Project:

Wrangle and Analyze Data

Wrangle Report

Data Analyst

<u>Presented by:</u>

Abdulaziz Alqumayzi

Table of Content:

1 Introduction

Wrangle and Analyze Data project goal is to wrangle **WeRateDogs** Twitter data to create interesting and trustworthy analyses and visualizations. Using wrangling phases which are gathering, assessing, and cleaning. This report is to show my effort in the wrangling of the data given by Udacity.

2 Wrangling Data

2.1 Gathering

Three files were given by Udacity:

1- image-predictions.tsv
2- twitter-archive-enhanced.csv
3- tweet-json.txt

The first file was collected programmatically using the requests library. The others downloaded manually from Udacity website. Unfortunately, I couldn’t use tweepy because twitter didn’t approve on my developer account.

2.2 Assessing

I assessed these files visually and programmatically.

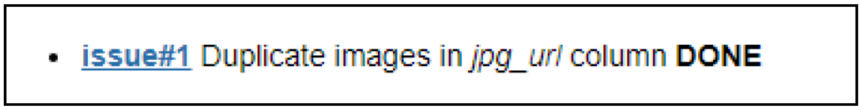
Visually I used Pandas functions (such as head and tail) and Excel for some values. Programmatically I used many Pandas functions (such as info, value_counts and query)

I found many quality and tidiness issues. 14 in total. 12 quality issues and tidiness issues. Some of the quality issues are duplicate images, many of variables (or columns) not be used, wrong data types and inaccurate data.

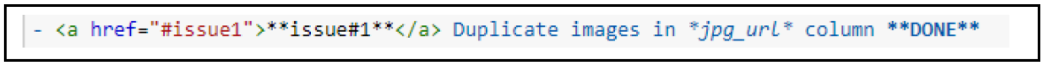
2.3 Cleaning

Before the cleaning phase, I made copies from the original files to work on. This step was very helpful. I made some mistakes in coding, but this step saved me a lot.

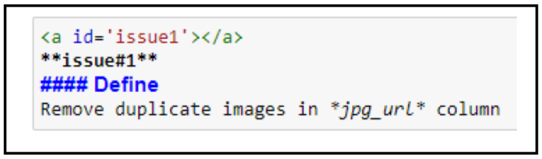
I assigned every issue with an issue number and made it easy for you by clicking the issue number it will get you to the issue Define , Code and Test procedure to the issue you clicked on.



- issue#1 Duplicate images in *jpg_url* column **DONE**



```
- <a href="#issue1">**issue#1**</a> Duplicate images in *jpg_url* column **DONE**
```



```
<a id='issue1'></a>
**issue#1**
#### Define
Remove duplicate images in *jpg_url* column
```

The most important issue is duplicates. I solved it first. Another issue I found is clearly wrong dog names (such as “None”, “such”, “not” and “this”), I used replace function to replace all wrong names to null.

The numerator and denominator ratings were important to my analysis, I did not delete wrong data values in numerator and denominator immediately. Even though they contain inaccurate data I tried to solve this problem. For the denominator, values should be 10. But I found a few records that are different from 10, so every denominator values above 10 were decreased to 10 and denominator values below 10 were deleted. For the numerator, I only change values that are above 10 to 10. Because it is possible that WeRateDogs makes higher rating for fun or joke. There were inaccurate numerators that’s have different values from text. In text most of them around 10 and above so decide to make them all 10.

3 Conclusion

Data wrangling is challenging phase in the Data Analysis process. It took long time, gathering was easy because the files were available. Assessing was the longest procedure, especially for a guy like me that are looking on everything to find quality issues. Cleaning some issues were hard and some very easy.