12-2021

Teaching Data Analytics Using Real Estate Data

Ting-Ting (Rachel) Chung

Teaching Data Analysis with Real Estate Data

Teaching Case

Tingting (Rachel) Chung
William & Mary
rachel.chung@mason.wm.edu

Abstract

Data analysis, commonly known as statistics, can be intimidating due to its jargon, symbols, and obscure terminologies. This teaching case demonstrates how to make these abstract and technical concepts much more concrete and approachable by having students participate in the process of putting together a real estate dataset. Data collected from real estate websites are excellent for teaching data analysis, including predictive modeling, because real estate data are easy to understand, produce strong and predictable results, and demonstrate core concepts of predictive modeling and tangible applications. This case provides an example data collection survey, in-class exercises, and templates for calculating statistics from the real estate dataset.

Keywords

Teaching data analysis; statistics; predictive modeling; real estate data

Introduction

Data analysis, commonly known as statistics, is an increasingly important skill for most professions. It is, however, also an academic topic about which many people feel ambivalent. They are told it is important, and many are required to take a statistics course. The good students push themselves through a dreadful collection of Greek letters, memorize an array of formulas, and pass an exam. They may have even earned an A. How much they actually remember a few years later, not to mention how much they can actually apply their statistical knowledge personally or professionally, is another story.

Statistics is a branch of mathematics that derives its beauty from elegant and concise mathematical equations. However, these abstract representations can also be so terse that most people have trouble decoding them, if the concepts are presented as equations and symbols only. Is deviation the same as standard deviation $\boldsymbol{\sigma}$? How is standard deviation different from standard score Z? What about standard error? Is that a special type of error $\boldsymbol{\varepsilon}$? Isn't error also called residual? Why does everything also have a Greek letter name? Most importantly, what is the point of standard deviation? Do businesses actually make use standard deviation to make money? As a result, much of the effort in learning statistics may be spent on decoding and applying mathematical equations and very little on its practical use. This is unfortunate because statistics is such a practical tool used in so many ways. For example, quality control in manufacturing, Netflix recommendations, and Google ads are all made possible by statistical analyses.

Following recommendations by Gelman and Nolan (2017), I have designed classroom activities using real estate data from Zillow to illustrate core concepts of statistics. These activities have been quite effective for foundational Data Analysis courses in both Master in Business Administration (MBA) and Master of Science in Business Analytics (MSBA) programs. Most of these graduate students have taken some form of statistics course at the undergraduate or high school level. However, they usually have retained very little and struggle to articulate the definition or use cases of basic concepts such as standard deviation. Therefore, I have assumed no prior prerequisite knowledge when designing these activities. These students have been introduced to probability concepts prior to taking my courses. Activities described below are what I have designed primarily for a two-credit MBA Data Analysis course that meets three times a week over seven calendar weeks. Weekly topics are listed in Table 1.

Week	Topic	Week	Topic
1	Describe Single Variables • Central tendency • Dispersion • Distributions • Visualization	5	Make A/B Comparisons • T Test • ANOVA • Chi Square
2	Use a Sample to Make Inferences • Sampling • Statistical Inferences • Confidence Intervals	6	Advanced Topics • Regression Diagnostics • Logistic Regression
3	Relate Two Variables • Correlation • Simple Regression	7	Final Team Projects
4	Relate Two Variables • Correlation vs. Causation • Multiple Regression		

Table 1. Weekly Topics

Method

The first assignment in my MBA Data Analysis[1] and MSBA Machine Learning courses is always for each student to go house shopping by choosing a real estate property in the local area. Students enter data points about their chosen properties into a survey, which is available in Appendix 1. All of the student entries are compiled into a property dataset that we can then use for class discussions and exercises. Table 2 shows sample records from the most recent dataset[2]. The survey questions include both categorical variables (e.g., ZIP code) and numeric variables (e.g. Price). Price, like most financial data, is always skewed, which gives us the opportunity to discuss transformation. Therefore, we can use the same dataset to run a wide range of analyses by combining different categorical and numeric variables. We would spend the next month or so calculating numbers and analyzing the dataset in different ways. In fact, we can get through an entire semester of the introductory Data Analysis course using the same dataset and never run out of ideas.

FName	ZIP code	Property Type	Year	Parking	Miles	Price	Zestimate	Zillow Days	Taxes	BED	BATH
Alexis	23185	House	1999	4	4	355000	394200	7	2608	3	3
Allan	23188	House	1995	2	5.6	1300000	1283200	11	5007	4	4
Andi Muhammad Farid	23188	House	2017	1	3.8	349500	355300	105	2126	4	4
Andres	23188	House	2002	3	6.2	874000	882500	75	7011	3	4
Andrew	23185	House	1964	2	5.1	398000	428200	2	3087	3	3
Andy	23188	House	2002	2	5.8	899000	887400	4	6715	5	5
Antoine	23185	House	2006	4	3.5	895300	883700	19	4763	4	4

[1] The course materials are available on this site: https://buad5701.blogspot.com/

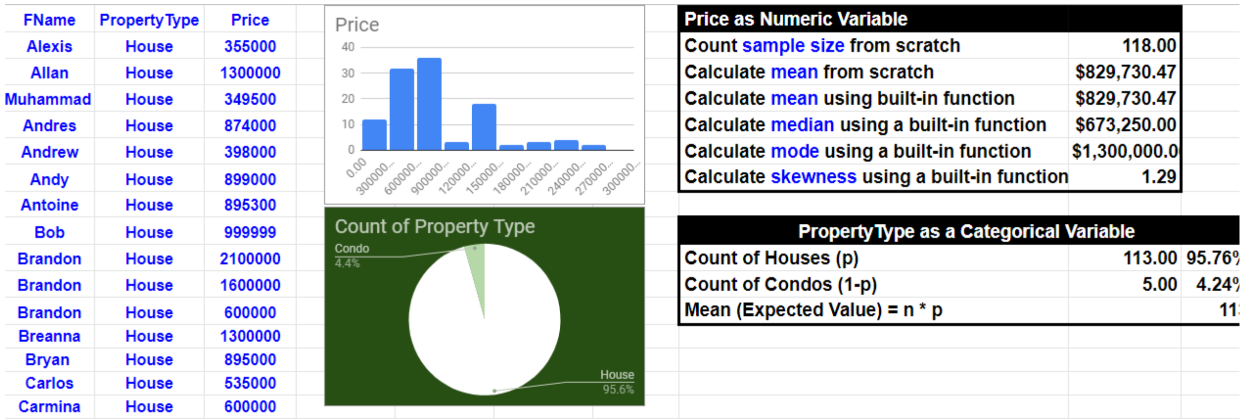
[2] The dataset and the associated analyses are available here: https://docs.google.com/spreadsheets/d/1zvbfHTq0nsCf5gvk0-Y7YidMk9dX1BC_ZgWoE1-VPKc/edit#gid=2029118503

Bob	23188	House	2005	2	6.1	999999	994900	47	7143	4	5
Brandon	23185	House	2015	3	6.8	2100000	2040200	115	12251	4	5

Table 2. Sample Data Records

Student names are always included in the dataset. Seeing their own names associated with data records in the class dataset allows students to make personal connections to data, and to have tangible references to otherwise very abstract concepts, such as deviations and Z scores.

During the first week, students are introduced to basic descriptive statistics of numeric variables. Figure 1 illustrates how calculations of central tendency are set up on a shared Google Sheets file. These calculations using formulas are complemented with an "arts and crafts" exercise. Students are given round-shaped sticky notes, and are instructed to write down their names, raw scores (e.g., property age), deviations, and Z scores (i.e., standard score.) The instructor can demonstrate how to compute these scores on the Google Sheets file, and then students can find their own scores and copy them onto their own "data dots." Questions to ask the class include: Who has a Z score of 1.96? Who is above 3? Below -3? Who has zero scores? Who is inside the 95% confidence interval? Who is outside? The Google Sheets file is accessible to all students (in view only mode) so they can follow along, inspect the formulas and consider how their own records contribute to the class statistics and models.



FName	PropertyType	Price
Alexis	House	355000
Allan	House	1300000
Muhammad	House	349500
Andres	House	874000
Andrew	House	398000
Andy	House	899000
Antoine	House	895300
Bob	House	999999
Brandon	House	2100000
Brandon	House	1600000
Brandon	House	600000
Breanna	House	1300000
Bryan	House	895000
Carlos	House	535000
Carmina	House	600000

Price as Numeric Variable	
Count sample size from scratch	118.00
Calculate mean from scratch	$829,730.47
Calculate mean using built-in function	$829,730.47
Calculate median using a built-in function	$673,250.00
Calculate mode using a built-in function	$1,300,000.0
Calculate skewness using a built-in function	1.29

PropertyType as a Categorical Variable		
Count of Houses (p)	113.00	95.76%
Count of Condos (1-p)	5.00	4.24%
Mean (Expected Value) = n * p		11

Figure 1. Central Tendency on Google Sheets

The instructor can invite the entire class to come up to a whiteboard and to construct a histogram together using their data dots (see Figure 2.) This exercise is dynamic, fun, and participatory. The physical experience of constructing a data dot individually, and then a histogram together as a class, reinforces basic but important statistical concepts discussed above. Students can clearly see where they are located on the distribution, how their positions on the histogram are reflected by their Z scores, and how deviations tell us about their relationships to the class mean. Together, the class can also compute variance, standard deviation, skewness, and other properties that belong to the entire class sample rather than individual records. This is particularly effective in helping students distinguish linguistically confusing terms, such as deviation (of a case) vs. standard deviation (of a sample), and standard score (of a case) vs. standard deviation (of a sample.) It also creates a deep appreciation of the relationship between a case, and a dataset. After the hands-on activity, the instructor can demonstrate how to create the histogram on the computer using a variety of software programs, such as Google Sheets, JASP, and Tableau. Students can then clearly see that the computer programs simply automated the manual process of putting together a histogram.

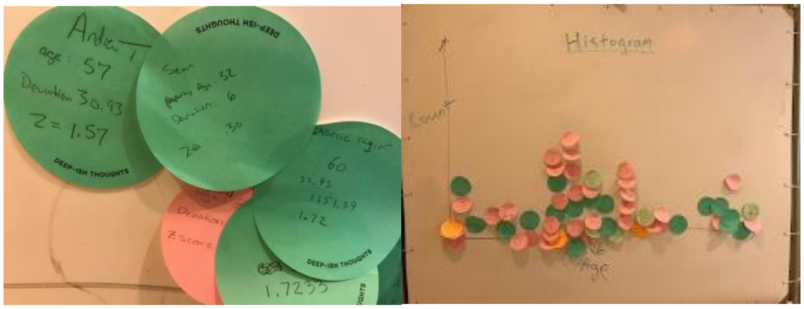


Figure 2. Histogram Exercise

When discussing correlation, the instructor can again ask students to make "data dots," but this time with their own data records for two, instead of one, variables (e.g., number of bedrooms, or "BED", and number of bathrooms, or "BATH.") The students would also write down their own Z scores for these two variables, and multiply them together. The class can then look at the Google Sheets to see how to add all of these scores up together, and divide the total by (sample size -1) to obtain Pearson's Correlation r. Students are invited to put up their data dots on the whiteboard to construct a class scatterplot together. We can then use the scatterplot to discuss regression, residuals, model fit, and other important concepts of predictive modeling (see Figure 3). Again, the instructor can follow up with a demonstration of calculating correlation and creating scatterplots using Google Sheets. See Figure 4 for an example.

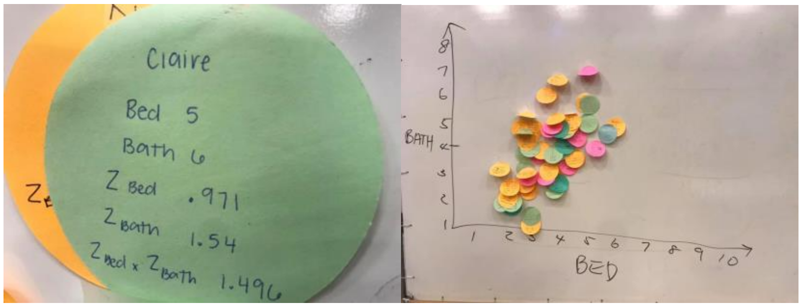
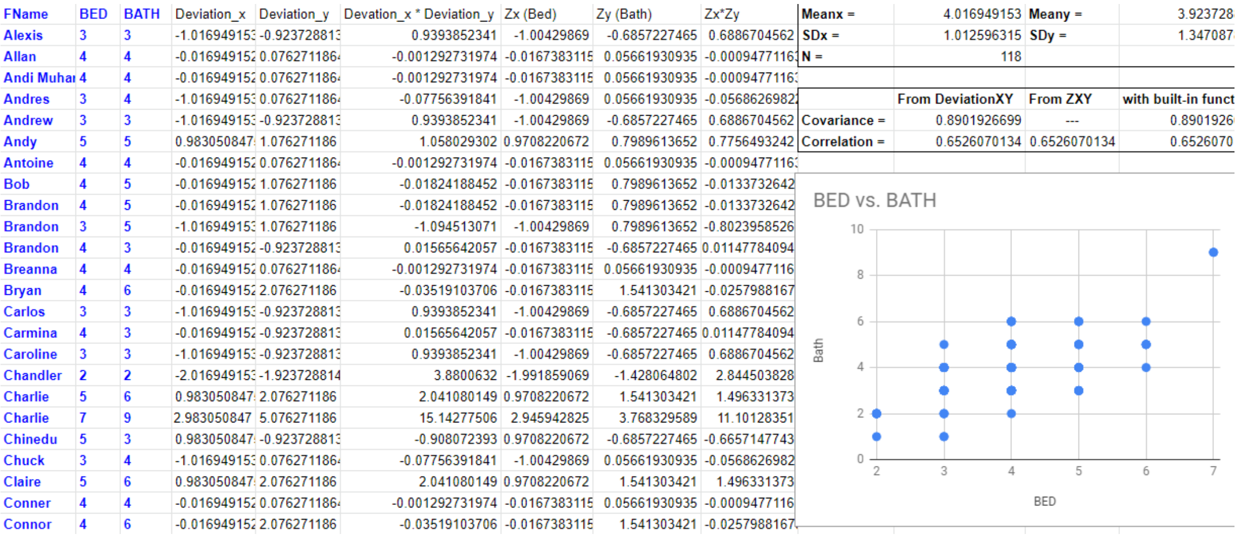


Figure 3. Correlation & Regression Exercise



FName	BED	BATH	Deviation_x	Deviation_y	Devation_x * Deviation_y	Zx (Bed)	Zy (Bath)	Zx*Zy
Alexis	3	3	-1.016949153	-0.923728813	0.9393852341	-1.00429869	-0.6857227465	0.6886704562
Allan	4	4	-0.016949152	0.0762711864	-0.001292731974	-0.0167383115	0.05661930935	-0.0009477116
Andi Muhai	4	4	-0.016949152	0.0762711864	-0.001292731974	-0.0167383115	0.05661930935	-0.0009477116
Andres	3	4	-1.016949153	0.0762711864	-0.07756391841	-1.00429869	0.05661930935	-0.0568626982
Andrew	3	3	-1.016949153	-0.923728813	0.9393852341	-1.00429869	-0.6857227465	0.6886704562
Andy	5	5	0.983050847	1.076271186	1.058029302	0.9708220672	0.7989613652	0.7756493242
Antoine	4	4	-0.016949152	0.0762711864	-0.001292731974	-0.0167383115	0.05661930935	-0.0009477116
Bob	4	5	-0.016949152	1.076271186	-0.01824188452	-0.0167383115	0.7989613652	-0.0133732642
Brandon	4	5	-0.016949152	1.076271186	-0.01824188452	-0.0167383115	0.7989613652	-0.0133732642
Brandon	3	5	-1.016949153	1.076271186	-1.094513071	-1.00429869	0.7989613652	-0.8023958526
Brandon	4	3	-0.016949152	-0.923728813	0.01565642057	-0.0167383115	-0.6857227465	0.01147784094
Breanna	4	4	-0.016949152	0.0762711864	-0.001292731974	-0.0167383115	0.05661930935	-0.0009477116
Bryan	4	6	-0.016949152	2.076271186	-0.03519103706	-0.0167383115	1.541303421	-0.0257988167
Carlos	3	3	-1.016949153	-0.923728813	0.9393852341	-1.00429869	-0.6857227465	0.6886704562
Carmina	4	3	-0.016949152	-0.923728813	0.01565642057	-0.0167383115	-0.6857227465	0.01147784094
Caroline	3	3	-1.016949153	-0.923728813	0.9393852341	-1.00429869	-0.6857227465	0.6886704562
Chandler	2	2	-2.016949153	-1.923728814	3.8800632	-1.991859069	-1.428064802	2.844503828
Charlie	5	6	0.983050847	2.076271186	2.041080149	0.9708220672	1.541303421	1.496331373
Charlie	7	9	2.983050847	5.076271186	15.14277506	2.945942825	3.768329589	11.10128351
Chinedu	5	3	0.983050847	-0.923728813	-0.908072393	0.9708220672	-0.6857227465	-0.6657147743
Chuck	3	4	-1.016949153	0.0762711864	-0.07756391841	-1.00429869	0.05661930935	-0.0568626982
Claire	5	6	0.983050847	2.076271186	2.041080149	0.9708220672	1.541303421	1.496331373
Conner	4	4	-0.016949152	0.0762711864	-0.001292731974	-0.0167383115	0.05661930935	-0.0009477116
Connor	4	6	-0.016949152	2.076271186	-0.03519103706	-0.0167383115	1.541303421	-0.0257988167

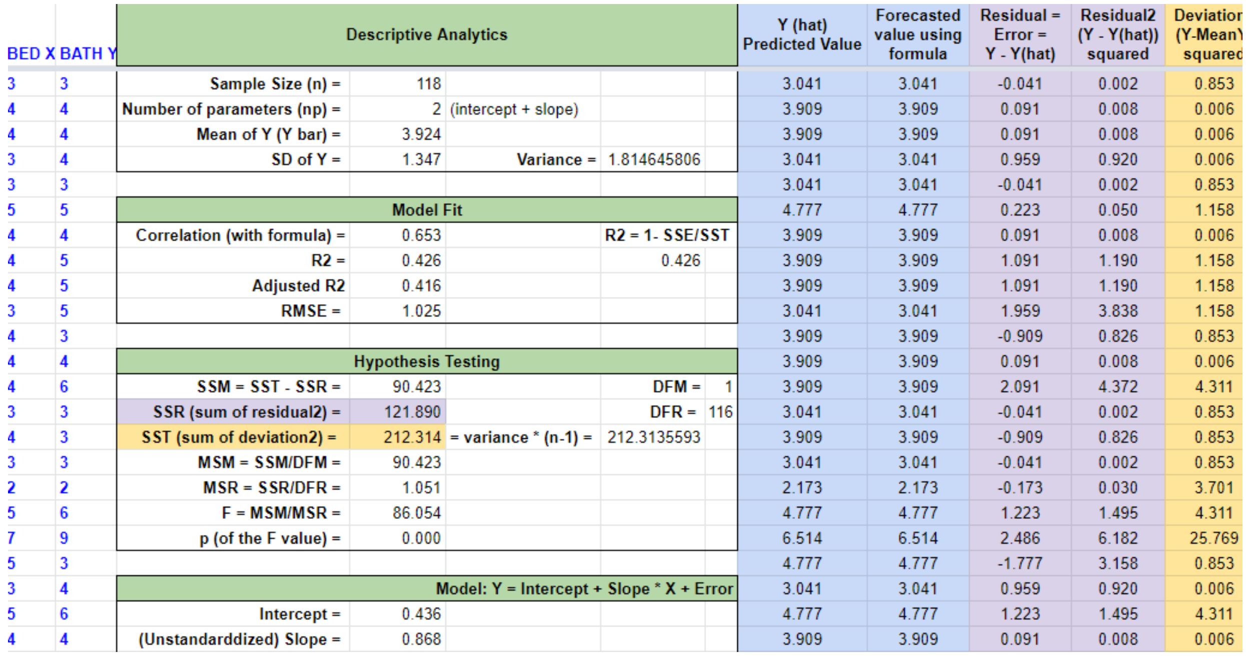
Meanx =	4.016949153	**Meany =**	3.923728
SDx =	1.012596315	**SDy =**	1.347087
N =	118		

	From DeviationXY	From ZXY	with built-in funct
Covariance =	0.8901926699	---	0.8901926
Correlation =	0.6526070134	0.6526070134	0.6526070

Figure 4. Correlation on Google Sheets

The real estate dataset is a good teaching tool for two key reasons. (1) The business domain knowledge is easily accessible to the general population. For example, most people intuitively understand that bigger houses (i.e., a larger square footage) will probably be more expensive, and adding a bathroom is probably going to increase the value of the property. Also, (2) relationships between the variables are usually very strong and predictable. For example, the number of bedrooms is always highly correlated with the number of bathrooms, and so we can always use them to discuss multicollinearity. Each additional bedroom is always going to add a significant chunk of value, and so we can always count on the beta coefficient of the regression model to be significant and positive. Therefore, instructors do not need to worry whether the statistical magic will work again in a new semester or not. The real estate dataset is also excellent for teaching prescriptive analytics. For example, students can determine which ZIP code has higher property values, by running T tests. Students can also determine which property type (house vs. condo) is older.

Most importantly, Zillow uses their dataset to produce Zestimate, a real-world application of predictive modeling, or machine learning predictions, for estimating property valuation (Schneider 2019). Conceptually, Zestimate is similar to the notion of home appraisals, which are intuitive for most people, including our students, to understand. However, few would immediately feel that they know how to create the estimates on their own. When the students realize that they can build their own Zestimate models after having learned predictive modeling, they often get a great sense of accomplishment. The class can build their own Zestimate equations, make their own predictions, and calculate individual students' residuals (i.e., the amount of mispredictions.) See Figure 5 for an example of simple regression model set up on Google Sheets. These models are very easy to interpret. For example, the beta coefficient shows how much the price would increase if the house had an additional bedroom, or how much the price would drop, if the house moves from one ZIP code to another. These simple, vivid, and intuitive exercises make the rather abstract concepts and Greek letters (e.g., **β**) easier to digest, remember, and apply.



BED X	BATH Y	Descriptive Analytics					Y (hat) Predicted Value	Forecasted value using formula	Residual = Error = Y - Y(hat)	Residual2 (Y - Y(hat)) squared	Deviatior (Y-MeanY squarec
3	3	Sample Size (n) =	118				3.041	3.041	-0.041	0.002	0.853
4	4	Number of parameters (np) =	2	(intercept + slope)			3.909	3.909	0.091	0.008	0.006
4	4	Mean of Y (Y bar) =	3.924				3.909	3.909	0.091	0.008	0.006
3	4	SD of Y =	1.347	Variance =	1.814645806		3.041	3.041	0.959	0.920	0.006
3	3						3.041	3.041	-0.041	0.002	0.853
5	5	Model Fit					4.777	4.777	0.223	0.050	1.158
4	4	Correlation (with formula) =	0.653		R2 = 1- SSE/SST		3.909	3.909	0.091	0.008	0.006
4	5	R2 =	0.426		0.426		3.909	3.909	1.091	1.190	1.158
4	5	Adjusted R2	0.416				3.909	3.909	1.091	1.190	1.158
3	5	RMSE =	1.025				3.041	3.041	1.959	3.838	1.158
4	3						3.909	3.909	-0.909	0.826	0.853
4	4	Hypothesis Testing					3.909	3.909	0.091	0.008	0.006
4	6	SSM = SST - SSR =	90.423		DFM =	1	3.909	3.909	2.091	4.372	4.311
3	3	SSR (sum of residual2) =	121.890		DFR =	116	3.041	3.041	-0.041	0.002	0.853
4	3	SST (sum of deviation2) =	212.314	= variance * (n-1) =	212.3135593		3.909	3.909	-0.909	0.826	0.853
3	3	MSM = SSM/DFM =	90.423				3.041	3.041	-0.041	0.002	0.853
2	2	MSR = SSR/DFR =	1.051				2.173	2.173	-0.173	0.030	3.701
5	6	F = MSM/MSR =	86.054				4.777	4.777	1.223	1.495	4.311
7	9	p (of the F value) =	0.000				6.514	6.514	2.486	6.182	25.769
5	3						4.777	4.777	-1.777	3.158	0.853
3	4	Model: Y = Intercept + Slope * X + Error					3.041	3.041	0.959	0.920	0.006
5	6	Intercept =	0.436				4.777	4.777	1.223	1.495	4.311
4	4	(Unstandarddized) Slope =	0.868				3.909	3.909	0.091	0.008	0.006

Figure 5. Simple Regression Model on Google Sheets

This manual process is not only critical for illustrating the inner works of predictive models, and the method of evaluating predictive model performance, it is also crucial for students to see how predictive models work in the real world, when we visit the Zillow Research site[3] together to study additional explanations of

[3] https://www.zillow.com/research/

Zestimate. Again students feel a great sense of satisfaction when they realize that they can actually understand all the technical jargon that explains how Zestimate works!

The same process can be used to demonstrate classification tasks. Using logistic regression, we can predict which ZIP code each property is located. Students then compare the predictions to the actual property locations to determine prediction accuracies (i.e., correct predictions vs. incorrect predictions.). By tallying prediction accuracies, students can create a confusion matrix together as a class, and compute the model's overall performance. This exercise also provides excellent opportunities to discuss business costs of different types of prediction errors – false positives vs. false negatives, and how these costs should be included when we select models.

Conclusion

Data analysis, commonly known as statistics, can be intimidating due to its numerous jargon, symbols, and obscure terminologies. This teaching case demonstrates how to make these abstract and technical concepts much more concrete and approachable by having students participate in the process of putting together a real estate dataset manually. Here is a sample of what students have said about using the real estate dataset for class exercises:

"The class dataset was extremely useful. It allowed us to follow along during class lectures and provided great study material."

"The class dataset was my favorite part. It was helpful to keep revisiting it each week. It was interesting to see how we built upon the concepts each week using the same dataset."

"The class dataset a fun way to go about teaching a difficult subject"

"I wanted to thank you for a great session in Data Analysis! Though the material was definitely challenging, I really enjoyed using the Zillow data set to work through each new concept. It was a great way to understand the material by practicing with real data!"

"I received my undergraduate degree in economics and spent a lot of time dealing with these metrics, and never did I receive such a clear and succinct explanation of the relationships between the raw data and these descriptive statistics as you gave us in class. Thank you!"

"I admire her for connecting theory to real–world scenarios and teaching in a way we can apply them."

"It was an entirely different way of approach statistics by using software versus hand calculations – very interesting!"

"As a student who learned statistics before, I found the class very insightful and I learned a lot of new and interesting concepts. The most important takeaway for me was how I can apply statistics to real–world applications."

To conclude, real estate data collected from real estate websites such as Zillow provide excellent opportunities for teaching data analysis because real estate data are easy to understand, produce strong and predictable results, and demonstrate core concepts of predictive modeling and tangible applications.

References

Gelman, A., and Nolan, D. 2017. *Teaching Statistics: A Bag of Tricks*, Oxford University Press.

Schneider, D. 2019. "Machine Learning Predicts Home Prices," *IEEE Spectrum* (56:1), pp. 42–43. (https://doi.org/10.1109/MSPEC.2019.8594795).

Appendix 1. House Shopping Survey[4]

This short survey is designed to collect a real estate dataset that we can use for class exercises. Please know that data you enter will be shared with the entire class publicly (in class only, not on Instagram or CNN ...), with your name, as a pedagogical tool to make learning about data more personal and concrete. If you feel uncomfortable answering any of these questions and sharing your answers with class, please feel free to contact the instructor via email.

1. My first name is (e.g. Rachel)
2. My last name initial is (e.g., C)

Visit Zillow.com and find a house or condo you'd like to buy (not rent) in the Williamsburg ZIP code 23185 or 23188. (No worries about commitment - you are not required to actually buy any properties!) If you are studying with other students, please make sure each of you pick a different property. Try to click through a few pages before picking a house. Don't just pick the first one you see. First, enter the address here (e.g., 2496 Sanctuary Dr)

3. ZIP Code: __ 23185 __ 23188
4. House Type: __ House __ Condo
5. Year the house was built (e.g. 2009): ____
6. Parking - Enter number of spaces: __
7. Using Google Maps, find out how far the house is from the Mason School of Business (in miles). Enter the mileage for the shortest route (numeric value only) (e.g., 7.8): __
8. Then, enter the listing price (not rent) of the house (numeric value only) (e.g., 599000): _______
9. Enter the Zestimate amount: __
10. Enter Time in Zillow (in Days) (e.g., 5): __
11. Now, go to the Price and Tax History section, and enter the most recent year's "TAX ASSESSMENT" value (numeric value only) here (e.g., 621400): _______
12. Now, enter the "PROPERTY TAXES" value (numeric value only) from the most recent year (e.g., 5922): _______
13. This house has __ bedrooms: __ 1 __ 2__ 3 __ 4 __ 5 __ 6 __ 7 __ 8 __ 9 __ 10 __ more than 10
14. This house has __ bathrooms: __ 1 __ 2__ 3 __ 4 __ 5 __ 6 __ 7 __ 8 __ 9 __ 10 __ more than 10
15. This house is ____ sqft (e.g., 4416)
16. The Walk Score of the property is ____ (Enter a number)
17. How much would you be willing to pay for this house/condo (enter a number that may or may not be different from the listing price) (e.g., 499000): _______

[4] The most recent version of the survey is available at https://docs.google.com/forms/d/e/1FAIpQLSfNTGRqqgxcDRhPAOtz4XF5BbHg7_ulOl1eVcdgor3qB5_Gyw/viewform