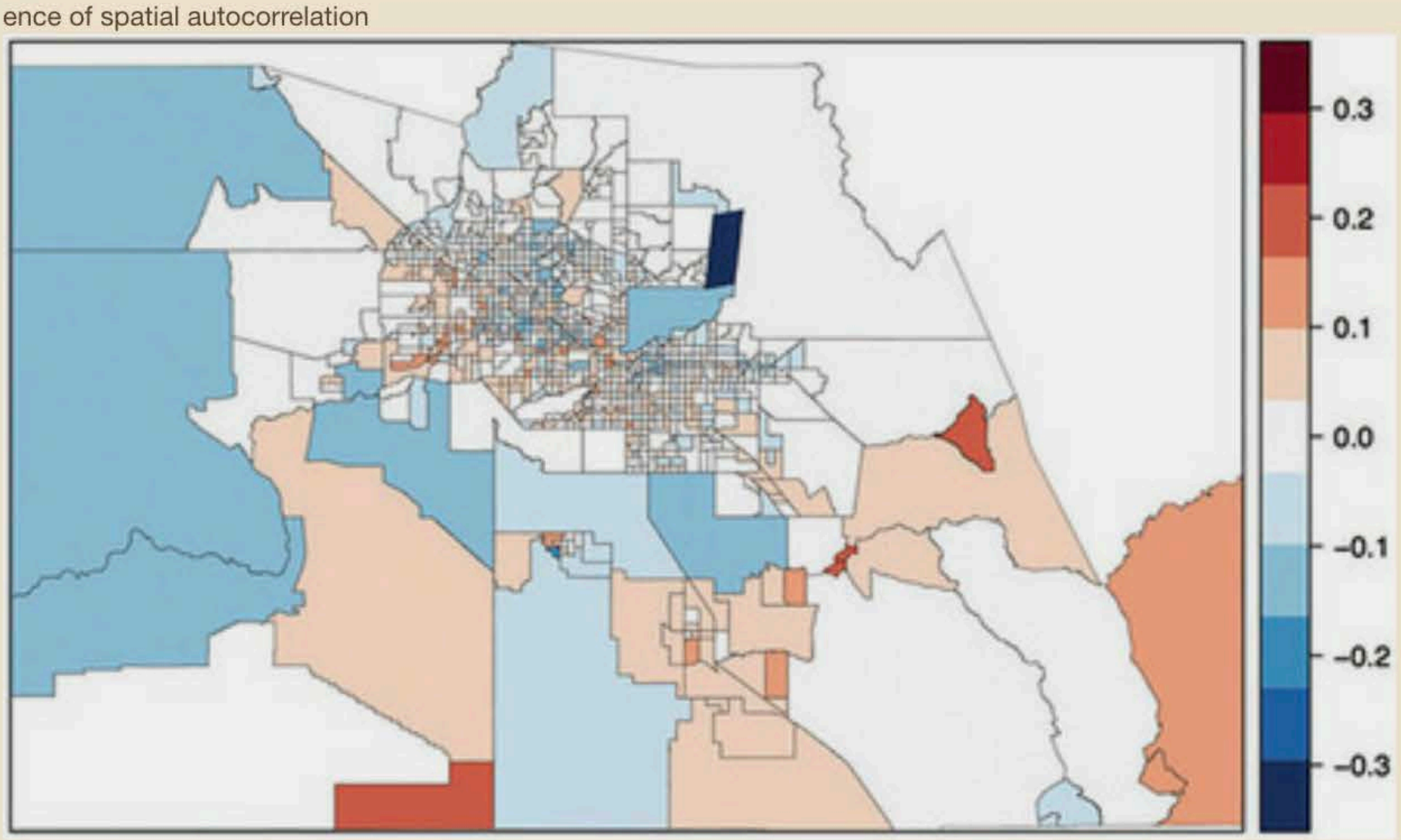
(3)

$$\varepsilon_i = \lambda \sum_j w_{ij} \varepsilon_j + u_i \quad (3)$$

where u_i is now the unexplained error in the model and λ describes the strength of the autocorrelation. This model is used when the autocorrelation is of the nuisance variety.

Looking back to the earlier example, we might expect there to be some nuisance spatial autocorrelation since Census Tracts likely do not align exactly with underlying characteristics of the population in connection with health insurance coverage. A map of the residuals (Figure 6.9) from the OLS model shows that many neighboring tracts have similar error. Table 6.3 shows the results for a spatial error model. The model results now show the spatial error parameter λ(lambda), which is statistically significant. The table also shows that percent white (pct_white) now has a statistically significant effect on insurance coverage: a higher white population corresponds to higher insurance coverage, holding the other variables in the model constant.

Figure 6.9 Map of OLS regression residuals for Phoenix health insurance coverage; clustering of similar values indicates the presence of spatial autocorrelation



Source: Authors' own

Table 6.3 Spatial error model results for health insurance coverage rate in Phoenix

	Estimate		*p*-value
(Intercept)	0.790672	***	< 2.2E-16
pct_unemp	–0.289507	***	4.44E-16
pct_mar	0.165052	***	< 2.2E-16
pct_65over	0.027081	.	0.07775
pct_white	0.085936	***	0.00003
pct_hisp	–0.171501	***	< 2.2E-16
Lambda	0.52429	***	< 2.2E-16

Significance codes: 0.001: ***; 0.01: **; 0.05: *' 0.1: ·

Source: Authors' own. Data from Minnesota Population Center: National Historical Geographic Information System (http://www.nhgis.org)

Significance codes: 0.001: ***; 0.01: **; 0.05: *' 0.1: ·
Source: Authors' own. Data from Minnesota Population Center: National Historical Geographic Information System (http://www.nhgis.org)

Explaining Cities

The city can be explained using many different approaches. Underlying all these approaches is the most important part of the modeling process: having a clear understanding of the research goal. Social science theory and EDA can help guide the modeling process, but these do not provide the substantive motivation for the work. Today, much of the data and modeling burden has been picked up by sophisticated software and hardware, but the urban analyst still drives the process from initial development to presentation of final results.

Questions

1. Consider urban trash collection. Provide descriptive, predictive, and explanatory research questions relating to this messy topic.
2. Some reading on exploratory data analysis will quickly reveal that this research framework is not embraced by all analysts. Write a 2,000-word essay explaining the arguments for and against this approach to research.
3. Correlation can be either negative or positive. Provide two examples of positive correlation and two of negative.
4. What is the difference between substantive and nuisance spatial dependence? Give an example of each.

Supplementary Reading

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*(3), 247–267.

This succinct article is one of the best summaries of how spatial data affects regression models. While it assumes some understanding of regression, it skips the formal proofs to explain spatial econometrics in plain language.

Greene, W. H. (2012). *Econometric analysis*. Harlow: Pearson Education.

This book is a comprehensive and detailed presentation of regression from an economics perspective. Although quite technical, it does a good job of linking various regression techniques to their statistical grounding and provides empirical examples.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

This extremely readable journal article disentangles the concepts of predictive and explanatory modeling.

Tufte, E. R. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

This classic book is an amazing compendium on the visual display of statistical data. It contains hundreds of graphics detailing strategies for presentation and analysis of information.