Assignment 3

STA 141A

Due **October 30 at 10:30am**.

You have more time for this assignment than the previous one, but the number of questions and page limit is about the same. Use the extra time to make sure your graphics are perfect and that your report has a clear narrative or theme.

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 10 pages** including graphics, but excluding code and citations. Think carefully about what information is important to include.

When you are finished, submit a digital copy on Canvas. The digital copy must contain:

- Your report as a PDF file, with your code in the appendix.
- Your code as 1 or more R scripts.

Your submission will be graded according to the STA 141A grading standards, which are available on Canvas.

Description

Craigslist is a website that allows people to post classified advertisements for free. These posts span a variety of subjects, including job postings, housing rentals, and item sales. Craigslist started in the San Francisco Bay Area and now serves 570 cities worldwide. Posts are grouped by city to make relevant ads easy to find.

For this assignment, you'll examine a data set based on recent Craigslist posts for apartment rentals in California. Craigslist has few rules about how posts should be formatted, so this data set is messy. In addition, some of the included variables were computed from the original post text and may not be accurate! Therefore it's important to be on the lookout for errors and suspicious results.

The data set is available on Canvas.

Questions

1. Give a big-picture overview of the data set. Some questions you might discuss here:
   - What is the unit of observation? How many observations are there?
   - What kind of information is recorded for each observation?
   - What time (or place) do the data span?
   - …these are just suggestions! You should also discuss any other big-picture properties of the data set you think are important or relevant.
2. Answer and discuss the following questions:
   1. Are apartments in suburbs more likely to be family-friendly (many bedrooms, pets allowed, etc) than apartments in major cities?
   2. Which adds more to rent: extra bedrooms, or extra bathrooms? How can you tell? Does the effect change as the number of bedrooms or bathrooms goes up?
   3. Do apartments in similar geographical areas tend to be similar? Discuss how they are or how they are not for 3 different cities or places.
3. (NO CODE) What kinds of questions can be answered with this data set? Come up with 10 questions about the data and explain why each question is meaningful (and to whom). Choose your questions carefully. They should be open-ended enough to support detailed data analysis, but specific enough to give you a starting direction for analysis.
4. Choose and answer at least 5 more questions about the data. These can be questions you brainstormed, or they can be follow-up questions from earlier questions. For each question, state the question, make a hypothesis (a guess) about what you will find, and then investigate using exploratory data analysis methods.
5. Finally, examine limitations of the data set. Discuss how the limitations affect your earlier results. Questions you might discuss:
   - Are there missing values? If there are, do they follow any patterns?
   - Are there outliers or anomalies? If there are, how can you account for them in your analysis?
   - Are there errors in the data set? What kinds of errors? How can you tell (provide evidence)?
   - How was the data set generated? Could the person(s) that created the data set have a bias? Does this affect the reliability of your results?
   - Are there enough observations to support the conclusions you made? Note that this may be different for each of your results.

- What are some features missing from the data set that might confound your results?
- Do your conclusions apply to new observations, or only the observations in the data set? How can you tell?
- ...these are just suggestions! You should also discuss any other limitations of the data set you think are important or relevant.

Data Description

Each row in the data set corresponds to a single post. The posts were downloaded on October 15th, 2018. Note that the San Francisco Bay Area Craigslist site is divided into several different subregions.

The original features for each post are:

Feature	Description
title	original title of the post
text	original text of the post
city_text	original city text of the post
date_posted	date and time the post was made
date_updated	date and time the post was updated, or NA if not updated
deleted	whether the post was deleted
craigslist	craigslist site where the post was made
latitude	latitude of the apartment
longitude	longitude of the apartment
price	monthly rent for the apartment, in dollars
sqft	size of the apartment, in square feet
bedrooms	number of bedrooms (0 means a studio apartment)
bathrooms	number of bathrooms (0.5 means a toilet with no bath)

In addition, the following features have been extracted from the text and may not be reliable:

Feature	Description
pets	pet policy
laundry	laundry
parking	parking policy
place	place (city, or name of rural area)
city	city
county	county
state	state