Project 4

Paul Ngouchet

During this project, I will downloading 3 Kaggle Data and analyzing them using Google Colab

1. Netflix Userbase data - Netflix_Userbase.csv
2. Anime Dataset - MAL-manga.csv
3. Instragam Threads Reviews - threads_reviews.csv

The files provided will 3 ipynp notebooks and the datasets used.

The best way to RUN those ipynp notebook is by using Google Colab with GPU enabled (T4 is free) but the fastest results are given when using V100 but you need to pay for it. Also before running each notebook, don’t forgot to upload the dataset.

Description of Dataset

Netflix Userbase

https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset

The dataset presents a snapshot of a sample group of Netflix users, offering a comprehensive view of their subscriptions, revenue, account particulars, and activity. Each entry in the dataset corresponds to a distinct user, uniquely identified by their

User ID. It encompasses valuable information, including the user's subscription level (Basic, Standard, or Premium), the monthly revenue derived from their subscription, the date of their Netflix enrollment (Join Date), the date of their most recent payment (Last Payment Date), and their country of residence.

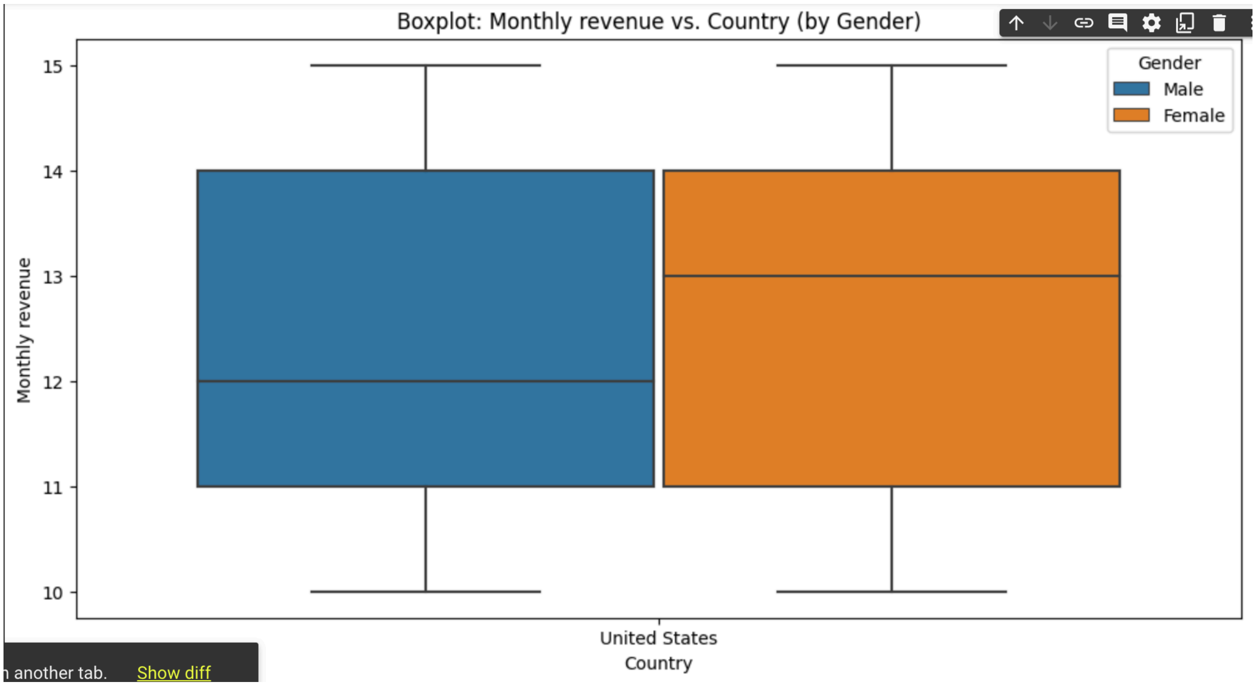
Furthermore, the dataset incorporates supplementary columns to provide insights into user behavior and preferences. These additional columns encompass Device Type (e.g., Smart TV, Mobile, Desktop, Tablet), Total Watch Time (measured in minutes), and Account Status (indicating whether the account is active or inactive). It is essential to note that this dataset is artificially generated and does not represent actual Netflix user data. Researchers and analysts can utilize it for analytical purposes and modeling to gain an understanding of user patterns, preferences, and revenue generation within a hypothetical Netflix userbase.

Libraries used

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from IPython.display import display
import ipywidgets as widgets
%matplotlib inline
```

In this visualization, I am comparing the money spent on Netflix by Man vs Women by country.

There are 2 global filters to automatically update the charts. Gender and Country. Using this Boxplot, we can clear see that Women spend more than men in terms of Median. ($ 13 vs $ 12)



Anime Dataset

https://www.kaggle.com/datasets/nikhil1e9/myanimelist-anime-and-manga

The data presented in this dataset was obtained by extracting information from MyAnimeList and encompasses comprehensive details about all the anime and manga currently listed on the platform. The dataset is divided into separate files, with one dedicated to anime and another to manga. Below are the explanations of the features included in the dataset.

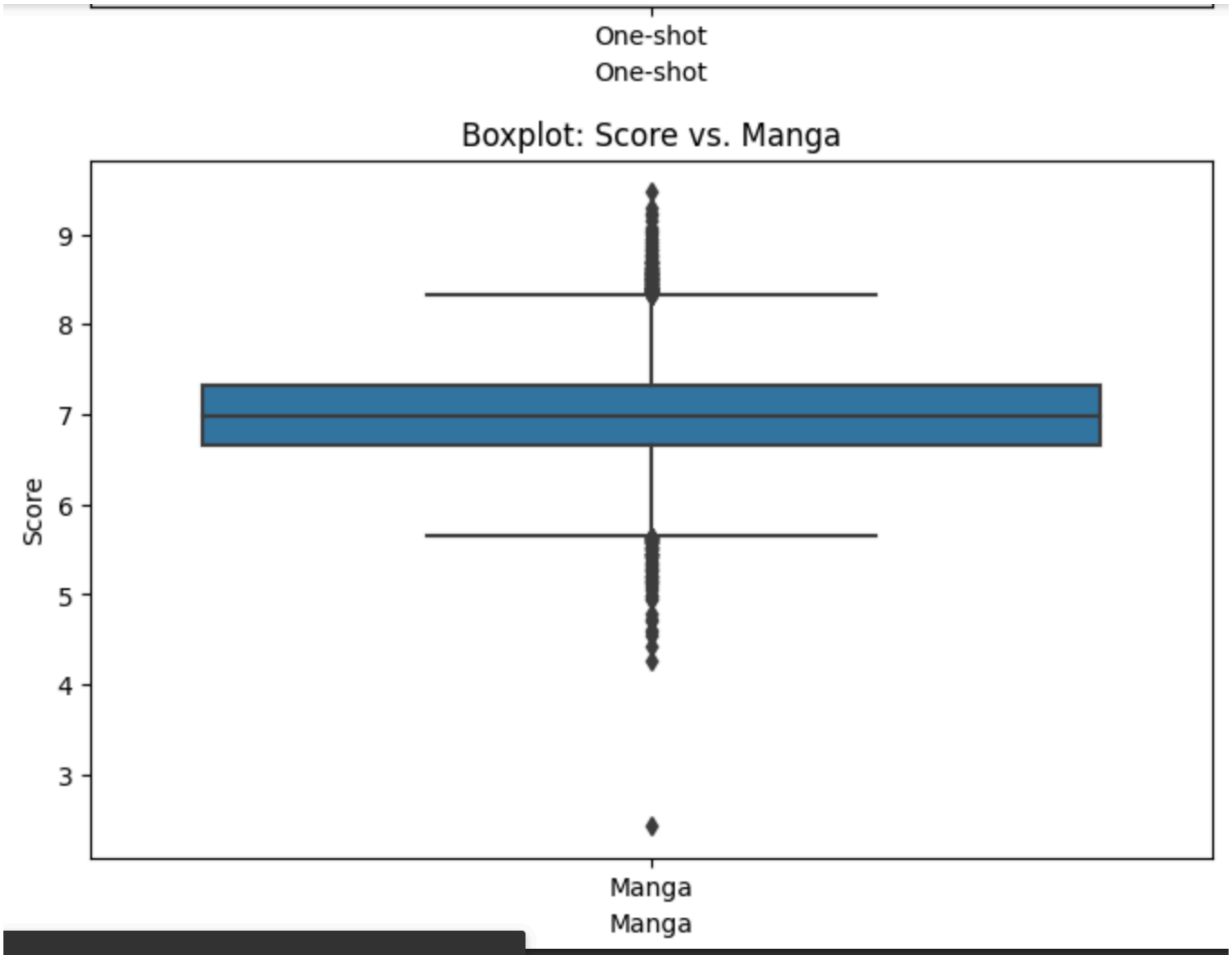
I am working with the manga dataset only.

Libraries used

import seaborn as sns

```
import ipywidgets as widgets
from IPython.display import display
import matplotlib.pyplot as plt
import pandas as pd

%matplotlib inline
```



In this visualization, I am using a Box plot to represent the score given to many type of anime like Manga and other types. The Global Filter is the type of anime watched which accounts to a good number. We can play around with this different types by using the ipynb notebook probvided.

Threads Reviews

https://www.kaggle.com/datasets/saloni1712/threads-an-instagram-app-reviews

The Threads, an Instagram App Reviews dataset, offers a wide-ranging compilation of user reviews extracted from the Threads mobile app on both the Google Play Store and App Store. It provides valuable insights into user sentiments and opinions. The dataset facilitates an in-depth analysis of user satisfaction, assessment of app performance, and the detection of emerging trends.
The data was obtained by scraping Threads App reviews from the Google Play Store and App Store.

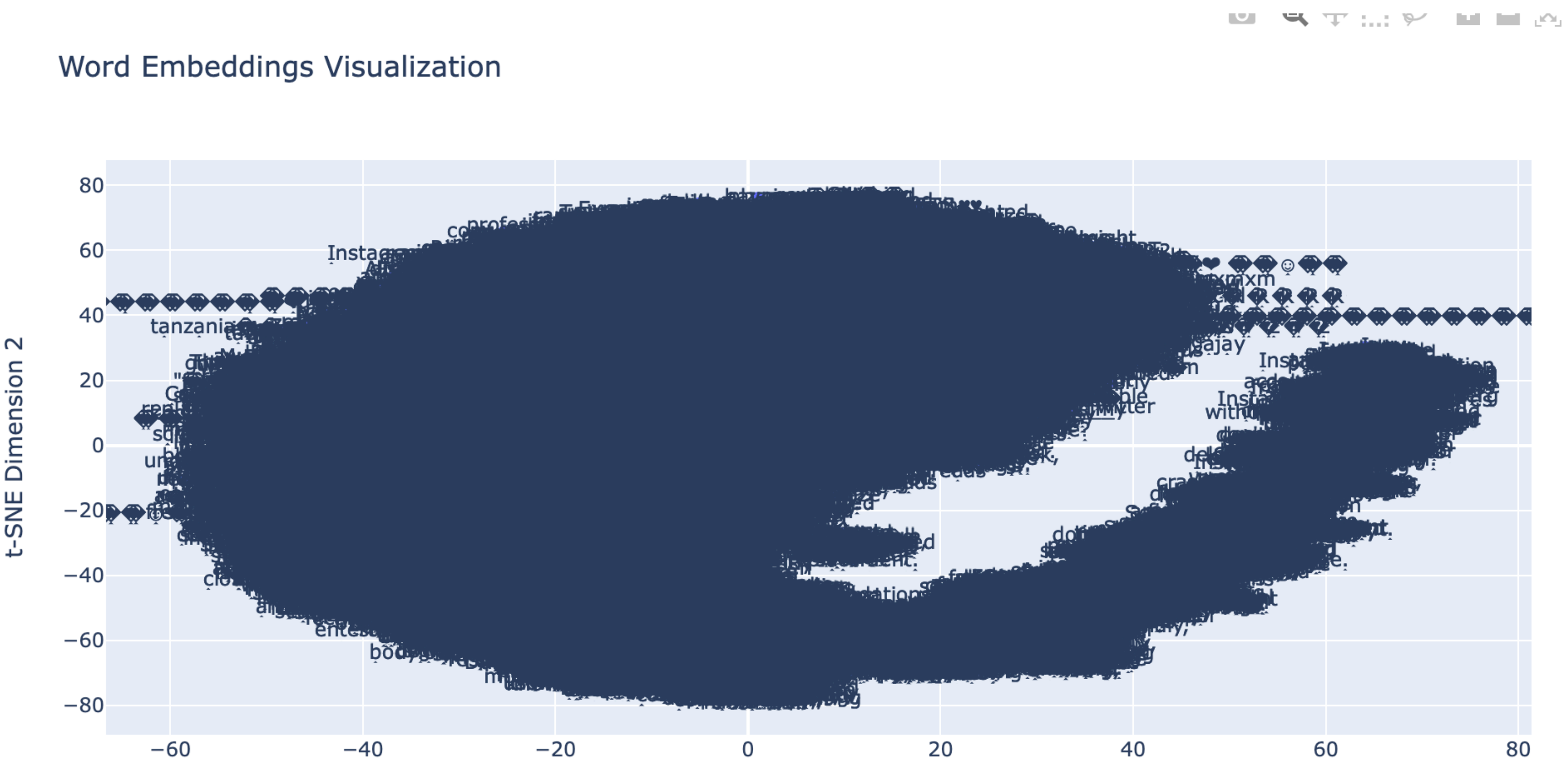
For this data I created a Word Embedding using Bigram (In the Bigram Language Model, we find bigrams, which are two words coming together in the corpus(the entire collection of words/sentences). For example: In the sentence, Expresso is awesome, and user-friendly the bigrams are : “Expresso is” “is awesome”)

and TSNE (t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.)

Librairies used:

```
import matplotlib.pyplot as plt
import scipy.sparse
import numpy as np
import ipywidgets as widgets
from IPython.display import display
import pandas as pd
from gensim.models import Word2Vec
from gensim.models.word2vec import LineSentence
from gensim.models.phrases import Phrases, Phraser
from nltk import bigrams
from sklearn.manifold import TSNE
import plotly.express as px
from gensim.models import Word2Vec
```

In this visualization I created a big word embedding of the new instragram app Threads review. I am providing a video in the project, so you can have a better understanding of how the graph. You can also come up with the same results by running the Ipynp notebook but since there a big amount of computation, it is better to use A GPU instance on Google Colab.



Conclusion:

As a data scientist, You need to look at type of datasets you have first. Then, After doing that you need to come up with a business story which makes sense depending on your use. Because the power of story telling is unlimited, you just need to guide your audience well.

References used:

https://stackoverflow.com/
https://seaborn.pydata.org/tutorial/categorical.html
https://colab.research.google.com/
https://matplotlib.org/
https://github.com/RaRe-Technologies/gensim#
https://www.nltk.org/
https://scikit-learn.org/stable/
google.com - In general to find insights how to build my different visualization