**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.

This dataset can be found on Kaggle, under the UNCOVER COVID-19 Challenge dataset (<https://www.kaggle.com/roche-data-science-coalition/uncover/tasks>). The UNCOVER COVID-19 Challenge dataset consists of many different datasets from a wide variety of sources. Under the Canada\_Hosp1\_COVID\_InpatientData.xlsx file (containing data about patients at a Canadian hospital), there is a sheet named Data-at-admission (the patients’ data recorded upon admission), which we separated out and saved as Canada\_Hosp1\_COVID\_Data\_At\_Admission.csv. We chose this dataset due to its high number of variables (54 columns) and decent mix of categorical and continuous variables.

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.

After cleaning, this dataset contains 508 samples with 40 columns with various data types. A complete listing is shown in Table 1. Before cleaning, the dataset contained 508 samples with 55 columns with various data types. We decided that any column with less than 40 non-null values were to discarded immediately, leading to us dropping columns *ethnicity\_other, received\_covid\_vaccine, covid\_vaccine, pao2, pao2\_fio2, ph, high\_senstivity\_cardiac\_troponin, esr, ferritin,* and *hs\_crp*. *id* was dropped as it was (for our purposes) simply an alternate index. *ethnicity* was dropped as its data was apparently scrubbed, consisting entirely of “[]”. *aptt\_aptr* was dropped as it consisted entirely of numbers but used the string data type, had only 288 non-null values, and did not seem useful after some research as to what it meant. *meds* was dropped due to having too many unique values to be of use, along with *comorb\_other.*

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

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| reason | nominal/object | 0% |
| age | ratio/int64 | 0% |
| sex | nominal/object | 0% |
| height | ratio/float64 | 46.46% |
| weight | ratio/float64 | 29.53% |
| comorb | nominal/object | 25.39% |
| smoke\_hist | nominal/object | 41.73% |
| year\_quit | interval/float64 | 91.54% |
| er\_2\_weeks | nominal/object | 0% |
| admission\_disposition | nominal/object | 0% |
| systolic\_bp | ratio/int64 | 0% |
| diastolic\_bp | ratio/int64 | 0% |
| heart\_rate | ratio/int64 | 0% |
| respiratory\_rate | ratio/int64 | 0% |
| oxygen\_sat | ratio/int64 | 0% |
| temp | interval/float64 | 0% |
| motor | nominal/object | .39% |
| verbal | nominal/object | .39% |
| eye | nominal/object | .39% |
| intubated | nominal/object | 0% |
| wbc | ratio/float64 | .39% |
| rbc | ratio/float64 | .39% |
| hemoglobin | ratio/float64 | .39% |
| hematocrit | ratio/float64 | .39% |
| mcv | ratio/float64 | .39% |
| mch | ratio/float64 | .39% |
| mchc | ratio/float64 | .39% |
| rdw | ratio/float64 | .39% |
| platelet\_count | ratio/float64 | .39% |
| pt | ratio/float64 | 37.99% |
| alt | ratio/float64 | 20.67% |
| ast | ratio/float64 | 20.67% |
| serum\_creatinine | ratio/float64 | .39% |
| sodium | ratio/float64 | .39% |
| potassium | ratio/float64 | .39% |
| total\_serum\_bilirubin | ratio/float64 | 20.67% |
| lactate | ratio/float64 | 28.74% |
| inr | ratio/float64 | 37.99% |
| d\_dimer | ratio/float64 | 50.39% |
| crp | ratio/float64 | 72.05% |

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* |
| age |  |  |  |  |  |  |  |  |
| height |  |  |  |  |  |  |  |  |
| weight |  |  |  |  |  |  |  |  |
| year\_quit |  |  |  |  |  |  |  |  |
| systolic\_bp |  |  |  |  |  |  |  |  |
| diastolic\_bp |  |  |  |  |  |  |  |  |
| heart\_rate |  |  |  |  |  |  |  |  |
| respiratory\_rate |  |  |  |  |  |  |  |  |
| oxygen\_sat |  |  |  |  |  |  |  |  |
| temp |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

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

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

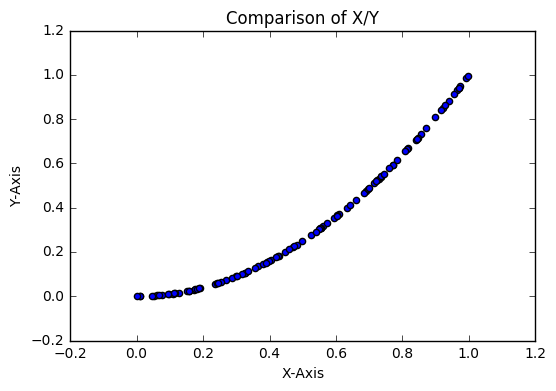
After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

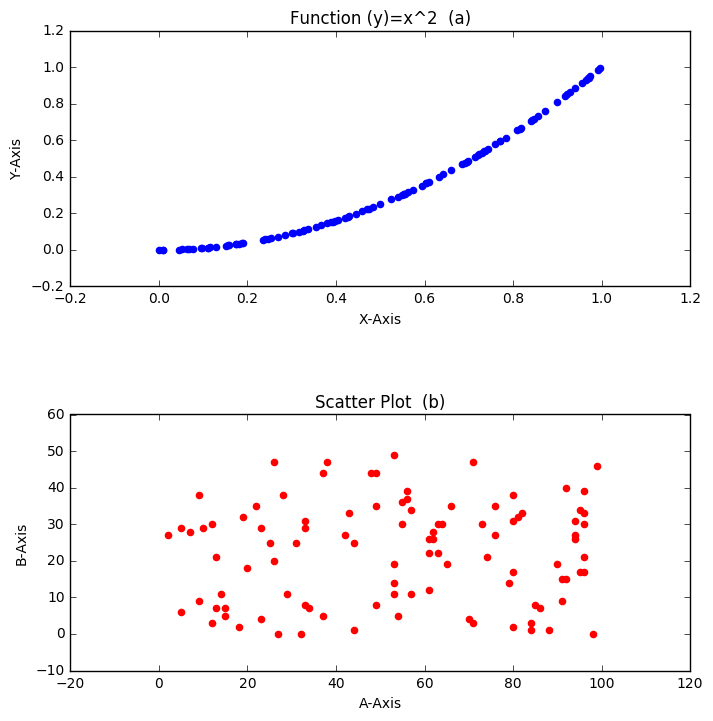
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

* 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.