IS457 Final Project - Spring 2019

April 15, 2019

This project asks you to apply the data science skills we have learned in this course to a business analysis scenario. As discussed in class, we will follow a *Lifecycle of Data Science*.

As always in this class, the work you prepare and turn in must be your own work. This is not a group project. All code used in your final project must be written by you and all analysis done by you.

We will use the Sydney Airbnb dataset to conduct analysis (make sure you download it from our course page).

Here are some details and documentation about the dataset you will be using:

variable name	**description**
id	the unique id number for each listing.
description	the brief description of the listing.
neighborhood_overview	overview of the neighborhood that the listing is in.
house_rules	some rules telling what is allowed and what is not.
host_id	the unique id number for each host.
host_since	the date that the host started to be a host in Airbnb.
host_response_time	the average time needed for the host to response.
host_response_rate	the average rate that the host responses.
host_is_superhost	whether the host is a 'superhost' (for details see Airbnb).
host_verifications	ways that the host is verified.
host_identity_verified	whether the host identity is verified.
city	the city where the listing is in.
zipcode	the zipcode for that listing.
property_type	type of the listing, for example, apartment or house etc.
room_type	room type, private room or entire room.
accommodates	the number of people allowed.
bathrooms	number of bathrooms.
bedrooms	number of bedrooms.
beds	number of beds.
bed_type	bed type. For example, real bed, pull-out sofa.
amenities	facilities like TV, WIFI.
price	price for one night.
cleaning_fee	cleaning fee per stay.

extra_people	the price for one extra people for one night.
minimum_nights	the minimum number of nights required for booking.
number_of_reviews	total number of reviews for the listing.
review_scores_rating	the overall scores rating for the listing.
review_scores_accuracy	how accurately did the listing page represent the listing.
review_scores_cleanliness	did the guests feel that the space was clean and tidy.
review_scores_checkin	is it convenient for guests when checking in.
review_scores_communication	how well did the host communicate with the guest before and during their stay.
review_scores_location	how good is the listing location.
review_scores_value	did the guest feel the listing provided good value for the price.
cancellation_policy	the cancellation policy for the listing.
reviews_per_month	the number of reviews received per month.

What you need to submit:

- An R script
- A PDF report

Note: You might need to try many R commands (especially in your exploratory analysis), but **only need to report meaningful results (business-actionable insights)**.

A few things to keep in mind as you proceed:

- Think critically, creatively;
- Think about how a current step affect all later steps;
- Explain your thoughts **clearly**. We will give as much partial credit as we can and it helps if we know how and why you are taking the decisions you are.

Part I: Data processing.

Q1

Import the data to R and **deal with missing values**.

1.1 What variables have missing values? What types/forms of missing values are they? (e.g blank, NA, N/A, -, etc.)

1.2 Please briefly describe how you deal will with these missing values and **justify why you chose these methods** (Hint: common imputation methods include impute by mean/median/mode, keep the NAs, drop the observations with NAs). You don't need to use methods beyond those we have discussed in class, however you should be thinking about the data and explain why you chose the steps you did based on observations about the data.

1.3 Describe how your choice of method to deal with missing values will affect your later analysis.

1.4 Implement your methods to deal with the missing values.

1.5 After dealing with missing values, show the dimensions of the data.

1.6 Comment on and explain any other data cleaning or preparation steps you think would be necessary from your inspection of the data (you do not need to carry them out).

Q2

Since it is hard to get an overview for large data sets, conduct a **preliminary exploration** to explore the variables of the Sydney Airbnb data by calculating some descriptive and distributional statistics.
Describe what you find that is unexpected or interesting.

Q3

To **understand large amounts of complex data**, it is helpful to use charts/tables and graphs to visualize the data. Here are general steps you can follow:

3.1 Think about the types of variables. Then choose appropriate graphs to find distributions and trends of multiple variables.

3.2 Compare different graph types to see which ones best convey trends, outliers, and patterns in the data.

3.3 Describe what you find from the graphs.

For Q2 & Q3, make sure you not only look at a few variables, but explore comprehensively.

Q4

Now look at the **relationships among several variables**.

4.1 For example, look at "review_per_month" and "number_of_reviews." They are both indicators of number of guests, but convey different information. What different information do they convey and what similar information? How are the two variables related? Answer this quantitatively.

(Hint: You can find the host ids by the top 100 "review_per_month," top 100 "number_of_reviews." Check how much they overlap (same id). Or use another approach that makes sense.)
Explain what you find.

4.2 Following this example, analyze at least three other groups of variables where you think there might be a potential relationship.

Q5

Now let's make some **(business) hypotheses** based on what you found in **Q1 - Q4**. These will serve as a guide for the rest of your analysis. Not all hypotheses turn out to be correct or business-actionable, but this is how you start.

Your task here is to **propose three different hypotheses** that will contribute to the business analysis **(think about practical business end goal)**. Consider carefully what variables you will use here and think about the exploratory analysis you have already done. You will assess your hypothesis in the rest of your project.

Explain briefly in words (the steps) how you plan to test your hypothesis without writing the code here.

Here is an example, I think being a superhost or not will affect the number of reviews and the overall ratings a host get, I will check for each category of "host_is_superhost," explore the relationship among "number_of_reviews," "review_scores_rating," and the six relevant variables "review_scores_xxx."

Part II: Data analysis.

Q6

Knowing how to **summarize different variables in a graph** is crucial to getting an overview of the data, and visualize **multiple variables and their relationships**.
(hint: functions in *lattice* and *ggplot2* packages might be helpful).

6.1 Reviews are an important factor when users choose listings. Users have different expectations for different types of property. For example, if a user chooses a house, he might expect the house to be more spacious than an apartment. Therefore, we would like to see if property type affects how users give their reviews.
Make ONE plot to visualize the relationship between "review_scores_rating" and "number_of_reviews" for all categories of "property_type." Explain what you find from your graph.

6.2 From the plot in (1) we see that some "property_type" have more listings with reviews, some have less. Now find out which types have more reviews.
Consider other variables that might affect user reviews. For those with same property types, they might have different room types, like whether it is the entire house/apartment or just a private room. Note that other attributes of the rental might be different as well, for example some provide real beds while others might have sofa beds.

To make sure there are enough samples to observe the patterns, focus on those categories of "property_types" with higher listing counts (for example "Tipi" only has one observation).
Find the top 10 categories of "property_type" by the number of listings in each category. Now subset the data so it only contains the listings in these top 10 categories of "property_type". **Make ONE plot** to show the relationship among property types, room types, bed types and reviews per month. Explain what you find from your graph.

6.3 Now make some plots (visualizing relationships between multiple variables) to explore your hypotheses in Q3. (Limit to 3 graphs per hypothesis)
Explain why you choose this kind of plot. Describe what you find that is meaningful or interesting.

Q7

The next step is **data manipulation**: **data cleaning, transformation, and aggregation.**

7.1 Before we do any further analysis involving calculations, we should first clean the data for mathematical operations. For example, the character "$" appears in the "price" column, making the data type of "price" character instead of numeric. Remove the "$" and "," in this column and convert the data type as numeric (modify the raw data).

7.2 The "amenities" column lists all the amenities provided by the host. What's the total number of amenities offered? Convert this to a numeric value that indicates the number of amenities provided. For example, if an instance of "amenities" is {TV,Internet,Wifi,Washer}, it should convert to 4. Add this as a column to the dataframe.

7.3 To understand how some of these variables behave, we can evaluate one variable by the categories of another variable.
For example, calculate the mean "review_scores_rating" according to the different kinds of cancellation policies. What do you find?

Perform more data manipulation as you see fit. **Explain what you did and why it is meaningful for the rest of your analysis.**
(Hint: check out the functions in package *dplyr* for more similar operations.)

Q8

After you manipulate the data into the format you want, you can **fit a simple linear model** to gain insights about how a certain variable might affect another variable.

For example, to determine if a listing is more popular among guests, we can use "number of reviews" or "reviews_per_month" as a primary indicator. (Recall what you did in Q4 that they convey different info.) Which one will you choose as the primary indicator? (you can only have one dependent variable in linear models.) Explain your choice.

What variables do you think that will affect the primary indicator you chose above? **Explain at least 10 candidate variables** from our data. Choose the one you think is the most promising and fit a simple linear model.
Is this model statistically significant?
Evaluate how the model fits the data.
Show some diagnostic plots for the fitted model, describe what you find meaningful from each plot.

Part III: Further analysis.

Now we conduct more in-depth analysis for business insights.
Suppose, from the perspective of a host, there are three key ways to make more money: **how many nights the listing was booked, the price, and the rating of the listings**. We will first explore what variables can increase bookings in **Q9**, then look at what factors affect the price and the rating of the listings in **Q10**.

Q9

"On average, Superhosts earn up to 22% more than other hosts. For some, that can mean as much as $1,250 in extra income. Airbnb says, "Superhosts will be featured to guests in search results, emails, and more. There's even a search filter to find Superhost listings. We'll also add a Superhost badge on your profile and listing so you can really stand out." (https://www.airbnb.com/superhost)

9.1 Explore whether there are any relationships among being a superhost ("superhost") and the following variables (and maybe more of your choice): "host_since," "host_response_time," "host_response_rate," "host_verifications," "host_identity_verified").
You may need to clean the data first (for example, converting the date variable "host_since" to numeric).

9.2 Create a mosaic plot for the "host_response_time" by whether the host is "superhost." What do you learn from doing this?

Q10

Another aspect of making more money is price.
Look at the "description" column to see whether any keywords suggest a relationship with the price of the listing.

10.1 Extract the unique words used in the "description" column and eliminate all the stop words (given in the list below).
Store the words and their frequencies in a dataframe and sort the dataframe by word frequency in decreasing order.
What do you infer from the words with top 10 frequency?

Stop Words *a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, is, to, too, was, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your.*

10.2 The word "beach" is one of the most frequent (your work in 10.1 should confirm this). Explore whether this element affects the price of a listing.

Note: the word "beaches" and "beach" should refer to the same thing. Find the listings in which "beaches" or "beach" are included in "description." What's the difference in average "price" between those with these words and those without?

Now explore multiple high frequency words. Write a general function to get the specific word frequency from some text (for example, input the text and the word of interest, it should return the word frequency.) Then calculate the word frequency for the "description" of each row. If the frequency is 0 then it means this word is not included in the "description" for this row.

10.3 Select other words from your dataframe and do similar analysis (at least 3 words).
What conclusions do you find?

Q10 (2)

In part (1), we analyzed an influential keyword –beach– from the "description," and explored how the "price" of the listing is affected.

The word beach gives information about the listing location and we also have two other variables with respect to location: "zipcode" and "city." In this question, we will look at whether location also has an impact on the ratings of the listings.

10(2).1
Choose one variable either from "zipcode" or "city" with which to explore location. Explain your choice. Calculate the number of listings for each "city" (or "zipcode") category.
Now look at the top 100 locations (by "city" or "zipcode") with the highest number of listings.
Explore whether these top 100 locations have higher ratings. Here ratings refers to **weighted means** of "review_scores_rating" (calculate their weights by their respective "number_of_review").

Choose a graph type to show your findings. (Hint: show the trend)
If you find there is no obvious trend using the top 100 locations, you can also look at the top 150 or top 200.
Describe what you find meaningful in the graph.

10(2).2 Choose two other aspects from "description" that you think will help improve the weighted mean of "review_scores_rating".

(Hint: look at other keywords with a high frequency of occurance in "description" and see whether these keywords indicate the relevance of other variables in data set. Perform similar analysis to check if these variables show any trends with "review_scores_rating").

Part IV: Your turn

Now, it is your turn to conduct further in-depth analysis to gain business actionable insights. You could follow the detailed example we gave in Part III to explore more on the hosts' side, or analyze other aspects such as from the users'/customers' end.

Part V: Conclusion

Summarize your finding in this project, translate the results from your analysis to business suggestions.

Part VI: Clearly the explain and describe the Lifecycle of Data Science for this project.

Bonus: Implement Your Analysis in WholeTale

The Whole Tale project is a research project to enable reproducible data science, located at wholetale.org. The platform supports RStudio and Jupyter notebooks that also support R. Implement your analysis from this project on the Whole Tale platform and include the resulting URL your submitted PDF file. The URL will point to your work, which Whole Tale calls a Tale. There is more documentation here: https://wholetale.readthedocs.io/en/stable/README.html

Using Whole Tale:

1. Go to the whole tale website https://wholetale.org/ and press the large button 'Access Whole Tale'
2. Select University of Illinois at Urbana-Champaign from the drop-down list of institutions and complete the Single Sign On page.
3. You should now be at the WT dashboard. Click on the 'COMPOSE' tab at the top of the page.
4. Enter a name for your tale in the 'Tale name:' text box in the 'Compose' window on the left of the page.
5. Select the 'Jupyter Notebook' entry under the 'Environments' window on the upper right of the page, or RStudio if you prefer.
6. Press the 'Launch New Tale' button, and wait for Jupyter to be launched.
7. Within the Jupyter window, we now will upload the necessary files. Click the 'Upload' button in the Jupyter window and select all necessary files for this assignment.

Note:

1. Use a common title for your tale: IS457_SP19_YourClassID;
2. Make your tale private;
3. If you all start doing your work right before the deadline, the system may crash.