Udacity - Data Analysis

Wrangle Report

All About My Wrangling Efforts

Introduction

When gathering data from any source, it is usually always messy. Data can be classified as messy in many different ways. Values can be duplicated, entries can be missing, there could be spelling mistakes or other formatting errors, and sometimes the data can even be outdated or incorrect. This is why it is important to assess the data that is collected, so that the errors are known before analysis. There is often no way around gathering messy data unless someone has already cleaned it. Cleaning data is a process used in data analysis to fix all the errors associated with messy data. Whether that be dropping a duplicated column, filling in missing entries, or correcting formatting mistakes, cleaning data is essential to data analysis.

Gathering Data

The first step in wrangling data is to gather the data. For my project, I needed to gather data from three different sources. The first source was relatively easy as it was a csv file that can easily be read using a library called pandas. Next, I needed to extract data from a url that was given. To do that, I used a library called requests which let me extract the data into a tsv file which could be read using the same function as the csv file before. Lastly, I needed to extract data from Twitter. First, I had to create a twitter developer account which had to be reviewed and authorized. Next, I received keys and tokens needed to access the archive. Lastly, after accessing the archive, I was able to create the last data frame needed for this project and was ready to move on.

Assessing Data

The next step in the wrangling process is to assess the data. After gathering all the data that is required, it is imported to assess the data to get a better understanding of how to

clean it and analyze later on. For my project, I looked at each dataset separately and took notes. I kept an eye out for spelling mistakes, formatting errors, and incorrect data types. I then looked for duplicates and null values, which in my case, there were not many. Lastly, I looked into any issues that were directly related to the dataset itself. For example, in my archive data frame, there were names that were incorrect and thus, needed to be removed. After I had fully assessed the data, I wrote out a bullet list of everything that needed to be fixed in the next step.

Cleaning Data

The last step in wrangling, is the process of cleaning the data. After the data has been assessed and the issues written out, it is time to do the work. This is the most time consuming part of data wrangling as there are often many steps to do. I tried to be most efficient in my cleaning by again separating the work and working on one dataset at a time. In this process, I changed incorrect data types such as switching from an integer type to a string type. I took four separate columns that classified four different types, and combined them to be under one column. I also removed certain columns that I deemed unnecessary for analysis. Additionally, I removed entries that were not along the guidelines for my project such as entries without images or retweets. After I completed my cleaning list, it was time to combine the three datasets into one to create a master dataset. I did this by merging them with a common column which was the twitter id. Concluding my efforts, the dataset was now complete and ready to move on to the analysis and visualization process.