- Steps Performed.
- Imported all necessary libraries
- Loaded the data file using pandas “pd.read_csv”
- Removed all the duplicates as part of data cleaning to increase accuracy using “df.drop_duplicates”
- Checked for null values in the data and displayed null values for each column using “df.isna().sum()”
- Dropped records with nulls from columns - 'Content Rating','Type','Android Ver','Current Ver' using “df.dropna”
- Replaced nulls from Rating column with "mode of rating" which is 4.4 “df["Rating"].fillna”
- 4.1 Extracted the numeric value from the column using expression pattern, extract method and Multiplied the value by 1,000 for size mentioned in Mb
- Dropped the Size column
- Renaming the new column to original using “df.rename”
- The Size column generated 1525 null values which were originally as string "varies with size", Replace null with mode of strings which is 11000.0 kb
- 4.2) Reviews is a numeric field that is loaded as a string field, Converted it to integer using “.astype”
- 4.3) Removed symbols from Installs column using “.str.replace”
- 4.3) Installs column converted to integer
- 4.4) Price field is a string and has a $ symbol. Removed ‘$’ sign using str.replace
- 4.4) Converted Price field to float using “.astype”
- all 10346 values in the rating field are within the specified range of >=1 and <=5
- 5.1) average rating is 4.2 with the range and all the value falls within the range, nothing to be dropped.
- there are 11 records where review is greater than installs
- 5.2) All the 11 records where the number of reviews greater than number of installs are dropped.
- 5.3) All 9579 free app are priced 0, so there is nothing to drop
- Created a function to calculate turkey's fence, IQR etc
- 7.8% values of price are outliers
- 5) boxplot of price using sns.boxplot
    - Univariate analysis of Price:
        - 1) Price range: 0 to 400.
        - 2) Majority of dataset: '0'.
        - 3) Tukey's fences calculation: Q1, Q3, IQR, lower, and upper fences all '0' due to dataset majority.
        - 4) Standard method for identifying outliers: Values below threshold are potential outliers.
        - 5) Understanding further requirements is necessary to identify outliers. Extremely high price could be considered an outlier with a defined threshold.

        - 6) Potential outliers: Sorted unique high prices - 400.0, 399.99, 379.99, 299.99, 200.0, 154.99, 109.99, 89.99, 79.99, 74.99, 46.99, 33.99, 30.99, 29.99, 28.99, 25.99.
        - 7) Outliers account for approximately 7.8% of the data.
- data of apps with rating higher than 116878. The values seems reasonable because they are less than the number of installs
- box plot of Reviews:
    - Univariate analysis of Review:
        - 1) Review has 5998 unique values, ranging from 0 to 78158306.
        - 2) Out of the total apps, 1866 (18.04%) have extremely high review values compared to the majority. This is determined based on IQR values and an upper fence value of 116878. The three highest ratings are 69119316, 78128208, and 78158306 respectively. Currently, 18.04% of the data are considered outliers.
        - 3) The high rating values appear reasonable as they are lower than the number of installs.
        - 4) The data might require normalization.
- 5) Histogram of Ratings
    - Univariate analysis: Histogram of Rating:
        - 1) The majority of the ratings are higher, with the highest count observed at 4.4 (2487 counts). Other significant counts include 4.3 (1016), 4.5 (976), 4.2 (887), and 4.6 (768).
        - 2) The bin ranging from 4.2 to 4.6 has the highest count according to the histogram, indicating it as the majority range.
        - 3) There are fewer counts towards the left side of the histogram.
- Histogram of Size
    - Univariate analysis: Histogram of Rating:
        - 1) The size field consists of 454 unique values, ranging from 8.5 to 100,000 kb.
        - 2) The histogram reveals that there is a higher number of apps with smaller sizes, and fewer apps towards the right side of the graph.
        - 3) The majority of the apps fall within the bin range of 10,000 to 20,000.
- All the columns, including Price, Reviews, Rating, and Size, exhibit outliers.
- 6.1.1)App with price above 200:
    - 1) Apps with prices above 200 raise suspicion as they all share the same name, "I'm Rich," which suggests they may be scam apps intended to deceive and defraud customers.
    - 2) The app "EP Cook Book" is priced at 200 but has no downloads or ratings. This app seems inappropriate to include as it is excessively expensive for a cookbook, and its categorization as "Medical" is also incorrect.
    - 3) It is recommended to remove these app data, as promoting such applications goes against ethical considerations.
- box plot of Reviews
- 6.3.2) Finding the cuttoff threshold for installs.

- In addition to the above:
- IQR = 999000.0
- Upper Fence = 2498500.0
-
- The cutoff threshold for installs is set at the upper fence value of 2,498,500.0 using Tukey's fence method.

- Q)7.1.1 What pattern do you observe? Does rating increase with price?
  - Scatterplot/joinplot observation of Rating vz Price:
    -
    - From the scatterplot and the bin plot from jointplot we can see that,
    - 1) Apps with ratings around 4.3-4.4 tend to have the highest prices, ranging from approximately 0 to 158. Additionally, this rating bin (4.3-4.4) contains the largest number of apps.
    - 2) The majority of apps are priced between 0 and 10, while the remaining apps are dispersed between 10 and 40. There are only a few apps (around 6) priced above 40.
      -
    - Conclusion: Although there is a trend suggesting that higher-rated apps may have higher prices in some cases, it's important to note that when comparing the price ranges of 4.4 and 5.0 ratings, the apps with a rating of 4.4 tend to have higher prices. Therefore, it is not necessary to conclude that apps with higher ratings are always more expensive. Thus, the observation indicates that rating does not necessarily increase with price.
- Q)7.2.1 Are heavier apps rated better?
  -
  - Scatterplot/joinplot observation of Rating vz Size:
  -
    - From the scatterplot and the bin plot from jointplot we can see that,
    - 1) The majority of apps fall into the size range of 0-10,000 kb, while the remaining apps are dispersed between 10,000-100,000 kb. The count of apps reduces as the size increases, indicating that there are fewer apps with larger sizes compared to smaller ones.
    - 2) Apps with ratings of 4.3-4.4 have the highest count, and they exhibit a wide range of app sizes, spanning from 0-100,000 kb. This rating bin (4.3-4.4) contains the largest number of apps.
    -
    - Conclusion: Although heavier apps are more common in higher ratings, it's important to note that lighter apps are also prevalent in the majority. The presence of heavier apps is relatively less compared to lighter apps in higher ratings. While the majority of heavier apps receive better ratings, not all of them do; some have average ratings. Therefore, it is not appropriate to conclude that heavier apps are consistently rated better than lighter apps. This is because

there are more lighter apps with higher ratings than heavier apps with high ratings.

- Q) 7.3.1 Does more review mean a better rating always?
    - Scatterplot/joinplot observation of Rating vz Reviews:
    -
    - From the scatterplot and the bin plot from jointplot we can see that,
    - 1) Apps with 0 reviews are uniformly distributed across the rating range of 1 to 5, indicating that there is no specific correlation between the absence of reviews and the app's rating.
    - 2) While the majority of apps with higher reviews tend to have higher ratings, it is important to note that apps with 0 reviews also have higher ratings. In fact, the count of apps with 0 reviews and higher ratings is greater than the count of apps with high reviews and higher ratings.
    -
    - Conclusion: It is evident that having more reviews generally contributes to a better rating, but this relationship is not true in all cases. There are instances where apps with fewer reviews still manage to receive better ratings in greater numbers. Therefore, while more reviews typically aid in achieving a higher rating, it is not a guarantee. Other factors may influence the rating of an app
- 7.4.1. Is there any difference in the ratings? Are some types liked better?
    - 1) There are notable differences in ratings across all content categories. Ranking the categories based on the median, we have: Adult only 18+ > Everyone > Everyone 10+ > Teen > Unrated > Mature 17+.
    -
    - 2) "Adult only 18+" content category stands out as it is more favored compared to other content types. It has a higher median rating and a smaller interquartile range (IQR) when compared to the "Everyone" category..
- 7.5.1) Which genre has the best ratings?
    - 1) The genre "Comics;Creativity and Board;Pretend Play" has the highest ratings among the genres considered. It showcases consistently positive ratings across various metrics.
    - 2) From the boxplot, it is evident that the category "ART and DESIGN" has the best ratings. It demonstrates higher values in the upper quartiles compared to other genres and boasts the highest maximum rating value, indicating its strong performance.
- Performed log transformation and dropped unnecessary variables.
- Performed the final steps, PFA in source code PDF
- Final Observations:
    - The R2 score of the training set is 0.116, indicating that approximately 11.65% of the variance in the target variable can be explained by the predictor variables in the model. This suggests a relatively weak relationship between the predictor variables and the target variable in the training data.

-
- Similarly, the R2 score of the test set is 0.072, indicating that approximately 7.21% of the variance in the target variable can be explained by the predictor variables in the model when applied to unseen test data.
-
- These results suggest that the model's performance on the test data is consistent with its performance on the training data, but still reflects a relatively weak predictive ability.
-
- Considering the correlation coefficients between the predictor variables and the target variable, none of the variables show a strong correlation. Therefore, it would be more appropriate to consider either using a different dataset or exploring alternative modeling techniques and evaluation metrics.
-
- Additionally, the correlation coefficient values for the target variable "Rating" are provided below:
-
- Correlation coefficients for target variable "Rating":
- 1.000000 -0.028403 -0.119659 0.028496 0.002698 0.028610 0.004383 0.017988 0.047664 0.009433 ... 0.004496 0.006606 -0.080340 -0.025298 -0.001833 -0.006362 0.004496 -0.045823 0.010178 0.007054

■