ETL Project

Apple App Store vs Google Play Store



Last Modified May 11, 2019

| Document Revision History | | | |
| --- | --- | --- | --- |
| Revision | Date | Rev. by: | Remarks |
| 1.0 | 05/06/2019 | Ellise Carpenter, Rohith Bhattaram, Jonas Haskins, Kavya Shabnavees |  |

**Table of Contents**

[1 Project Description 4](#_Toc8066597)

[2 Extract 5](#_Toc8066598)

[3 Transform 6](#_Toc8066601)

[4 Load 8](#_Toc8066603)

[5 Analysis 9](#_Toc8066605)

# Project Description

When it comes to mobile ecosystems, there are two giants locked in a battle, not only for revenue, but also for the hearts and minds of developers and consumers alike.

The purpose of our project is to extract, transform and load data sets from Google’s Play Store and Apple’s App Store. Ultimately, the transformed data set will allow for better understanding of the relative variances between vendors.

# Extract

## We used 2 datasets from Kaggle. Both data sets were the most recent available and were originally obtained from via web scraping. The specific datasets used for the project are as follows:

[Kaggle - Google Play Store Stats](https://www.kaggle.com/lava18/google-play-store-apps)

[Kaggle - Apple App Store Stats](https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps)

## Convert CSV to Pandas Dataframe



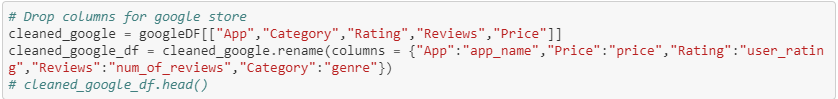
# Transform

## File Clean-up with selected columns

Drop columns for Apple store



Drop columns for Google Store

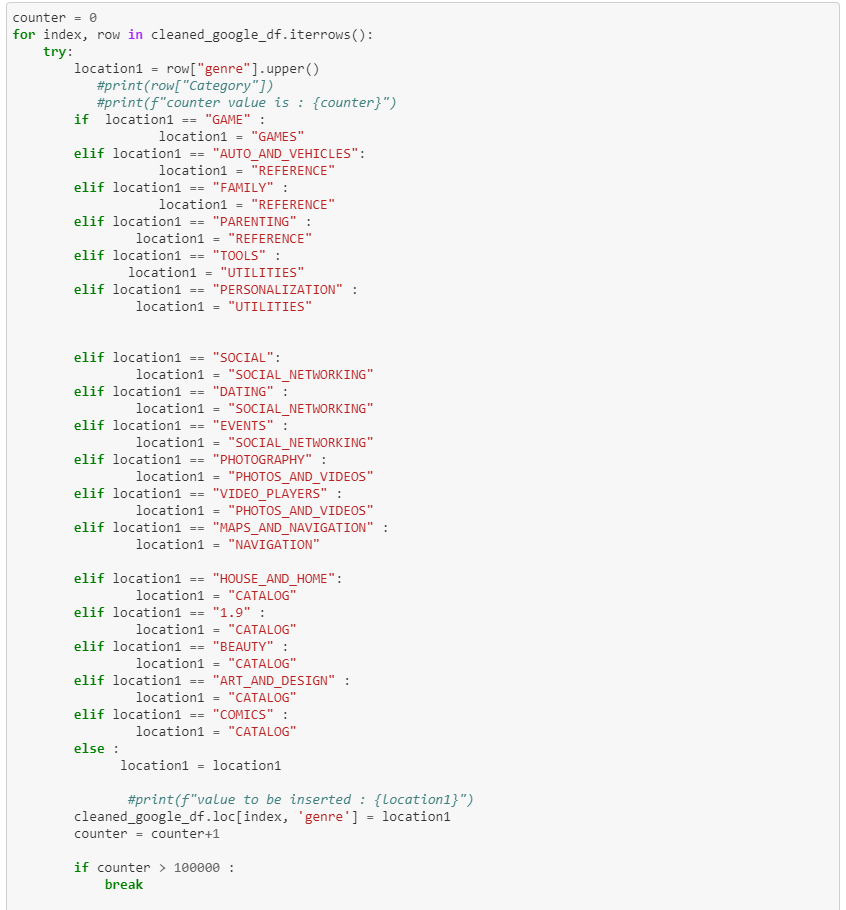


Removing special characters





Normalizing Google Categories

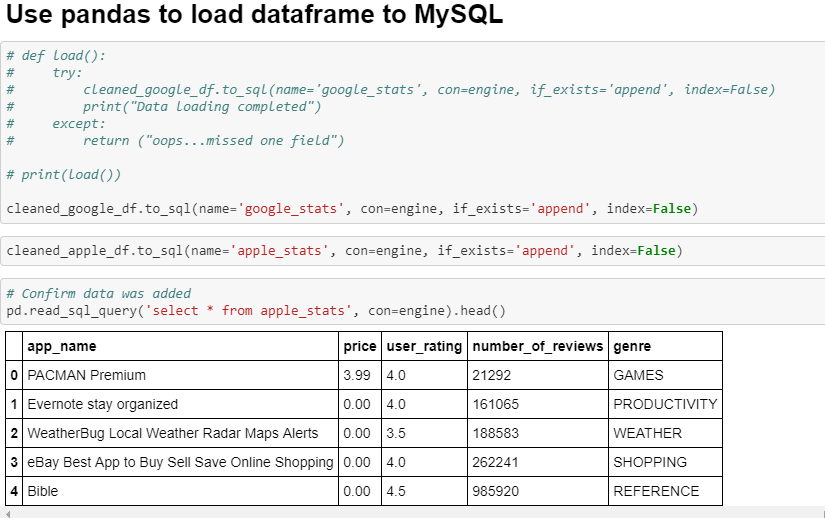


Normalizing Apple Categories



# Load

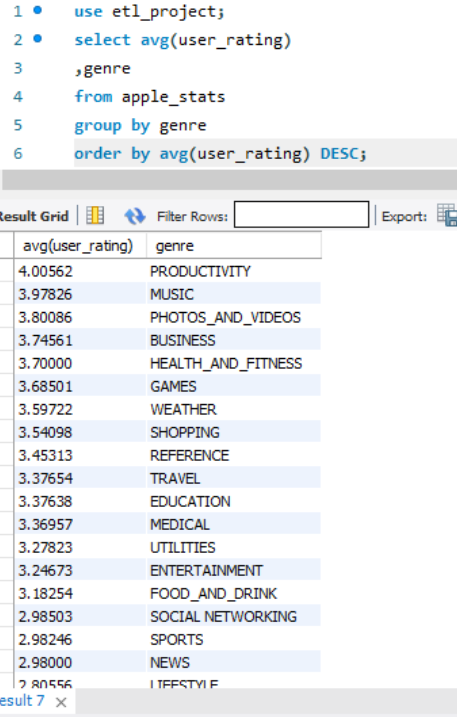
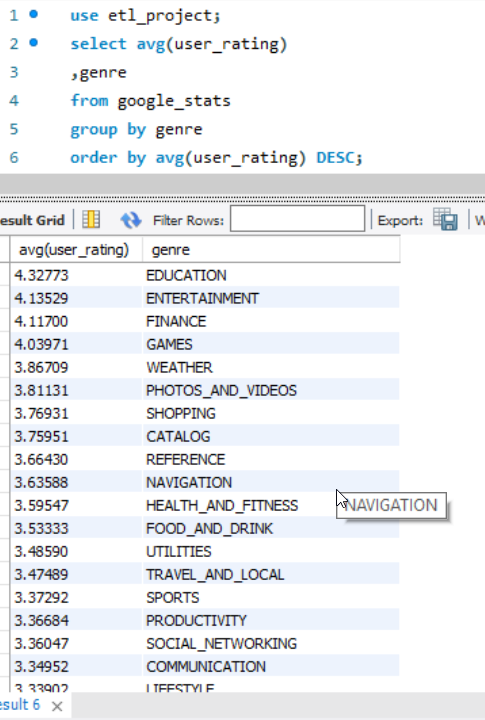
## The last step was loading our final data frames into Database. We created a MYSQL database and respective tables to match the columns from the final Panda’s Data Frames and then connected to the database via SQLAlchemy.



# Preliminary Analysis

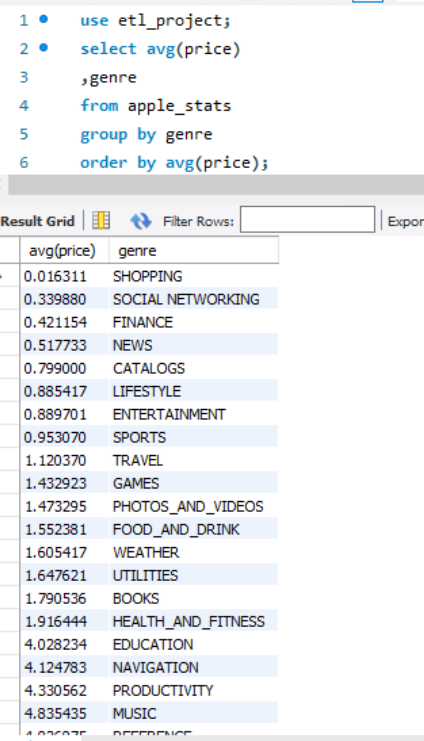
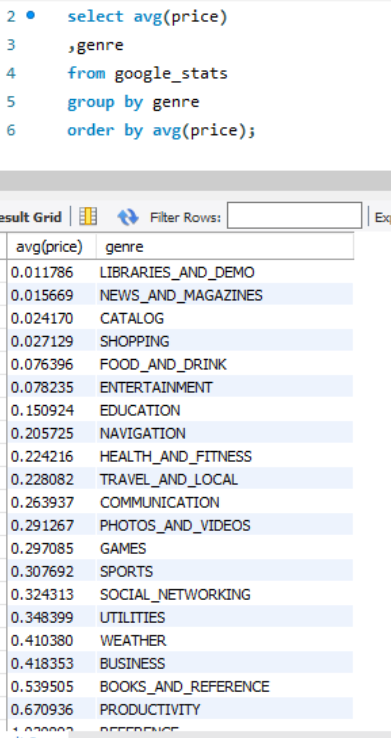
**Genre Rating (Avg)**

Apple Google

**Genre Price (Avg)**

Apple Google

**Highest Rated Apps**

Apple Google

