Julia Revier

Data Analysis Bootcamp

2 March 2020

ETL Project Technical Write-Up

1. Extract:

I used and manipulated Olist store files found on Kaggle. I used two distinct datasets; one tracking marketing data and one composed of Ecommerce transactions. All of the files were CSV files and were relationally linked by various ids. I chose to manipulate the closed deal file from the marketing funnel dataset and the order and review files from the Brazilian Ecommerce dataset. These two datasets are relationally linked through a seller id. My goal was to merge and streamline the datasets so that a marketing team could view a relationship between review scores and different business segments.

1. Transform:

Once I uploaded the files, I read them into dataframes using pd.read\_csv. There was not a ton of cleaning that needed to be done, but I did remove unnecessary columns from the data frames and removed duplicate ids. Next, I joined the review dataframe with the orders data frame so that I could link the two on order ID and obtain the seller ID from the order csv file. Then, I reset the index on the review data frame to the review ID. I also reset the index on the closed deals csv to the seller ID.

1. Load:

I decided to load my data frames into SQL. First, I made tables in SQL for the review data and the order data. I only created columns that I extracted during the transformation phase of this project. Then, I established a SQL connection in my Jupyter notebook and exported each of the dataframes into their respective table. Finally, in SQL I merged the two tables on seller id and extracted the columns that interested me; the business segment and the review score. This way, a firm can run analyses on correlations between different business segments and customer review scores.