Data Wrangling Report

1 Table of Contents

2 Introduction

2.1 Background

Real world data is messy and untidy with no real structure or schema. Data wrangling is a vital step in the data analysis process as it heavily impacts the results of the analysis and visualization.

This projects intent is to simulate in between where real world data has been structured to suit the provider's needs, gathering to put to practice that which has been learnt academically.

2.2 Purpose

The purpose of this project is to practically apply the data wrangling process onto a real world application.

This report will analyse the data of the raw files prior to and after cleaning to signify the importance of data cleaning.

3 Exploratory Data Analysis (EDA)

3.1 Analysis

Understanding the data is critical in the data analysis process as it sets the pace for the remainder of the project. Knowing what to present in the visualizations will provide a pathway in how the cleaning process will be undertaken, the priorities, etc.

Instead of manually visualising the cleaned data, I set out to work smarter and looked for automated visualization libraries to assist in the EDA process.

My findings led me to the discovery of **SweetViz** and **Pandas Profiling**.

3.2 Visualization

3.2.1 Pandas Profiling

This visualization report provides a basic analysis of the data frame and is suited to both the data assessing and the start of the EDA process.

The missing values matrix diagram is quite useful, as it gives a rough indication of which part of the data frame have missing values.

For categorical and/or string datatypes, it provides summary of distinct as a number and percentage, similarly with Missing as well as the memory capacity required to hold the values within the column/series.

Depending on the quantity of unique values, a chart is provided showing the distribution, for instance the source_app column is shown as a pie chart.

Number data type reports are similar to categorical/object data types, however the results also include top and bottom 5 results.

3.2.2 SweetViz

The next visualization library provides a report that is extensive and slightly interactive.

With regards to formatting, this report is quite straight forward, space is better utilised allowing for more data to be displayed.

With text columns, there are string size limitation when displayed.

4 Findings

From the data cleaned, the reports highlight the following:

4.1 Twitter Data

- Between the periods of 2015 to 2017, the majority of the original tweets originate from an iPhone user/s with Tweet Deck being the least popular Platform.

4.2 Image Predictions

- Of the top 3 predictions made, the AI program was confident in the first image of four images provided. This is supported by the histogram provided, where around 140 entries were considered to be 100% in its evaluation confidence.
    - From the first image, 1532 were considered to be dogs.
- Across the three predictions, retrievers were the most reoccurring dog identified.
- Prediction 2 resulted with more dog classifications than either P1 or P2.
- P1 is the only left skewed histogram, whereas P2 and P3 are right skewed.

5 Conclusion

The approach undertaken to produce the visualization reports and findings is influential to future data analysis and science projects as it will save enormous amounts of time and effort on coding to purely focus on the analysis portion.

In hindsight, as this was a learning opportunity, the visuals were probably best approached manually to emphasis the learning of code to:

- Produce the visuals,
- Trial and error of code written to better understand the syntax
- Use the appropriate data types,
- Become more familiar with python the libraries and there functions.

Sweetviz provides a significant potential for quick analysis, if the columns were identical in both quantity, naming and datatypes after cleaning. We would have been able to do a comparison on the one report instead of the present workaround of comparing the raw data input vs the cleaned data input.

There are many libraries readily available and created by others that are able to assist us in during the entire data analysis process. Combining the knowledge obtained from the course to be able to understand what tools we require, know to implement them to our projects, and keeping up with the pace of new technology improve as well as working smarter is the general workflow I wanted to demonstrate in order work more efficiently.

There are many libraries readily available and created by others that are able to assist us in during the entire data analysis process. Combining the knowledge obtained from the course to be able to understand what tools we require, know to implement them to our projects, and keeping up with the pace of new technology improve as well as working smarter is the general workflow I wanted to demonstrate in order work more efficiently.