Assessment cover

Module No:	DALT7016	**Module title:**	Data Visualisation
Assessment title:	Assignment		
Due date and time:		Friday 15th December 13:00	
Estimated total time to be spent on assignment		30 hours	

LEARNING OUTCOMES

On successful completion of this assignment, students will be able to achieve the module following learning outcomes (LOs):
Critically analyse data visualisation approaches with respect to human sensory modalities
Create appropriate visualisations for temporal, dynamic, and high dimensionality data
Devise methodologies for data interaction to facilitate exploratory data analysis

Engineering Council AHEP4 LOs assessed (from S2 2022-23)		**Met? (Y/N)**
M1	Apply a comprehensive knowledge of mathematics, statistics, natural science and engineering principles to the solution of complex problems. Much of the knowledge will be at the forefront of the particular subject of study and informed by a critical awareness of new developments and the wider context of engineering	
M2	Formulate and analyse complex problems to reach substantiated conclusions. This will involve evaluating available data using first principles of mathematics, statistics, natural science and engineering principles, and using engineering judgment to work with information that may be uncertain or incomplete, discussing the limitations of the techniques employed	
M3	Select and apply appropriate computational and analytical techniques to model complex problems, discussing the limitations of the techniques employed	
M17	Communicate effectively on complex engineering matters with technical and non-technical audiences, evaluating the effectiveness of the methods used	

Statement of Compliance *(please tick to sign)*

I declare that the work submitted is my own and that the work I submit is fully in accordance with the University regulations regarding assessments "http://www.brookes.ac.uk/uniregulations/current"

DALT7016 Data Visualisation Assignment Report

Part 1: Context and EDA

1. Context and Source Description:

In this era of Mobile and Technology, we are surrounded by apps. These apps are easy to create and can generate good profits. Because of these two factors, more and more apps are being developed. The data we have selected is Google Play Store data and we will comprehensively analyse this data by comparing over nine thousand apps in Google Play across different Genres and Categories. We will explore the relationships between Apps and other variables such as Categories, Ratings, and Installs and will identify patterns and trends based on our observations. We'll look for more insights into the data to derive strategies to enhance growth. The dataset has been used from Kaggle from the below link.

https://www.kaggle.com/datasets/lava18/google-play-store-apps/data

Fig. https://s1.eestatic.com/2015/03/11/elandroidelibre/el_androide_libre_17258747_179231872_1706x960.jpg

2. Dataset Description and Summary Statistics:

In this dataset we have 9657 rows and 11 attributes [9657 x 11]; and the size of the exported csv file (googleplaystore.csv) is 946 KB. The five features that we will be working with most frequently henceforth are Category, Installs, Size, Rating and Reviews Let's see the datatypes of the attributes in the dataset along with their description.

Column Name	Datatype	Description
App	chr	The name of the application
Category	chr	Category of the application
Rating	dbl	The rating of the app on Google PlayStore
Reviews	dbl	Number of reviews given on the app
Size	dbl	Size of the app in MB
Installs	nbr	Number of downloads of the app
Type	chr	The type of the app whether it is free or paid
Price	dbl	Price of the app in US dollar
Content Rating	chr	Content rating of the app
Genres	chr	Under which genre the app falls
Last Updated	chr	Last Updated date of the app

The steps taken to pre-process and clean the dataset:

1. Identifying Data Type Issue
2. Detecting/handling Missing Values in Columns
3. Converting String Values to double/numeric
4. Renaming Columns for Clarity

In the dataset, the Installs column data was in character which was then converted to numeric to carry out statistics based on it. Also, the price of the data was string represented as 9.99$, 4.99$ which was then converted to numeric by removing the '$' and renaming the column name as 'PriceInUSD' for better understanding. The same goes for the 'SizeMB' column in which the data was character which was converted to double.

3. Missing data

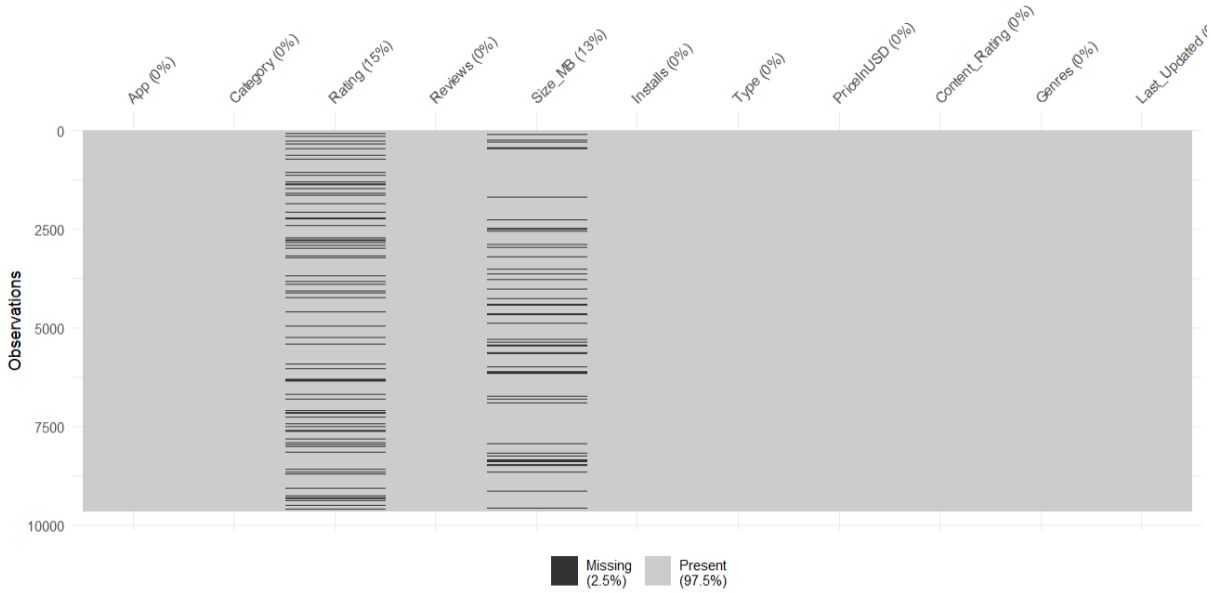

Fig 2: *Missing data*

Total 2.5% of the data is null or NA or missing data out of 100%. Most of the missing data is from the column Ratings and Size of the app. As we cannot put a value 0 in Rating column as

well as Size column, we will keep the value as NA. Replacing it to 0 can cause incorrect data analysis.

Part 2: Design

The insights of these plots collectively provide an understanding of user preferences, engagement levels, and the diversity of apps within different categories, which will benefit strategic decision-making for developers and stakeholders in the app market.

1. Purpose:

The Category distribution of the Count/ Average-size-Installs-reviews plot gives us an overview of the app's statistics in terms of average which will be related to the total number of apps to gain valuable insights and understand the trends and demands in the market and allows app developers to take decisions accordingly. Have you ever wondered why we get the "Rate this App" popup after every app use? And What will be the use of this rating data? By plotting the graphs let's discuss what is the aim main outcome behind this. The "Distribution of Ratings Across App Categories" plot aims to provide insight into how the user ratings are distributed across different categories of mobile applications on the Google PlayStore.

2. Description of the Audience:

The target audience for this visualization includes app developers, marketers, and analysts seeking to understand the distribution of Average Installs, Size, Reviews, and user ratings in various app categories. Decision-makers in the mobile app industry can use this plot to identify which categories are downloaded more/less, reviewed more/less, and receive higher or lower ratings, helping them make informed decisions about app enhancements, marketing strategies, and user engagement. These graphs serve as a valuable tool for developers, stakeholders, and users alike, aiding in the understanding of storage implications and preferences across various Applications.

3. Design Plan:

a. Static Design:

This design plan aims to convey insightful information about category-wise count, Average Installs, Average Reviews, and Average Size of the apps interactively and engagingly. Bar plots are used to display this information. We have Category on the x-axis for all the plots and Count, Average Size, and Average Installs on the y-axis respectively in different plots. The scale is in millions for Average Reviews and Installs. We have plotted four different graphs and then clubbed them together for better understanding.

b. Interactive Design:

In the Interactive Design we will see more about the impact of user rating on different categories of apps. The design plan aims to convey insightful information about category wise rating of the app's in an interactive and engaging manner.

Box-plot is used to display the summary statistics (min value, max value, median, quartile) of ratings in each category. The 33 categories in the data and the rating-wise distribution of each category are visualized. We have taken Rating on the y-axis (0-5) and Category on the x-axis.

The below count plot gives the count of apps within each category based on their respective ratings providing additional details on Categories and their ratings.

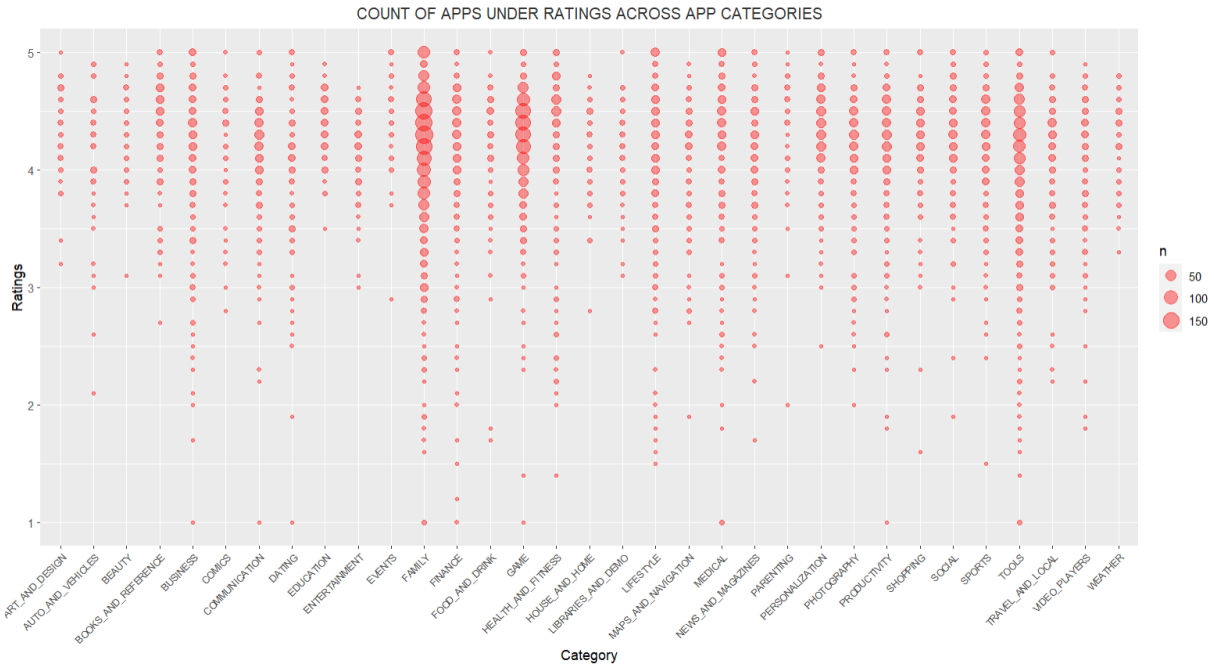

Fig 3. Count plot of Ratings vs Categories

The density of the points at different rating levels gives us an overview of the distribution of apps within each category. Stakeholders with the help of this plot can assess the distribution of app ratings, identifying categories with a higher concentration of apps at specific rating levels. This plot along with the interactive plot shows how many apps come under the rating range and how many are the outliers.

Part 3: Final Visualisations

1. Visualisation 1 Commentary

The bar-graph "Category wise count" shows a visual landscape where count of apps across each category. The bar-graph "Category-wise average installs" gives valuable insights into user preferences and popularity across different app categories. As visualized, the ‘Communication’ category stands tall with the highest average installs, implying its dominance in user engagement and demand in the market. ‘Social’ and ‘Productivity’ categories also show significant average installs, indicating their importance in the app market. ‘Social’ and ‘Communication’ having a smaller number of apps than ‘Family’ but stand high in most installs tells us about the influence of communication and social media on this generation. On the other hand, categories like ‘Medical’ and ‘Events’ show lower average installs. In the category-wise average reviews of apps. The ‘Social’ category apps have the most reviews followed by

the ‘Communication’ category, which previously demonstrated high average installs, suggesting not only the most engaging but also positive user experiences, ‘Game’ categories also garner substantial average reviews. Conversely, ‘Events’ and ‘Medical’ continue to display lower average reviews, which indicates there might be some areas for improvement. Size the ‘Libraries and Demo’ category apps have the highest average size followed by ‘Game’ and on the other hand ‘Beauty’ and ‘Art_and_Design’ have the lowest app size average.

2. Visualisation 2 Commentary

As visualized for all of the Categories the median line lies in the range of 4-4.5, but the ones showing narrower interquartile ranges as seen for Communication, Education, Entertainment, Personalization, Shopping, and Social show consistency in Ratings which suggests user satisfaction while the ones showing wider spread indicates a more diverse user experience. Many Categories have several apps with below-average ratings and those are visualized as dots(outliers) in the plot such as Family, Tools, and Finance. These outlier apps may need further investigation and enhancements to catch up to the user expectations.

Viewers can quickly grasp the variability in ratings across different categories because of the simplicity of the box plot and also it becomes easy to analyse it interactively. It gives the stakeholders and app developers proper idea about how different category apps are behaving in the market. The plot effectively visualises the requested information without unnecessary complexity.