**Vishwakarma Government Engineering College**

****

**Data Science Project Report**

**Data Science - 3151608**

**Information Technology**

**Semester 5**

**180170116046 – SUTHAR NEEL J**

**180170116029 – PATEL JAY A**

**PROJECT TITLE: -**

**ANALYSIS OF GOOGLE PLAY STORE DATA**

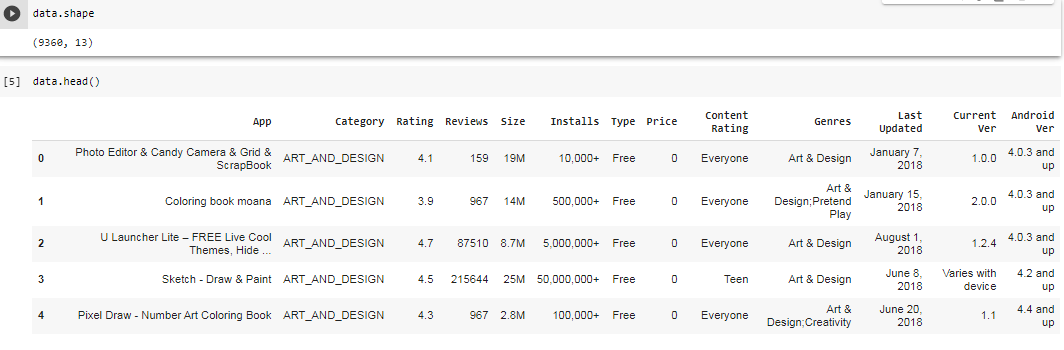
**AIM:-**

**The Internet is a true gold mine of data. E-commerce and review sites are brimming with a lot of untapped data with a prominent potential to convert into meaningful insights that can help with robust decision making. Here, we explore using data science and machine learning techniques on data retrieved from one such avenue on the internet, the Google Play Store.**

**The dataset that is used for this project can be found at: -**

<https://drive.google.com/file/d/1C7hS3iKxUUTs6DANzlRmD9eL-ENlEME7/view?usp=sharing>

<https://drive.google.com/file/d/1K9DLzf2ln1RTyr0A1LpSRUzZM72_db64/view?usp=sharing>



**Variables:-**

To make the understanding and manipulating of the dataset easier, several variables were used which are mentioned as follows:

1. **Total Variables that exist in the dataset:**
2. App
3. Category
4. Rating
5. Reviews
6. Size
7. Installs
8. Type
9. Price
10. Content Rating
11. Genres
12. Last updated
13. Current version
14. Android version
15. Sentiment

**2) Variables used from dataset:**

1. App
2. Category
3. Rating
4. Reviews
5. Size
6. Installs
7. Type
8. Price
9. Content Rating
10. Genres
11. Android version

**3) List of independent variables:**

1. App
2. Category
3. Rating
4. Install
5. Last Updated
6. Current Ver
7. Android Ver

**4) List of dependent variables:**

1. Reviews
2. Size
3. Installs
4. Type
5. Price
6. Content Rating
7. Genres

**5) Model used for machine learning:**

1. We clean the data by filtering columns.
2. Data visualization
3. Data preprocessing
4. In Rating, we used three models: decision tree, linear regression and random forest regression for predict which is best for our data.

**Steps Implemented with little description and code:-**

1. **Importing Libraries**

* Pandas – To read data and manipulate in different forms.
* Numpy – To manipulate numbers and arrays easily.
* matplotlib.pyplot – Used for plotting graphs.
* Sklearn – Scikit-learn is used for various algorithms like support vector machine (svm), k-neighbors etc.
* Seaborn – To visualize data.
* Plotly – To plot interactive graphs.
* os – To collect data from local directories, files, etc.

1. **Importing the dataset** – googleplaystore.csv and reading it

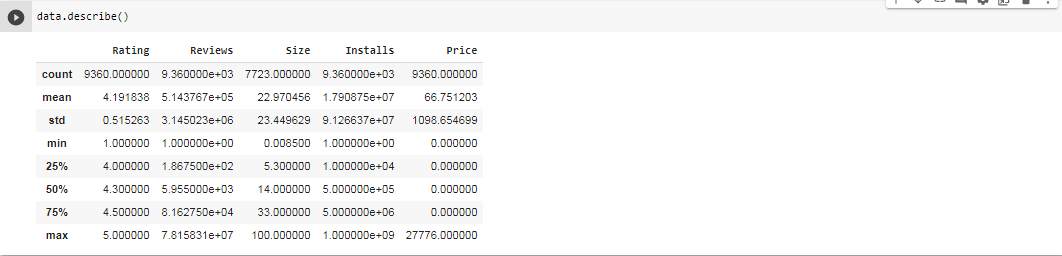
* googleplaystore.csv was imported and using functions like head ().We get shape of dataset by data.shape function.

1. **Determining the dependent and independent variables:-**

* Independent variables are those whose values are not decided by any other variable. Its values exist on its own uninfluenced by other variables. A list of the independent variables used in this project is given above on page 3.
* Dependent variables are those variables whose values depend on other factors such as independent variables. They are influenced by all the changes in the independent variables. The dependent variable used in this project is ‘target’.

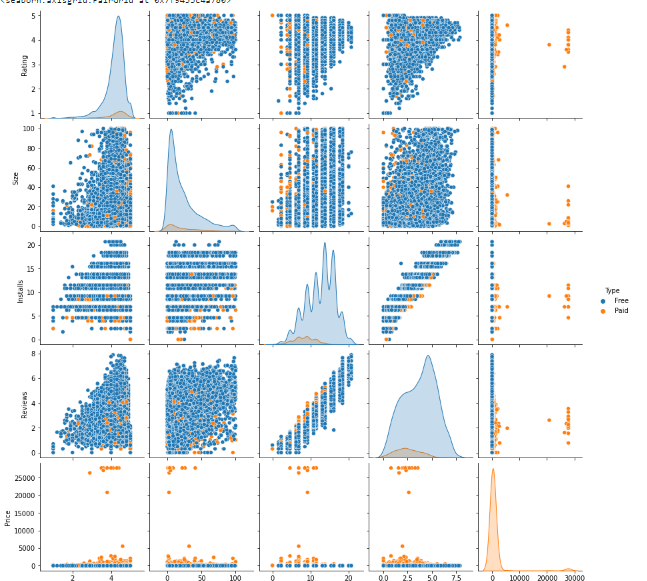
1. **Collecting general information about our data:-**

* Information about data is collected using commands such as:
* Size – Our data has 9360 entries in it.
* Shape – Our data has 9360 rows and 13 columns.
* Info - Information like class, range, columns, data-type, and null/not-null and memory usage is mentioned.
* Describe – Statistical functions like count; mean, max, min, standard-deviation, quartiles etc. are counted for each column and displayed.

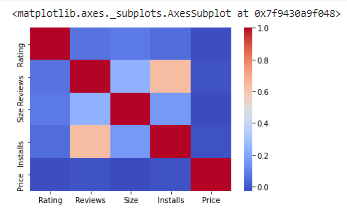


**5) Visualizing the data using graphs:-**

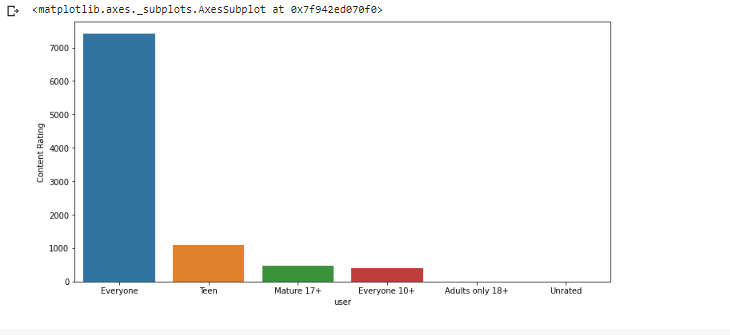
**1. pair-plot:-**

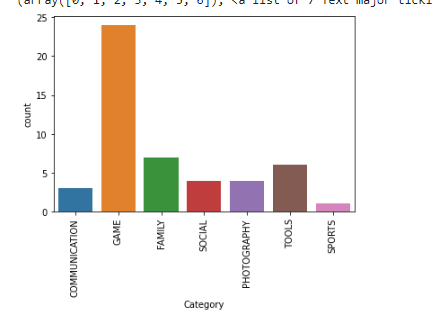
****

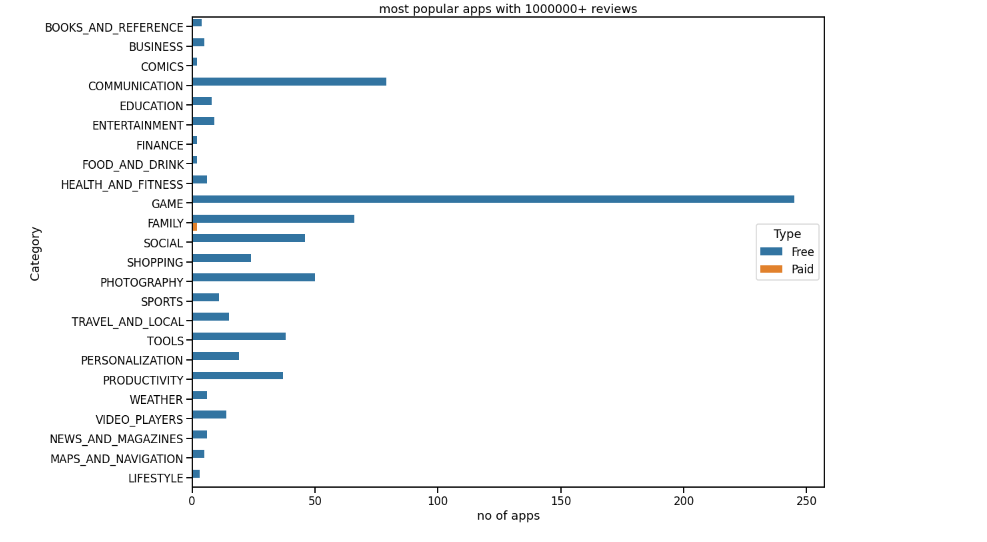
**2. correlation-matrix(heatmap):-**

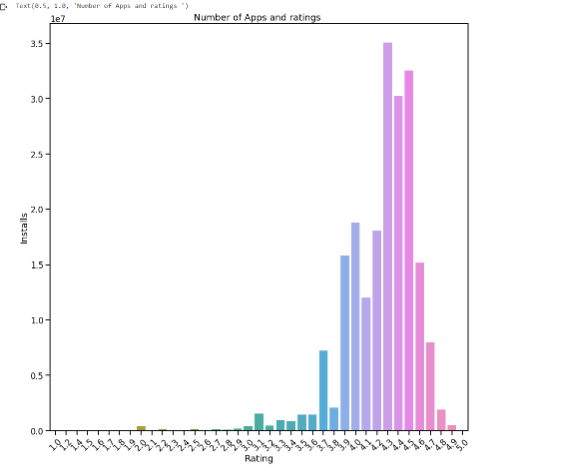
****

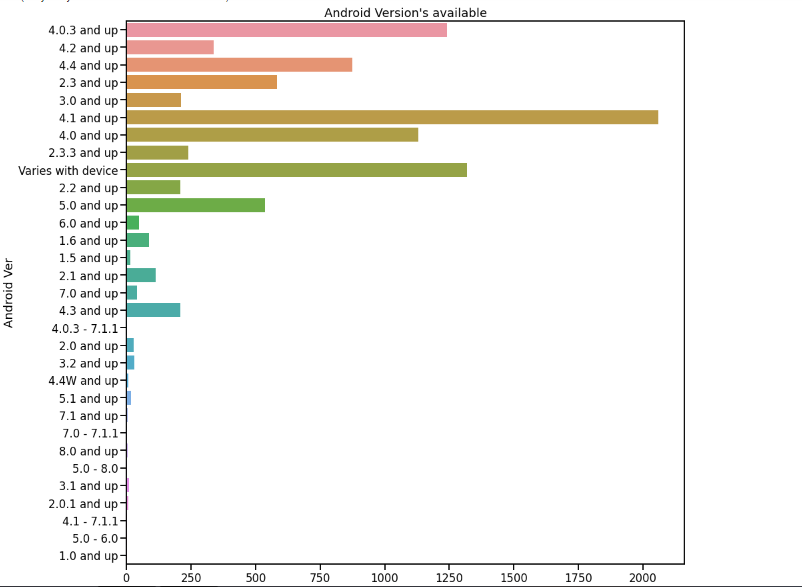
**3. bar plot:-**

****

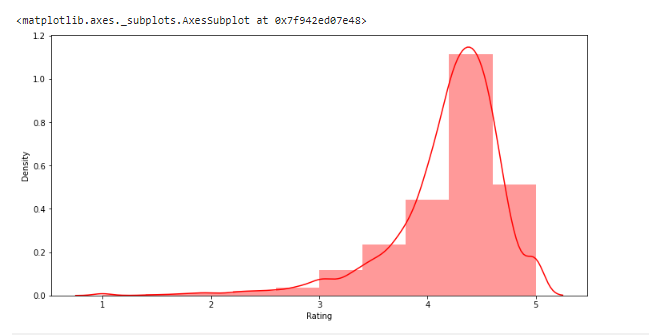
****

****

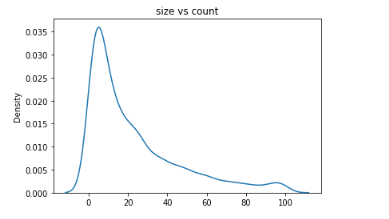
****

****

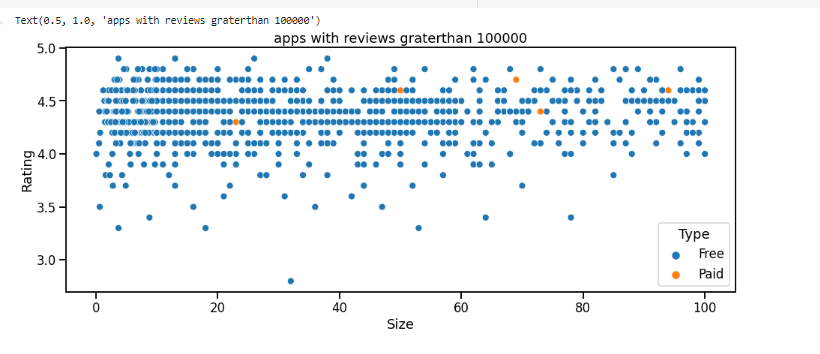
**4. histogram:-**

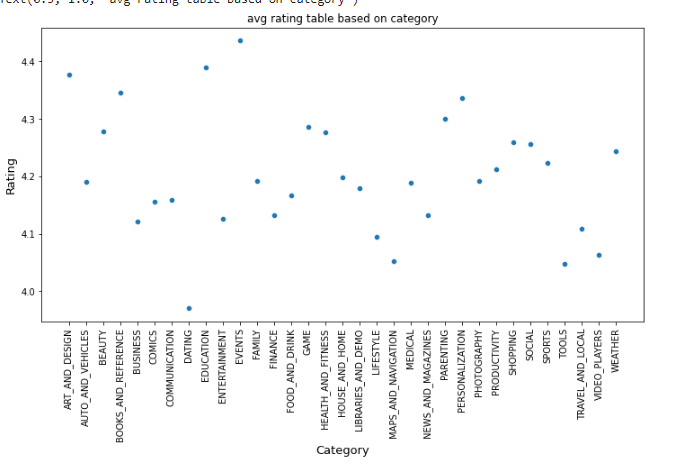
****

**5. kde plot:-**

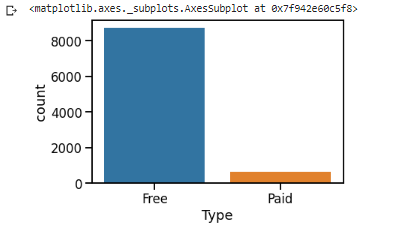
****

**6. Scatter plot:-**

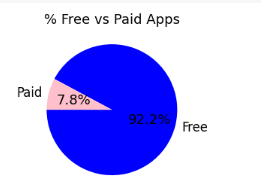
****

****

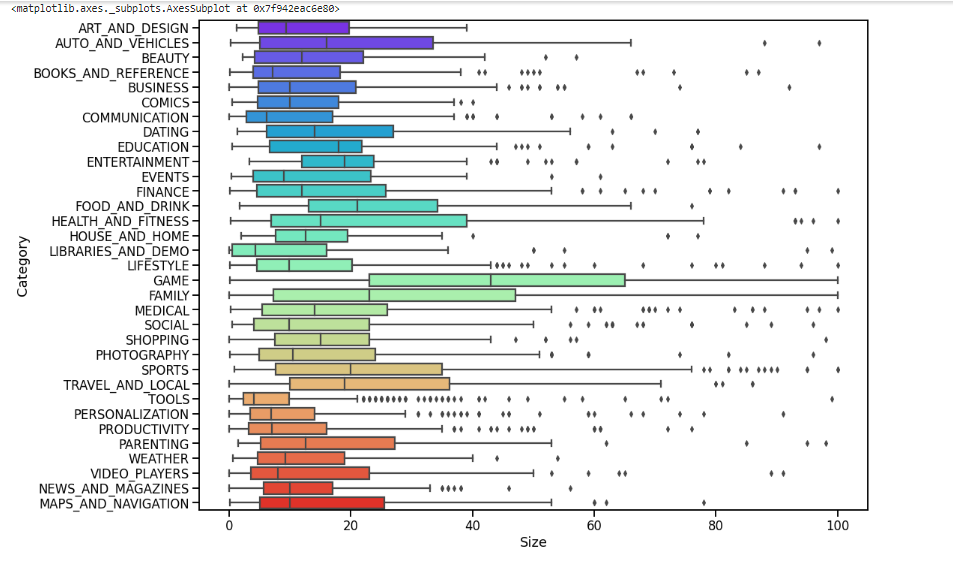
**7. Count plot:-**

****

**8. Pie chart:-**

****

**8. Box Plot:-**

****

**6) Pre-processing the data and making 2 different datasets to isolate dependent and independent variables:**

**Pre-processing the data and making 2 different datasets to isolate dependent and independent variables.**

**7) Training and Testing the data:**

**We take 30% of random data as x\_test and y\_test (our test datasets).And using RandomForest regression we find the prediction on rating of app.**

**8) Applying Models:**

* **cleaning of data**
* **data visualization**
* **data preprocessing**
* **randomforest regression,linear regression,decision tree**

**9) Making Prediction and check accuracy:-**

**First we loaded the data and analyzed it. We found various rows of data which was null or not appropriate for model. So first we prepared the Data. We cleaned the data and divided it according to their attribute. After that we applied feature engineering and data visualization to decide the feature based on which we can train model accurately.**

**Now our dataset has continuous value. So we find prediction for rating column, which is changes with app.**

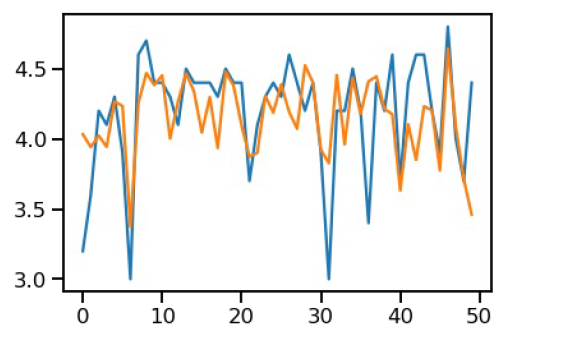
**Then we plotted the graph between the actual and predicted rating of app.**

**10) CONCLUSION:-**

**For this project, we took the Google Play Store Data sets and analyzed and processed the data. After the data was transformed into a usable set, we used plots and functions to understand the correlations between features. We then used this knowledge to build the best model we could for finding ratings based on the cleaned data set.**

* **Exploratory data analyis of the Google play store data was done.**
* **Top Apps that are most downloaded is found out.**

**Thus, we can find rating of app using randomforest regression.The predicted rating and actual rating graph is as follows:-**

****

**THANK YOU!!**