

**Programming Basics and Data Analytics with Python**

**Course-End Project: 1**

**App Rating Prediction**

**Submitted By**

**Aruna Rajeswari Sivaperuman**

1. Load the data file using pandas.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Check for null values in the data. Get the number of null values for each column.

A computer screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Drop records with nulls in any of the columns.

A white screen with black text

AI-generated content may be incorrect.



1. Variables seem to have incorrect type and inconsistent formatting. You need to fix them:
2. Size column has sizes in Kb as well as Mb. To analyze, you’ll need to convert these to numeric.
   1. Extract the numeric value from the column

A screenshot of a computer code

AI-generated content may be incorrect.

A white background with black text

AI-generated content may be incorrect.

* 1. Multiply the value by 1,000, if size is mentioned in Mb

A screen shot of a computer code

AI-generated content may be incorrect.

A number with numbers on a white background

AI-generated content may be incorrect.

1. Reviews is a numeric field that is loaded as a string field. Convert it to numeric (int/float).

A screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Installs field is currently stored as string and has values like 1,000,000+.
   1. Treat 1,000,000+ as 1,000,000
   2. remove ‘+’, ‘,’ from the field, convert it to integer

A screenshot of a computer code

AI-generated content may be incorrect.

A computer screen shot of a number

AI-generated content may be incorrect.

1. Price field is a string and has $ symbol. Remove ‘$’ sign, and convert it to numeric.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

5. Sanity checks:

1. Average rating should be between 1 and 5 as only these values are allowed on the play store. Drop the rows that have a value outside this range.

A computer code with colorful text

AI-generated content may be incorrect.

1. Reviews should not be more than installs as only those who installed can review the app. If there are any such records, drop them.

A screenshot of a computer

AI-generated content may be incorrect.



1. For free apps (type = “Free”), the price should not be >0. Drop any such rows.

A screenshot of a computer code

AI-generated content may be incorrect.



5. Performing univariate analysis:

* Boxplot for Price
  + Are there any outliers? Think about the price of usual apps on Play Store.

A screenshot of a computer code

AI-generated content may be incorrect.

A graph of a box plot

AI-generated content may be incorrect.

* Boxplot for Reviews
  + Are there any apps with very high number of reviews? Do the values seem right?

A screen shot of a computer code

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

* Histogram for Rating
  + How are the ratings distributed? Is it more toward higher ratings?

A screenshot of a computer program

AI-generated content may be incorrect.

A graph of a distribution of app ratings

AI-generated content may be incorrect.

* Histogram for Size

Note down your observations for the plots made above. Which of these seem to have outliers?

A screen shot of a computer code

AI-generated content may be incorrect.

A graph of a distribution of apps

AI-generated content may be incorrect.

6. Outlier treatment:

1. Price: From the box plot, it seems like there are some apps with very high price. A price of $200 for an application on the Play Store is very high and suspicious!
   1. Check out the records with very high price
      1. Is 200 indeed a high price?

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A white screen with black text

AI-generated content may be incorrect.

* 1. Drop these as most seem to be junk apps

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

1. Reviews: Very few apps have very high number of reviews. These are all star apps that don’t help with the analysis and, in fact, will skew it. Drop records having more than 2 million reviews.

A computer screen shot of text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Installs:  There seems to be some outliers in this field too. Apps having very high number of installs should be dropped from the analysis.
   1. Find out the different percentiles – 10, 25, 50, 70, 90, 95, 99

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* 1. Decide a threshold as cutoff for outlier and drop records having values more than that

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

7. Bivariate analysis: Let’s look at how the available predictors relate to the variable of interest, i.e., our target variable rating. Make scatter plots (for numeric features) and box plots (for character features) to assess the relations between rating and the other features.

1. Make scatter plot/joinplot for Rating vs. Price
   1. What pattern do you observe? Does rating increase with price?

A screen shot of a computer code

AI-generated content may be incorrect.

A graph of blue dots

AI-generated content may be incorrect.

1. Make scatter plot/joinplot for Rating vs. Size
   1. Are heavier apps rated better?

A computer code with many colorful text

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

A graph of a number of blue dots

AI-generated content may be incorrect.

1. Make scatter plot/joinplot for Rating vs. Reviews
   1. Does more review mean a better rating always?

A screenshot of a computer code

AI-generated content may be incorrect.

A graph of a number of reviews

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.

1. Make boxplot for Rating vs. Content Rating
   1. Is there any difference in the ratings? Are some types liked better?

A screenshot of a computer program

AI-generated content may be incorrect.

A graph of a box diagram

AI-generated content may be incorrect.

1. Make boxplot for Ratings vs. Category
   1. Which genre has the best ratings?

A screenshot of a computer code

AI-generated content may be incorrect.

A graph with blue and black rectangles

AI-generated content may be incorrect.

For each of the plots above, note down your observation.

8. Data preprocessing

For the steps below, create a copy of the dataframe to make all the edits. Name it inp1.

1. Reviews and Install have some values that are still relatively very high. Before building a linear regression model, you need to reduce the skew. Apply log transformation (np.log1p) to Reviews and Installs.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Drop columns App, Last Updated, Current Ver, and Android Ver. These variables are not useful for our task.

A screenshot of a computer program

AI-generated content may be incorrect.

A white background with black text

AI-generated content may be incorrect.

1. Get dummy columns for Category, Genres, and Content Rating. This needs to be done as the models do not understand categorical data, and all data should be numeric. Dummy encoding is one way to convert character fields to numeric. Name of dataframe should be **inp2**.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

9. Train test split  and apply 70-30 split. Name the new dataframes df\_train and df\_test.

A computer code with text

AI-generated content may be incorrect.

A close up of words

AI-generated content may be incorrect.

10. Separate the dataframes into X\_train, y\_train, X\_test, and y\_test.

A screenshot of a computer code

AI-generated content may be incorrect.

A number and text on a white background

AI-generated content may be incorrect.

11 . Model building

* Use linear regression as the technique
* Report the R2 on the train set

12. Make predictions on test set and report R2.

A screenshot of a computer program

AI-generated content may be incorrect.

A computer code with text

AI-generated content may be incorrect.

A black text on a white background

AI-generated content may be incorrect.