Final Project Data Memo

```r
getwd()
```

```
## [1] "/Users/joshfreitas/Desktop/pstat 131/PSTAT131"
```

```r
housingData <- read.csv("/Users/joshfreitas/Desktop/pstat 131/PSTAT131/clean_2000_2018.csv")
head(housingData)
```

```
##          post_id     date year   nhood    city  county price beds baths sqft
## 1 pre2013_134138 20050111 2005 alameda alameda alameda  1250    2     2   NA
## 2 pre2013_135669 20050126 2005 alameda alameda alameda  1295    2    NA   NA
## 3 pre2013_127127 20041017 2004 alameda alameda alameda  1100    2    NA   NA
## 4  pre2013_68671 20120601 2012 alameda alameda alameda  1425    1    NA  735
## 5 pre2013_127580 20041021 2004 alameda alameda alameda   890    1    NA   NA
## 6 pre2013_152345 20060411 2006 alameda alameda alameda   825    1    NA   NA
##   room_in_apt address lat lon
## 1           0    <NA>  NA  NA
## 2           0    <NA>  NA  NA
## 3           0    <NA>  NA  NA
## 4           0    <NA>  NA  NA
## 5           0    <NA>  NA  NA
## 6           0    <NA>  NA  NA
##                                                                               title
## 1                                 $1250 / 2br - 2BR/2BA   1145 ALAMEDA DE LAS PULGAS
## 2 $1295 / 2br - Walk the Beach! 1 FREE MONTH + $500 TRADER JOES SHOPPING CERTIFICATE
## 3                                                              $1100 / 2br - cottage
## 4                  $1425 / 1br - 735ft² - BEST LOCATION SOUTHSHORE GARDENS APARTMENTS
## 5              $890 / 1br - Classy "Painted Lady" VICTORIAN - Top Floor w/ Sunporch!
## 6                                                    $825 / 1br - Bayview Apartments
##   descr details
## 1  <NA>    <NA>
## 2  <NA>    <NA>
## 3  <NA>    <NA>
## 4  <NA>    <NA>
## 5  <NA>    <NA>
## 6  <NA>    <NA>
```

Dataset Overview

- This data set provides information about Craigslist rental listings in the Bay Area from 2000 to 2018. This dataset was created as the result of undependable housing information from non-government sources, especially in a region widely known for its competitive and unjust housing market. While it doesn't include information about if the property was actually rented or not, it does provide information about the location of the rental, list price, year, title of the listing, and more.

- I found this dataset from the FiveThirtyEight Data is Plural newsletter, and it is from an independent website created by one woman who personally scraped the data. https://www.katepennington.org/data . She works at the US Census Bureau in the Center for Economic Studies and provides the raw and cleaned data for these craigslist listings, alongside her methods of scraping and cleaning the data. I will be using the cleaned data set, and possibly a random sample of it because it does have a lot of observations and I do not know if it will be too many for my computer to handle.
- There are 200,796 observations and 17 predictors, although I am not sure how useful some of them will be.
- I will be working with both numeric and character data variables, especially dates and the listing titles will be important, but there are no boolean values that I will work with in the data set.
- There are quite a bit of missing data from certain variables, but because so much of it is missing I don't know if I will use them at all. In particular, I probably will not use the `address`, `lat`, `lon`, `descr`, or `details` variables because the actual data is sparse. If I find I want to use any of them, I will ignore the observations which are missing a certain variable, but be explicit with the amount of missing observations from any EDA or model-building.

An Overview of Potential Research Questions

- While I think there could be a number of predictor variables, I think price would be the most insightful and broad choice. I think a lot of EDA would go towards the evolution (or not) of the rental housing market in the Bay Area in terms of pricing, which could then predict the future listing prices of homes. I understand that the data set's scope only goes to 2018 and we are well-beyond that, but it might even be interesting to go back and compare what the model had predicted with actual data from today and see how well it might have held up. Although I don't think this is the main intention of the data set, I would be interested in also doing some EDA about characteristics of titles of listings and see if it has a strong relationship to any of the listings. In particular, I thought it might be interesting to look at which titles have non-English characters in them and see if there is a difference in their listing price or neighborhoods, or really any other variables. Even some of the titles are much more enthusiastic and descriptive than others, and it makes me wonder if there is a relationship to how they will price the rental comparatively to other cities in the area.
- As I am thinking of approaching my data now, I would say my main response variable would be `price`, but I would also like to see if I could make `title` into a response variable based on things like square footage and location. I would like to see even if the title is longer than a certain amount of characters and fit it into a sort of classification model. I think the main predictors I would like to work with are `nhood`, `city`, `county`, `title`, `year`, `beds`, `baths`, and `sqft`. All of these are important factors for someone buying a home and since craigslist allows individuals to put these rentals up for sale and not just always a leasing company, it might be interesting to look at variability in list prices.
- I think a classification approach using `title` as the response could be fun but probably trickier, and a regression approach would likely be best for using `price` as the response. I think more EDA will likely lead me to pick one or the other.
- I think `beds`, `sqft`, and `nhood` will all be very useful predictors as those are usually the main selling points of a house or apartment.
- I think my model will be a combination of predictive and inferential. It will be predictive in the sense that I want to see if other factors will influence the title or pricing of a Craigslist listing and see what the price would be, and it would be inferential as we can see which factors will be best in seeing if it falls into a certain price or title category.

Proposed Project Timeline

- Since the data is already cleaned for the most part (the data was said to be cleaned to the best of her ability but I still would like to go through it myself and see), I want to begin EDA within the next

week and a half. I am not totally sure how long all of this will take, but I see many ways that this data set can take me and would like to reserve a lot of time for EDA which will help me decide on a model.
- I think I would like to take a few weeks doing EDA and getting more familar with my data set and then spend the rest of the quarter working on model building. I'm not totally sure what that will look like since I do not have a lot of experience with model building, so all of this could be a terrible timeline. I want to try to be done with everything a week before it is due so I can spend the last week doing final touches and making the data look pretty and organized in its html file.

Questions or Concerns

- I think I am just slightly concerned about not having enough predictor variables, because I would have liked to have had more descriptions, details, and square footage, but I do enjoy the other information that the data set provides.
- Also, I was wondering how set in stone we need to be of our research questions, because I wanted to spend a bit more time learning the data set than be stuck exploring a part of the data set of which I may not be that interested in.