Introduction

Wine is a popular alcoholic beverage made from fermented grapes or other fruits. Different types of wine are produced using different grape varieties, fermentation techniques, and aging processes. The four main types of wine are red, rose, sparkling, and white wine. Red wine is made from red or black grapes, which are crushed and fermented with the grape skins, giving it its distinctive red color. Red wine is typically served at room temperature and pairs well with red meat and rich, hearty dishes. Rose wine, also known as blush wine, is made from a combination of red and white grapes. The skin of the grapes is left in contact with the juice for a short period of time, giving it a pinkish color. Rose wine is usually served chilled and is a popular choice for summer drinking.

Sparkling wine, as the name suggests, is carbonated and has bubbles in it. It is typically made using the same methods as champagne, which is a type of sparkling wine that comes from the Champagne region of France. Sparkling wine is often associated with celebrations and is a popular choice for toasting. White wine is made from white or green grapes, which are pressed and fermented without the grape skins. White wine is usually served chilled and pairs well with fish, chicken, and other light dishes. Understanding the differences between these four types of wine can help you choose the right wine for different occasions and meals. A dataset of these wines could provide insights into their production, popularity, and consumption patterns, which could be useful for wine producers, sellers, and enthusiasts alike.

Dataset

Initially, there are four types of datasets of wine each for each category containing the columns:

- Name
- Country
- Region
- Winery

- Rating
- Number of Ratings
- Price
- Year

All dataset contains same columns but each data set has different category i.e., red, rose, sparklin g and white. The Shape of four dataset is

- (8666,9) for red
- (397,9) for rose
- (1007,9) for sparkling
- (3764,9) for white

We concat the four datasets into one dataset by adding one column named ‘Wine-Category’ whic h hold the information of wine means it will tell whether the wine is of red, rose, sparkling or whi te category. The Shape of resultant dataset is (13834,10) containing 10 columns with 13834 insta nces of that columns.

Data Analysis

Data analysis refers to the process of inspecting, transforming, and modeling data with the aim of extracting useful information that can be used to make informed decisions. The goal of data analysis is to find meaningful patterns, relationships, and insights from a given data set.

Information of Dataset

Figure 1 shows the datatype of column of dataset as can be seen that only rating and number of rating column are in integer and float format other all are of object type.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13834 entries, 0 to 1006
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Name             13834 non-null  object
 1   Country          13834 non-null  object
 2   Region           13834 non-null  object
 3   Winery           13834 non-null  object
 4   Rating           13834 non-null  float64
 5   NumberOfRatings  13834 non-null  int64
 6   Price            13834 non-null  float64
 7   Year             13834 non-null  object
 8   Wine_Category    13834 non-null  object
dtypes: float64(2), int64(1), object(6)
memory usage: 1.1+ MB
```

Figure 1 Datatype of all columns

Non-vintage wine is wine that is made by blending grapes from different years or vintages, rather than from a single year. The purpose of blending different vintages is to create a consistent and reliable taste year after year, regardless of the variability of the grapes in a particular season. Non-vintage wine is typically produced using a combination of different grape varieties, which are often sourced from different vineyards or regions.

The labeling of non-vintage wine typically does not include a year, as it is not made from grapes harvested in a single year. Instead, the label may indicate the winery or brand name, the grape varietals used, and the region where the grapes were grown. Non-vintage wines are most commonly found in sparkling wines, such as Champagne, as well as in fortified wines, such as Port or Sherry. In this example, we convert the 2050 year to non-vintage.

Head of Dataset

The head of the dataset is shown in figure 2

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Wine_Category
0	Pomerol 2011	France	Pomerol	Château La Providence	4.2	100	95.00	2011	red
1	Lirac 2017	France	Lirac	Château Mont-Redon	4.3	100	15.50	2017	red
2	Erta e China Rosso di Toscana 2015	Italy	Toscana	Renzo Masi	3.9	100	7.45	2015	red
3	Bardolino 2019	Italy	Bardolino	Cavalchina	3.5	100	8.72	2019	red
4	Ried Scheibner Pinot Noir 2016	Austria	Carnuntum	Markowitsch	3.9	100	29.15	2016	red

Figure 2 Head of resultant dataset

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyze and summarize data sets to uncover patterns, relationships, and anomalies in the data. The primary goal of EDA is to understand the data and its characteristics, identify patterns and trends, and explore the relationships between variables.

EDA is typically performed before the formal modeling stage, as it helps to identify the relevant features and variables to include in the model. EDA involves using various techniques to examine the data visually and numerically, such as histograms, scatter plots, box plots, and summary statistics. If we look at countries that have highest volume of wine entries that it can be seen in figure 3 that Italy has highest volume of wine entries around 4000 while France with 3500 entries has second number in this race. On the other hand Australia has lowest entries among top 10 countries.

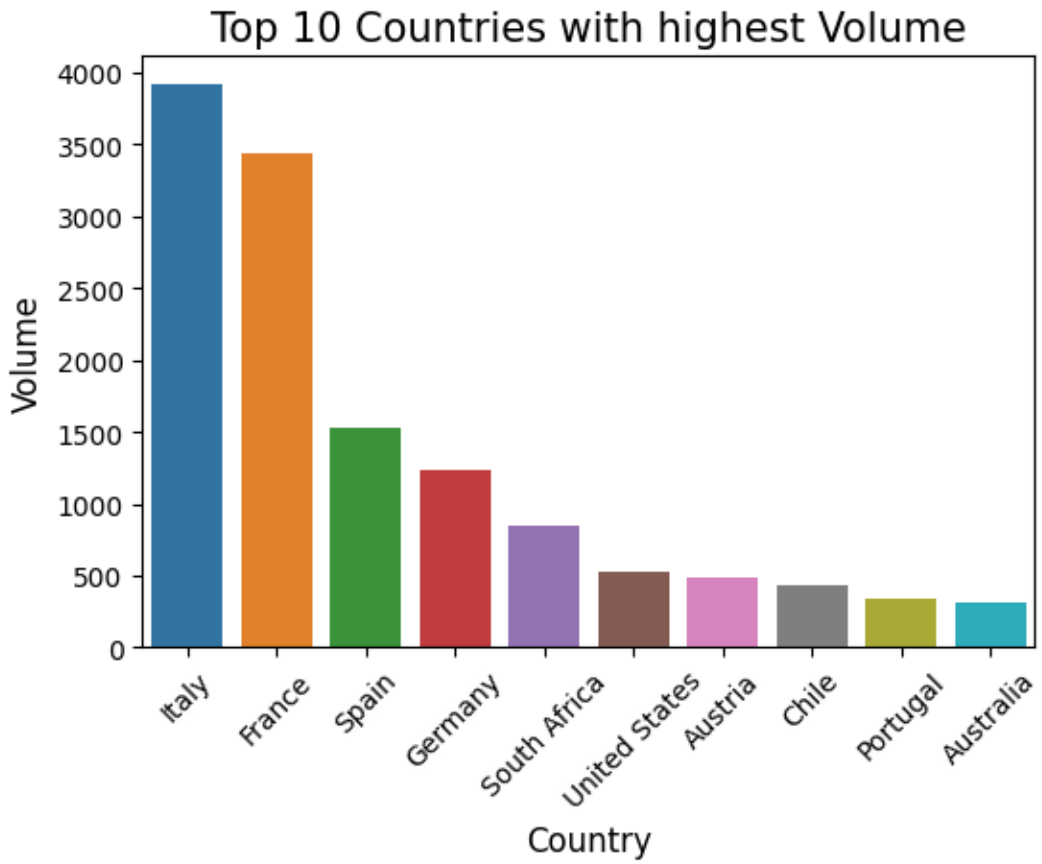

Figure 3 Volume of Countries

If we look at regions which has highest volume then we can see in figure 4 that Rioja is the region which has more than 350+ entries while Barolo has the last frequency then but if we look at year-wise record than it was found that 2018 year has most record and the year wise record show that 2011 year has lowest among the top ten-year record. While It can be seen that 2018 and 2017 year has most sale compare to other year as shown in figure 5

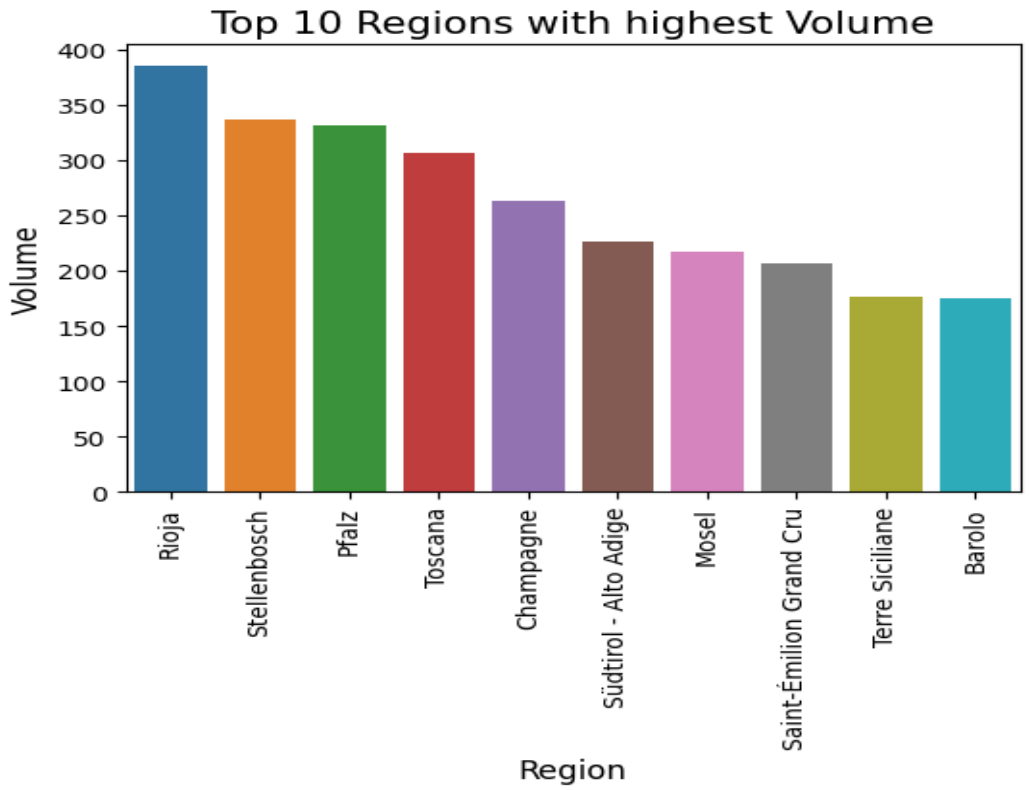

Figure 4 Volume of Top ten Region

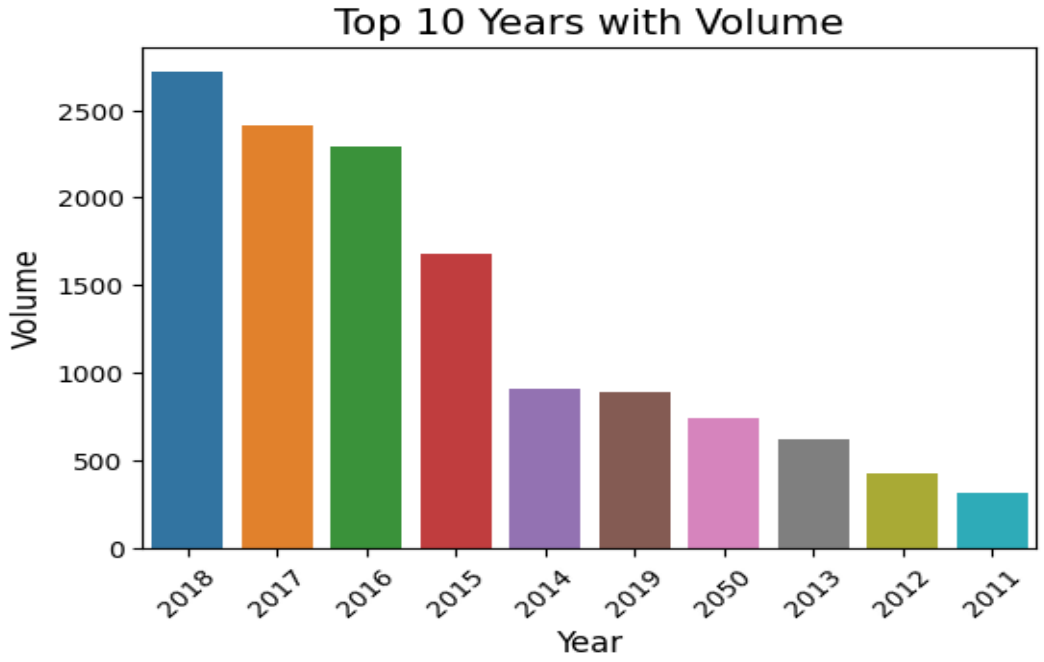

Figure 5 Volume record of top ten year

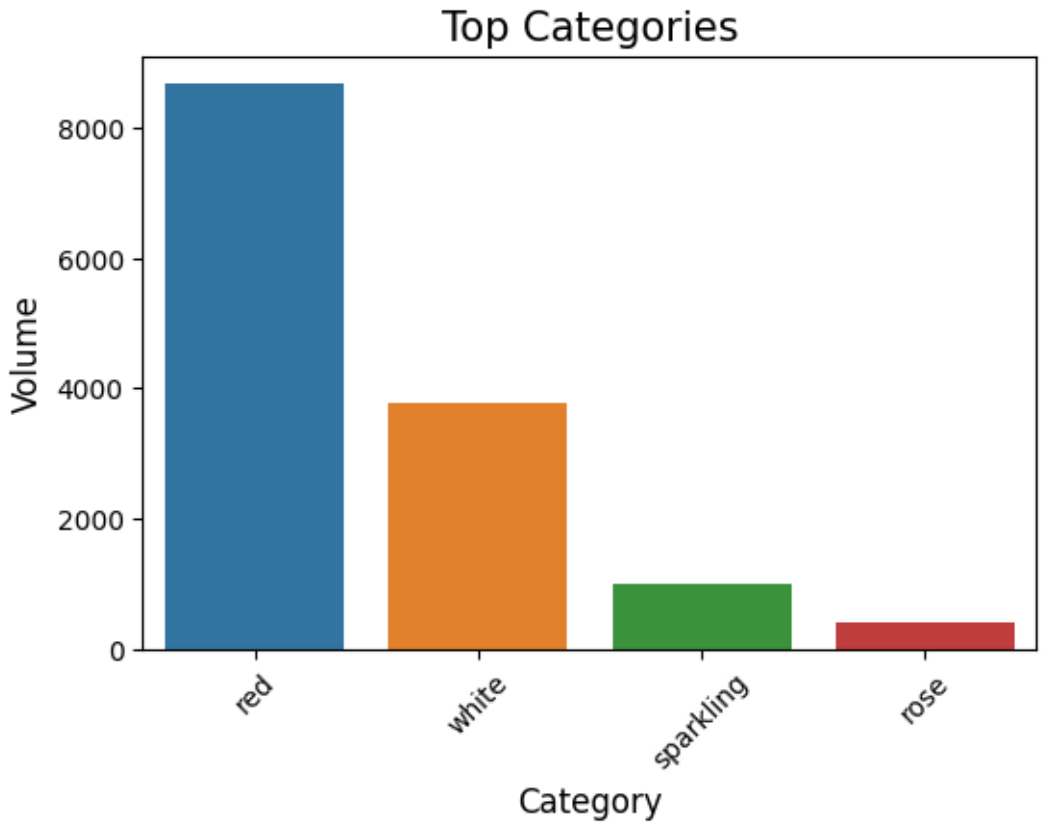

Figure 6 Top Category of Wine

Figure 6 shows the volume of the wine according to their category and it can be seen in graph that red wine has most record (more than 8000+) and rose wine has least amount of the wine compare to other 3 categories. If we look at distribution of wine rating then it can be seen in figure 7 that most of the people give rating between 3.6 to 4.2 rating and very few people give 5+ rating and low rating and one interesting pattern the rating distribution shows that the distribution is quite similar to normal distribution. Figure 8 shows Number of rating distribution and it can be seen in figure that most of the people give 3 to 5 rating in records.

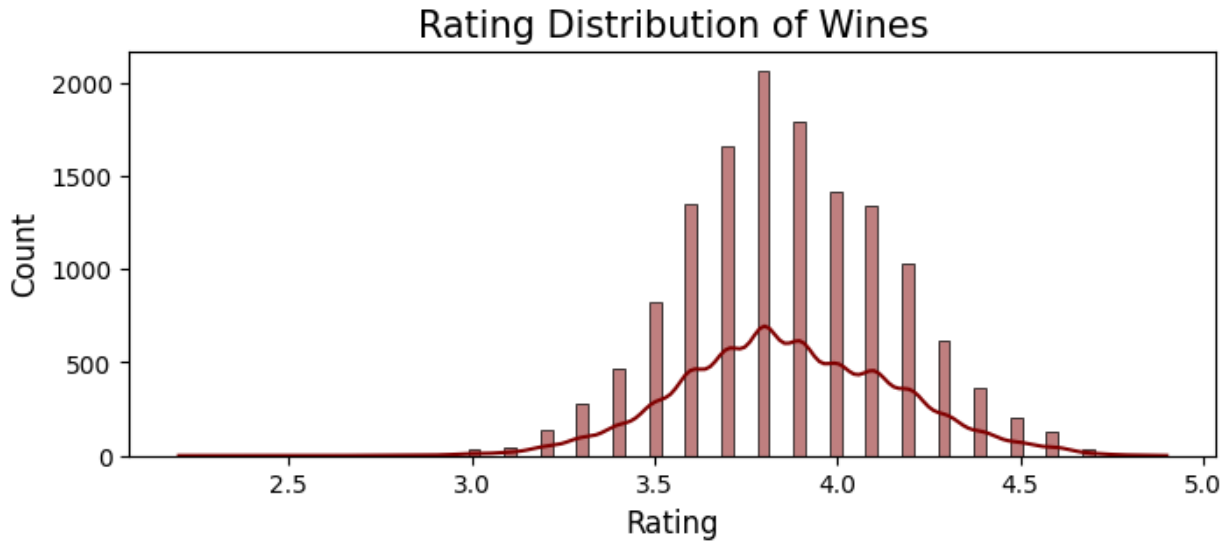

Figure 7 Rating distribution of the wine

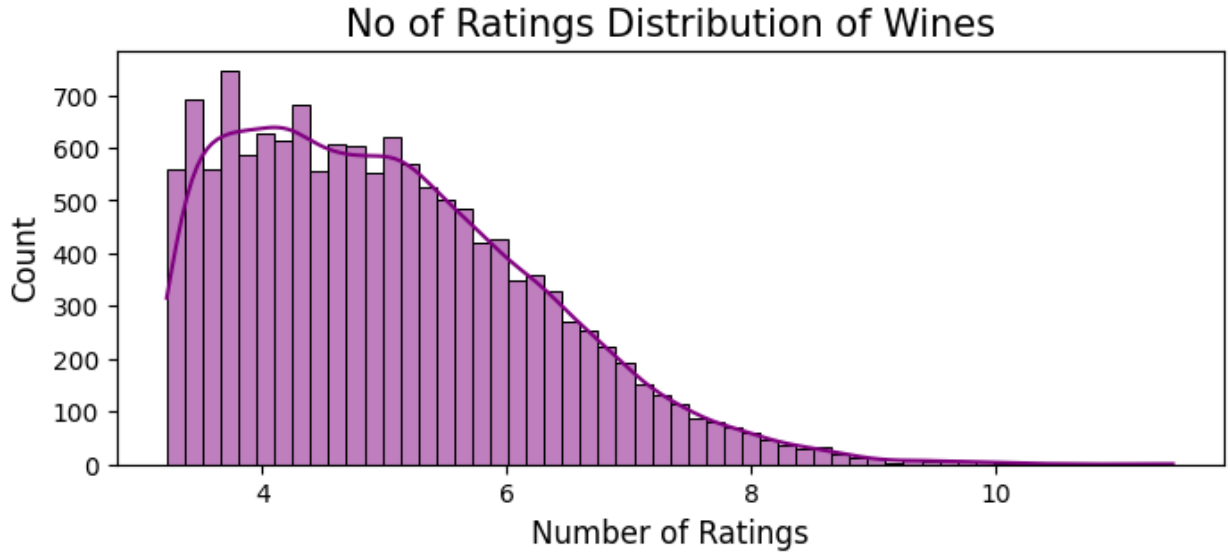

Figure 8 Number of Rating distribution of wine

If we look at rating according to countries it can be seen that China, Mexico and Turkey are the countries which give poor rating while Moldova and Lebanon and Croatia are the countries that have best rating as shown in figure 9. Figure 10 shows that UK, France and US have highest price for wine while Mexico and Bulgaria are the countries which has lowest price of wine in their countries.

	Country	Rating
0	China	3.200000
1	Mexico	3.400000
2	Canada	3.433333
3	Brazil	3.510000
4	Turkey	3.660000

	Country	Rating
0	Moldova	4.175000
1	Lebanon	4.137500
2	Croatia	4.083333
3	Czech Republic	4.050000
4	United Kingdom	4.033333

Figure 9 Countries with top 5 best and poor rating

	Country	Price
0	United Kingdom	57.770000
1	France	55.539331
2	United States	43.719170
3	Lebanon	36.887500
4	Australia	36.781132

	Country	Price
0	Mexico	8.650000
1	Bulgaria	10.150000
2	Romania	12.816216
3	Slovenia	13.050000
4	Hungary	13.366842

Figure 10 Countries with prices

If we look at number of rating column with respect to year then it can be seen in figure 11 that Year 2050 have highest number of ratings but as we know that we put 2050 to that value when the 2 years combined for that wine. So, we can say that 1989 have the Highest Number of Ratings while if we look at lowest it can be seen that year 2020 has lowest number of rating record in it.

	Year	NumberOfRatings
0	2050	2125.002688
1	1989	1075.000000
2	1993	886.500000
3	2002	834.000000
4	2008	629.386139

	Year	NumberOfRatings
0	2020	61.500000
1	1996	104.000000
2	1995	156.000000
3	2019	189.458007
4	1988	203.000000

Figure 11 Number of rating year-wise

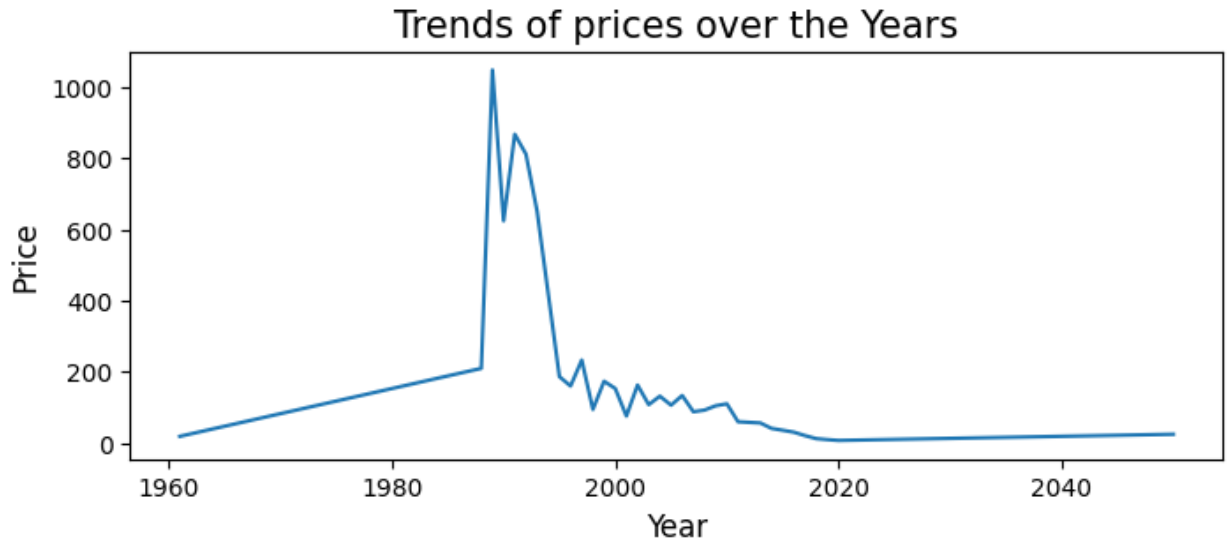

Figure 12 Trends of Prices over the years

Figure 12 shows the trend of prices over the years it can be seen that the hike to the graph is seen between the year 1990 to 1995 while it lower the price between 1998 to 2008. Figure 13 shows the trends of prices with rating and it can bee seen as rating goes higher the price also goes rises which shows the direct relation between them.

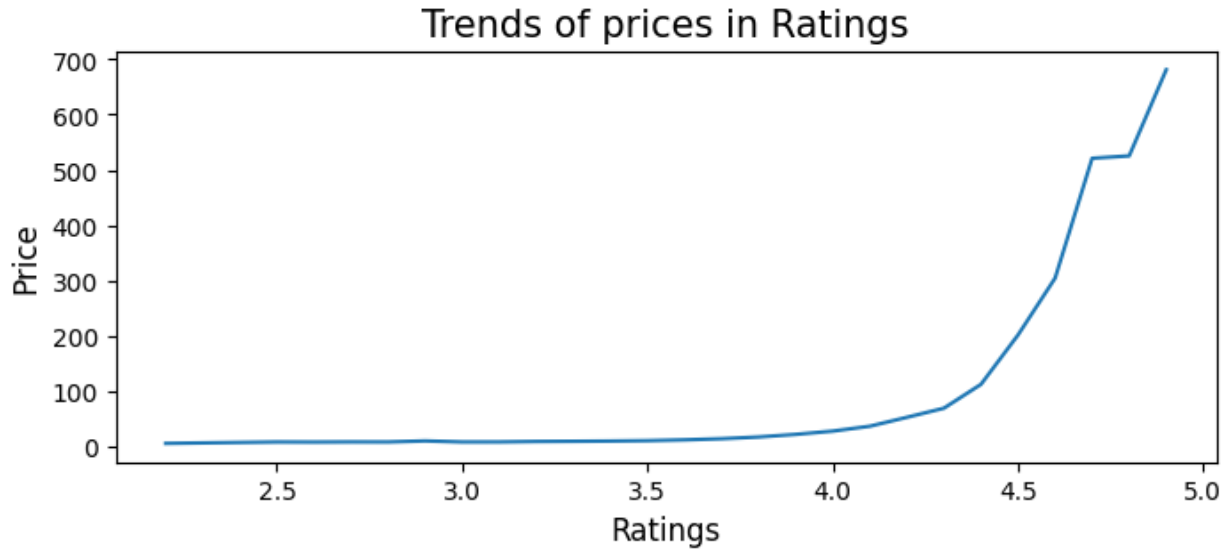

Figure 13 Trends of rating over the years

Correlation Graph

A correlation graph is a graphical representation of the strength and direction of the relationship between two variables. In this graph, the values of one variable are plotted on the x-axis, and the values of the other variable are plotted on the y-axis. Each point on the graph represents a pair of values for the two variables.

The correlation between the two variables can be determined by calculating the correlation coefficient, which is a measure of the linear relationship between the two variables. The correlation coefficient can range from -1 to 1, where a value of -1 indicates a perfect negative correlation (as one variable increases, the other decreases), a value of 0 indicates no correlation, and a value of 1 indicates a perfect positive correlation (as one variable increases, the other increases).

The correlation graph can be used to visualize the relationship between two variables and to identify any patterns or trends in the data. It can also be used to identify any outliers or unusual observations that may be affecting the correlation coefficient. Figure 13 shows that correlation graph between different columns shows that price and rating has highest correlation between them which is 0.45 while rating and year has lowest correlation between them which is -0.17.

Figure 14 Correlation graph of dataset

Label Encoding

Label encoding is a process of transforming categorical data into numerical values. In this method, each unique category is assigned a numerical value, usually starting from 0 or 1 and continuing sequentially. The resulting numerical values are then used in place of the original categorical values for further analysis or modeling.

For example, consider a dataset containing a categorical variable "color" with categories "red", "green", and "blue". Using label encoding, we can assign the values 0, 1, and 2 to these categories, respectively. In this example, the label encoding is applied to encode all the column of the dataset.

Machine Learning Algorithm

Linear regression is a type of statistical analysis that is used to predict the relationship between two variables, typically a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and uses the least squares method to estimate the coefficients of the regression equation. Linear regression is widely used in various fields such as finance, economics, and social sciences.

Lasso regression, on the other hand, is a method of regression analysis that is used to perform variable selection and regularization. It is particularly useful when dealing with datasets that have a large number of variables, as it can help to identify the most important variables and reduce overfitting. Lasso regression achieves this by adding a penalty term to the linear regression equation that encourages the coefficients of certain variables to be reduced to zero.

Ridge regression is similar to Lasso regression in that it is also used for variable selection and regularization. However, Ridge regression uses a different penalty term that adds a squared term to the coefficients of the variables, rather than an absolute value term like in Lasso regression. This makes Ridge regression better suited for situations where there are many small coefficients rather than a few large ones.

Overall, these three methods of regression analysis have their own strengths and weaknesses and are useful in different situations. Linear regression is a good starting point for most regression analyses, while Lasso and Ridge regression can be used to address specific issues such as variable selection and regularization. Result shows that mean square error of all the three algorithm is 0.08.

```
Linear Regression:
Mean squared error: 0.08
R2 score: 0.06
Lasso Regression:
Mean squared error: 0.08
R2 score: 0.11
Ridge Regression:
Mean squared error: 0.08
R2 score: 0.06
```

R-squared of 0.06 is not a good fit for the data and suggests that there is little to no relationship between the independent variable(s) and the dependent variable. It is important to note that a low R-squared value does not necessarily mean that the independent variable(s) are not significant or that the model is useless, but it does suggest that the model is not a strong predictor of the dependent variables.