Gideon Blinick
Springboard Capstone 1

Data Wrangling

For this capstone, the dataset of Default Payments of Credit Card Clients in Taiwan from 2005 was used. The dataset can be found here (https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home) and here (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients).

Cleaning Steps

Minimal data wrangling was required on this dataset. The data was loaded into a Pandas Dataframe directly from a CSV file and analysis proceeded from there.

Missing Values

No Null or Missing values appeared in any of the columns. Thus, no decisions had to be made and no actions needed to be taken to fill in or remove this missing data. The lack of missing values could be seen from a simple call to the info method on the dataframe:

```
In [48]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 1 to 30000
Data columns (total 24 columns):
LIMIT_BAL                     30000 non-null float64
SEX                           30000 non-null int64
EDUCATION                     30000 non-null int64
MARRIAGE                      30000 non-null int64
AGE                           30000 non-null int64
PAY_0                         30000 non-null int64
PAY_2                         30000 non-null int64
PAY_3                         30000 non-null int64
PAY_4                         30000 non-null int64
PAY_5                         30000 non-null int64
PAY_6                         30000 non-null int64
BILL_AMT1                     30000 non-null float64
BILL_AMT2                     30000 non-null float64
BILL_AMT3                     30000 non-null float64
BILL_AMT4                     30000 non-null float64
BILL_AMT5                     30000 non-null float64
BILL_AMT6                     30000 non-null float64
PAY_AMT1                      30000 non-null float64
PAY_AMT2                      30000 non-null float64
PAY_AMT3                      30000 non-null float64
PAY_AMT4                      30000 non-null float64
PAY_AMT5                      30000 non-null float64
PAY_AMT6                      30000 non-null float64
default.payment.next.month    30000 non-null int64
dtypes: float64(13), int64(11)
memory usage: 5.7 MB
```

Or by summing the Null values in the dataframe and finding that they add to 0:

```
In [49]: data.isnull().sum().sum() # Number of NULLs in the whole Data Frame

Out[49]: 0
```

Outliers

For this step, each of the columns/features were examined individually. For each column, the values were plotted with both a box-and-whisker plot and a histogram, and the important numerical data (count, mean, standard deviation, quartiles, min, and max) was obtained using the describe method. What was found was that there were no outliers in the dataset that needed to be removed. For this insight I relied on an article on the website *The Analysis Factor* about dropping outliers (found here: https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/).

The article stipulates that an outlier should be dropped only if:

a) "it is obvious that the outlier is due to incorrectly entered or measured data"
b) "the outlier does not change the results but does affect assumptions"
c) "the outlier creates a significant association"

In the case of my dataset, a and b are almost certainly not true and c remains to be analyzed in later work. Hence, no values needed to be dropped.

It is important to emphasize that most of the columns did contain statistical outliers, defined as being any value greater than the third quartile value by 1.5X the interquartile range, or any value less than the first quartile value by 1.5X the interquartile range (defined here: http://mathworld.wolfram.com/Outlier.html).

These values made sense in context, however, and removing them would result in significantly different (and likely incorrect) conclusions. For example, when examining the age column, it was found that ages over 60.5 years were found to be outliers. About 1% of the values in the dataset were thus classed as outliers by age alone. Removing these values simply because they are outliers would be wrong as no justification could be given for how the removal improves analysis. The same reasoning was applied to outliers in other columns.

Also worth mentioning is that some of the columns had values that they should not have been able to have by the dataset descriptions. For example, some columns had a repayment status of -2, which is not coded in the description. Further research is needed to clarify these values.