EDA Capstone Project:

Play Store App Review Analysis

By- Ranajay Biswas

Data science trainee

AlmaBetter, Bangalore

Abstract:

Exploratory Data Analysis (EDA) is an approach of analysing datasets to summarize their main characteristics, often with visual representations, in order to have a better understanding of the data and make meaningful decisions. EDA often involves many steps such as –

1. Data collection
2. Descriptive Statistics
3. Data Cleaning
4. Data Pre-processing
5. Graphical Representation & Data Visualization

In this experiment, we tried to understand the android app market by doing EDA on two different datasets. We explored and analysed the data to discover key factors responsible for app engagement and success.

Keywords: Exploratory Data Analysis, Dataframe, Cleaning, Visualization, Features, Summarization.

Problem Statement:

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.

Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.

The objective is to explore and analyse the data to discover key factors responsible for app engagement and success.

Steps Involved:

- **Importing Libraries :** We used 'Pandas', 'Numpy', 'Matplotlib', and 'Seaborn' libraries for doing most of our exploratory data analysis. Hence, we imported these four libraries as the very first step.

- **Collecting Data :** For any exploratory data analysis, the first obvious step is to collect the data. In this case, the data we had, were two CSV files. In order to be able to work with the data, we had to load them into our python notebook. For that, we used Panda's read_csv() function.

  We used the df.head() method to check the first 5 rows of our data. It gave us a brief understanding of the columns and how our data looks at a glance.

- **Descriptive Statistics:** Using the .describe() method, we got a table showing the different values in our columns, their quartile range, maximum and minimum values, etc.

  A boxplot gave us an idea about the outliers that were present in our data. So in the next step, we treated the outliers. Using the 'Seaborn' library, we plotted a distribution plot to check the distribution of the data.

- **Data Cleaning & Imputation:** This process of data cleaning and pre-processing involves ---

  - ***Null values identification:*** Using the .isna() method, we identified the null values that were spread throughout the data. In different columns, we had different numbers of null values present and since the data types varied column-wise, we also had to be careful about treating those values. This brings us to the next point.

  - ***Null values Treatment :*** The 'Rating' column in our data had numerical values. It was also skewed. So, we used the median value to replace all the null values in the column instead of the mean.

'Type', 'Current Version', and 'Android Version' columns had categorical data types. For that reason, we had to use the mode values in their respective columns to fill in the null values.

We also checked if any of these columns had more than one mode before doing the imputation. Lucky for us, each column had only one mode.

Data Transformation & Processing : Upon doing further inspection, we find that we can change the data types of some of the columns from categorical to numerical, so that we can perform statistical operations and visualizations on these column values.

- We removed the '**$**' sign associated with the values in the '**Price**' column and then applied the lambda function to make the values numeric.

- '**Installs'** column was having values like '100,000+'…So, we got rid of **'+'** and ',' associated with the number characters and turned the values into integer data types.

- '**Reviews'** column values were simply passed through the pandas to_numeric() function. And we got them transformed into numerical values.

- **'Size'** column was having either **'M'** or **'K'** at the end of each value. We figured that this 'M' and K' probably stand for Megabytes and Kilobytes. So, we just removed the 'M' wherever we encountered it. And for the values having 'K', we had to remove it & also divide the values by 1000 to get those values in megabytes format.

We checked the data to see the invoked changes. Now that we have made the changes, we were all set to do visualizations.

- ## Graphical Representation & Data Visualization :

Visualization is an important part of the EDA process as we try to figure out trends and patterns and behaviours in our data with the help of graphs and charts.

We used different charts and plots to visualize our data for different business aspects.

- ○ We used Scatterplot for the Rating vs Category graph. It showed us how the app ratings were scattered among different categories.
- ○ A histogram showed us the app size distribution in our data.
- ○ The line plot tells us about the no. of reviews in each category.
- ○ We used bar charts to show **prices** & **number of apps** in each category.
- ○ With a bar-plot, we understood the proportion of free and paid apps in each of our app category.
- ○ Again a scatterplot was used to visualize the number of app installations for different app categories.

In our 2nd data set, we used the bar charts to visualize the top most reviewed apps, and also apps with the top-most positive and negative sentiment received from the users.

- ○ With the help of boxplots, we visualized Sentiment Polarity and Sentiment Subjectivity.

Sentiment Analysis Model :

The 2nd dataset having the user sentiments data, compelled us to build a basic Sentiment Analysis model for this user reviews dataset. We used the 'Sentiment' and 'Translated Reviews' columns as our main features for this model training. But, before building the model, we needed to import the necessary libraries and packages.

Nltk (Natural Language Toolkit) and Sklearn were the two most important modules used in the process.

Steps :

- ❖ Transformed all the text reviews to lower case.

- From the text of the reviewers, we removed all the punctuations and stop words.
- Used the lemmatize method to clean the texts furthermore for better model accuracy.
- Split the cleaned dataset into train and test sets with 75% data going for the model training.
- ‘Reviews’ were used as independent variables to study the ‘Sentiments’ as a dependent variable.
- Trained the model with the training data and got approximately 90% accuracy.
- Finally, tested the model on some of the random reviews picked from the data and the results came back pretty satisfactorily.

Conclusion: With that, we have reached the end of this exercise. Starting with the dataset loading and understanding our problem statement, we went on to do the EDA process which included null- values treatment, other data cleaning, pre-processing, and many visualizations.

Finally, we implemented a very basic Sentiment Analysis model that gave us good enough results.

This exercise really helped with the understanding of doing EDA from a business perspective based on real-life data.

For our references, we used –

- AlmaBetter
- Towardsdatascience
- Analytics Vidhya
- Kdnuggets