**Name: Pratik Premkishor Mantri**

**NUID:001353214**

**Final Milestone project**

**Appendix:**

|  |  |
| --- | --- |
| Index | Page no |
| Summary | 2 |
| Data analysis | 3-11 |
| References | 12 |

**Summary of exploratory data analysis:**

**Part1**

* ‘maximum price of wine belongs to Germany which is around $750. From the above graph is crystal clear that from the old wines (France and Italy) France is on the 2nd highest expensive in terms of wines, but Italy does not look that much expensive. There are many wines from ‘new wine category which are more expensive than the Italy (which belongs to Italy).
* ‘Chardonnay’ is the expensive among all the wines. Moreover, we can say that Rating for Muscat is highest among all.
* The country who is highest in buying wines is nothing, but USA and the least won is CANADA
* we can say that rating is highest when the price is between $300-$400. The most wines comes in the range of $0 to $100.
* Muscat has the highest rating of 100 which belongs to country Australia, with winery ‘chambers rosewood vineyards and taster name, Joe Czerwinski.

**Part-2**

* **Statistical hypothesis testing:**
* Hypothesis testing is a type of inferential statistics which helps us draw us conclusions about the entire population sample which is based on a sample. We get tremendous benefit from this as it is very difficult to observe data of entire population to understand its properties. The only way to take random samples and use statistics to analyses it.
* **Introduction to One-sample T-test in R:**
* **What is one-sample T-test:**
* One sample T-test is nothing but used to compare **the mean** of a sample to a known standard mean (the theoretical mean)
* In the one sample-T-test we have solved the average rating of the wine is 76, where our assumption came as false, that is the average rating of the wine was not 76 but 86.43.
* **P-value:**
* p value is probability of obtaining the results which are as extreme as the one which are observed. There are two concepts in P-value analysis. **“Likely”** and “**unlikely”.** If the **p value >0.05**(which is level of confidence), then we can say its likely that we will **reject the Null hypothesis** **and p value is significant to reject the null hypothesis**. If the **(P ≤ 0.05**) then we can say it **“unlikely” to reject the null hypothesis and it is not significant to reject the null hypothesis**
* **Two-sample T-test:**
* It is applied to compare whether the average difference between two groups is significant or it is just due to random chances.

we want to compare the differences between the average prices of wine between the country US and Spain, which was false assumption and selected the alternative hypothesis that the average prices of wine in US are statistically different than the average prices of wine in Spain.

**Regression of data and partial F-test**

In this example I want to see whether or not there is relation between price and rating.

Dependent variable is Price

Independent variable is Rating

Lets check if the data between two variable is linear or not:

Text

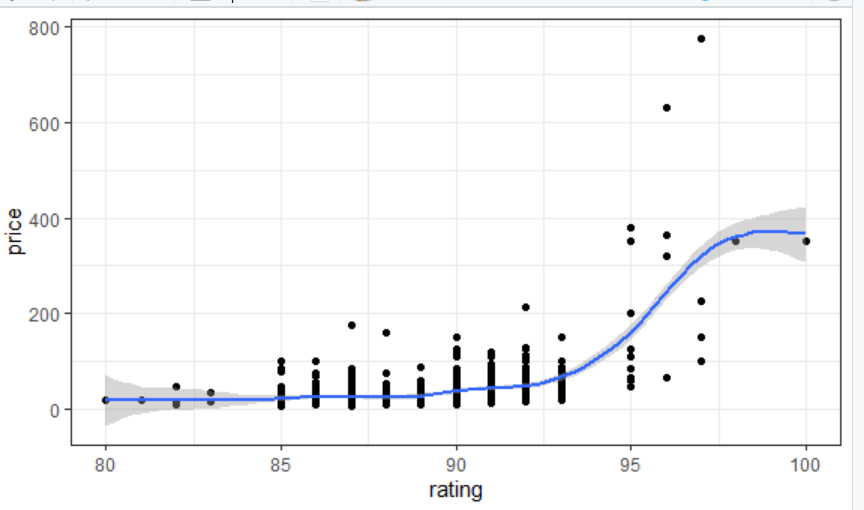
Description automatically generated

Here, the R squared value is very low which 0.2232. Which shows us there is little dependency among the variable.

In this example we will be looking for two regression models, and comparing them

1. linear regression model

2. Polynomial non-linear model



The above scatter plot shows that a non-linear relationship between the two variables which is rating and price.

The **RMSE** represents the model **prediction error**, that is the average difference the observed outcome values and the predicted outcome values. The **R2 represents the squared correlation between the observed and predicted outcome values.** The best model is the model with the lowest RMSE and the highest R2.

Let’s build the model.

Graphical user interface, text, application

Description automatically generated

**For linear model**

**RMSE:76.63**

**R2: 0.29**

The RMSE which represent the **prediction error is too high so let’s build non-linear model** for comparison.

**visualize the data**

Graphical user interface

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

Sometimes, the true relationship between the outcome and a predictor variable may not be linear.

Sometimes we use non-linear regression for that:

* **Polynomial regression**. Used to build a non-linear relationships. It add polynomial or quadratic terms to the regression.
* **Spline regression**. There are series of polynomial segments which fits the smooth curve. The spline segments are called **Knots**.
* **Generalized additive models** (GAM). Fits spline models with automated selection of knots.

Polynomial Regression:

The polynomial regression adds polynomial or quadratic terms to the regression equation, which as follow:

price=b0+b1∗rating+b2∗rating

Lets now build the polynomial model to see the difference between linear and non-linear

Graphical user interface, text, application

Description automatically generated

Here we get two outputs that contains two coefficients associated with rating : one for the linear term (rating^1) and another for the quadratic term (rating^2).

Lets now compute six-order polynomial fit

Table

Description automatically generated with medium confidence

* **Estimate:** the intercept (b0) and the beta coefficient estimate which are associated to each predictor variable
* **Std.Error**: the standard error of the coefficient estimates. This shows the accuracy of the coefficients. The larger the std error, the less confident we are about the estimate.
* **t-value**: the t-statistic, is a coefficient estimate divided by the standard error of the estimate.
* **Pr(>|t|):** The p-value which is corresponding to the t-statistic. The smaller the value of p, the more significant the estimate is.

To investigate the relationship between variables we use correlation. The correlation coefficient is denoted by **' r ' (between -1 and +1)** and quantifies the linear association's direction and strength between the variables. The correlation can be positive or negative between two variables. The sign of the 'r' indicates the direction of the association, and the magnitude of the 'r' indicates the strength.

Its value ranges between -1 (perfect negative correlation: when x increases, y decreases)

A value **closer to 0 suggests a weak relationship between the variables.**

The R squared value is 0.43.

If the r^2 is close to 1 tells us that the actual values and the predicted values are close to each other. In our example is almost 43% close to each other.

In contrast, a low value lets says 0.004 tells us the distance between predicted and actual values is higher. Moreover, is r^2=1, we can say we can predict any value of y for any given value of x.

In that, polynomial terms beyond the fifth order are not significant.

So, just create a fifth order polynomial regression model.

Text

Description automatically generated

**For non-linear model (polynomial model)**

**RMSE: 58.61 R2: 0.71**

**Visualize the data**

**Chart

Description automatically generated**

From above two models, we can say that,

* For linear model: RMSE=76.63 and R2= 0.29
* For polynomial model: RMSE: 58.61 and R2: 0.71

**We know that the best model is that model with the lowest RMSE and the highest R2.**

From the above values we can conclude that the **polynomial regression model, outperform the linear regression model.**

**Using partial F-test we can compare two models (linear and polynomial)**

**Null hypothesis**: There is no significant difference between two models.

**Alternative hypothesis**: The full model (with rating^2) is significantly better.

Text

Description automatically generated

We have a **p-value <0.05(level of significance)**

Hence, we can **reject** the Null hypothesis and select the alternate hypothesis which is The full model (with rating^2) is significantly better.

Lets now predict the data:

Text

Description automatically generated

So here we can say that, From that we can take references as for wines with the ratings of 90 the price would be around 33.25$ and for the rating of 100 the price would be approximately 350$.

We want to see is their any relation between rating and price of country Spain?

Here we will take.

price as a **dependent variable** and

Rating as the **independent variable**.

**Null hypothesis**: There is no relation ship between price and rating of a country Spain.

**Alternate hypothesis**: Price increases with rating for country Spain.

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

We can graphically say that the price is increases with the rating

Linear regression:

**Linear regression:**

  It is the most simple and popular technique for predicting a **continuous** variable. It shows a linear relationship between the outcome and the predictor variables.

It can be written as y = b0 + b\*x + e, where:

* b0 is the intercept,
* b is the regression weight or coefficient associated with the predictor variable x.
* e is the residual error.

Lets now plot the linear regression for this

Text

Description automatically generated

* **Estimate:** the intercept (b0) and the beta coefficient estimate which are associated to each predictor variable
* **Std.Error**: the standard error of the coefficient estimates. This shows the accuracy of the coefficients. The larger the std error, the less confident we are about the estimate.
* **t-value**: the t-statistic, is a coefficient estimate divided by the standard error of the estimate.
* **Pr(>|t|):** The p-value which is corresponding to the t-statistic. The smaller the value of p, the more significant the estimate is.

**A low correlation (-0.2 < x < 0.2) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x)**

**Here the R squared is 0.32 so we can use linear model.**

**We know that:**

R-square value gives us estimation of variation by our model. So, **0.32** R-square means that **32%** of variation within the data. The bigger the R-square the better the model. Whereas p-value tells you about the F statistic hypothesis testing of the model. So, if the **p-value(8.692e-14)** is less than the significancelevel (usually 0.05) then it fits the data well.

There are **four** scenarios:

**1) low R-square and low p-value (p-value <= 0.05):**

model does not say much of variation of the data, but it is significant.

**2) low R-square and high p-value (p-value > 0.05):**

: model does not say much of variation of the data and it is not significant (worst scenario)

**3) high R-square and low p-value:**

explains a lot of variation within the data and it is significant (best scenario)

**4) high R-square and high p-value:**

model explains a lot of variation within the data, but it is not significant (model is worthless)

Lets now predict the data:

Text

Description automatically generated

So here we can say that, we can take references as for wines with the ratings of 90 the price would be around 41.857$ and for the rating of 100 the price would be approximately 91.69$. So, the price increases as we increase the rating for the wine in country Spain.

Hence the p-value is less than 0.05(level of significance) we can reject the null hypothesis and select the alternate hypothesis which says, Price increases with rating for country Spain.

**References:**

Kassambara, J., Pant, T., Gooch, S., & Mann, T. (2018, March 11). Nonlinear regression essentials in r: Polynomial and spline regression models. Retrieved April 08, 2021, from <http://www.sthda.com/english/articles/40-regression-analysis/162-nonlinear-regression-essentials-in-r-polynomial-and-spline-regression-models/>

Rato, S., Rausell, A., Muñoz, M., Ciuffi, A., Duraffour, S., Mertens, B., . . . Doan, T. (2017, December 18). What is the relationship between r-squared and p-value in a regression? Retrieved April 08, 2021, from <https://www.researchgate.net/post/What_is_the_relationship_between_R-squared_and_p-value_in_a_regression>