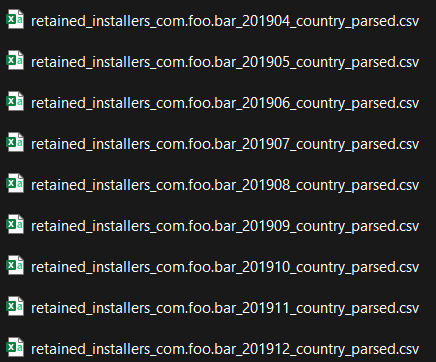
IN498 Unit 2 Assignment

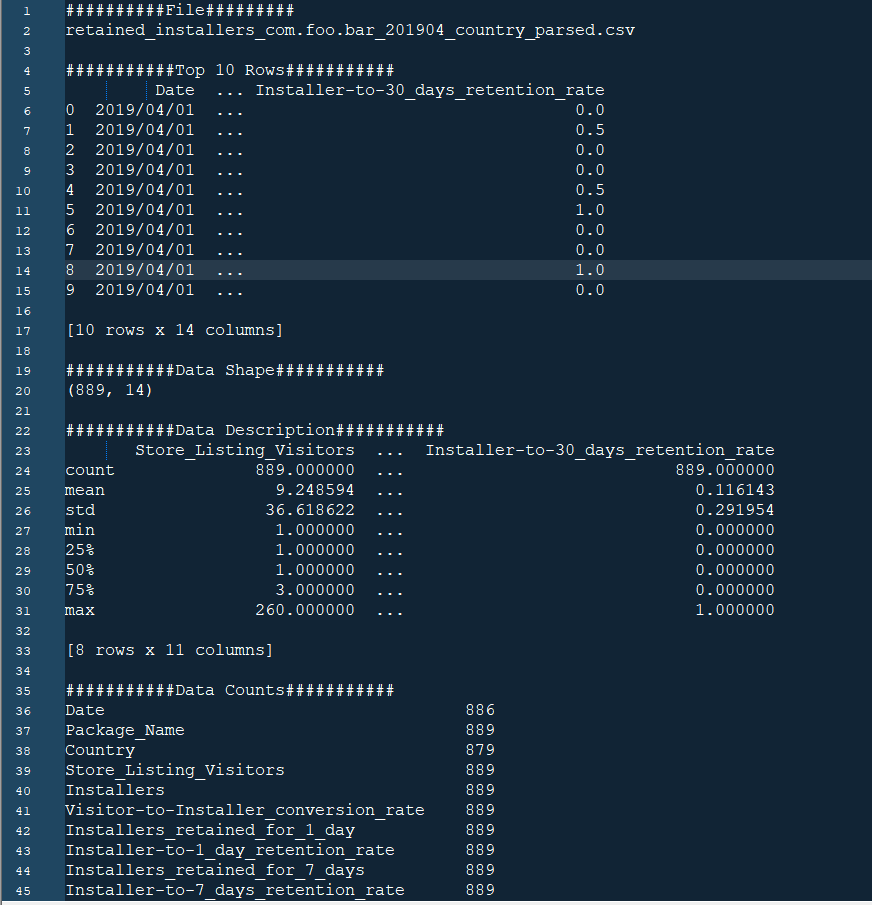
Laurence Burden

1. **Initial Files**

****

1. The files were parsed to a true CSV file using Python and the csv package. Within the code, the files are opened individually and each line is then read as a row and added to the file list object for saving as the previous filename plus “\_parsed.csv” on the end.
2. Subset of code and execution files from the initial explorer files.





1. **Dataframe shapes:**

1. 889, 14

2. 895, 14

3. 790, 14

4. 727, 14

5. 778, 14

6. 560, 14

7. 582, 14

8. 595, 14

9. 601, 14

**Missing data:** The data missing includes 14 dates, 59 country listings, and 8 each of the 30-day install and retention rates.

**Means and Modes:**

##########File#########

retained\_installers\_com.foo.bar\_201904\_country\_parsed.csv

Installers retained 1 day mean: 1.455568053993251

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.1361079865016872

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 0.9651293588301463

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.8098987626546682

Installers retained 30 days mode: 0 0

dtype: int64

##########File#########

retained\_installers\_com.foo.bar\_201905\_country\_parsed.csv

Installers retained 1 day mean: 1.5329608938547485

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.2089385474860335

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 1.0324022346368715

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.8621076233183856

Installers retained 30 days mode: 0 0.0

dtype: float64

##########File#########

retained\_installers\_com.foo.bar\_201906\_country\_parsed.csv

Installers retained 1 day mean: 1.5531645569620254

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.240506329113924

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 1.0658227848101265

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.8949367088607595

Installers retained 30 days mode: 0 0

dtype: int64

##########File#########

retained\_installers\_com.foo.bar\_201907\_country\_parsed.csv

Installers retained 1 day mean: 1.453920220082531

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.1031636863823935

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 0.9463548830811555

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.814404432132964

Installers retained 30 days mode: 0 0.0

dtype: float64

##########File#########

retained\_installers\_com.foo.bar\_201908\_country\_parsed.csv

Installers retained 1 day mean: 1.551413881748072

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.199228791773779

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 1.0398457583547558

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.8586118251928021

Installers retained 30 days mode: 0 0

dtype: int64

##########File#########

retained\_installers\_com.foo.bar\_201909\_country\_parsed.csv

Installers retained 1 day mean: 1.3625

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.1089285714285715

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 0.9821428571428571

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.8392857142857143

Installers retained 30 days mode: 0 0

dtype: int64

##########File#########

retained\_installers\_com.foo.bar\_201910\_country\_parsed.csv

Installers retained 1 day mean: 1.8109965635738832

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.3865979381443299

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 1.2010309278350515

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 1.0309278350515463

Installers retained 30 days mode: 0 0

dtype: int64

##########File#########

retained\_installers\_com.foo.bar\_201911\_country\_parsed.csv

Installers retained 1 day mean: 1.5680672268907563

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.2436974789915967

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 1.1126050420168068

Installers retained 15 days mode: 0 0

dtype: int64

Installers retained 30 days mean: 0.9193277310924369

Installers retained 30 days mode: 0 0

dtype: int64

##########File#########

retained\_installers\_com.foo.bar\_201912\_country\_parsed.csv

Installers retained 1 day mean: 1.4276206322795342

Installers retained 1 day mode: 0 0

dtype: int64

Installers retained 7 days mean: 1.1164725457570714

Installers retained 7 days mode: 0 0

dtype: int64

Installers retained 15 days mean: 0.956738768718802

Installers retained 15 days mode: 0 0

dtype: int64

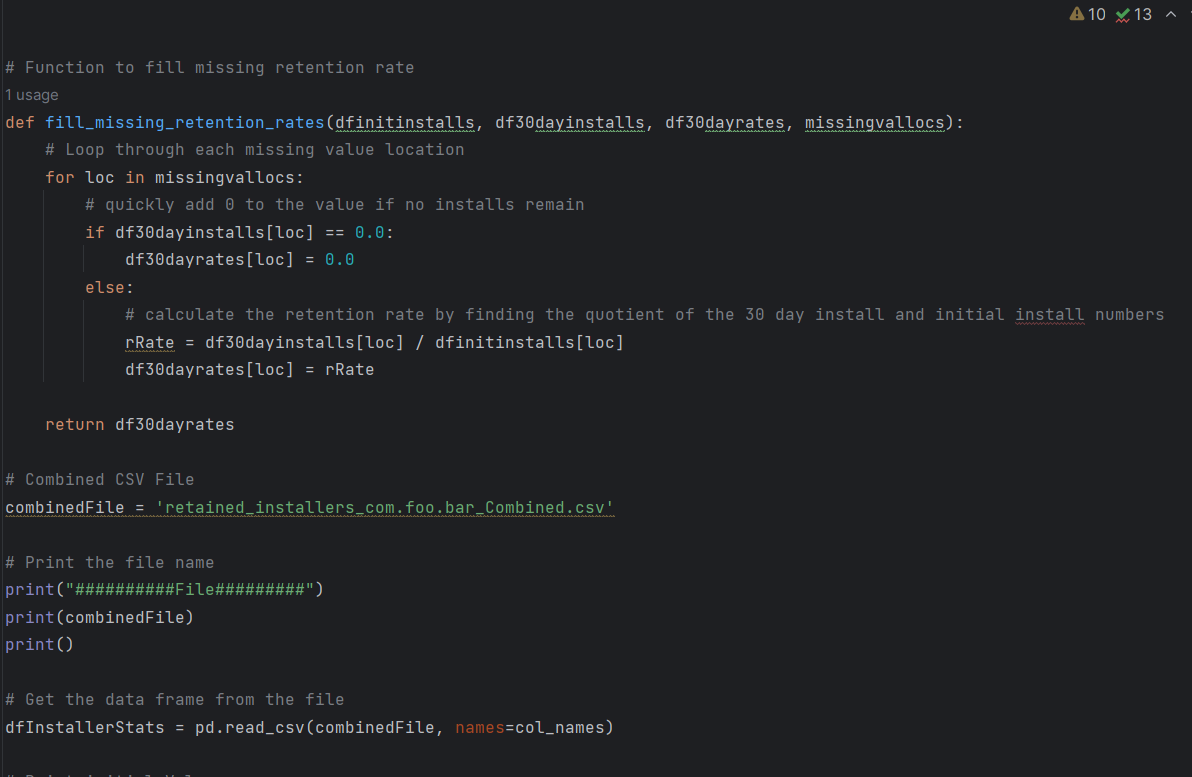
Installers retained 30 days mean: 0.7920133111480865

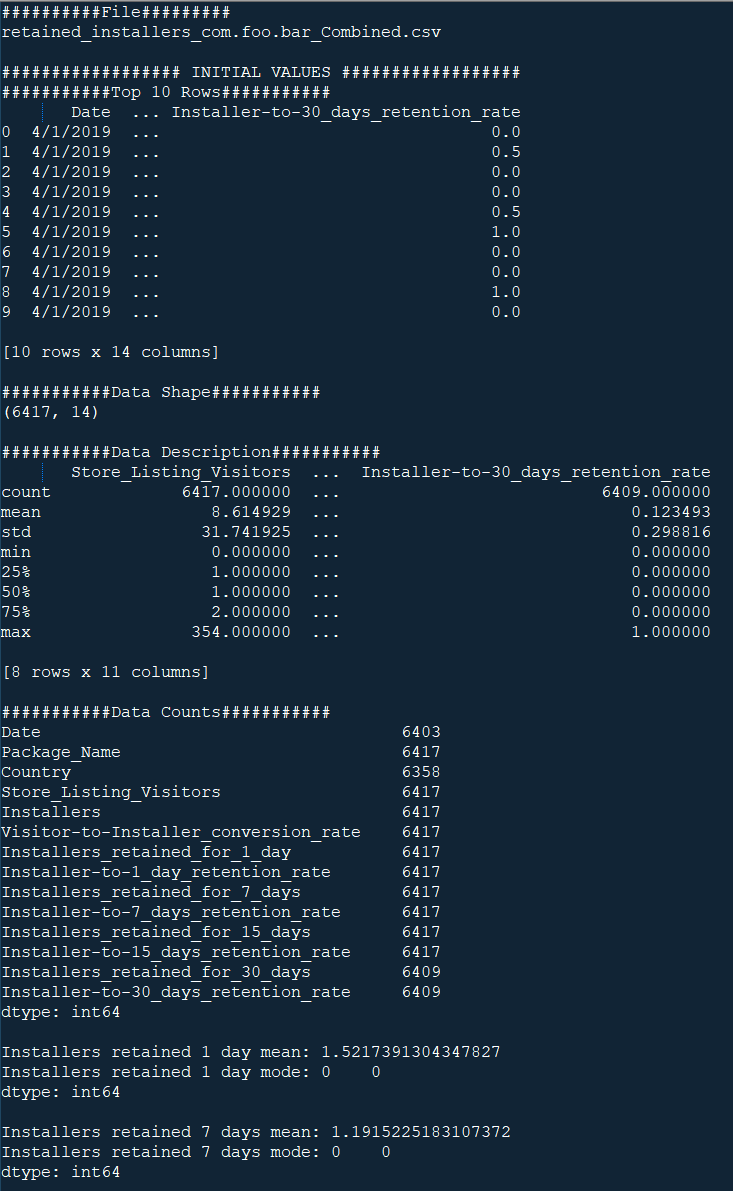
Installers retained 30 days mode: 0 0

dtype: int64

**Conclusion:** The initial data contains many missing pieces of information that will affect the exploratory statistics. The average retention of the application is below 2 with many timeframes having less than 1 as the retention.

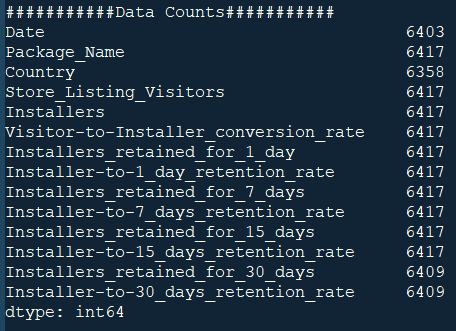
1. **Combined Files**
2. **Subset of code and execution files from the final explorer files.**



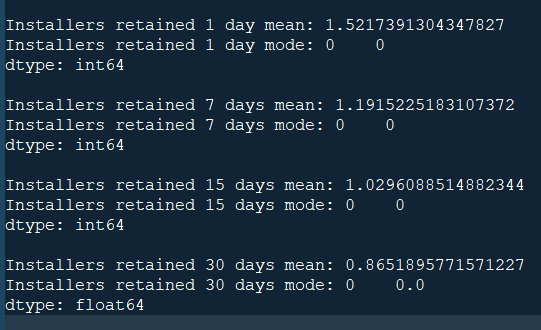


1. **Dataframe Shape:** 6417, 14

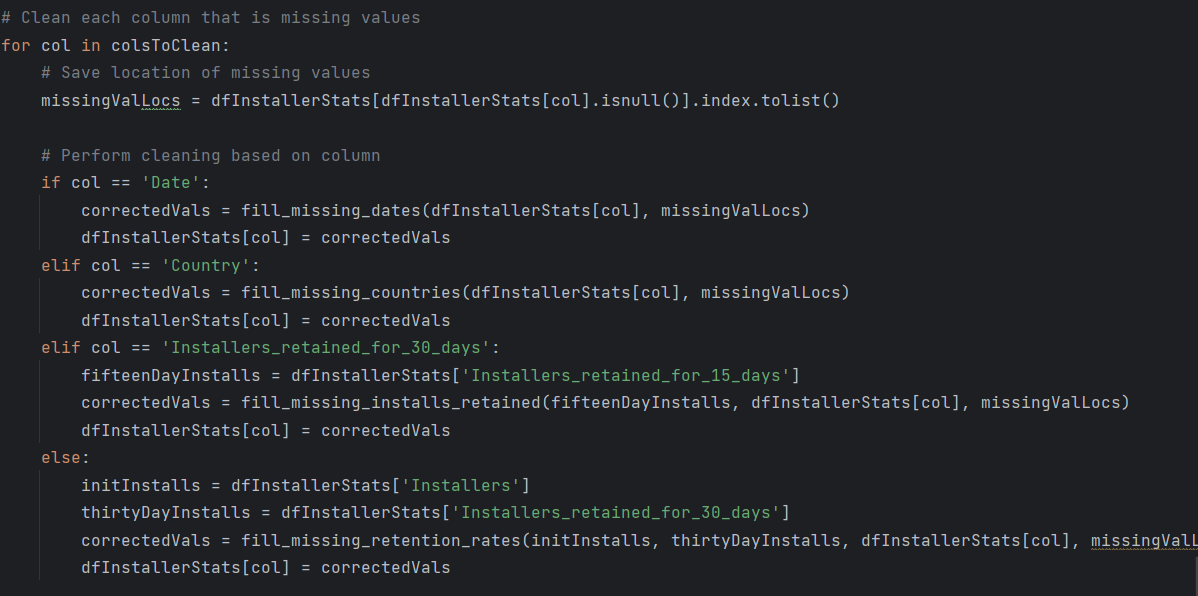
**Data Counts:**

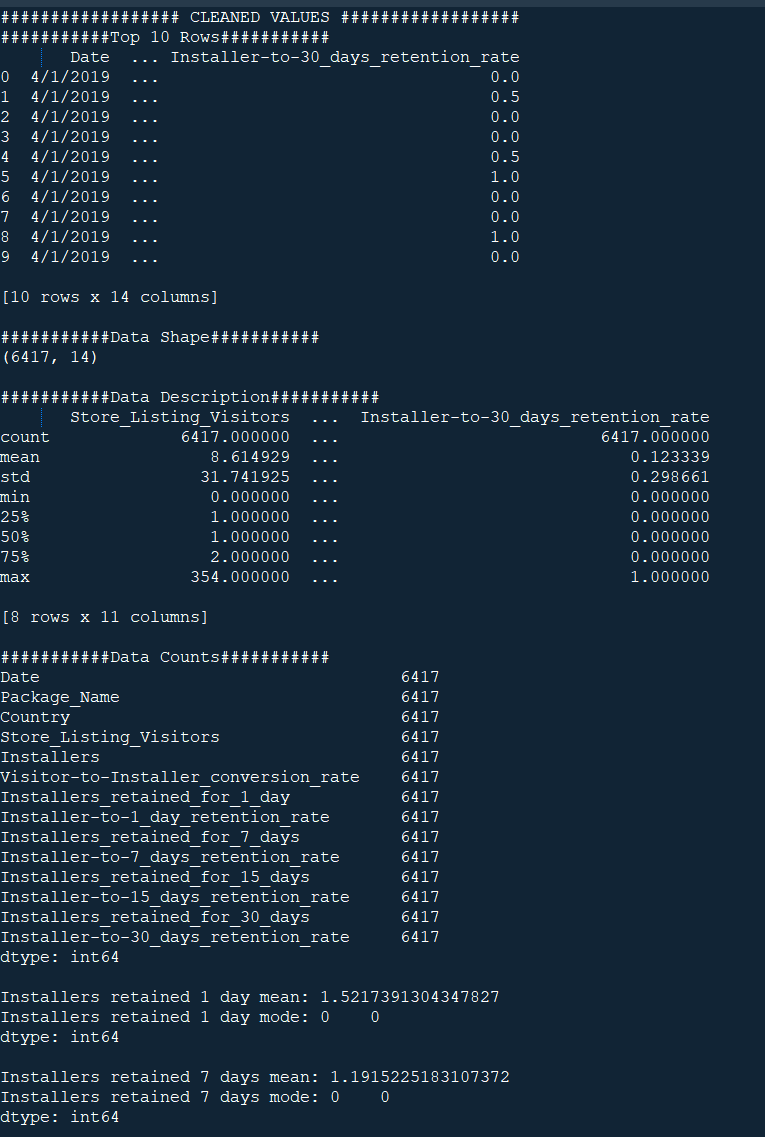
****

**Means and Modes:**

****

1. **Data Cleaning Code Snippet and Output**

****

****

1. **Data Cleaning Process**

The process I took includes finding the number of missing values for each column and then looping through the columns that are missing data. Four functions are created to handle filling in the data for each of the four columns that have missing information.

For the date column, the previous and next dates are found. If the two dates match, then that date is used in the missing location. This fulfills the needs for this dataset, but more logic would be needed if any of the dates were at the end or beginning of a month.

For the country column, an initial list of all countries is built first. The function then finds the location of the missing value and finds the country name in the previous location. The index of that country is found in the overall country list and then the next country is used to fill in the missing value of the data frame.

The missing installs retained for 30 days uses two options to fill in the data. First, if the amount of retained installs at 15 days is 0, then that number is used. This is because the number cannot go back up from 0. If that isn’t the case, then the mean of the non-missing values within the data frame is used as the replacement value.

The missing values in the 30-day retention rate column are found by calculating the quotient of the 30-day install retention and the initial install number for that row.