Pattern Recognition Report

Team Id: CS_19

Id	الاسم
20191700732	هاني محمد سيد أحمد عبد الهادي
20201700540	عمر علاء السيد محمد
20201700099	ادهم احمد سامي عنتر
20201700983	يحيى مجدي عبد الوهاب بكري
20191700795	إبتهال ملك الناصر سيد أحمد
20191700300	سميره يحيى خليل على

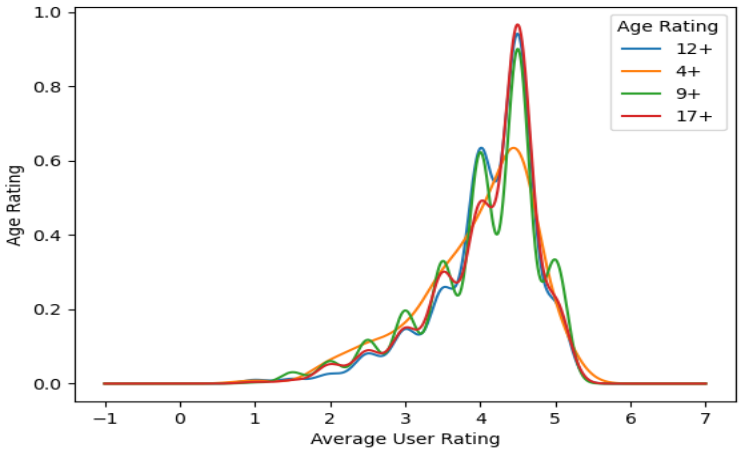
Preprocessing

- **Drop:** we dropped columns that have unique values (values that isn't repeated) (Id and Name columns).
- **Dropped columns** that have unique values with percentage 55% and we dropped columns that have missing values with percentage 35%.
- **Dropped column** (primary genres)? as that column has the same values of the column genres.
- **Split data into 80% training and 20% with shuffling testing then applies preprocessing.**

- **Apply preprocessing both on training and test:**

    - **Replace missing values** in columns (languages, genres, User Rating Count, Price, Age Rating) with Mode value (most their repeated values) and columns (Original Release Date, Current Version Release Date) with median value.

    - **Apply one hot encoding** in some categorical columns (Languages, Genres, Age Rating).

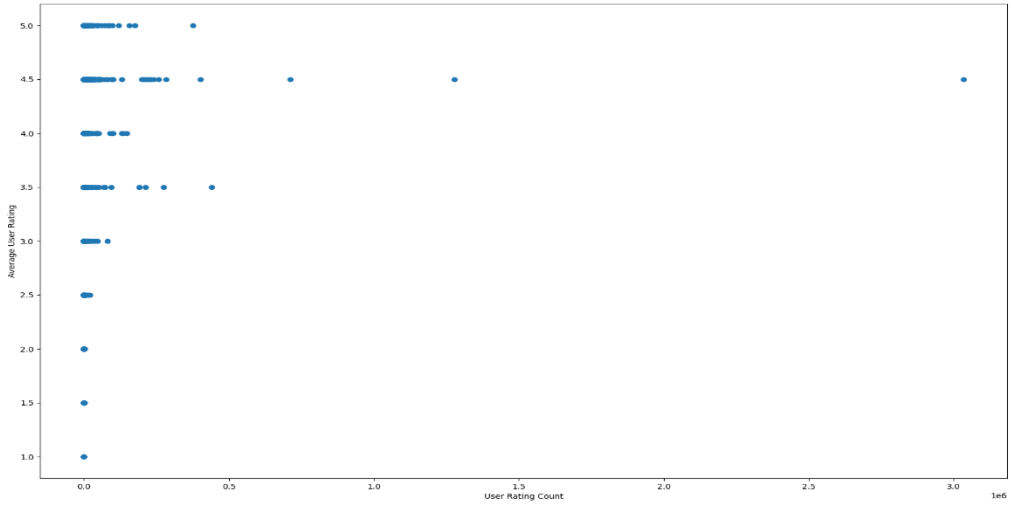


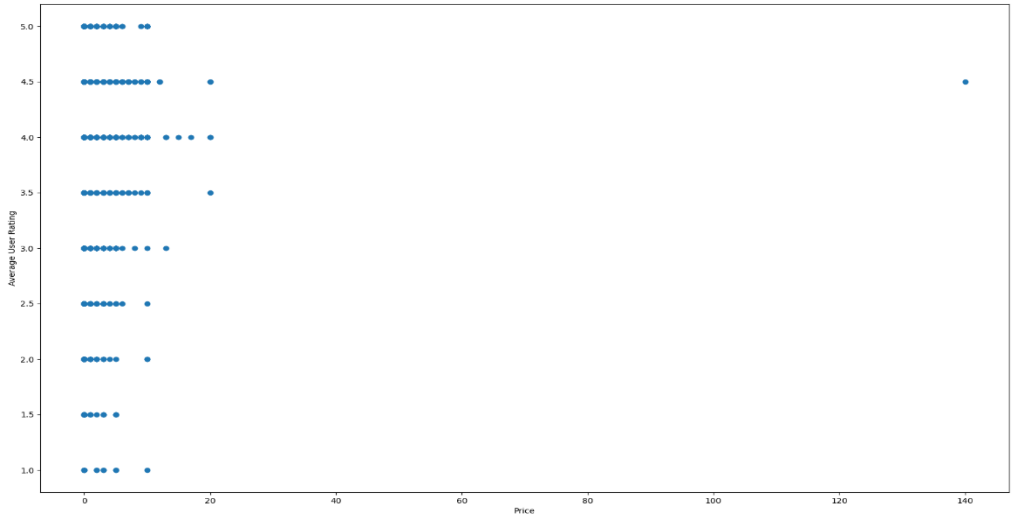
- **Date column** converts it to numerical data? how? We subtract it from the actual date (today's date).

- Drop columns that have **zero standard deviation** (has only one value in column).

- Visualize data to detect outliers and understand distribution of data.

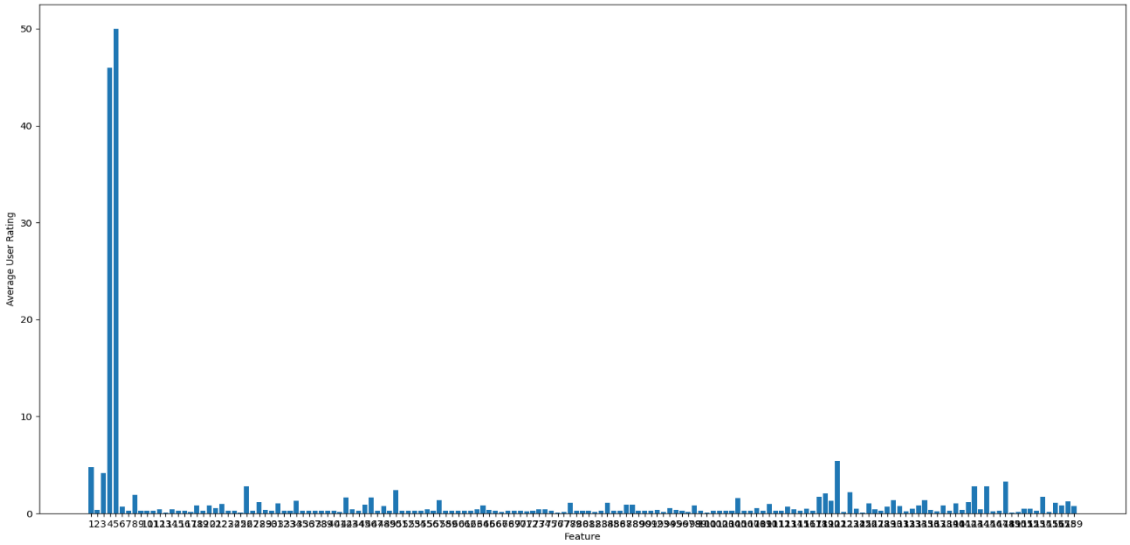
- **Handle outliers** in some columns (user rating count and price) by finding z score and get rows that z score is greater than 5 and replace it with the next or previous value in row.





Feature selection:

- Feature selection using Anova selector technique. (Because we have continuous data).
- After Appling more experiments we get the most affected features on target column (34 columns) based on max p values.



This clarifies the effect of features with the Average User Rating

Feature scaling:

- Apply feature scaling in chosen features by StandardScaler because we want each feature to have zero-mean, unit standard-deviation and doesn't affect if there is another outlier in data.

Regression techniques:

- Linear regression.
- Lasso and Ridge regression.
- Support Vector Regression.

The differences between each model:

- **Linear regression:** to find the best-fitting line that describes the relationship between the independent and dependent variables.
    - **Mean Square Error of training Model: 0.520684934841523**
    - **Mean Square Error of testing Model: 0.44534058571422047**

- **Lasso and Ridge regression:** is a type of linear regression that uses regularization to minimize the coefficients of some variables towards zero.
    - **Mean Square Error of training Ridge Model: 0.5206849348758612**
    - **Mean Square Error of testing Ridge Model: 0.4453407303388169**

- **Support Vector Regression**: based on support vector machines and tries to find a hyperplane that maximizes the margin between support vectors (the closest data points).
    - **Mean Square Error of training Model: 0.472381381794463**
    - **Mean Square Error of testing Model: 0.44999676759323515**

improving the results:

We will try to improve the preprocessing by adding new features and model by applying cross validation.

Conclusion:

In this project, it is filled with many categorical values and outliers that must be dealt with in specific way, and relies very heavily on preprocessing, so for this we have put great attention to preprocessing and also to models, but the increased attention is paid to preprocessing, and each of the models gave very good results, but Support vector regression Model excelled over the others.

regression line plots: