# Stats 10 Final Project

Abhishek Devarajan (106072269)

Unlike an English report, which is expected to flow smoothly from beginning to end, math and stats reports are often highly segmented. This makes it easier for readers to quickly determine where each topic begins and ends. The following paper provides an example of what your Final Project report can look like. Please remember that this not a definitive guide and that you can choose to format your report in a different way as long as it's easily understandable and cohesive. Throughout the document, I will include notes like this (in blue) to give you insight into what's going on.

## 1. Introduction:

In this section, you can introduce your dataset and talk about some of the goals of your analysis. Keep this section as a brief, high-level overview of the project. An example introduction is shown below.

In this paper, we will use real-estate data to analyze the relationship between housing prices and various features of houses. These features include the year the houses were built, their square footage, and the number of bathrooms, among other things. We analyze the data by using graphs such as histograms and boxplots to estimate the distributions of different variables, and by using linear regression to assess the relationship between numerical variables.

Your intro can be shorter or longer than this but I wouldn't make it too much longer. As long as you inform the reader about what they're going to see within your report, you're good to move on.

## 2. Detecting and Handling Missing Data:

In this section, you can discuss the analysis you did for the first task of the project. Make sure you explain the steps you took and include any graphs or tables you made. The following paragraph is an example of what you might write in this section.

In its initial form, the dataset contains many missing values in the form of NAs. By checking if each element in the dataset is an NA, and then summing the number of NAs by column, we can see how many missing values there are for each variable. This information is summarized in Table 1. (This passage is just an example and does not come from one of the project tasks!)

| ID | YearBuilt | SqFt | Story | Acres | N_Baths | Fireplace |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 2 | 1 | 2 | 1 |

| TotalPrice | LandPrice | BuildingPrice | Zipcode |
|---|---|---|---|
| 0 | 1 | 2 | 2 |

Table 1: NA Values by Variable

Notice that I included the output of my R code directly in the report with a small caption for clarity. I've also referenced the table I made within my writing. This is a good idea when writing your project because it makes it easy for the reader to quickly understand and access the visualizations/information you're referring to. Generally, we number tables and figures separately, so the first graph I include will be labeled Figure 1.

Useful Code/References:

- Functions:
  - is.na() to detect NA values
  - rowSums() to sum each row of a data frame or matrix
  - na.omit() to remove NA values from a data frame
- References:
  - Lab 2/Assignment 2
    - Deals with subsetting vectors and data frames
    - Discusses missing values and how to deal with them

### 3. Variable Summarization:

Sections 3-6 will follow a similar structure to Section 2. You should write about the steps you took in your analysis and show any relevant graphs and tables. Make sure to include an interpretation in your own words about the results of your analysis.

Useful Code/References:

- Functions:
  - summary() to create a 5-number summary for a particular variable
  - boxplot(), histogram(), barplot(), etc. to visualize the distribution of variables
  - tally() to find the counts/proportions for categorical variables
  - **Note:** tally() and histogram() come from the mosaic package so make sure you have that installed.
- References:
  - Lab 1/Assignment 1:
    - Discusses all of the functions listed above

        - Includes examples for creating stacked histograms, two-way tables, and side-by-side boxplots.

4. **Price Comparison Between Houses with and without Fireplaces:**

Useful Code/References:

- Code:
    - Subsetting dataframes using logical conditions ( == , <=, >=, etc.)
    - histogram() or boxplot() for plotting side-by-side or stacked graphs
    - tally() for creating two-way tables
- References:
    - Lab 2/Assignment 2
        - Provides examples for subsetting data frames
    - Lab 1/Assignment 1
        - Shows code for creating stacked graphs, side-by-side graphs, and two-way tables

5. **Numerical Relationship Exploration:**

In this section, you have to identify continuous numerical variables. A common technique is to assume that just because a variable has a decimal means that it's continuous. I've often used this criterion in class to determine whether a variable is continuous or discrete. However, there is a bit more nuance to this than just looking for decimal values. Discrete random variables can only go up or down by a fixed amount. For example, the variable N_baths always goes up by 0.5. This is because you can build or take away a bathroom or a half-bathroom in a house but you cannot build a 0.1 (or any other fractional amount) of a bathroom. A true continuous variable, on the other hand, can be increased or decreased by any amount. The price of a house, for example, could go up or down by $1, $1000, $0.33333, or any other quantity.

Useful Code/References:

- Functions:
    - plot() to show the relationship between two variables
    - cor() to calculate the correlation between two variables
- References:
    - Lab 2/Assignment 2
        - Discusses plotting two variables against each other
        - Includes practice with using graphics parameters (color, shape, etc) to make relationships more visually obvious

6. **Linear Regression Analysis:**

Useful Code/References:

- Functions:
  - lm() and summary() to construct and summarize the linear regression model
  - plot() and abline() to visualize the regression and the residuals
- References:
  - Lab 3/Assignment 3
    - Shows how to construct a linear model, plot the residuals, and analyze the conditions

7. **Conclusion:**

In this section, you should briefly summarize your findings from the previous section. If you set out particular goals in the Introduction section, you can call back to those right now. An example is shown in the paragraph below.

Ultimately, we found that houses with fireplaces are more expensive, on average, than houses without fireplaces. Our analysis also suggests that there is a statistically significant linear relationship between price and square footage. (I made this stuff up so please don't copy this unless you have the work to back it up!)

Your conclusion can be longer or shorter than this, as long as you summarize your findings accurately.

8. **Appendix:**

All you have to do in the Appendix is copy and paste either screenshots of your code or your code itself. If you choose to copy and paste the code itself, please make sure that the formatting is correct. Also make sure your screenshots are in order so that I can easily tell which code you used for which parts. Add comments to make things easier to understand. In the end, the easier it is for me to understand what you did, the easier it will be for you to get full points. An example is below.

```
#Checking NA Values per Column
colSums(is.na(data))
```

One miscellaneous thing is the POV that I wrote the example statements in. You'll notice that I used sentences like "we found that…" or "we can see…". This first-person plural POV is how all math and stats papers are written. Feel free to use whatever POV you want, but this is considered the standard.