wrangle_report

July 22, 2020

1 Wrangle Report

1.1 Introduction

This project was based on wrangling a dataset and the dataset was WeRateDogs data from twitter. In order to performing wrangling of data, three main steps are involved 1. Gather Data 2. Assess Data 3. Clean Data

Once these steps are performed, it makes it very easy to analyse and visualize the dataset.

1.1.1 Gather Data

There were 3 different kinds of datasets involved. 1. Twitter Archive - An archived data file as a .csv available for download on the Udacity database server. This file had to be downloaded and read as a .csv file.

2. Image Prediction - A Url link was provided where the image prediction .tsv file was available. This file had to be programmatically downloaded with the help of python request libraries

3. Twitter API - A file with the favorite and retweet counts had to be gathered from twitter. This was done with the help of twitter API to query the favorite counts and retweet counts for the tweet IDs that we had in the archive .csv file

1.1.2 Assess Data

In the second step of assessing the data, the action was performed in two ways:

1. Visual Assessment - Here, the datasets were queried with the help of .head(), .tail() and their data was assessed. A lot of quality and tidiness issues were found and noted. After performing the visual assessment, the second kind of assessment was carried out
2. Programmatic Assessment - The datasets were again looked into to find their datatypes with the help of .info() and .describe(), if there were duplicates with .duplicated() and value counts of certain columns with .value_counts() With the help of programmatic assessment some more quality issues were found and they were noted.

Note - Not all issues identified in this step have been cleaned in the next step.

1.1.3 Clean Data

Cleaning again involves three steps, | 1. Define - Here, the issue identified in the assessment is defined 2. Code - The code that addresses the issue is executed 3. Test - Verification if the desired result is achieved While performing cleaning, the idea is to first address missing data, followed by tidiness issues and then quality issues.However, sometimes tidiness and quality issues had to be addressed together. To clean the code .merge(), .melt(), .split(), .extract(), .to_datetime, .astype(), .drop() have been used

My Experience While performing data wrangling, I felt it was faily easy to gather and assess data but not very easy to clean the data. This is because while cleaning the data, I wanted to ensure not to lose a lot of data due to quality issues and wanted to fix as many as possible. The more data available the more accurate the findings will be. Secondly, as I performed data cleaning, I felt the need to assess the data more and clean it more. I am aware that my cleaned dataset is not cleaned perfectly and there are quality issues that can be cleaned further. Having said that, I have cleaned more than 8 quality issues and 2 tidiness issues

1.2 Conclusion

With quite a clean data, I had enough data entries to carry out analysis and visualization and draw some conclusions.