Declaration on Plagiarism

Name/s:	Prem Chandramohan Mehta, Dhruv Parimal Bhai Patel
Student Number/s:	22263922, 22264919
Programme:	MSc in Computing (Data Analytics)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	06/12/2022
Module Coordinator:	Dr Suzanne Little

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Prem Chandramohan Mehta Date: 05 December 2022

Name: Dhruv Parimal Bhai Patel Date: 05 December 2022

Most streamed song on Spotify

Group: HY

Abstract:

Spotify is one of the most used streaming services used to stream songs. So, we decided to find out which was the most streamed song on spotify, and which region/country streams that song the most. We used Tableau for the visualization. The visualization highlights that Shape of You is the most streamed song across all regions, followed by Blinding Lights with very low difference and Dance Monkey on the third position. It also revealed that the United States of America is the country with the highest number of Shape of You streams.

Dataset:

Link to the dataset: https://www.kaggle.com/datasets/dhruvildave/spotify-charts.

The dataset was available to download in csv format, so we did not have much to do get the dataset.

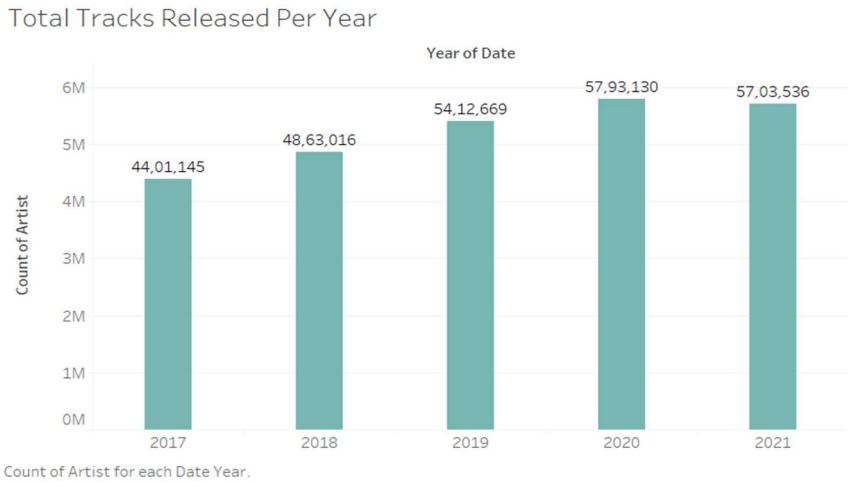
We went for the first V of big data, which is Volume. The size of our data is of 3.25 GB. It consists of 1,64,807 rows and various attributes.

The attributes are title, rank, date, artist, URL, region, chart, trend, and streams. The data types are string, integer, date, and category.

The dataset we used has only the top 200 and viral 50 ranking categories, so we had limited data. But this was the best data we found on spotify. Our dataset has data form 2017 to 2021.

Data Exploration, Processing, Cleaning, and/or Integration:

Data Exploration:



We made this chart for the purpose of exploring the data. This chart shows how many total songs were released by artists in a single year. We found that there were less songs put out by artists in 2021 than in 2020, other than that, it followed a trend of increasing songs released every year. We also tried a couple of a circle graphs while exploring the data, but it was too cluttered and not easily understandable, so we skipped it.

Data Processing and Cleaning:

What did you need to do to prepare the dataset(s) to create your graph/chart?

After we got our dataset we filtered the attributes, we only kept the column we needed for our visualization.

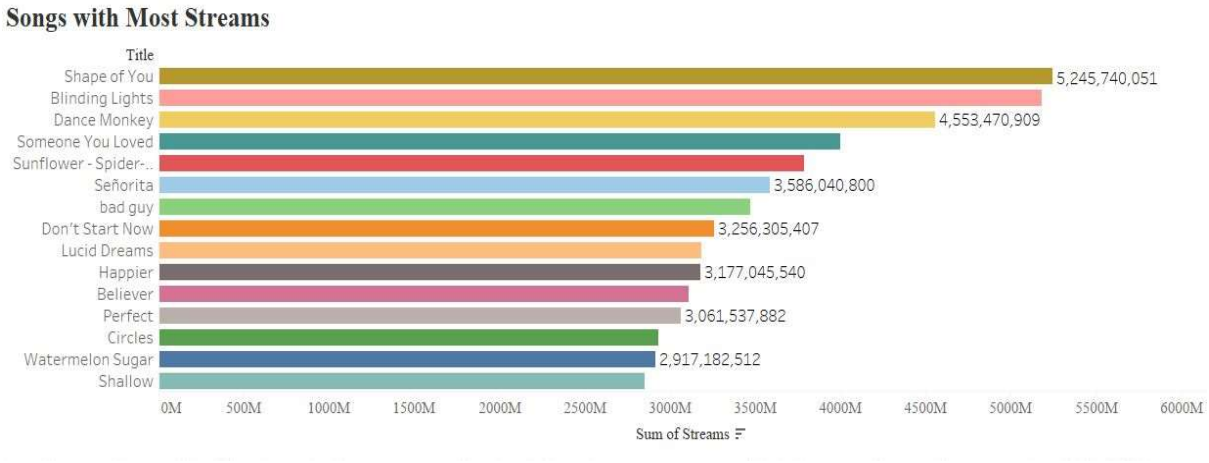
As we got our dataset from kaggle it was pretty much clean. But all the columns in our dataset were not necessary for visualization. For cleaning our dataset we used Jupyter Notebook, we read our csv file and drop all the unnecessary columns that we didn't need as it decreased the size of dataset and hence decreasing the processing time. We also searched for duplicate and null values but we didn't find any. We then exported the processed csv file and used it for visualization.

How did you choose the attributes and data subset to visualize?

As we are trying to find the most streamed song on Spotify we decided to choose streams as it will help in finding the most streamed song, we also chose the region attribute to see where it was most streamed.

Visualization:

The first graph shows top 15 tracks with most overall streams.



For designing choice, we thought that bar chart would be very appropriate to show this data, as it makes it easier to spot the most played track. We also tried line chart, but there was no relation between these songs, so the 'angle' point of view of line chart, showing increment and decrement, was not useful here.

For colour palette, we have used Tableau 20, which is available as a default option. We used that instead of Tableau 10, because we wanted to show top 15 tracks and tableau 10 would repeat colours.

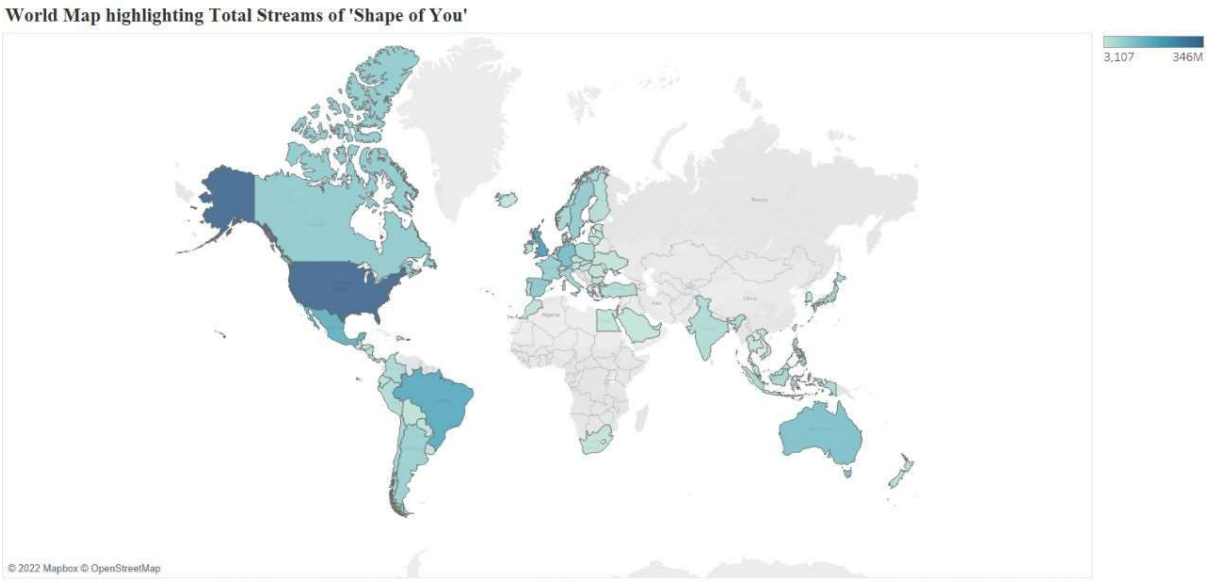
Also, we have sorted the graph in increasing order of number of streams, as it achieves the task we wanted to show.

The second graph is used to show which region has the highest listens to the most played song.

The font we used for both these chart is Times New Roman for the main tittle, and axes data and we used Tableau Lite for the Label. We used times new roman because it is the most clear font to read - https://blog.datawrapper.de/fonts-for-data-visualization/

For this graph, we used a map chart. The attributes we used for this are region, and streams. We filtered this data by Title and selected only the song, 'Shape of You' manually. We also filtered this by region.

For colour palette, we used Blue-Teal shade combination. We can clearly see that United States was the most listener to this song.



Technology we used and why:

We used Tableau for Visualisation because is a pretty default choice of technology to use, also we planned from the very beginning to use a map graph, which is very easy to implement in Tableau.

For cleaning, we originally planned to use OpenRefine, it being simple to use, but then as out data was big, we faced an issue creating project on OpenRefine, so we went ahead with using jupyter notebook with python as coding language.

Conclusion:

In the first graph, we wanted to use same colour for two songs by the same artist, such that tracks like, Shape of You, Perfect, and Happier all by Ed Sheeran, would use the same colour. But we could not achieve this.

Similarly, because of our data being limited to only top 200 and viral 50 ranking, we cannot show overall listens for all countries, as shape of you wasn't in all the top 200 lists.

We thought of scraping the web for the data we wanted, but then we found difficulty in scraping, so decide to use this particular dataset after all.

We worked as a team for this assignment. We both looked up for dataset on various sites. We sometimes divided the task amongst each other so as to avoid repletion, and on the basis of who being the best of us to do that task. We shared the effort for research and implementation of what steps should be taken in the task ahead.

For visualization, Dhruv and I discussed on what colour should we pick and what type of chart type to use and we did what we thought was the best, for project.

References:

https://blog.datawrapper.de/fonts-for-data-visualization/

https://youtu.be/X0n0s0gzcvE

https://www.kaggle.com/datasets/dhruvildave/spotify-charts

We also referred to the Example Visual Report available on the loop while writing report.