**Data Set Title**

**Exploratory Analysis**

Naat Ambrosino, [nambrosino@bellarmine.edu](mailto:nambrosino@bellarmine.edu)

Jardin Dantzler, [jdantzler@bellarmine.edu](mailto:jdantzler@bellarmine.edu)

1. **INTRODUCTION**

Short description of the data set including a reference to where it can be found and why you chose it.

1. **DATA SET DESCRIPTION**

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data. NOTE: Mention here about dropping mostly empty columns.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| V1 |  | 0% |
| V2 |  | 0% |
| V3 |  | 0% |

1. **Data Set Summary Statistics**

Narrative introduction to the section.

**Table 2: Summary Statistics for XXX (name of dataset)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
|  |  |  |  |  |  |  |  |  |

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
|  |  |  |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 4: Correlation Table/Tables

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

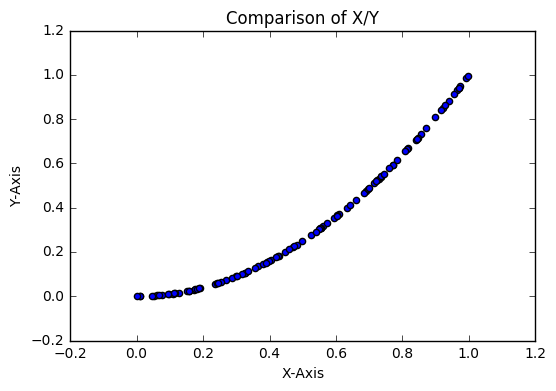
After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

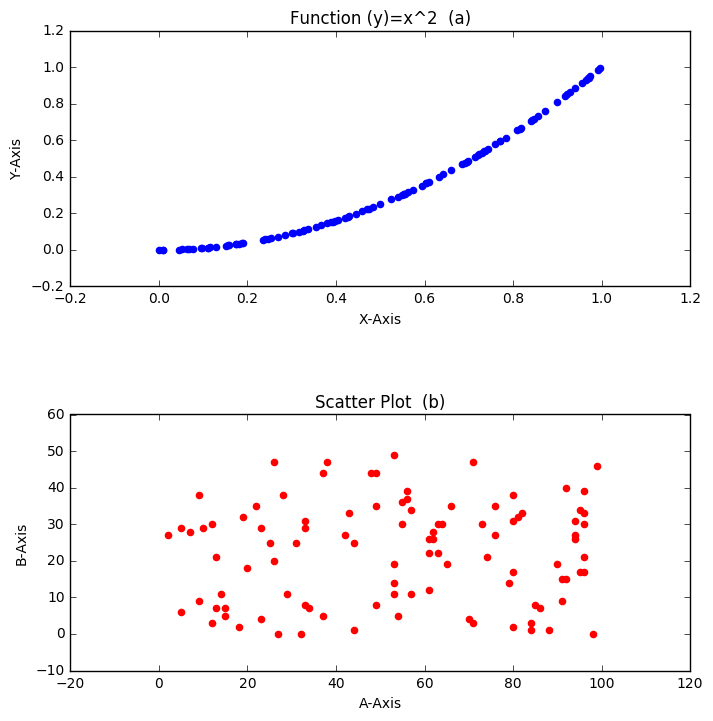
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

* 1. *Distributions*
  2. *ScatterPlots / Pairwise Plots (continuous variables)*
  3. *Barcharts (categorical variables)*
  4. *Other Plots - don’t skimp – there are likely other plots that would be useful that I haven’t already specified. Include those in this section.*

All figures should be cited formatted like this and mentioned in the text.



**Figure 1: Comparison of X/Y from dataset (single plot) (8 pt)**



**Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt)**

1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.