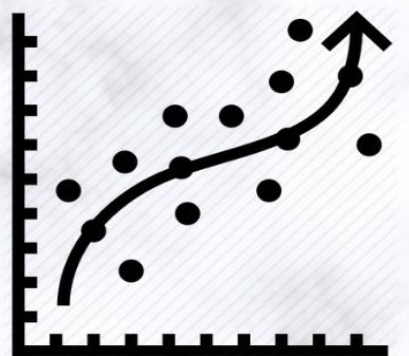


BEER CONSUMPTION



LINEAR REGRESSION ANALYSIS



Beer Consumption

Linear Regression Report

Quang Huynh

BS20DSY027

Bachelor of Data Science 2020

S P Jain School of Global Management

Statistical Data Analysis

Pro. Suchismita Das

May 6, 2021

Contents

Acknowledgement	3
1. Introduction	4
1.1 Report Objectives	4
2.2 Problem Statement	4
2. Overview	5
3. Dataset Dictionary	5
4. Univariate Analysis	6
4.1 Median Temperature	7
4.2 Minimum Temperature	8
4.3 Maximum Temperature	9
4.4 Precipitation	10
4.5 Weekend Day	11
4.6 Beer Consumption	11
5. Bivariate Analysis	12
5.1 Median Temperature	13
5.2 Minimum Temperature	14
5.3 Maximum Temperature	15
5.4 Precipitation	16
5.5 Weekend Day	17
6. Multivariate Analysis	18
6.1 Describe the regression	18
6.2 Analyse Residual	19
6.3 Checking significance of model	20
6.4 Estimating error variance	20
6.5 Parameter Evaluation	21
6.6 Conclusion	22
7. Reduced Model	23
8. Conclusion	24
Reference	25
Appendix	25

Acknowledgement

Special thanks are given to professor Suchismita Das for helping me with guidelines, choosing a dataset and assisting in my research. This report is developed for project of subject Statistical Data Analysis

1. Introduction

1.1 Report Objectives

The data set contains records of 365 samples corresponding to number of days in Year 2015 about beer consumption that were collected in São Paulo – Brazil, in a college area from which students aged between 18 and 28 years old (average) were involved. The paper will focus on the summary of each statistics and pay heed to correlation between variables, develop and evaluate some linear regression models.

To that end, we will deploy the dataset and find relationship between the target and the five other independent attributes. Then, some hypotheses will be put forward and lots of model evaluation tool will be utilised.

2.2 Problem Statement

Beer is one of the most democratic and consumed drinks in the world. The truth is that around the world, millions of people consume beer regularly and in different situations. But have you ever stopped to think about which countries are the most consuming beer in the world? The first one that come to your head is Germany.

The answer to that question is surprising: the Czech Republic. According to research by the Japanese beverage company Kirin, the country has topped the per capita beer drinking table for 23 consecutive years. In 2015, the Czechs drank 142.4 litres per person. That's the equivalent of 250 pints—or one every 35 hours.

For national total consumption, China remained the largest beer-consuming country in the world for the 16th consecutive year; however, consumption decreased in by 2.0% year-on-year. Mexico, in fourth place with 5.3% increase in beer consumption, maintaining growth for two years in a row. Besides Mexico, other countries among the world's top 10 beer-consuming countries which witnessed an increase were Brazil, Germany, the United Kingdom, Vietnam, and Spain.

This dataset was built with the purpose of finding factors that affect beer consumption, then help the beer companies make up the best strategy and be more competitive in the alcoholic drink market.

Moreover, our observation will be put forward with the help of two tools: Python for data cleaning and JMP Pro 15 as statistical tools. Finally, various hypotheses would be tested to predict relevant data for the whole.

2. Overview

As stated before, dataset is developed for beer consumption at a specific region in São Paulo and provide the general ideas of factors on this market to help beer company increase the most profit.

Hence, this statistical report includes the five considerable aspects of weather conditions: Median Temperature, Maximum Temperature, Minimum Temperature, Precipitation, Weekend Day. Five of them are corresponding to each day in 2015 and used to analyse beer consumption and evaluate what is the most crucial factor. To that end, some analysis methods from univariate method, bivariate to multiple regression would be used to summarise the data statistics, its distribution as well as finding predictive model.

This dataset is inspired by the Consumo Cerveja Dataset and is owned by Alexandre George Lustosa.

3. Dataset Dictionary

Normally, the meteorologists record the temperature many times in a day from 12.00 am to 12.00 pm to find the range and the median temperature.

Maximum Temperature (°C): The highest environment temperature that record in a day.

Minimum Temperature (°C): The lowest environment temperature that record in a day.

Median Temperature (°C): The middle value temperature through many times of records in a day.

Precipitation (mm): amount of atmospheric water vapour that falls from the clouds to the ground due to condensation. This can be any form including rain, snow, hail, graupel, sleet, drizzling, etc.

Weekend Day: is a dummy value for 1 is weekend day (Saturday, Sunday) and 0 for others.

Beer Consumption (litres): the amount of beer consumed in the area examined.

Day: this variable only plays as index corresponding days throughout the year. Hence, we will not consider it in our report.

4. Univariate Analysis

This analysis focuses on one variable at a time. It doesn't mention the relationship between variables; instead, its major perspective is describing how the variable is distributed, taking summary, and finding patterns by using some descriptive statistics and some charts like Box plot, Histogram, Normal Quantile Box Plot.

i. Histogram: A histogram is a graphic presentation of a univariate dataset. Based on its shape, it is used to find out the distribution of variables, the frequency of its occurrence and categories.

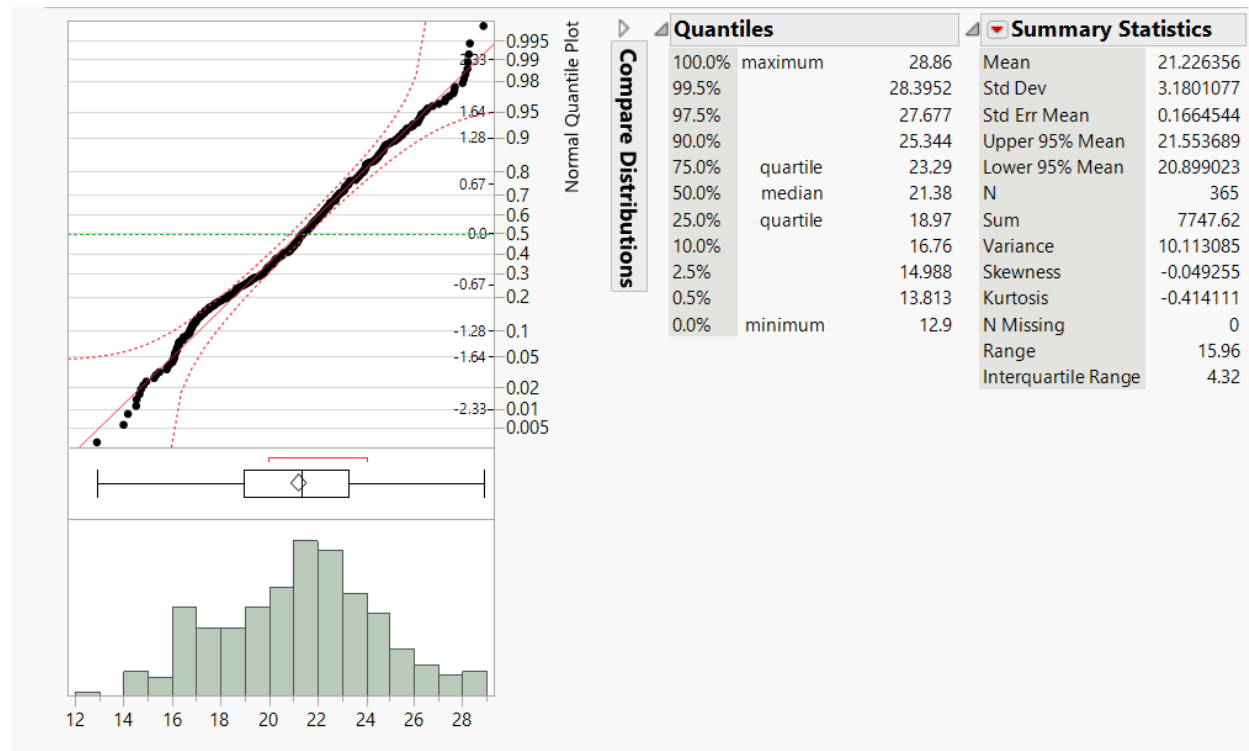
ii. Box plot: is a method for depicting distribution of variables from five values of a dataset: minimum, maximum, first quantile, median, third quantile. Using box plot, the data which is not included in between whiskers, called outliers, can be detected easily

iii. Normal Quantile Plot: (also known as a quantile-quantile plot or QQ plot) is a graphical way of checking whether the data are normally distributed.

Below is univariate analysis of seven features including:

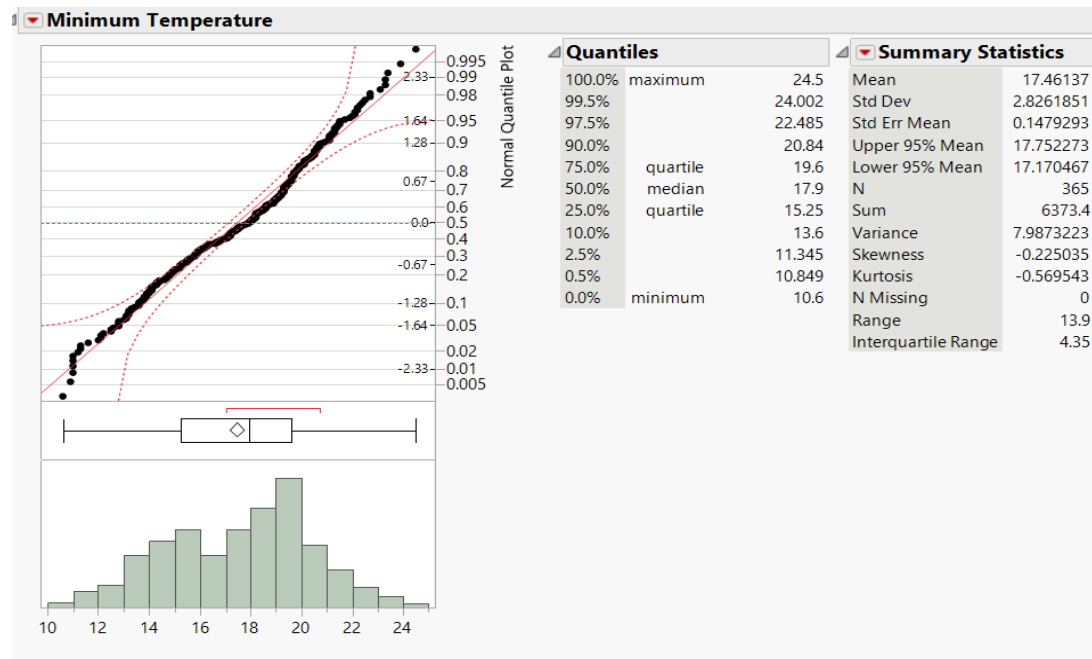
- Summary Statistics
- Quantiles
- Histogram
- Normal Quantile Plot
- Box Plot

4.1 Median Temperature



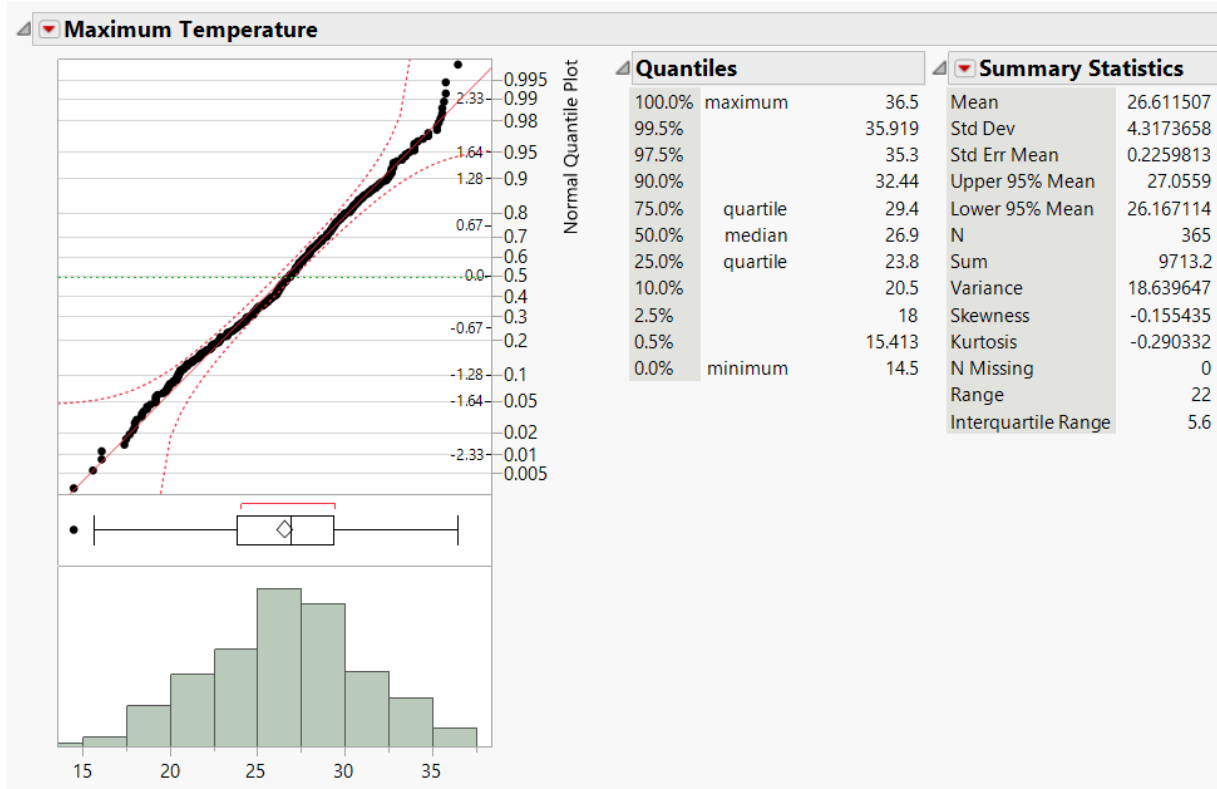
- Lengths of the two whiskers are quite equal and most of points form a diagonal in QQ Plot, mean and median is nearly equal (around 21.3), which indicates this distribution is relatively normal.
- No outliers are detected. The values fall in range: 12.9 – 28.86 (°C).

4.2 Minimum Temperature



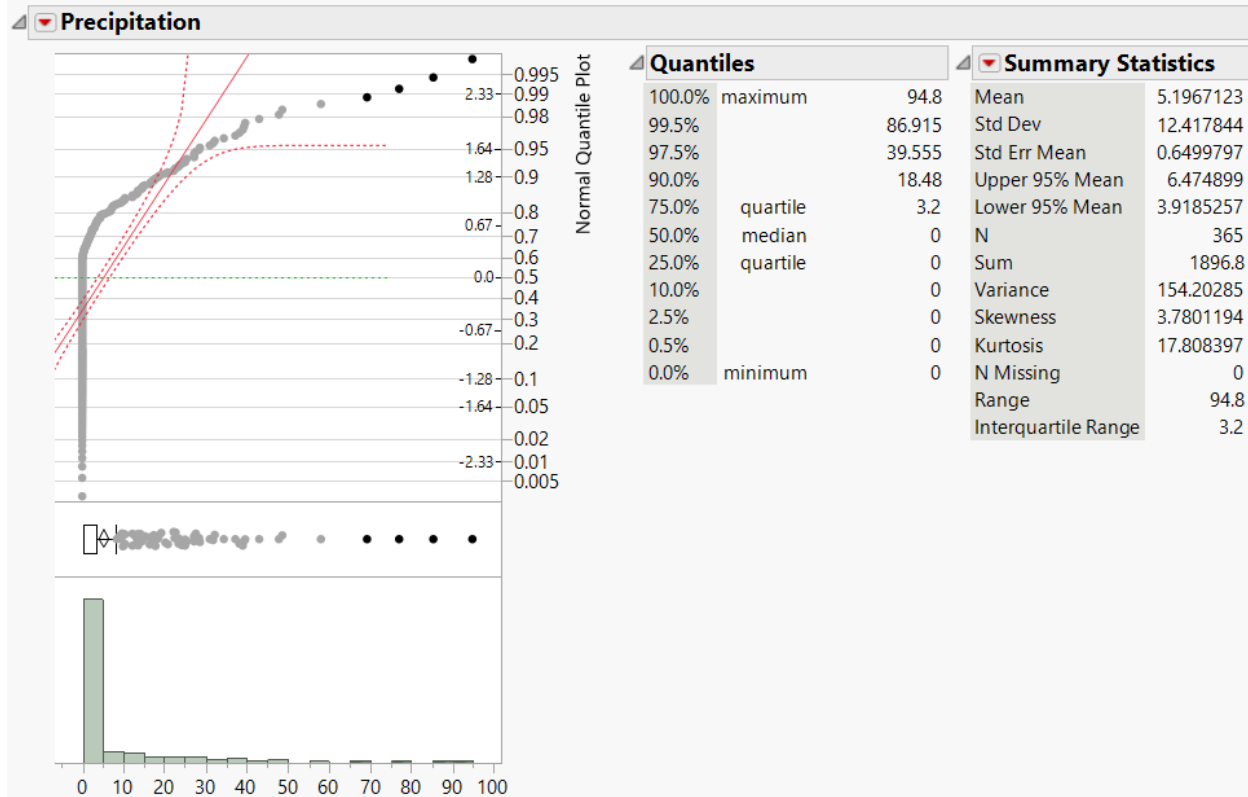
- Mean and median is quite different, some points in QQ plot exceed the red border, which indicates this distribution is quite left skewed.
- No outliers are detected. The values fall in range: 10.6 – 24.5 (°C).

4.3 Maximum Temperature



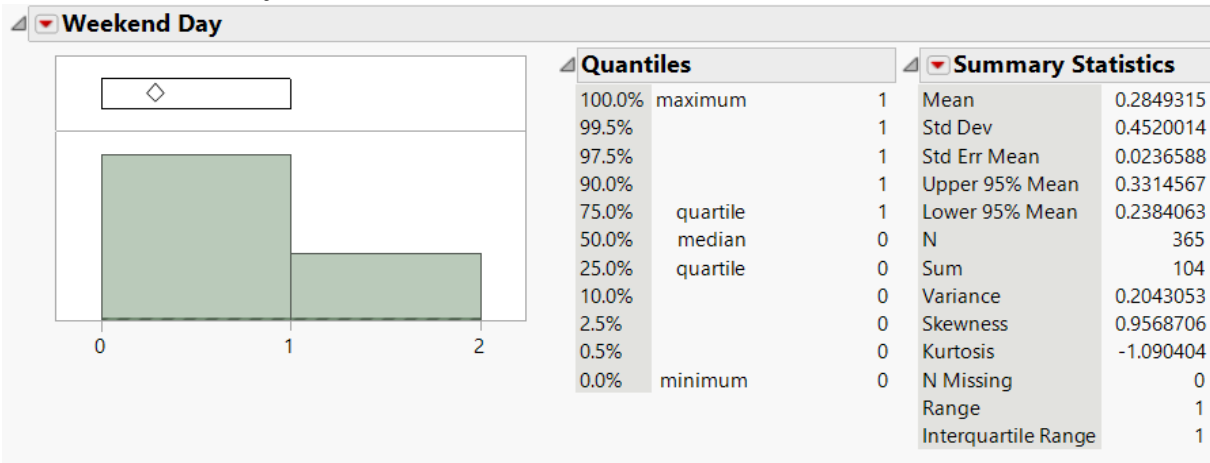
- The distribution is almost perfectly normal as all point in QQ Plot fall totally in the red border although mean and median is about 0.3 (°C) different.
- One outlier is detected. The values fall in range 14.5 – 36.5 (°C).

4.4 Precipitation



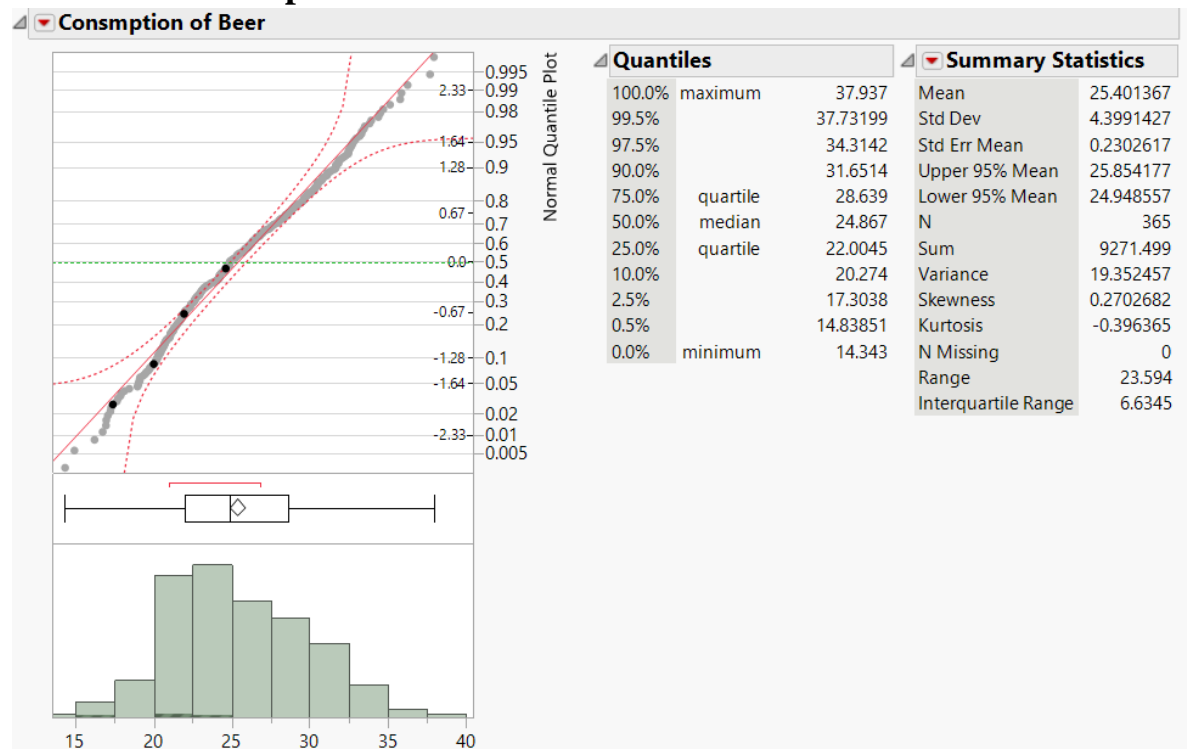
- The distribution is terribly right skewed with many outliers can be detected (skewness is up to 3.78).
- The values fall in range 0 – 94.8 mm. However, 90% of them fall between 0 and 18.48. Median (0 mm) is largely smaller than mean (5.2).

4.5 Weekend Day



- It can be understood that number of values 0's (weekday) is about three times larger than that of 1's values (weekend). Their sum forms 365 days in a year.

4.6 Beer Consumption



- Distribution is roughly right skewed as some points in QQ Plot exceeds the red boundary
- The values fall in range 14.3 to 38 (litres). No extreme values are detected. The mean (25.4) and the median (24.9) is not largely different
- This is the target we are going to predict in the two next sections.

5. Bivariate Analysis

This analysis studies the relation between two variables and uses some charts like scatter plot, comparative box plot, etc ... to present different statistics and their relationships including correlation, covariance. Below are some specific terms in JMP for regression analysis:

Scatter plot: is a type of mathematical diagram using values of two variables to observe and identify the relationship between two numeric variables.

R Square/R Square Adj: coefficient of determination which shows how well the line fit the data. The model is good when these values are high. Formula for $RSquare = 1 - \frac{SSResid}{SSTotal}$. In this case R Square is better enough. R Square Adj is only used when we have data from entire population.

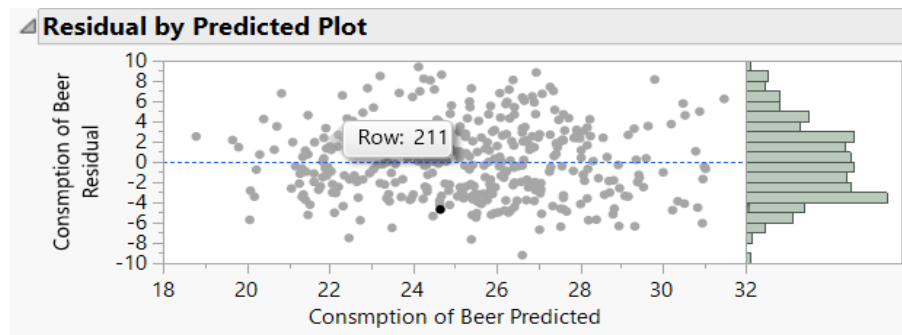
Root Mean Square Error: determined by $MSE = \sqrt{\frac{1}{n-2} \sum (y_i - \hat{y}_i)^2}$. The smaller this value, the better the model fit.

Analysis of Variance: The method to test the statistic $\frac{MSReg}{MSRedid}$ (*F distribution*) to check the significance of our model.

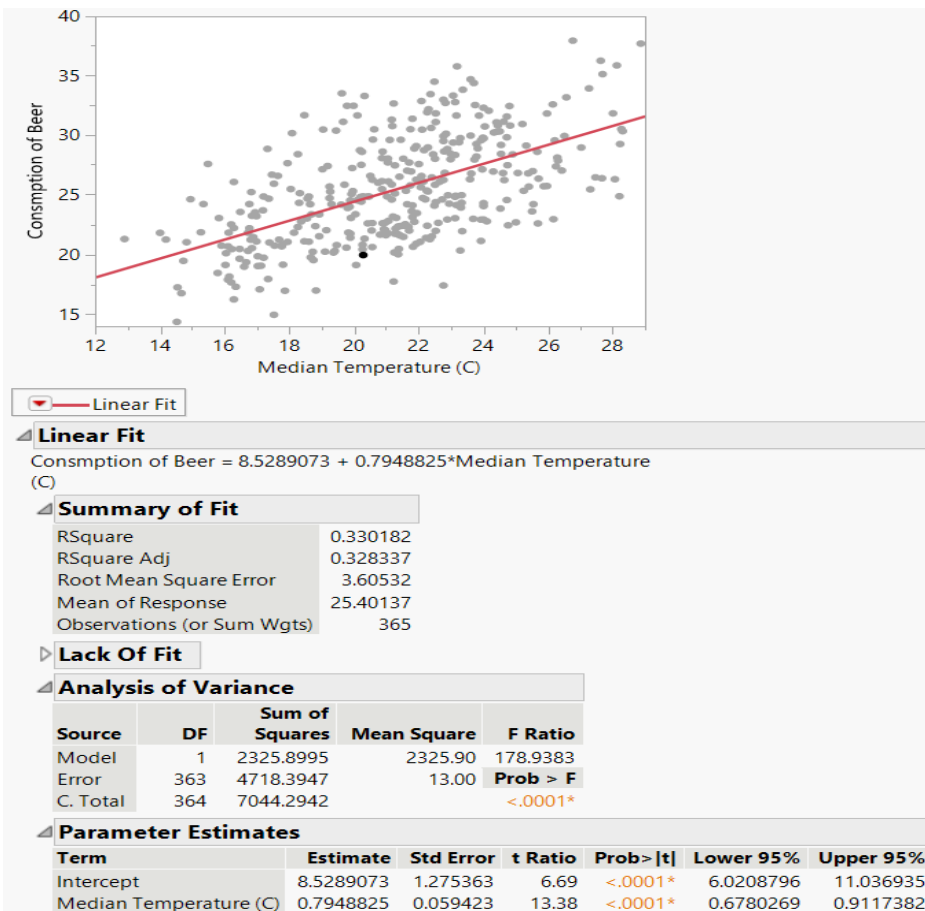
Parameter Estimates: Analysing the importance of the variable in that model by scrutinising its *confidence interval* and its *t - statistic*.

In regression analysis, no cause – effect relationship is implied. Also, all the statistic below are two-tailed hypothesis testing and take level of significance $\alpha = 0.05$ as well as 95% confident interval are applied.

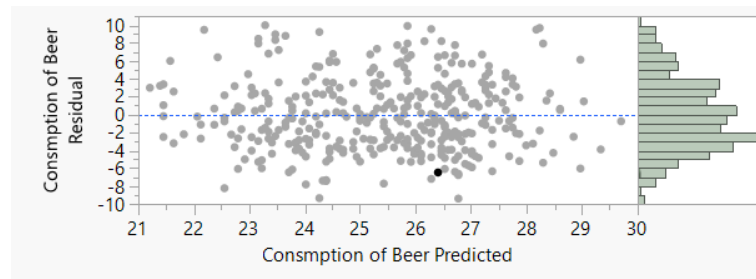
5.1 Median Temperature



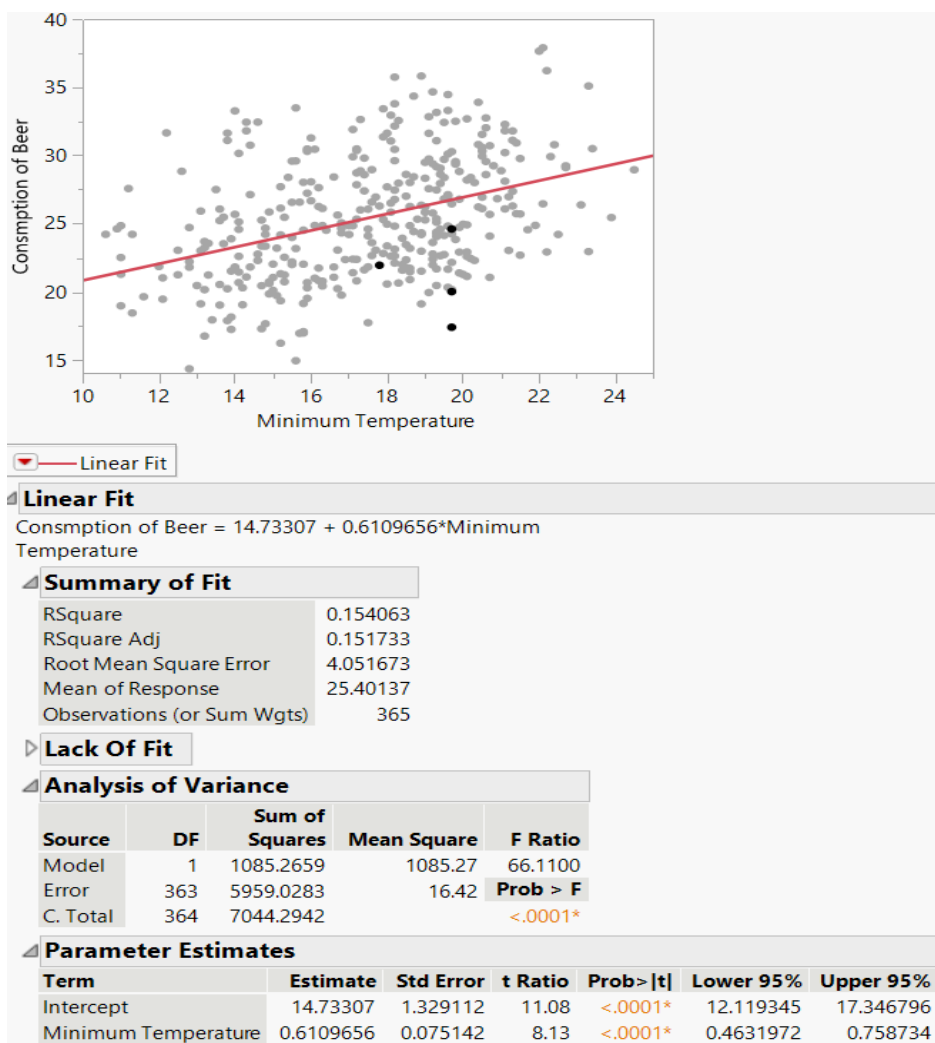
- $RSquare = 0.33$, which means 33% of variation can be explained by this model. There is a **moderate** positive linear relationship between medium temperature and consumption of beer. The *RMSE* or Error is quite high comparing to the value y . The residuals follow a relative normal distribution and randomly distributed. Overall, the model is significant as **$p\text{ value} < 0.0001$** ($F\text{ ratio} = 178.9$) and the parameter $b = 0.794$ still has an effect on consumption of beer ($t\text{ Ratio} = 13.38$)



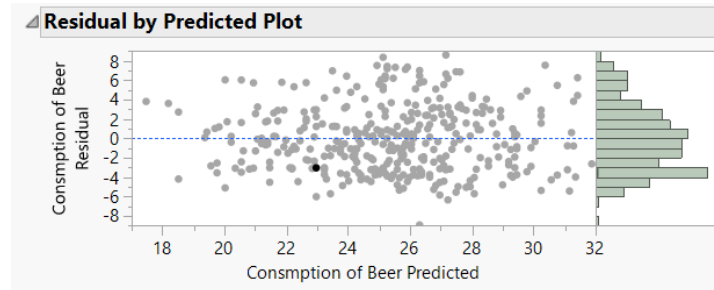
5.2 Minimum Temperature



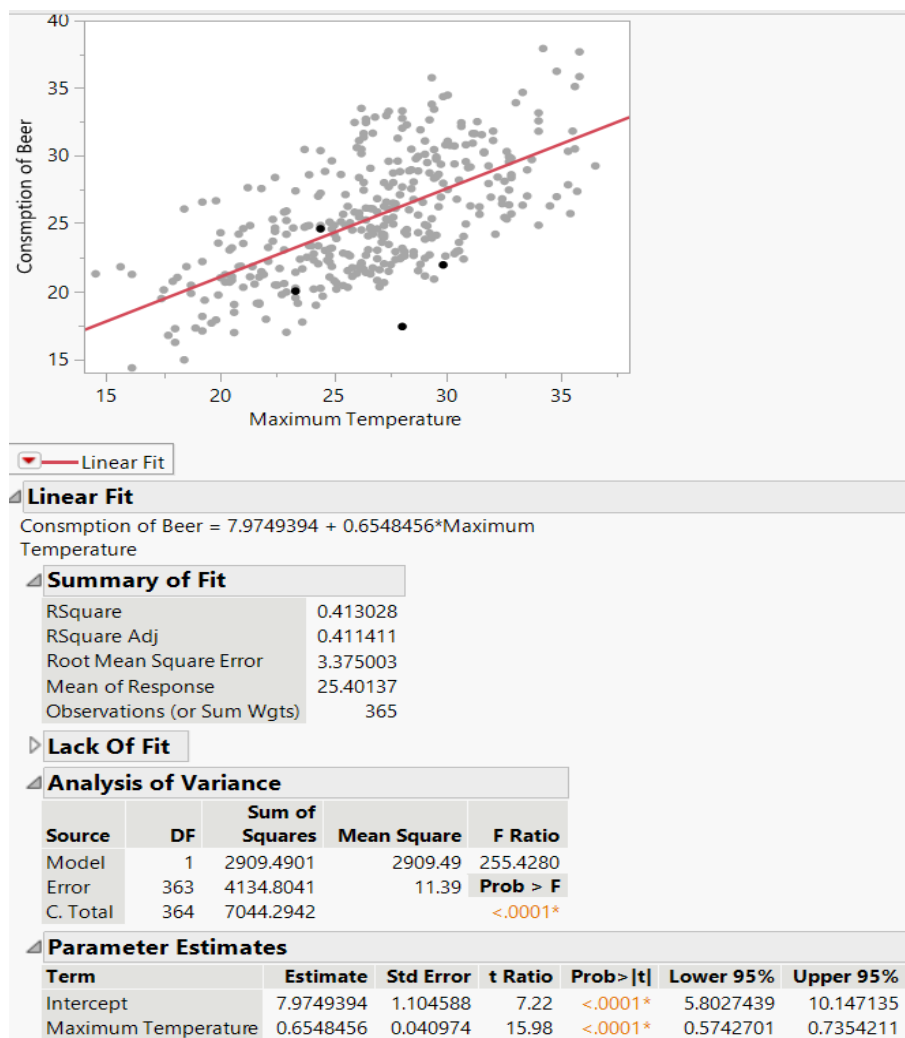
- $RSquare = 0.15$, which means 15% of variation can be explained by this model. There is a **weak** positive linear relationship between minimum temperature and consumption of beer. The *RMSE* or Error is high comparing to the value y .
- The residuals follow a relatively normal distribution and randomly distributed.
- Overall, the model is significant as **$p\text{ value} < 0.0001$** ($F\text{ Ratio} = 66.1$) and the parameter $b = 0.62$ still has an effect on consumption of beer ($t\text{ Ratio} = 8.13$).



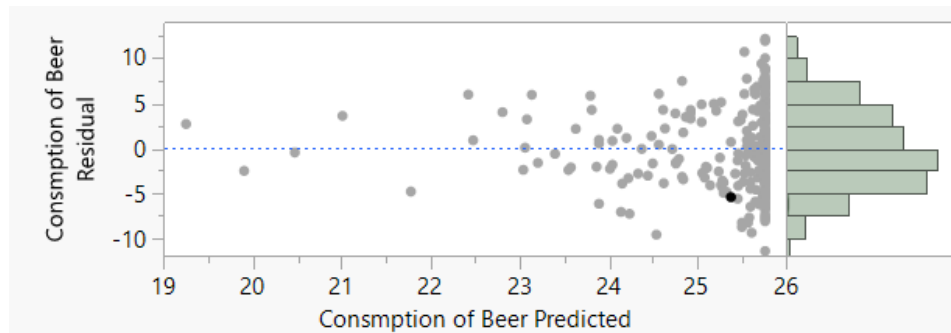
5.3 Maximum Temperature



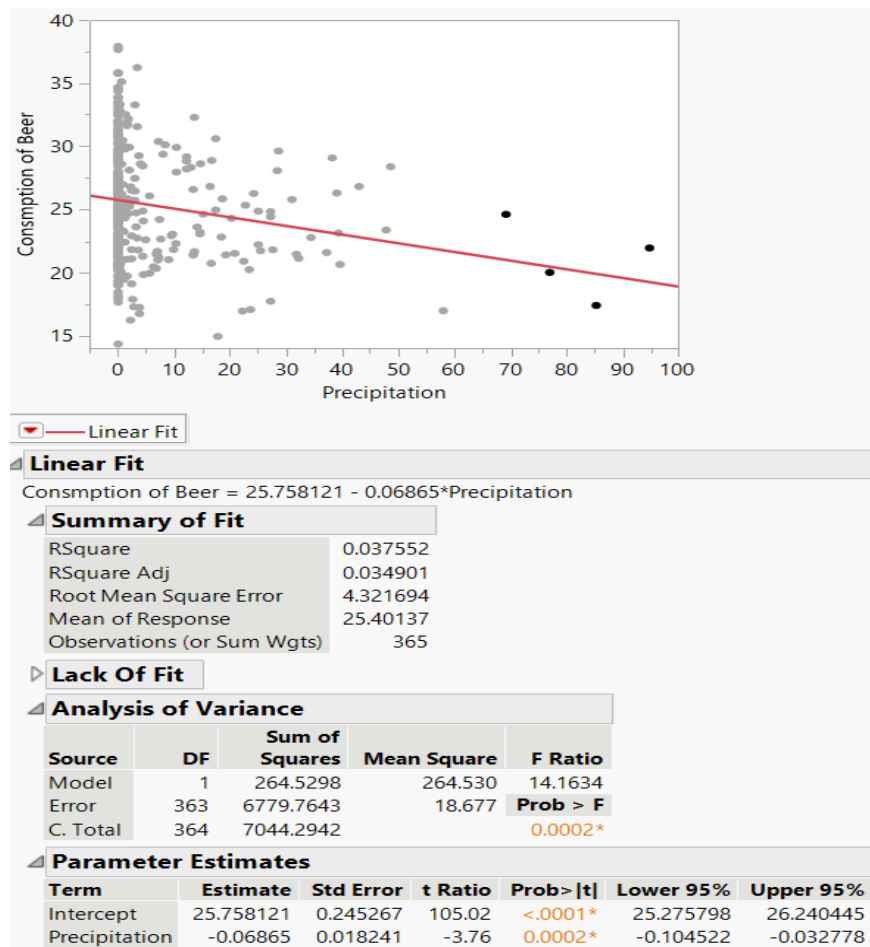
- $RSquare = 0.413$, which means 41.3% of variation can be explained by this model. There is a **moderate positive** linear relationship between maximum temperature and consumption of beer. The *RMSE* or Error is quite high comparing to the value y
- The residuals follow a relative normal distribution and randomly distributed
- Overall, the model is significant as **$p\text{ value} < 0.0001$** ($F\text{ Ratio} = 255.4$) and the parameter $b = 0.65$ still has an effect on consumption of beer ($t\text{ Ratio} = 15.98$)



5.4 Precipitation

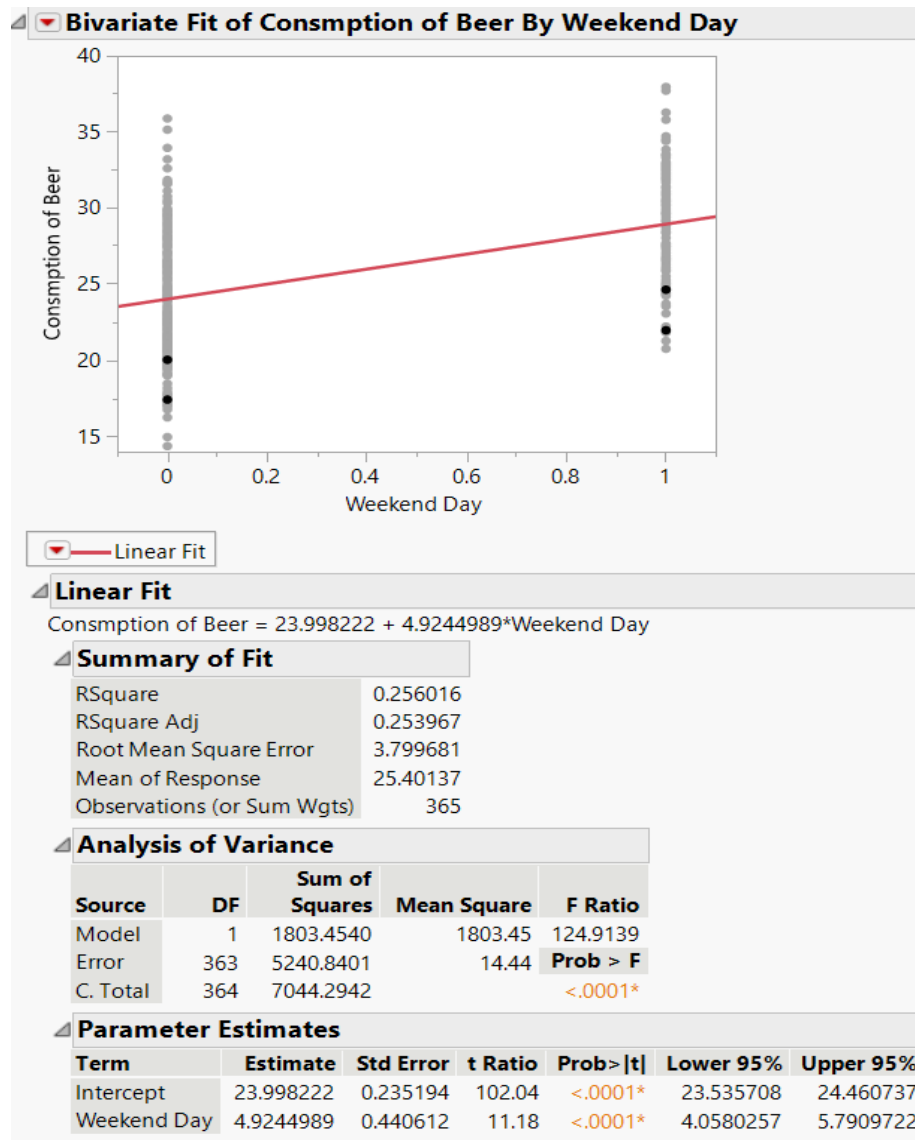


- $RSquare = 0.037$, which means 3.7% of variation can be explained by this model. There is almost **no** linear relationship between precipitation and consumption of beer (although some evidences show that it is negative relationship). The *RMSE* or Error is high comparing to the value y
- The residuals follow a relative normal distribution and randomly distributed, and many **influential points** can be detected because precipitation follows a right skewed distribution
- Overall, the model is significant as **$p\text{ value} = 0.0002$** and the parameter $b = -0.069$ still has an effect on consumption of beer ($t\text{ Ratio} = -3.76$)



5.5 Weekend Day

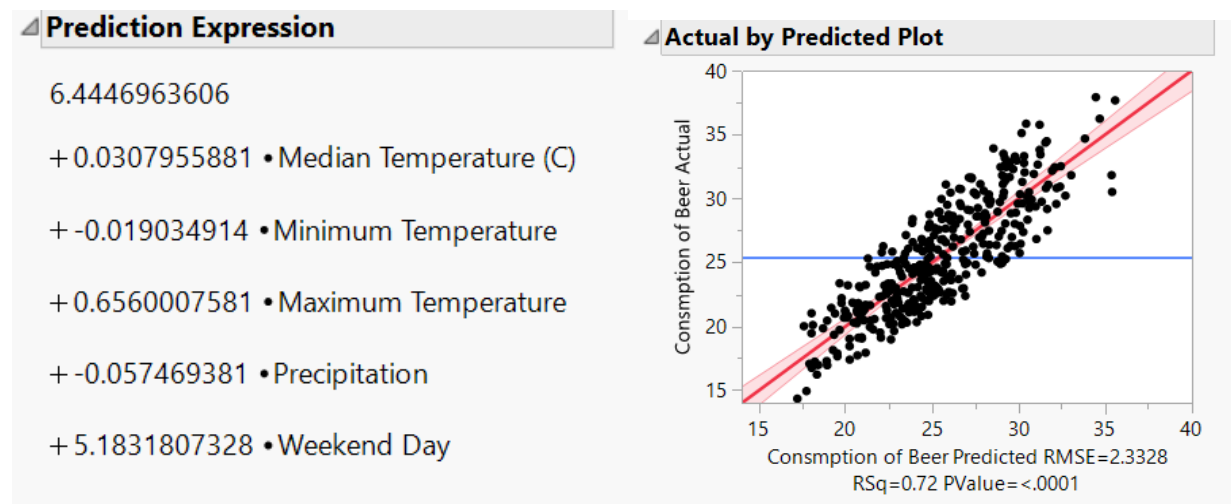
- $RSquare = 0.25$, which means 25% of variation can be explained by this model. There is a **moderate** positive linear relationship between weekend day and consumption of beer. The *RMSE* or Error is quite high comparing to the value y
- Overall, the model is significant as ***p value* < 0.0001** and the parameter $b = 4.9$ still has an effect on consumption of beer (*t Ratio* = 11.18)



6. Multivariate Analysis

As bivariate only takes one paired of variables at the time, and every predictor variable are assumed independently, we cannot see its mutual relationship in affecting the target variable. Hence, multivariate analysis is considered to solve this problem. This method examines several variables at a time to give the best regression model and to understand better the relationships between variables and their relevance to the problem studied.

6.1 Describe the regression



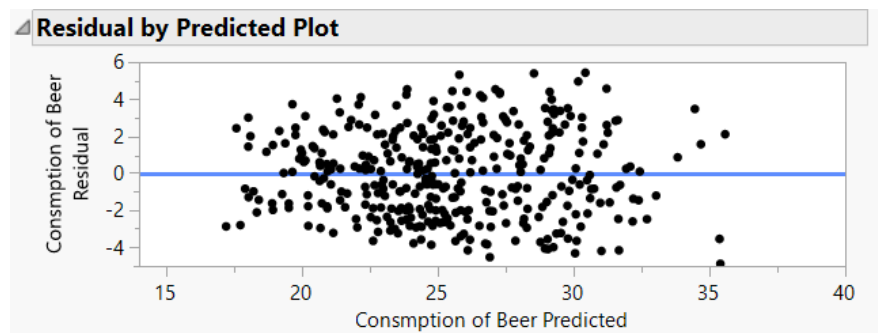
The regression has target y - consumption of beer which is modelled by five predictor variables at a time namely median, minimum, maximum temperature, precipitation, and weekend day.

We generalise the equation in mathematics as below:

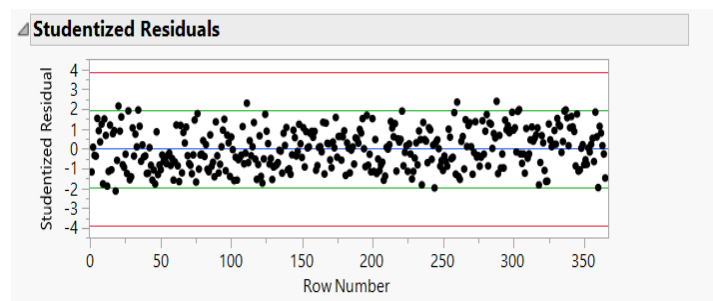
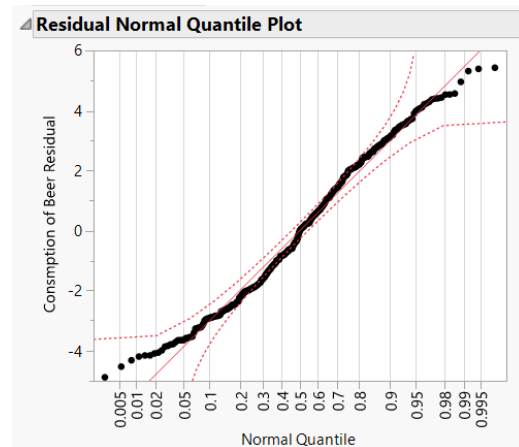
$$y = 6.445 + 0.031x_1 - 0.019x_2 + 0.656x_3 - 0.057x_4 + 5.183x_5$$

Every variable is positively related to the target variable, except precipitation (-0.057). Without deep analysis, we can see weekend day (5.183) is the most affect to beer consumption and Minimum Temperature (-0.019) contributes the least.

6.2 Analyse Residual



The analysis of residual plays an important role in checking basic assumptions of linear regression model. As can be seen from the “**Residual by predicted Plot**”, the mean residual (blue line) is around value zero, and they are randomly distributed (no patterns when consumption of beer increase).



Looking at “**Residual Normal Quantile Plot**”, most of the point fall within the red border and form a diagonal line. Also, inferring to “**Studentized Residuals**”, just a little number of residuals exceeds $\pm 2\sigma$. Hence, we may conclude the residual follows distribution normal and they are independent of one other and the same for any values of x .

As a result, four assumptions for linear regression model are satisfied.

6.3 Checking significance of model

Summary of Fit		Analysis of Variance				
RSquare	0.72265	Source	DF	Sum of Squares	Mean Square	F Ratio
RSquare Adj	0.718787	Model	5	5090.5575	1018.11	187.0785
Root Mean Square Error	2.332844	Error	359	1953.7367	5.44	Prob > F
Mean of Response	25.40137	C. Total	364	7044.2942		<.0001*
Observations (or Sum Wgts)	365					

Based on the summary above, 72.265% of variation of data can be explained by this model, which is higher than any simple linear model in bivariate analysis. In addition, the model is significant as $RMSE = 2.332$, which is quite small comparing to y -beer consumption value.

Finally, the table of **Analysis of Variance (ANOVA Test)** shows the model utility test for multiple regression. We can see the $\frac{\text{Mean Square Model}}{\text{Mean Square Error}} = \text{F Ratio} = 187.0785$, making the $p - \text{value} < 0.0001$. This is one evidence showing that this multiple model is significant, or **at least** one parameter is different from 0 ($\beta_i \neq 0$, for at least one i).

6.4 Estimating error variance

From the table Analysis of Variance in part 6.3, we easily see that:

$$SS_{\text{Resid}} = \sum_{i=1}^{365} (y_i - \hat{y}_i)^2 = 1953.7367$$

$$\hat{\sigma}^2 = \frac{SS_{\text{Resid}}}{365 - 6} = 5.44$$

Choosing significant value $\alpha = 0.05$, we get $\chi_{359,0.025}^2 = 413.3861$ and $\chi_{359,0.975}^2 = 308.4009$

Then 95% Confidence Interval for σ^2 is:

$$\frac{(n-k)\hat{\sigma}^2}{\chi_{n-k,\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-k)\hat{\sigma}^2}{\chi_{n-k,(1-\frac{\alpha}{2})}^2}$$

$$\Leftrightarrow \frac{359 \times 5.44}{413.3861} \leq \sigma^2 \leq \frac{359 \times 5.44}{308.4009}$$

$$\Leftrightarrow 4.724 \leq \sigma^2 \leq 6.333$$

6.5 Parameter Evaluation

After checking the significance of our model, below are some tables used for evaluating each parameter as well as their interactions.

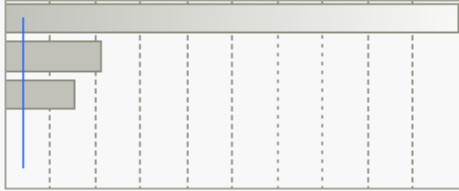
Correlation of Estimates							
Corr	Intercept	Median Temperature (C)	Minimum Temperature	Maximum Temperature	Precipitation	Weekend Day	
Intercept	1.0000	-0.0617	-0.1588	-0.1038	-0.0296	-0.1444	
Median Temperature (C)	-0.0617	1.0000	-0.8422	-0.9132	-0.0577	-0.0029	
Minimum Temperature	-0.1588	-0.8422	1.0000	0.6194	-0.0475	0.0265	
Maximum Temperature	-0.1038	-0.9132	0.6194	1.0000	0.1166	0.0023	
Precipitation	-0.0296	-0.0577	-0.0475	0.1166	1.0000	-0.0074	
Weekend Day	-0.1444	-0.0029	0.0265	0.0023	-0.0074	1.0000	

Looking at the **Correlation Matrix**, we easily spot correlations between maximum, minimum and median Temperature having high absolute values (red and blue ink). These indicates that the three temperature values are not independent; they interact each other.

Besides correlation, **VIF** (variance inflation factor) is another index to examine variable interactions. It quantifies the severity of multicollinearity in an ordinary least square regression analysis. Normally, good parameter has **VIF** < 10. Hence, looking at this table, median and maximum temperature are variables causing multicollinearity in the regression with VIF is 23.9 and 11.3 respectively

Parameter Estimates							
Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%	VIF
Intercept	6.4446964	0.845035	7.63	<.0001*	4.7828551	8.1065376	.
Median Temperature (C)	0.0307956	0.188	0.16	0.8700	-0.338924	0.4005154	23.907247
Minimum Temperature	-0.019035	0.110358	-0.17	0.8632	-0.236064	0.1979938	6.5063267
Maximum Temperature	0.6560008	0.095148	6.89	<.0001*	0.4688829	0.8431186	11.286737
Precipitation	-0.057469	0.010036	-5.73	<.0001*	-0.077207	-0.037732	1.0389042
Weekend Day	5.1831807	0.271007	19.13	<.0001*	4.65022	5.7161415	1.0036228

Also, in **Parameter Estimates** table we also use statistic $\frac{b-\beta}{s_b}$ to check their effect on model (model utility t test). Looking at the table weekend day, precipitation and maximum temperature are important values having $p - value < 0.001$ while median and minimum temperature has high $p - value$ (0.87 and 0.8632 respectively).

Effect Summary			
Source	LogWorth		PValue
Weekend Day	55.997		0.00000
Maximum Temperature	10.612		0.00000
Precipitation	7.663		0.00000
Minimum Temperature	0.064		0.86315
Median Temperature (C)	0.060		0.86998

Effect Summary table is also useful for us to check parameter effects using **LogWorth** values. It is a $p - value$ transformation based on Pearson Chi-Squared Test. The higher the absolute LogWorth value is, the more significant that parameter is.

6.6 Conclusion

Many statistics from RSquare to Root Mean Square Error show that our model is significant. However, we can see the multicollinearity stills exists due to the interaction between the temperature's values. These factors can make our model overfit and result in unprecise prediction. Hence, the next part will solve this problem.

7. Reduced Model

For improving our performance of model, we remove median and minimum temperature variables and having new model with three independent variables as below:

Beer Consumption

$$= 6.432 + 0.669 \cdot \text{Maximum Temperature} - 0.057 \cdot \text{Precipitation} + 5.184 \cdot \text{Weekend Day}$$

Source	LogWorth		PValue
Maximum Temperature	74.519		0.00000
Weekend Day	56.393		0.00000
Precipitation	7.951		0.00000

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	5090.3892	1696.80	313.4971
Error	361	1953.9050	5.41	Prob > F
C. Total	364	7044.2942		<.0001*

Parameter Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	6.4320852	0.773978	8.31	<.0001*	.
Maximum Temperature	0.6685427	0.028301	23.62	<.0001*	1.0040619
Precipitation	-0.057489	0.009832	-5.85	<.0001*	1.0024371
Weekend Day	5.1840827	0.269998	19.20	<.0001*	1.0016235

As can be seen from three tables, this model has improved better. All the p – value for each parameter are less than 0.0001, which make them more significant. Besides, the F ratio is extremely high (313.5), which means the model is important. No multilinearity can be detected as all VIF values are around 1.

8. Conclusion

In short, this report has successfully made some important summary statistics and put forward numerous regression models from bivariate to multivariate to predict beer consumption based on weather conditions. Some have positive linear relationship, negative, or even have no linear effects.

However, as we state before, no cause-effect relationship can be implied in regression model. That means we cannot state that, for instance, precipitation or temperature are reasons for increase or decrease in beer consumption. They can be related to each other due to **lurking variables (confounding)**.

When building regression model, we should consider interactions between explanatory variables to prevent **multicollinearity** and **overfitting**. Hence, we must pick features having better $p - value$ and remove those having worse $p - value$. In our case, median and minimum temperature are considered to remove as they have strong linear relationship with maximum ones.

Our data set is collected at a certain region in São Paulo, examined on people aged for 18 to 28 and consider only six variables. This should not be generalized in other regions or other age intervals because explanatory variables may be different. We can only use the model for predict our target in that region and this specific age interval only.

Reference

Lustosa, A. G. (2019). *Beer Consumption - Sao Paulo Dataset*. Kaggle. Retrieved May 01, 2021, from <https://www.kaggle.com/dongearge/beer-consumption-sao-paulo>

Roxy Peck, C. O. (2010). *Introduction to Statistics and Data Analysis - Fourth Edition*. BROOKS/COLE CENGAGE Learning. Retrieved May 1, 2021

Appendix

Please refer these files below for:

1. Beer Consumption.csv for dataset
2. Beer Consumption.jmp for statistics summary and model building