IT 350 Data Analytics
Lab 1: Descriptive Data Analysis

P Akshara – 181IT132

Descriptive statistics analysis helps to describe the basic features of dataset and obtain a summary of the data. Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

In this study, the following real-world datasets have been analyzed:

1. Spotify 1921-2020 music tracks
2. Covid-19 global vaccination progress, Indian lockdown trend
3. Netflix movies & TV shows up to 2019.

The attached Jupyter notebooks contain the detailed analysis in Markdown cells corresponding to the frequency tables, and interactive plots. Here only a summary is stated.

Spotify (Spotify-EDA.ipynb)

1. It contains 170,000 songs with their audio characteristics and popularity.
    a. Acousticness range = [0,1], mean = 0.502, median = 0.516, 25^{th} percentile = 0.102, 75^{th} percentile = 0.89, std = 0.37
    b. Danceability range = [0,1], mean = 0.537, median = 0.548, 25^{th} = 0.42, 75^{th} = 0.67, std = 0.17
    c. Energy range = [0,1], mean = 0.48, median = 0.47, 25^{th} = 0.25, 75^{th} = 0.7. The max value = 1, std = 0.27
    d. Duration range = [5.1s,90min], mean = 3.8 min, median = 3.4 min, 25^{th} = 2.8min, 75^{th} = 4.3min, std = 2min
    e. Liveliness range = [0,1], mean = 0.2, median = 0.13, std = 0.17
2. The distribution and box plot of these characteristics indicates the presence of outliers in energy, duration (presence of albums along with single tracks), and danceability.
    a. Speechiness, liveliness, instrumentation have a highly skewed (right) distribution. Loudness is left skewed.
    b. Rest has a flat symmetric dist.
    c. The plot of duration with popularity (target) depicts a Poisson dist.
    d. Tempo is a normal dist. With most values around 120bpm. Gap in (0,30) bpm range. Less than that could mean podcasts and above that songs or jingles.
3. Audio features exhibit good correlation among each other.
    a. (valence, danceability) & (energy, loudness) are positively corr.
    b. (energy, acousticness) are negatively correlated.
4. These characteristics have a significant impact on the track's popularity. Louder and livelier music seem more desirable than acoustics.

5. Significant change in these over the years indicating that people enjoy more thumping, lively music these days.
    a. Acoustic average values have drastically declined.
    b. Loud, tempo mean values have risen.
6. A large no of songs has 0 percentile popularity. On analysis these mainly belong to the 1920 – 1950 period and have high acoustic values.
7. Most popular songs trending is Dakiti and Mood, followed closely by Blinding lights, Holy.
8. Most popular artists/bands are bad bunny, 24kgoldn.
9. Length of song 's name has 4 words median, with outliers having 30+ words.
10. The most commonly occurring words in songs are life, love, remaster.

COVID-19 (Covid-19-EDA.ipynb)

The 1st part analyzes the current global standing in terms of vaccination with data up to 19 Jan.

1. Across various countries, daily vaccination count, mean = 41897, median = 5375.5, max= 913912, min = 51, std = 108639. This shows that many countries have not yet actively started vaccination.
2. There is an apparent flattening in the no. of new cases globally from Dec-Jan possibly with the administration of vaccine.
3. The world and tree maps indicate the presence of 9 types of vaccines.
    a. Moderna/Pfizer: Prominently used in US, Germany, Spain, Canada
    b. Sinovac: In China
    c. Pfizer/BioNTech: Israel, France, Poland, Italy, UK
    d. Sputnik V: Russia, Argentina
    e. Covaxin/Covishield: India
    f. Sinovac: Turkey
4. The bar plots show that 18M people have taken the Pfizer, 15M Sinovac.
5. US, China, UK, Israel have actively begun administering the vaccine.
    a. Highest percentage of population who have been is Israel at 38%! Average daily vaccine count is 158K.
    b. India stands at 0.08%. Daily 191K vaccines are being given on an average.

Next is the analysis of Indian scenario during March-Aug 2020.

1. Starting with just 3 affected states in early March, by end of March 28 had been affected!
2. The worst affected states were Maharashtra, Tamil Nadu, Delhi, Andhra, Karnataka
3. The tree map shows the district level analysis.
4. June 16 witnessed max no. of deaths in 2004!
5. Jun-Aug witnesses a sharp drop in mortality rates and steep increase on recovery.
6. Of the available data, the most affected age group is 25-35.
7. Almost double the male population has been affected.
8. The mortality rate is highest for 55-75 age groups.

9. Analyzing the transmission type, most common is local spread through touch, next is imported from abroad. Others are unreported.

Netflix (Netflix-EDA.ipynb)

Explained in notebook.