Course Project - Airport Restaurant Data & Dataset (Module 2)

Shaun Pritchard

Rasmussen College

QMB4000

Benjamin Tasker

November 22, 2020

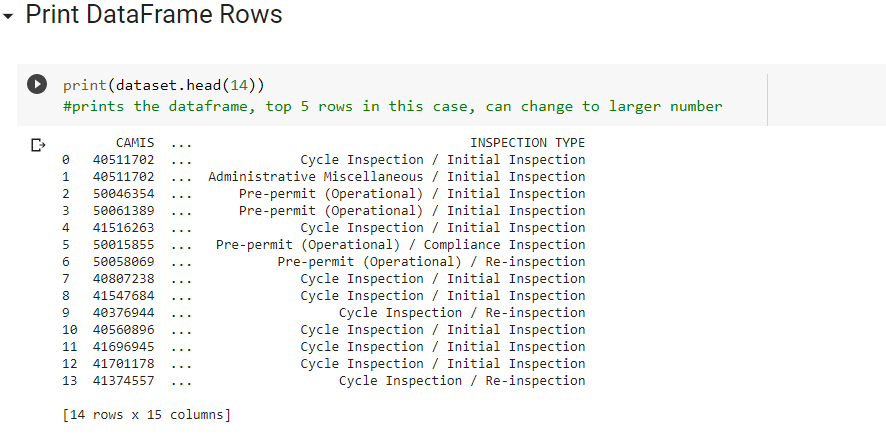
Course Project - Airport Restaurant Data & Dataset

In the airport restaurant that has I found that there were many no values and missing data values than the columns of data. This is the strategy I used to implement correcting that data with pythons’ tools and libraries (NumPy, Pandas, and Scikit learn) for imputing and validating training data based on which data types, columns, and cleaning of null values (missing data).

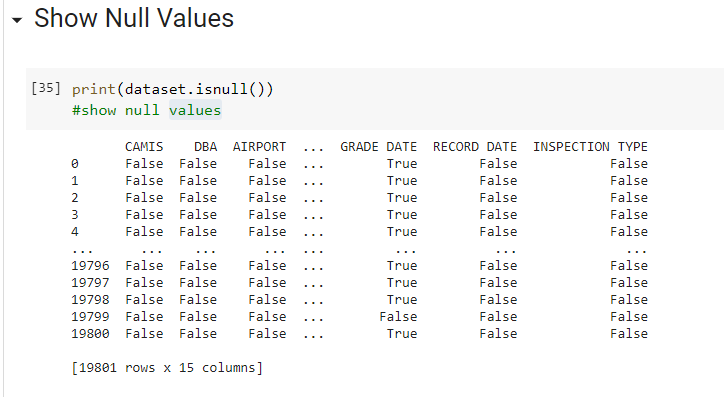
The first thing I did was to import the panda’s libraries for Python and import the data set for the flight restaurant data.



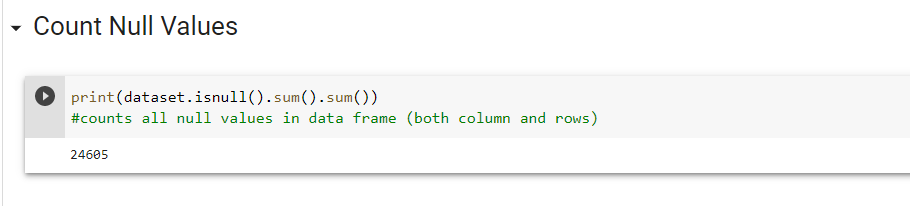
After the Imports, I created the data frames which brought the data into my working environment.



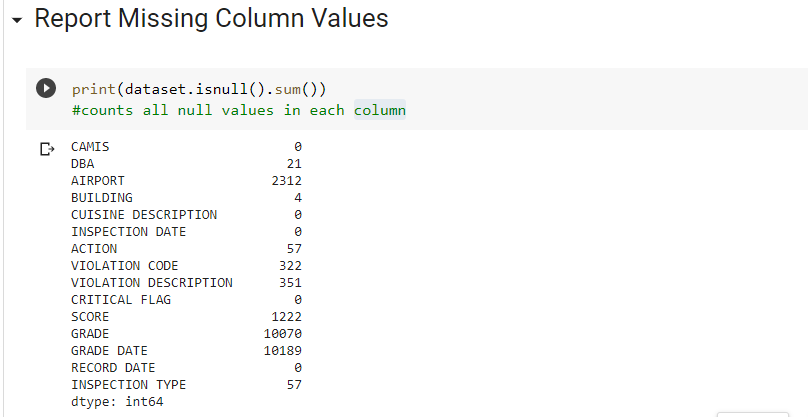
I then transcribed and ran through all of the columns of data to find any null values.



I then calculated all the null values in the dataset



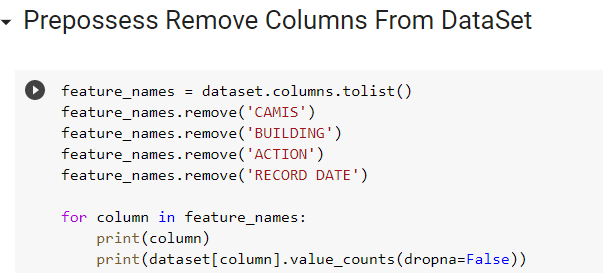
I then ran a report for missing values using the current context reporting this data below.



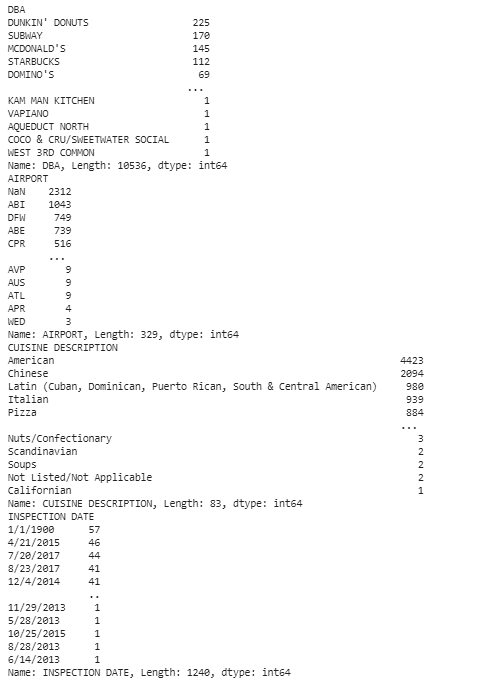
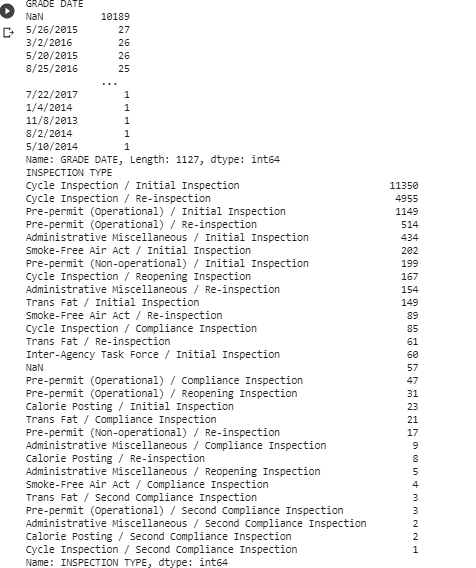
This is a breakdown of how I decided to fix the missing and NULL values. I have highlighted the necessity of the data in the actions to be taken upon the data and further steps.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Null values** | **Actions** |
| CAMIS | 0 | Not viable to use for calculation DELETE |
| DBA | 21 | Categorical under 10% replace with most (most-frequent) |
| AIRPORT | 2313 | Categorical under 10% replace with most (most-frequent) |
| BUILDING | 4 | Categorical under 10% replace with most (most-frequent) |
| CUSINE DESCRIPTION | 0 | No missing data |
| INSPECTION DATE | 0 | No missing data |
| ACTION | 57 | Not viable to use for calculation DELETE |
| VIOLATION CODE | 322 | Categorical under 10% replace with most (most-frequent) |
| VIOLATION DESCRIPTION | 351 | Categorical under 10% replace with most (most-frequent) |
| CRITICAL FLAG | 0 | No missing data |
| SCORE | 1222 | Numerical under 10% replace with most (mean) |
| GRADE | 10070 | Categorical under 10% replace with most (most-frequent) |
| GRADE DATE | 10189 | Numerical under 10% replace with most (mean) |
| RECORD DATE | 0 | No missing data DELETE |
| INSPECTION TYPE | 57 | Categorical under 10% replace with most (most-frequent) |

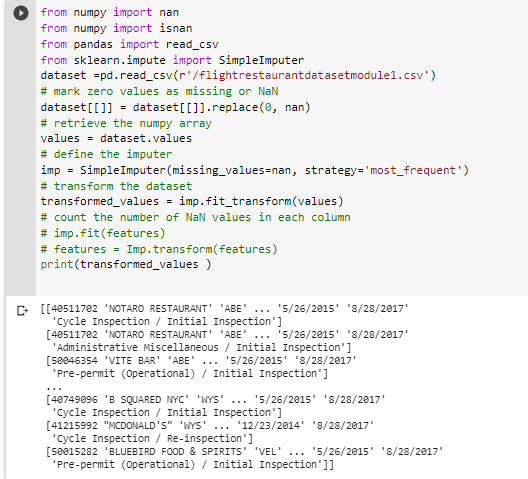
There are several Columns of data I needed to remove they were redundant and held no viable information that could be used to predict Analytics from this dataset.



I then looped through all the existing rows of data that where not removed in the data frame to calculate feature vectors.

I then ran actions to fill and no values on numerical data when to replace Common Core ask Oracle. That was under 10% or less missing from the dataset columns. I've been imported sci-kit learn simple cheer library to run invitation on the datasets to replace the missing values based off most frequent values per column that were missing I've been transformed that data into a data set array vector and output each row of data based on the new vector columns of data with the replaced null and missing values.



**Choosing a strategy:**

I was able to pick the strategy I was going to implement based on the amount of categorical data that I found relevant to the analysis the airport CEO suggested or possibly needed. Basically, determine the viability of independent factors and dependent variables that could be applied for analytics and removing redundancy. Many columns of data were just redundant and serve no purpose and serving metrics other Columns of data or categorical and can be used to serve as the dependent variables in future analysis.

There are 14 columns of data, I first took it upon myself to visually review the data. I found the CAMIS, BUILDING, ACTION, and RECORD DATE columns served no purpose I have no idea what CAMIS is besides an identifier for the rows of data. while BUILDING data would not be relevant given that we have the airport the rest exist in. the data in the ACTION column only provided 2 data types a categorical data answer.

Essentially, saying that they were violations when in retrospect every restaurant on the page exists in the dataset because they had a violation. I felt this information was redundant. I remove the column. RECORD DATE was also redundant there is no reason to keep the date that the incident was recorded when there was an actual date for the violation and grade for the restaurant in question.

I also contemplated removing either the score or grade column but I thought you going to be valuable including the fact that scores numerical data which there is not that much numerical data in the data set. Did I felt that grade to be a dependent variable used to categorize the violations at different stores to compare their average cumulative grade or to use with the score, to know which scores accumulate to being a specific passing grade or not?

For the rest of the data, I used mean-variance calculations to fill in the missing data, or I used the most frequent value for categorical data to replace the missing data. Missing values can be replaced by the mean, the median, or the most frequent value using the strategy hyper-parameter. The median is a more robust estimator for data with high magnitude variables which could dominate results *(Kaggle, 2020).*

**Normalized Dataset:**

The normalized data set that I selected was from a Kaggle challenge called Communities and Crime Data Set (Normalized). The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units. Data is described below based on original values. and there are missing null values. So yes, they do exist and depending on the data type and analysis being completed on the data they stop the trained model I am how you deal with these values. Some data could be removed while other data could simply handle multivariate imputation. using the statistics (mean, median, or most frequent) of each column in which the missing values are located *(Kaggle, 2018).*

**Non-Normalized Dataset:**

For this data set, I chose the set called Single non-normalized data of electron probe analyses of all glass shard samples from the Seward Peninsula and the Lipari obsidian reference standard. the set is needed to be used to infer reconstruct temporal patterns of early to late Holocene permafrost and thermokarst dynamics, site-specific on palaeo-records. After reviewing the data set there is missing value and no values in the data set very few compared to a normalized database, I would use the same methods of multi variance imputation column row removal and pre-processing techniques *(Mirror of Pangaea, 2019).*

# References

Kaggle. (2018, Dec 18). *Communities and Crime Data Set (Normalized)*. Retrieved from https://datasetsearch.research.google.com/: https://datasetsearch.research.google.com/search?query=Normalized%20Dataset%20with%20missing%20values&docid=7IoXMgpPB3exmv3gAAAAAA%3D%3D

Kaggle. (2020). *Handling Missing Values (Imputation Example)*. Retrieved from https://www.kaggle.com: https://www.kaggle.com/mrshih/handling-missing-values-imputation-example

Mirror of Pangaea. (2019, Nov 23). *Single non-normalized data of electron probe analyses of all glass shard samples from the Seward Peninsula and the Lipari obsidian reference standard*. Retrieved from https://datasetsearch.research.google.com/: https://doi.org/10.1594/PANGAEA.859554