**Play Store App Review Analysis**

**Saurabh Shinkar**

**Data science trainee,**

**Alma Better, Bangalore**

**Abstract:**

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.

We will analyse each feature and understand about the different factors in the data**.**

* **Problem Statement**

Explore and analyze the data to discover key factors responsible for app engagement and success**.**

**About play store and user reviews data**:

* **App:** Application Name
* **Category: Category to Rating:** Rating of the Application
* **Review:** Review of the Application
* **Size:** Size of the Application
* **Installs:** No. of installs of the Application
* **Type:** Whether the Application is free or paid
* **Price:** Price of the Application
* **Content Rating:** For whom the content has been created
* **Genres:** Genres to which Application belong
* **Last updated:** Application recently updated Date
* **Current Ver:** Current version of application
* **Android Ver:** Android version of application
* **Translated Review:** Review in Words
* **Sentiment:** An attitude, thought, or judgment prompted by feeling. It is Positive, Negative or Neutral in our Data.
* **Sentiment Polarity:** It is the expression that determines the sentimental aspect of an opinion
* **Sentiment Subjectivity:** Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

## 

### 

* **Steps involved:**
* **Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. That’s what we have tried to done.

* **Null values Treatment**

Our dataset contains a large number of null values which might tend to disturb our analysis hence we replaced those nulls which are in large amount with proper values and dropped those who have negligible nulls at the beginning of our project in order to get a better result.

* **Making data in proper format**

We removed signs from those string values which can be convert in integer or float type. Later we converted those string types into integer or float according to the respective data. We also made every feature in one proper format like made all sizes in MB.

* **Analyzing each feature separately**

For numerical feature we look at the distribution of each feature, checked for its skewness, calculated its mean. For categorical features we computed their value counts. Also plotted different diagrams to understand them better.

* **Analysis of how different features impacted on apps**

Our main motive through this step was to analyze how different features impacted on installation of apps. By looking into the different features at the same time we inspected how they impact on app engagement and success.

* **Fitting different plots**

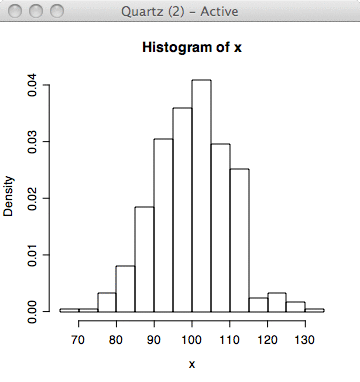
For modelling we tried various classification algorithms like:

1. **Histogram**
2. **Bar plot**
3. **Pie Chart**
4. **Density Plot**
5. **Correlation Heatmap**
6. **Regression Line**
7. **Word Cloud**

**3. Plots:**

1. **Histogram:**

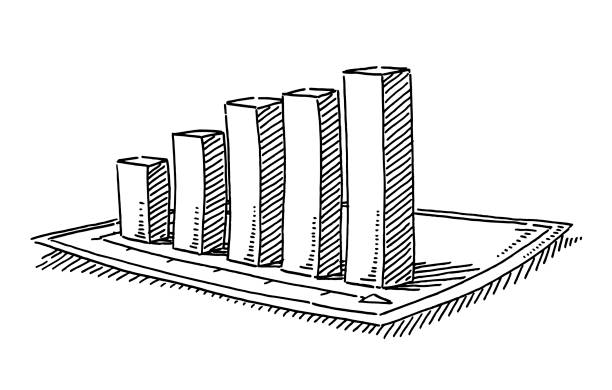
A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.



1. **Bar Plot:**

A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars. Bar plots include 0 in the quantitative axis range, and they are a good choice when 0 is a meaningful value for the quantitative variable, and you want to make comparisons against it.

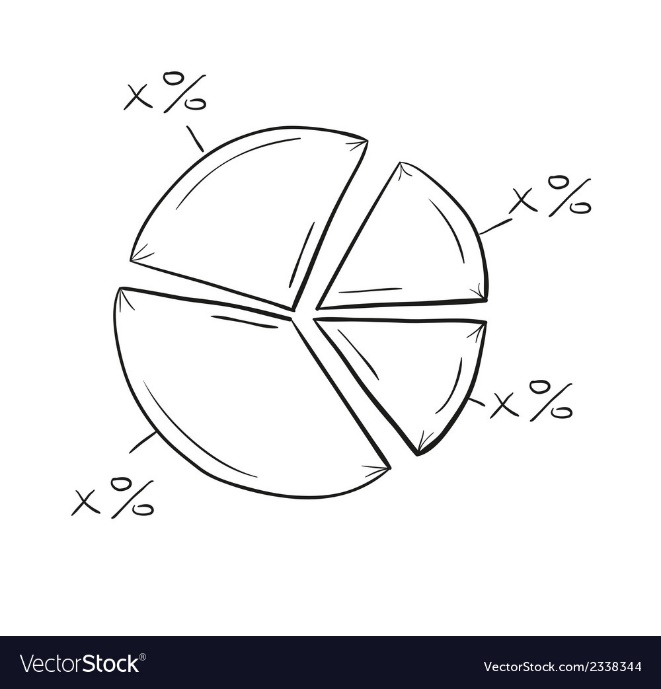
For datasets where 0 is not a meaningful value, a point plot will allow you to focus on differences between levels of one or more categorical variables.



We use hinge loss to deal with the noise when the data isn’t linearly separable.

Kernel functions can be used to map data to higher dimensions when there is inherent non linearity.

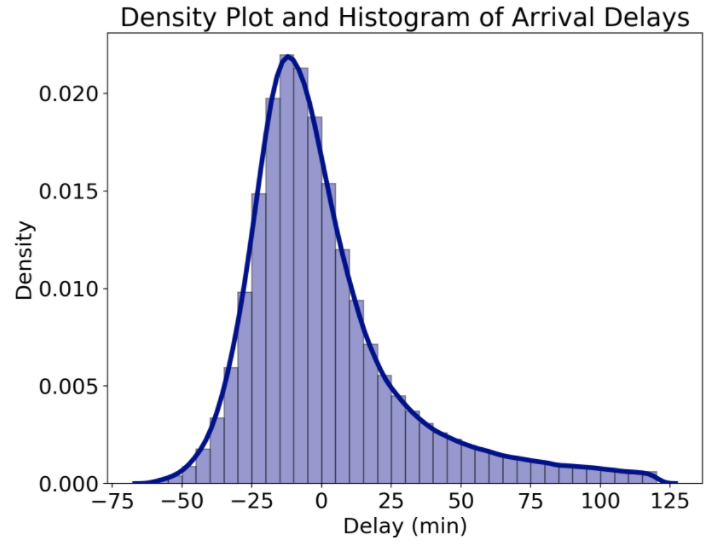
1. **Pie Chart:**
2. A **pie chart** (or a **circle chart**) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.



1. **Density Plot-**

Density Plot is a type of data visualization tool. It is a variation of the histogram that uses ‘kernel smoothing’ while plotting the values. It is a continuous and smooth version of a histogram inferred from a data.

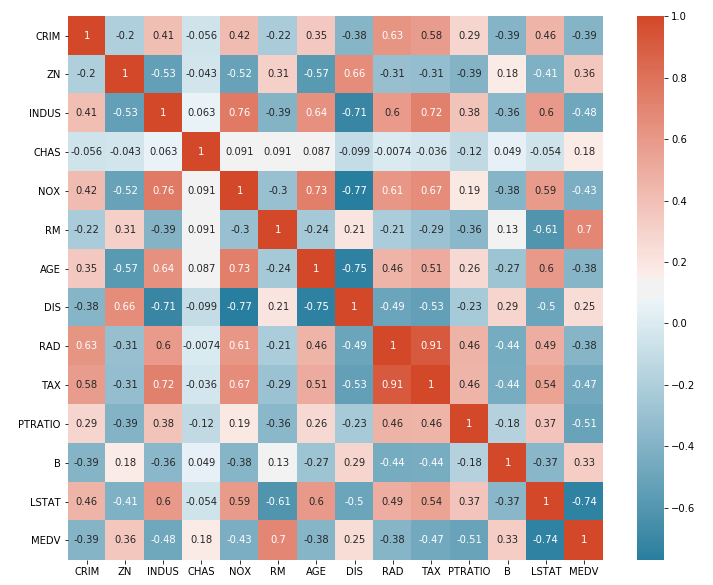
Density plots uses Kernel Density Estimation (so they are also known as Kernel density estimation plots or KDE) which is a probability density function. The region of plot with a higher peak is the region with maximum data points residing between those values



1. **Correlation Heatmap-**

Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. A correlation plot typically contains a number of numerical variables, with each variable represented by a column. The rows represent the relationship between each pair of variables. The values in the cells indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship. Correlation heatmaps can be used to find potential relationships between variables and to understand the strength of these relationships. In addition, correlation plots can be used to identify outliers and to detect linear and nonlinear relationships. The color-coding of the cells makes it easy to identify relationships between variables at a glance. Correlation heatmaps can be used to find both linear and nonlinear relationships between variables.

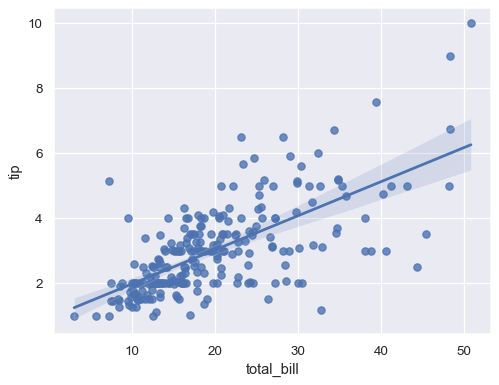
Here is a sample correlation heatmap created to understand the linear relationship between different variables in the housing data set. The code is discussed in the later section.

****

1. **Regression Line-**

Two main functions in seaborn are used to visualize a linear relationship as determined through regression. These functions, [**regplot()**](https://seaborn.pydata.org/generated/seaborn.regplot.html#seaborn.regplot) and [**lmplot()**](https://seaborn.pydata.org/generated/seaborn.lmplot.html#seaborn.lmplot) are closely related, and share much of their core functionality. It is important to understand the ways they differ, however, so that you can quickly choose the correct tool for particular job.

In the simplest invocation, both functions draw a scatterplot of two variables, x and y, and then fit the regression model y ~ x and plot the resulting regression line and a 95% confidence interval for that regression:



1. **Word Cloud-**

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analysing data from social network websites.

For generating word cloud in Python, modules needed are – matplotlib, pandas and word cloud.



* **Conclusion’s:**
* Most of the app's belongs to Family category. They are approximately 19% among all categories.
* Ratings are negatively skewed with mean rating of 4.19
* Sizes are positively skewed with median size of 12 MB.
* Prices are positively skewed with mean price of 1,median price of 0 and max price of 400.
* Most of The Applications Are Created For Everyone
* Approximately 92% apps are free.
* Tools, Entertainment and Education are the top 3 genres.
* Approximately 64% sentiments are positive, 22% are negative and 14% are neutral.
* Mean Sentiment Polarity is 0.180904
* Mean Sentiment Subjectivity is 0.493767
* There are 271 apps which have 5-star rating. In which 67 apps belong to the Family category.
* Among Top 50 install apps 22% belongs to the communication category and 16% belongs to the Game category.
* Events category have the maximum mean rating of 4.395313
* Genres of Communication, Tools, Productivity and Social have the maximum no. of installs.
* Apps belonging to Finance category have the highest mean price (8.408203) among all categories.
* Family category has the highest no. of paid apps.
* Game And Family Category Have The Highest No. Of Installs Of Paid Apps.
* Installs are positively correlated with reviews with correlation value of 0.63.

