**Statistics with R**



* **Name: Rohit Pravin Sapkal**
* **Course: Data Science For Business**
* **Subject: Statistics with R**
* **Roll Number: 3014304**

Table of Contents

* **Problem Statement**
* **Exploratory data analysis (EDA)**
* **Hypothesis Testing**
* **Modelling**
* **Prediction**

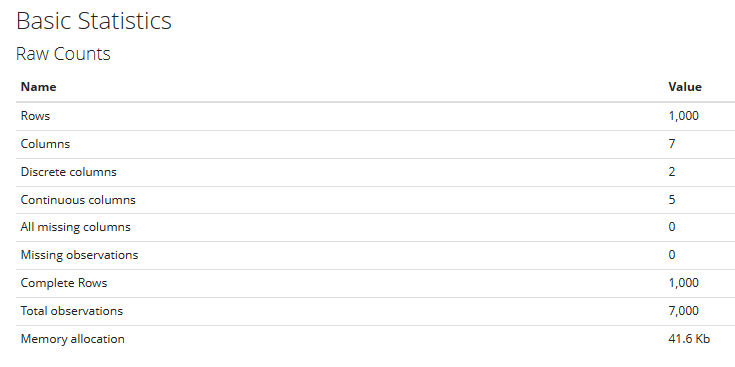
**Problem statement:**

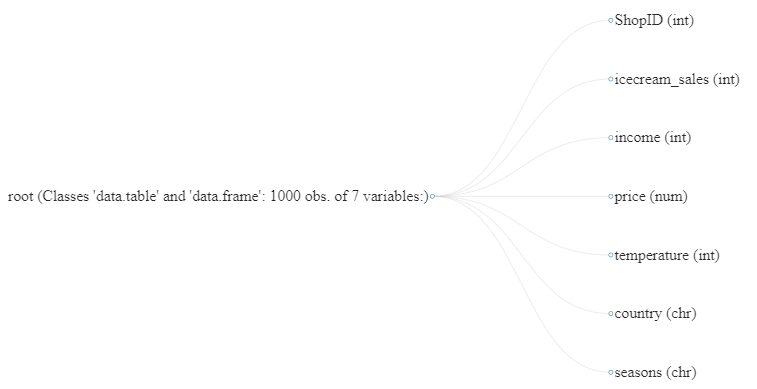
We get the data of Ice-cream sales from different cities of two countries and they want me figure out how they will increase their sales moreover, she is also interested to gather some new idea to boost their ice-cream sales?

**Exploratory data Analysis**

We'll use EDA (Exploratory Data Analysis) to find out all we can about our datasets, such as the mean, mode, and median, as well as how many missing values are there. We also amass complete information about each and every column.

Simple Statistics:





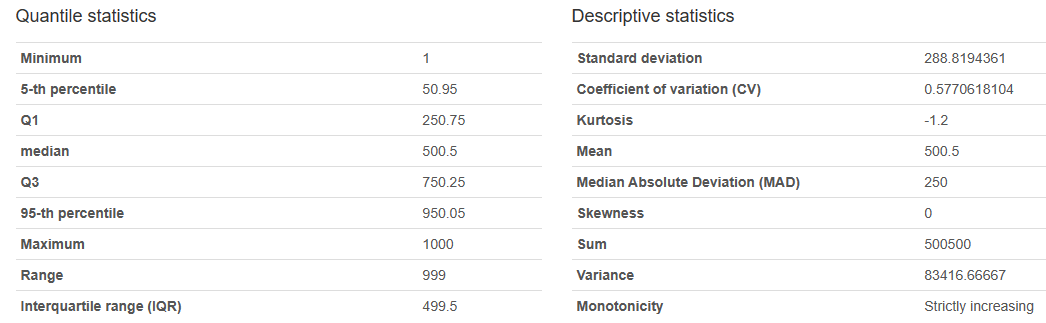
**Variables:**

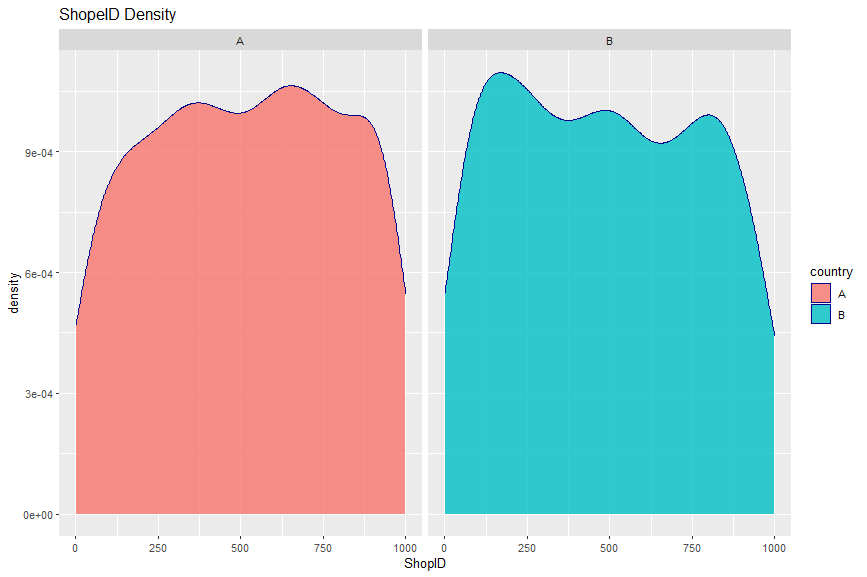
**Shop ID:**

* Type: Identifier variable (Numeric).
* Description:

Shops ID are distinguished by their uniqueness. This table has not more impact on model building.

* Statistics:



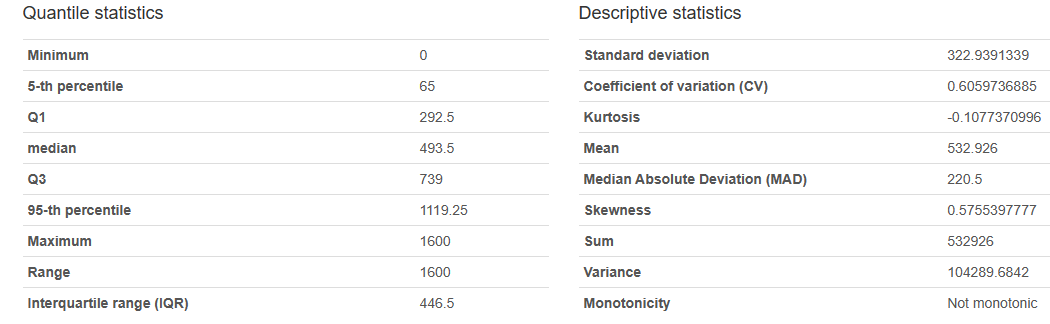


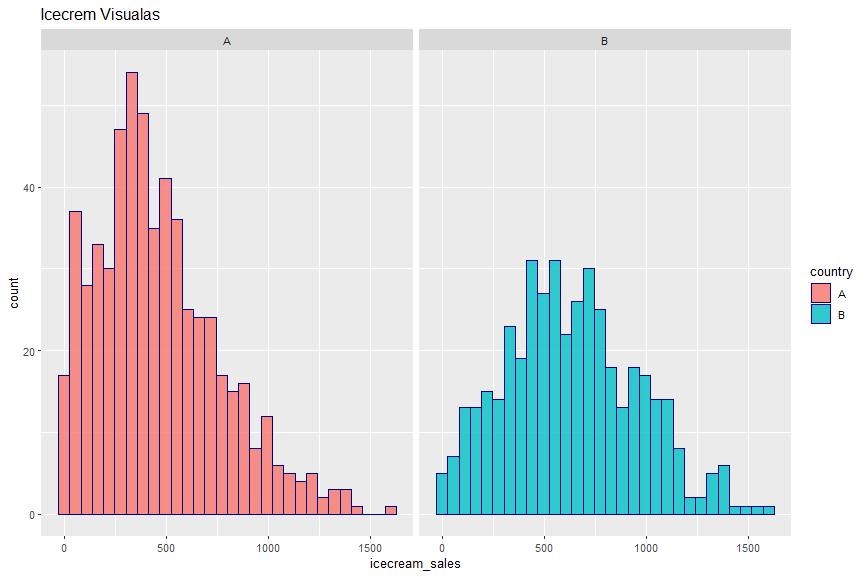
**Ice-cream-sales:**

* Type: Continuous Variable.
* Description:

This column is the on the day the data is gathered, ice cream sales at the store (£). This variable is highly correlated with season’s column.

* Statistics:



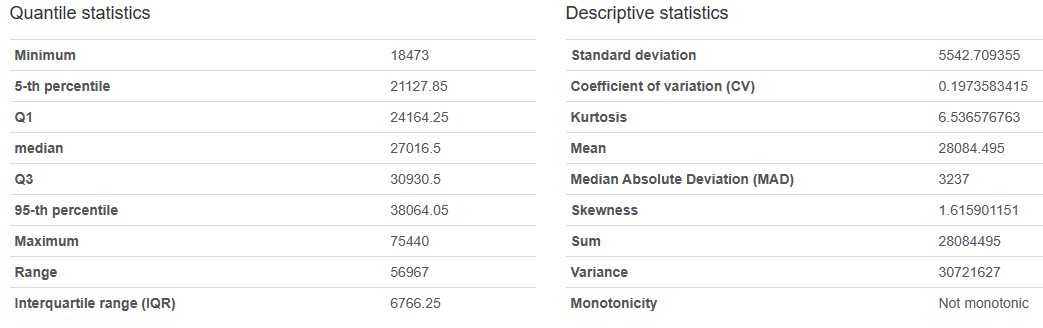


**Income:**

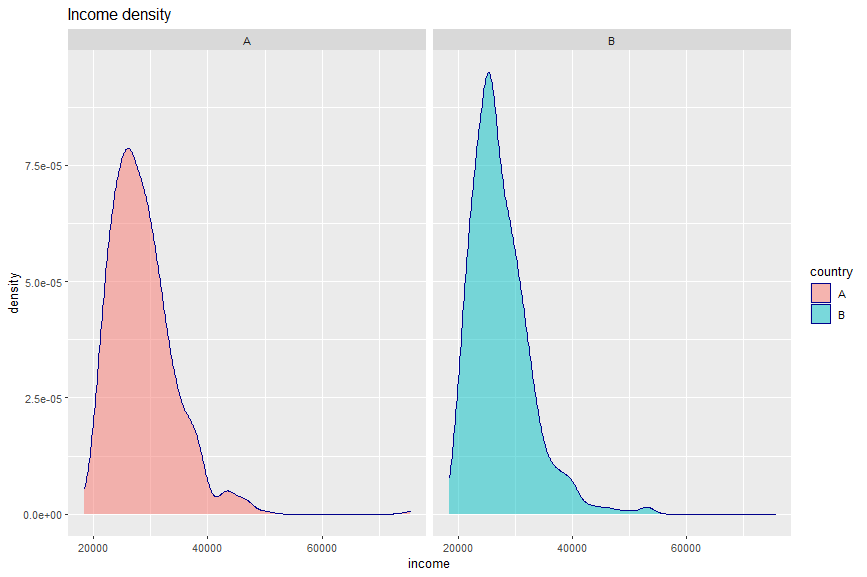
* Type: Continuous Variable.
* Description:

This column is the average annual income of the individuals in the shop's region (in pounds).We needs this column for the model building.

* Statistics:



Income column with compared with Country

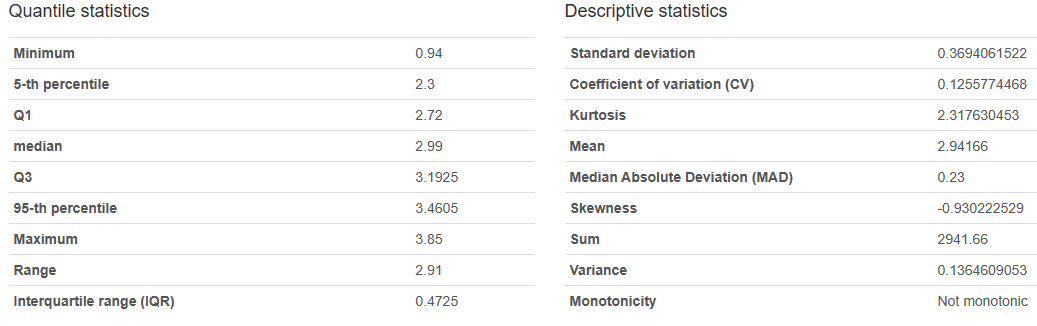


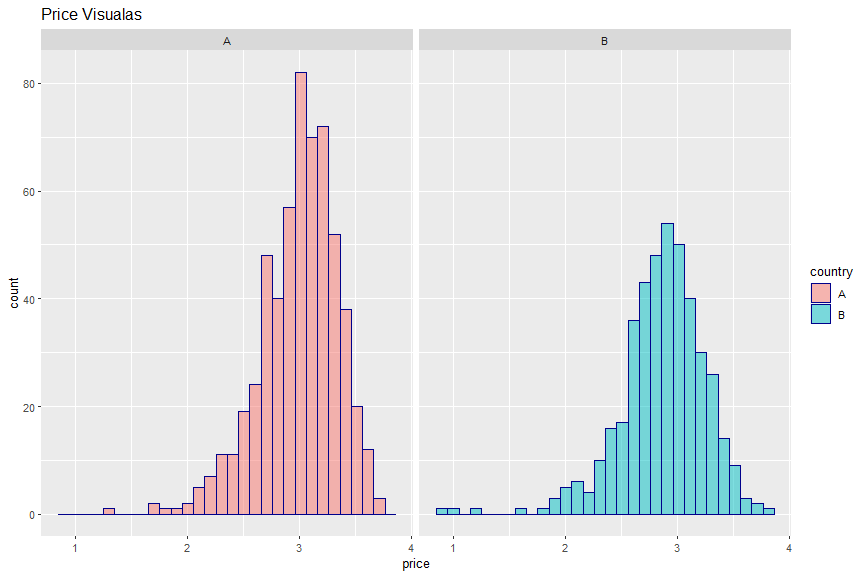
**Price:**

* Type: Continuous Variable.
* Description:

This is the column is on the day the data is taken, the average price per serving of ice cream in the shop (£).This variable is required for model construction.

* Statistics:



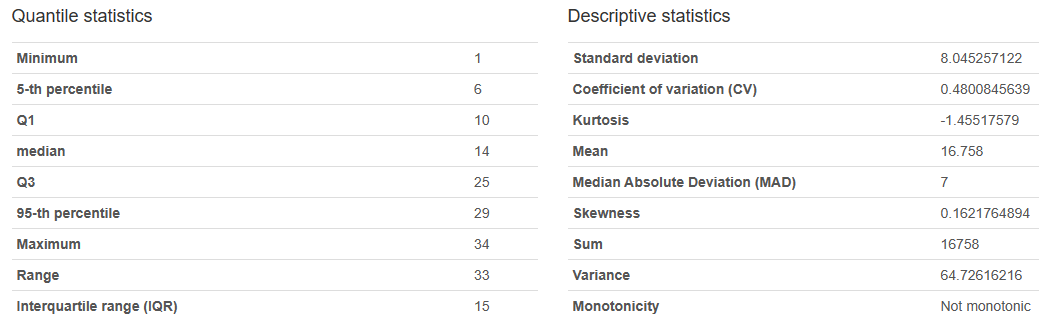


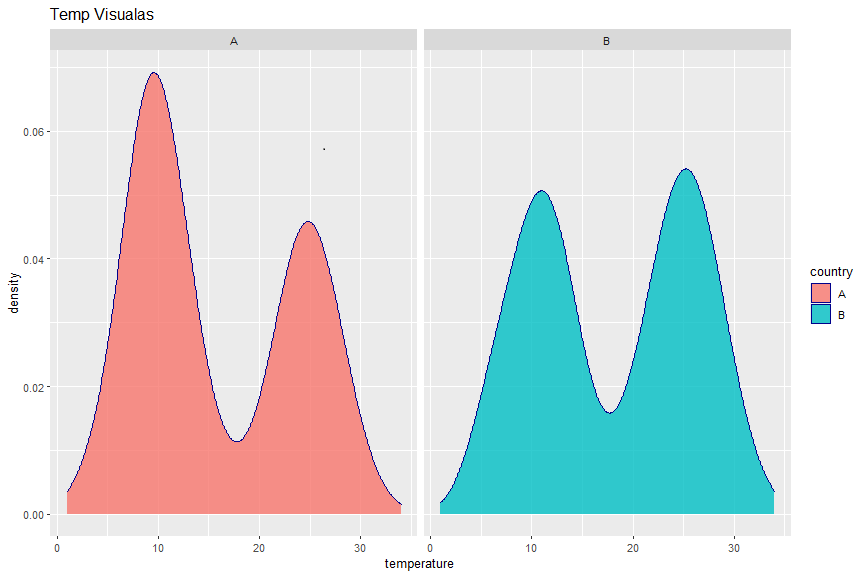
**Temperature:**

* Type: Continuous Variable.
* Description:

This variable is on the day the data is gathered, is the temperature in the region where the shop is located (Celsius). Highly correlated with the season’s table.

* Statistics:



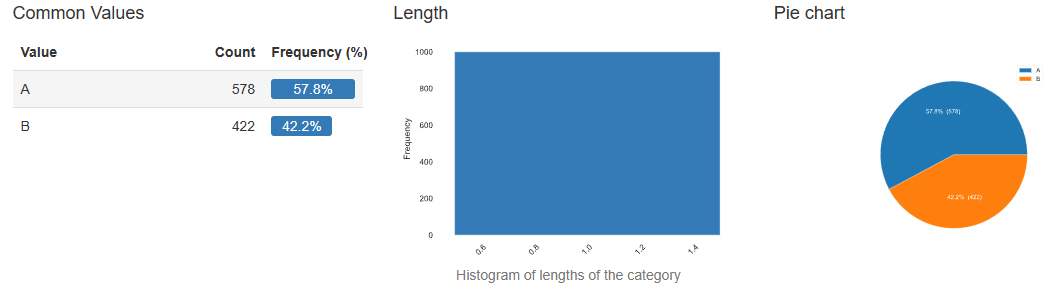


**Country:**

* Type: Categorical Variable
* Description:

This column has a two country which we can collect the data of ice-cream sales. We need this column to compare the sales in each country.

* Categories:

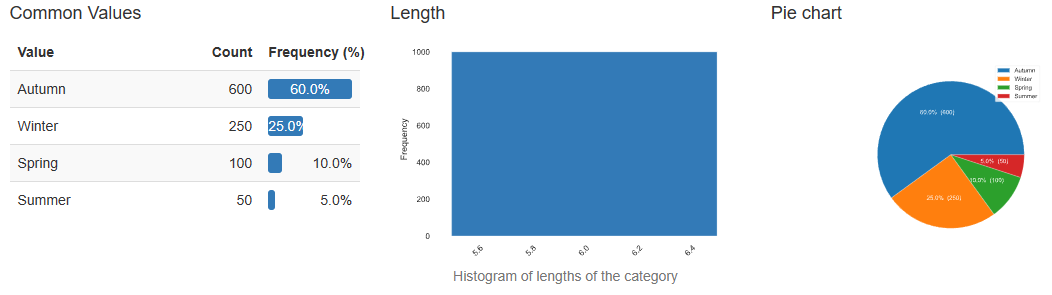


**Seasons:**

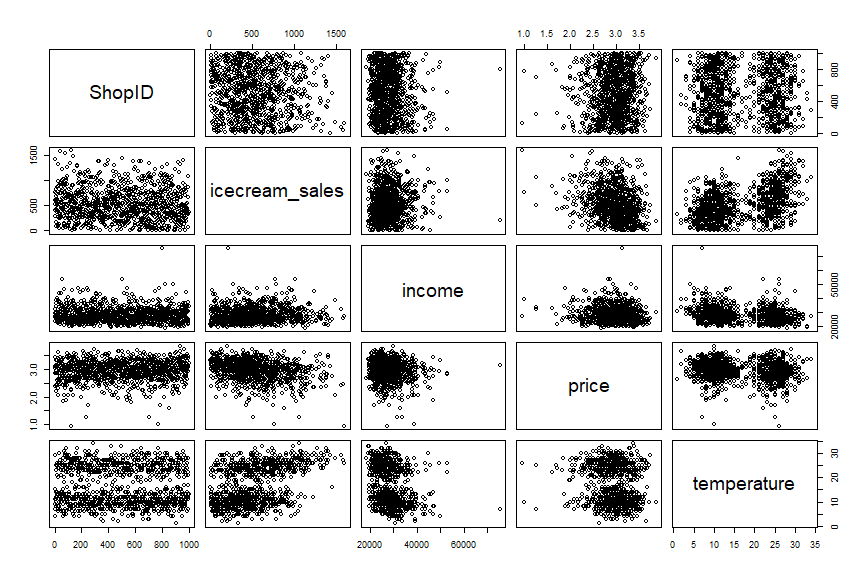
* Type: Categorical Variable
* Description:

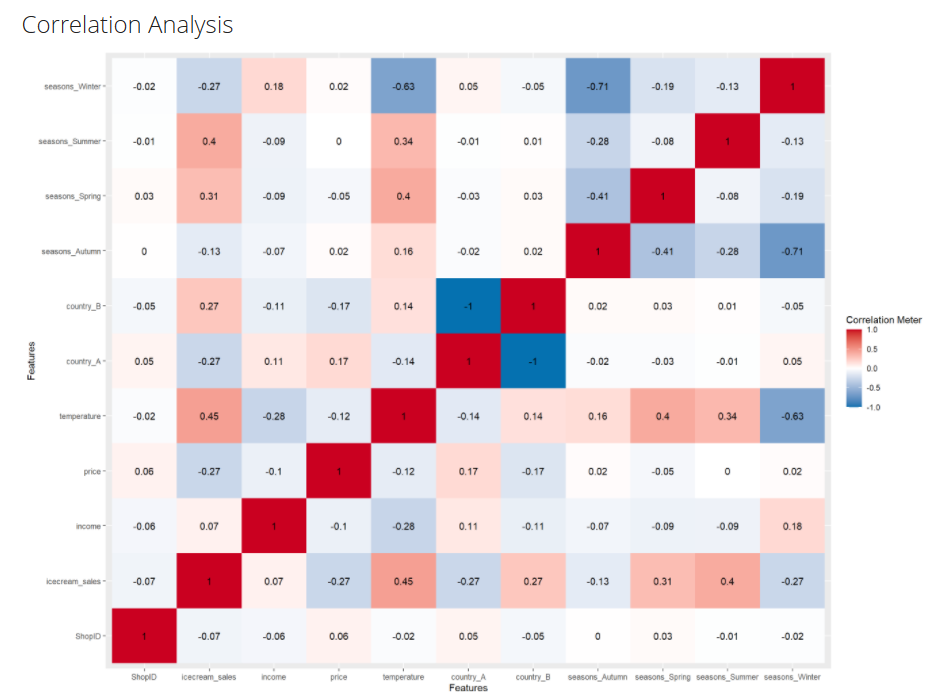
These columns contain 4 type of variable and this column is highly correlated with ice-cream sales and temperature.

* Categories:

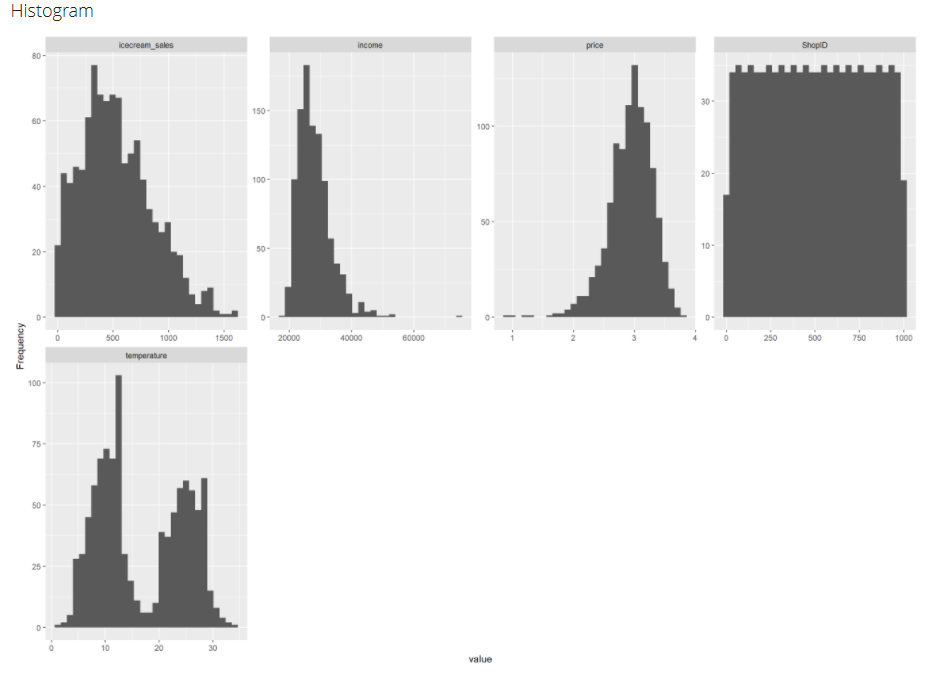


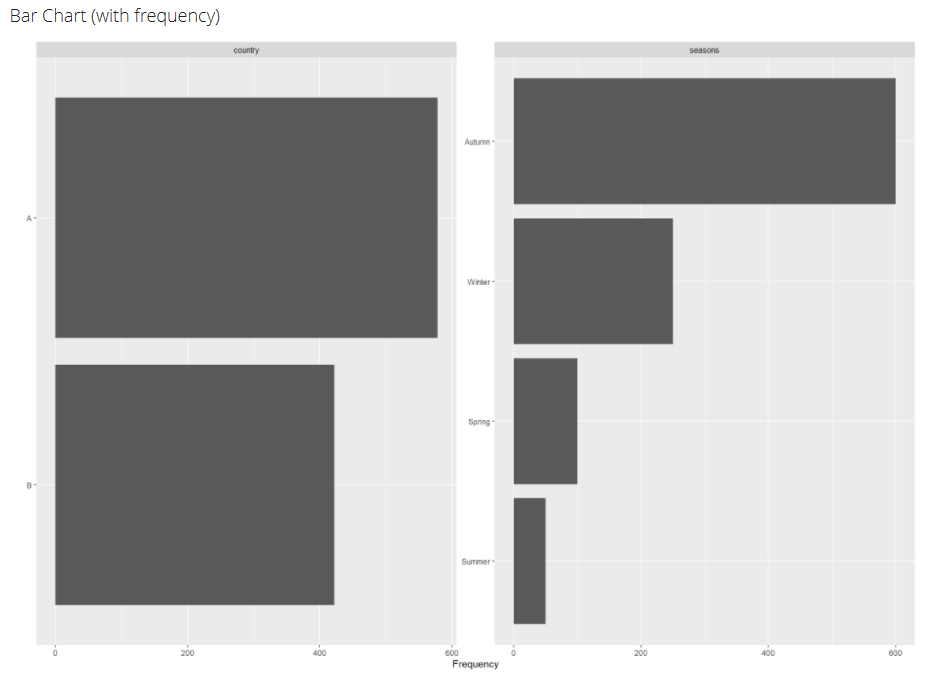
Scattered plot related to the country A and B

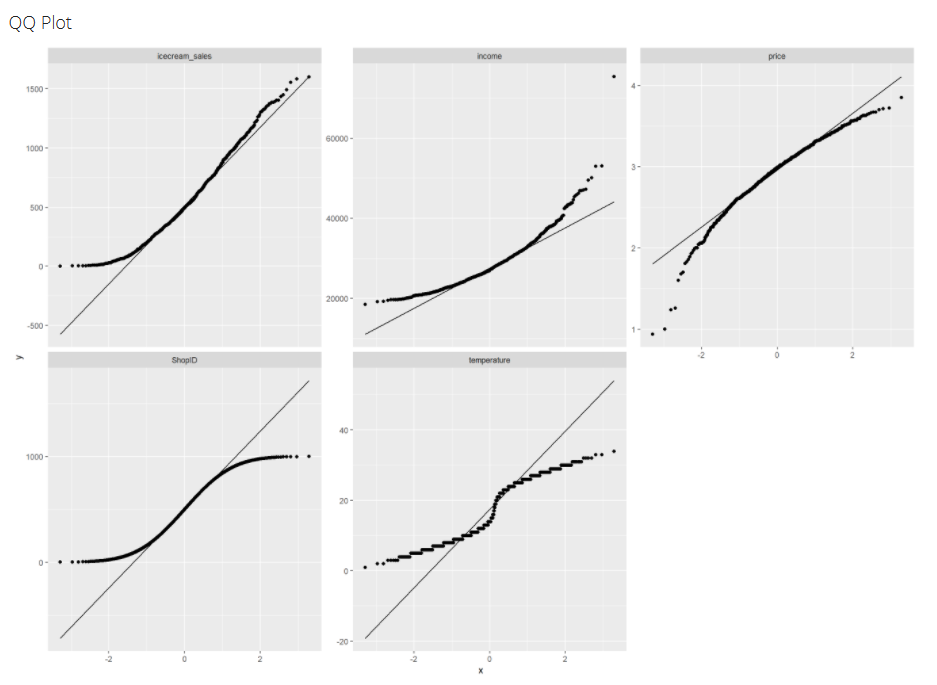




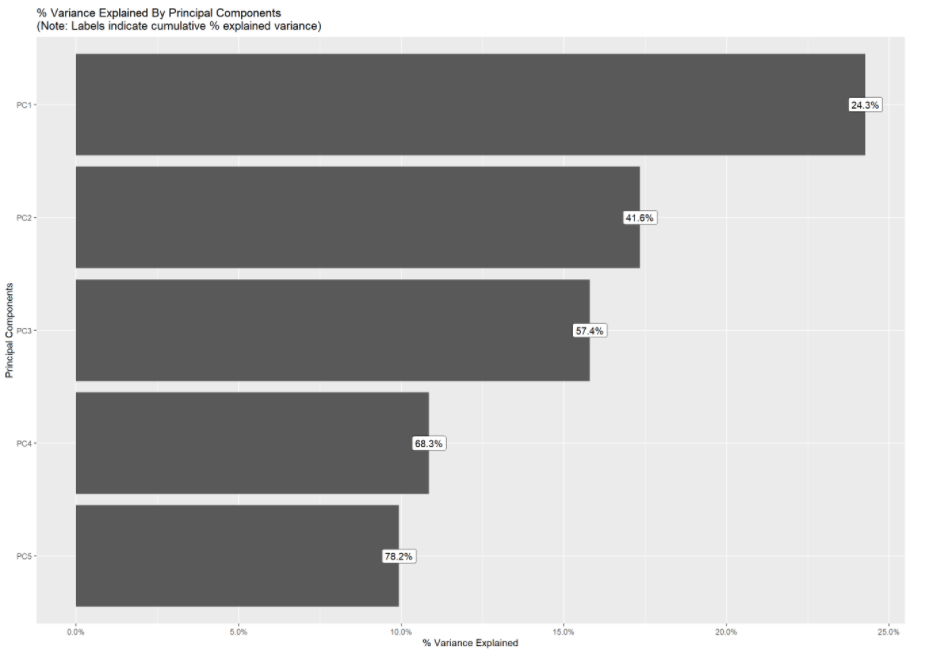
Univariate Distribution

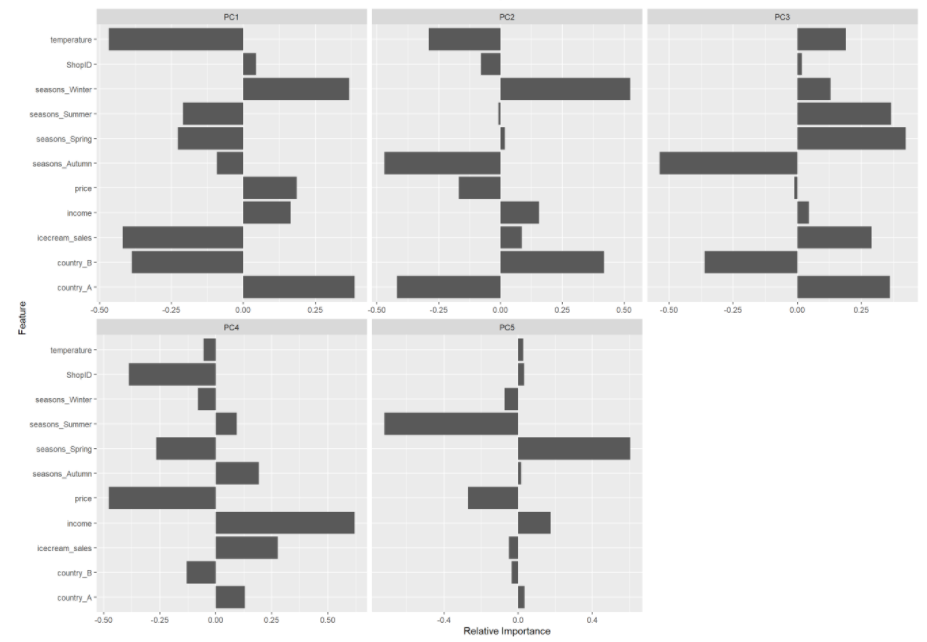




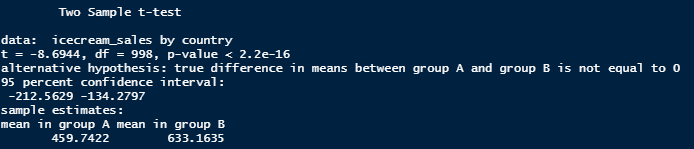


Principal Component Analysis





**Hypothesis testing**

****

Two Sample paired t-test

The country A and country B ice-cream sales are independent

* data: ice-cream sales by country

t = -8.6944, df = 998, p-value < 2.2e-16

* alternative hypothesis:

The true mean difference between groups A and B is not equal to zero.

* 95% confidence interval:

-212.5629 -134.2797

* Sample estimates:

Mean in group A mean in group B

459.7422 633.1635

T = -8.6944

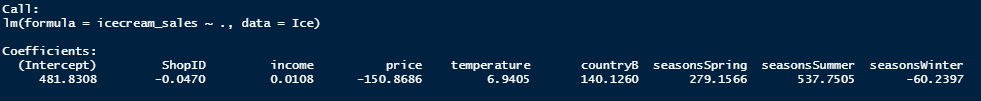
We reject the null hypothesis that there is no difference in mean ice cream sales between nation A and country B since our P value is 2.2e-16, which is less than 0.5, and the degree of freedom is 998.

**Modelling:**

As per the problem statement ice cream sales is the dependent variable and other variable like temperature, price, Income, season, country is independent variable. Now we can fit this data in multiple linear regression models.

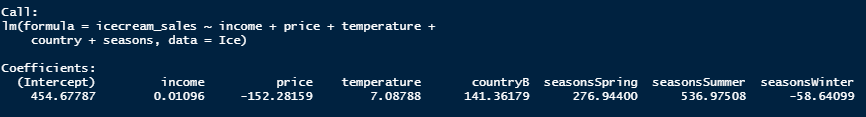
**Formula generation with shop ID:**

**Ice-cream sales = 481.8308+ (-0.0470)\*ShopID+0.0108\*Income+ (-150.8686)\*Price+ 6.9405\*Temperature+140.1260\*CountryB+279.1566\*SeasonsSpring+537.7505\*SeasonsSummer+ (-60.2397)\*Seasons winter**



Without Shop ID:

**Ice-cream sales = 454.67787+0.01096\*Income+ (-152.28159)\*Price+ 7.08788\*Temperature+141.36179\*CountryB+276.94400\*SeasonsSpring+536.97508\*SeasonsSummer+ (-58.64099)\*Seasons winter**



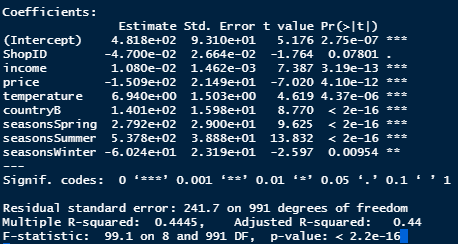
In both nations, income has the highest influence on ice cream sales, According to the P values coefficient, price and temperature are followed by price and temperature. Furthermore, the P-values reveal that income, price, temperature, spring, and summer and Country B are all statistically significant. Most crucially, ice-cream sales are influenced by price and income. From the Formula it is clear that the overall ice cream sales will increased by 0.0108 when income increased by 1GBP and declined by 150.8686in case of selling price increased by 1GBP Moreover, Ice-cream sales increased by 6.9405 units in case of temperature increased rise by 140.1260 in country B as well as sale rise by 279.1566 during Spring season and 537.7505 during Summer season and decreased by 60.2397 units in Winter seasons.

The variance between a response variable's actual value and its predicted value obtained from regression line.

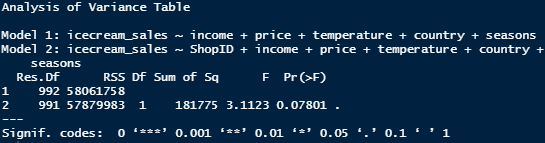
C:\Users\ROHIT\Desktop\Stat with R\graphs\residuals.png

According to data that we generated, it may possible that we get the prediction error it means there is difference between observed and predicted values that are not same. There is a slight difference between them.

The Most important from the below is the P-value and the star means they have high effect on ice cream sales such as Income, Price, Temperature, CountryB, Seasons (spring, summer) on the other hand, Shop ID has no impact on the ice cream sale moreover, Seasons Winter has slight low impact on ice cream sales as compared to the other variables. Multiple R-Squared values mean the variable in the model contribution percentage in this case we have 44.45% to the overall variability.



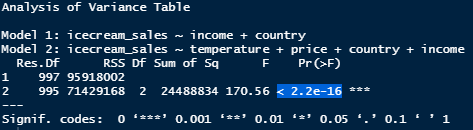
From the below, Adding this extra variable which is Shop ID has 93% confidence level which has not affecting the model so that we can remove this variable.



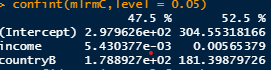
All else being equal, the predicted difference between ice cream sales in a location in Country A with an average income of £20,000 and a location in Country B with an average income of £30,000 is 235.5666. Although, When we take average temperature value which is 16.75(Celsius), Price is 2.94 GBP. On the other hand the difference between the ice cream sales in a location in Country A with an average income of £20,000 and a location in Country B with an average income of £30,000 is 250.9925.

All else being equal, the predicted change in ice cream sales if the price goes up by £0.50 which is 3.44 and temperature goes up by 2 degrees which is 18.75 at the same time is 241.7328.

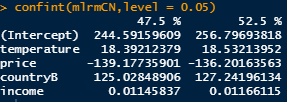
The percentage of variance of these two model is less than 0.5 P-value (< 2.2e-16) which means it has more than 95% confidence level which is very good Moreover, adding this extra variable we get the significant information. Moreover, both models are statistically significant at a 5% significant level.



Model1:



Model2:



With the 90% confidence interval we have only 10% chance to get the wrong prediction it would be narrower for instance plus and minus 2.5%.

For the predicted confidence interval of model 1 which has country A we get fitted value 354.4025 and lower limit for this is 316.3101 and upper limit for this is 392.4949 which means model 1 value lies between these two values.

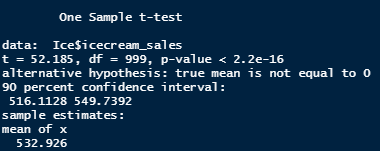
C:\Users\ROHIT\Desktop\Stat with R\graphs\modelBfit.png

For the predicted confidence interval of model 2 which has country B we get fitted value 596.1393 and lower limit for this is 557.0696 and upper limit for this is 635.201 which means model 2 value lies between these two values.

C:\Users\ROHIT\Desktop\Stat with R\graphs\modelA fit.png

The confidence interval on ice cream sales with 90% of confidence level using t-test we get 16.8131.

Confidence interval = 549.7392 - 532.926 = 16.8131. So that the margin of error is 16.8131.



For testing multi linear regression we chose the two shop with Ids 532 and 689. Ice cream sales and the country where this shop located is 308, A and sales take place in autumn for the shop ID 532. And for the shop ID 689 ice cream sale 416, country B and sales take place in autumn. Substituting all this figures in this model to identified weather this data meet the regression model or not.

**Ice-cream sales = 454.67787+0.01096\*Income+ (-152.28159)\*Price+ 7.08788\*Temperature+141.36179\*CountryB+276.94400\*SeasonsSpring+536.97508\*SeasonsSummer+ (-58.64099)\*Seasons winter**

Shop ID: 532

308 = 454.67787+0.01096\*23848+ (-152.28159)\*3.04+7.08788\*26+141.36179+276.94400+536.97508-58.64099

**308 < 1334.041**

Shop ID: 689

416 = 454.67787+0.01096\*28243+ (-152.28159)\*3.27+7.08788\*10+141.36179+276.94400+536.97508-58.64099

**416 < 1233.779**

The results of the two tests show that the data set does not meet the regression requirements.

**Prediction:**

The Prediction value of ice cream sales in

Country =A,

AVG Income= 30,000 GBP,

Temperature = 23 degree,

Season = spring,

Price = 3 GBP

Confidence interval = 95%

**The Predicted ice cream sale is 766.6907 GBP moreover its value lies between 716.1015 (lower limit) and 817.2799 (upper limit).**

**Full EDA report**

****