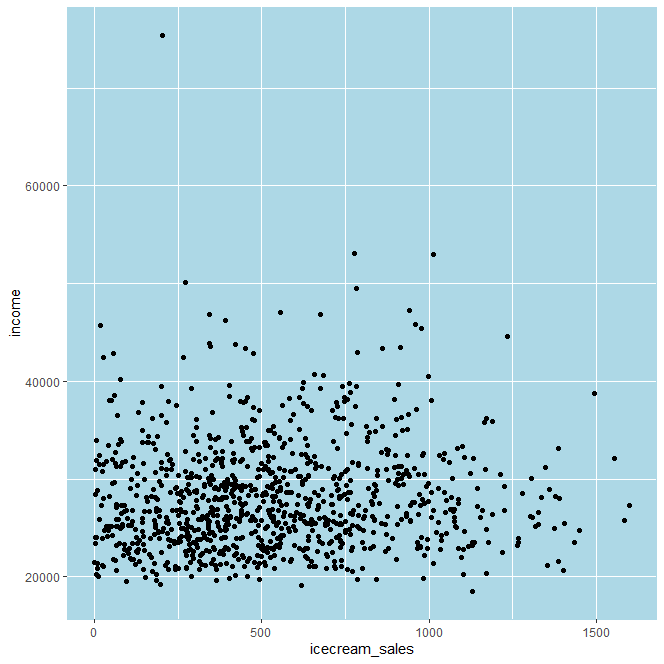
**A. Exploratory data analysis (EDA)**: (30 points) Perform exploratory data analysis (EDA) and explain your variables numerically and graphically. Please do not replicate the same investigation in numeric and visual explorations. A brief interpretation should accompany your R output and plot. (575 words max)

**Exploratory Data Analysis: Please refer Lines (28 to 81) R Code**

* Using git-hub link data loaded into R-Studio
* It has 4 seasons and 2 countries (Countries = [“A”,” B”] & Seasons = [“Summer”,” Winter”,” Autumn”,” Spring”]

**Scatter Plot between Ice Cream Sales Vs Income to understand if the income has any effect over ice cream sales**



**Conclusion:**

*we can say there is not much impact of employee salary over ice cream sales.*

*As we see from the graph, we can say even with lower salaries there has been high collection in ice cream sales and very few employees who are receiving high salaries recorded somewhat higher collection in ice cream sales*

**Scatter Plot between Ice Cream Sales Vs Income Over each country - to understand their distribution separately**

Chart, scatter chart

Description automatically generated

**Conclusion:**

*For Country A, we can say there is not much impact of Income over Ice cream Sales based on scatter plot but when it comes to Country B there is little impact of Income over Ice Cream sales*

**Scatter Plot between Ice Cream Sales Vs Income to understand if the income has any effect over ice cream sales Over Seasons**

Chart, treemap chart

Description automatically generated

**Conclusion:**

*All we can say is data is too scattered and happened more in Autumn > Winter > Spring > Summer*

*Above viz. says that income has less effect over the ice cream sales across the seasons.*

**About Ice Cream Sales:**

**Jitter Plot over Box-Plot - Ice Cream Sales Over Country**

Chart, scatter chart

Description automatically generated

**Conclusion:**

*We have more records for country A comparatively to Country B*

*Mean line (Average) of ice cream sales of Country B is higher than ice cream sales of Country A*

*We can also see most the ice cream sales happening better in B over A*

**Density Plot - Ice Cream Sales Over Country**

Chart, line chart

Description automatically generated

**Conclusion:**

*We can see most of the ice cream sales from Country A recorded with lower ice cream sales compared to Country B*

**Jitter Plot over Box-Plot - Ice Cream Sales Over Seasons**

Chart, scatter chart

Description automatically generated

**Conclusion:**

*From the above jitter plot over Box-plot we can say 2 statements*

* *Either data collection has done more in Autumn > Spring > Winter > Summer*

*Or More sales happened in Autumn > Spring > Winter > Summer*

*Also, data is evenly distributed for all the seasons (not much skewness)*

**Density Plot - Ice Cream Sales Over Seasons**

*Chart, line chart

Description automatically generated*

***Conclusion:***

*There exists some skewness for the summer season (sales with higher magnitude happened more)*

*For the rest of the seasons – Autumn, Spring, Winter we can say data is almost evenly distributed (Ice Cream Sales)*

**Jitter Plot Over Box-Plot - Ice Cream Sales Over Country + Seasons**

Chart, scatter chart

Description automatically generated

**Conclusion:**

*From the above viz. we can say more data under winter, also sales showing higher in magnitude in terms of value for Summer > Spring > Winter > Autumn and important thing is that Country B is leading in Ice Cream Sales in terms of its value*

**Density Plot - Ice Cream Sales Over Country + Seasons**

Chart

Description automatically generated

**Conclusion:**

*From the density visualization Country B is ahead in all the cases in terms of Ice Cream Sales value but we values are from Country – A (Quantity wise)*

**About Ice Cream Average Price**

**Jitter Plot Over Box-Plot - Ice Cream Average Price Over Country**

Chart, scatter chart

Description automatically generated

**Conclusion:**

*Jitter plot over Box plot – From the viz we can say both ice cream average price is almost same (mean line) and Country B has recorded least average ice cream price in more number than Country A but slightly higher in A compared to B*

**Density Plot - Ice Cream Average Price Over Country**

Chart

Description automatically generated

**Conclusion:**

*We see greater number of ice creams sold at higher price in Country A when compared to Country B (but in both countries A & B a greater number of ice creams are sold at higher price as per the below density plot)*

**Jitter Plot Over Box-Plot - Ice Cream Average Price Over Seasons**

Chart, scatter chart

Description automatically generated

**Conclusion:**

*Average price of ice cream is almost same across the seasons*

*Also, the median line for price distribution for Autumn and Winter – most of the records are lying about it as per the above viz.*

**Density Plot - Ice Cream Average Price Over Seasons**

Chart, line chart

Description automatically generated

**Conclusion:**

*We Clearly see that distribution of price is almost same across all the seasons and bit skewed towards**higher price*

**Jitter Plot Over Box-Plot - Ice Cream Average Price Over Country + Seasons**

Chart

Description automatically generated

**Conclusion:**

*we can say average ice cream price is slightly higher for Country – A when compared to Country – B across all the seasons*

**Density Plot - Ice Cream Average Price Over Country + Seasons**

Chart, diagram

Description automatically generated

***Conclusion:***

*Country-A is slightly ahead of Country-B for a particular price of ice cream across all the seasons*

*And for both countries data is skewed towards high ice cream price across all the seasons*

**Multi Linear Regression Model Summary (Please refer lines (88 to 110) R Code)**

**C. Modelling: (40 points) Develop a multiple linear regression model to predict ice cream sales using all explanatory variables. The outcome variable and the explanatory variables can be existing variables in the dataset or new variables you create based on existing variables. (1,150 words max)**

**Note: Please refer Line – 121 to 138 – For Data Transformation, regression model Building**

**Specifically, answer the following questions:**

1. **What is your regression equation?**

**Regression Equation:**

Country feature has 2 categories is converted to dummy one hot encoding and same has been applied for season feature which has 4 categories

When model is trained with data using multi linear regression, we will be getting coefficients for each feature except for one category each from the dummy one hot encoding features obtained from country & season features.

So, we will have NA for Country\_B and Season\_Winter. Their effect will be included along with intercept value obtained in the model

Below are the coefficients & intercept

|  |  |
| --- | --- |
| **Feature** | **Coefficient** |
| Intercept | 537.39 |
| Income | 0.01096 |
| Price | -152.281 |
| Temperature | 7.087 |
| Country\_A | -141.361 |
| Country\_B | NA |
| Season\_Autumn | 58.6409 |
| Season\_Spring | 335.5849 |
| Season\_Summer | 595.616 |
| Season\_Winter | NA |

Below is the equation for regression model trained

Ice\_cream\_sales (target) =

Intercept **+** 0.01096 **x** Income **+** (-152.281) **x** Price **+** 7.087 **x** Temperature **+** (-141.361) **x** Country\_A **+** NA **x** Country\_B **+** 58.6409 x Season\_Autumn + 335.5849 **x** Season\_Spring **+** 595.616 **x** Season\_Summer **+** NA **x** Season\_Winter

1. **What are the interpretations of all your coefficients?**

* Based on the coefficient magnitudes of all the features obtained after training model, we can say Price, Country\_A, Season\_Spring, Season\_Summer has reasonable impact over ice cream sales.
* Among them Price & Country\_A have strong negative impact over Ice Cream Sales i.e., with the increase of Price and from Location then there will be high chances that Ice Cream Sales will drop
* Season\_Spring and Season\_Summer has positive impact over Ice Cream Sales i.e., with the presence of Spring and Summer there are high chances that Ice Cream Sales can increase.
* Income and temperature have got least impact over Ice Cream Sales
* Temperature has got moderate impact over Ice Cream Sales

1. **All else being equal, what is the predicted difference between ice cream sales in a location in Country B in Spring compared to Autumn? How would the predicted difference between ice cream sales in Spring and Autumn be in Country A?**

With all else being equal

Ice Cream Sales comparison in Country – B between Spring & Autumn with all others being equal

Using the coefficients obtained above (Coefficient values mentioned in Question - 1)

Spring will have higher Ice Cream Sales compared to Autumn by 277 units

**Explanation:**

Ice Cream Sales for B in Spring = Constant + 335.5849

Ice Cream Sales for B in Autumn = Constant + 58.6409

Difference among them ~ 277 units(approx.)

It will be same for Country – A with else being equal

Which is 277(approx.) units different with higher in Spring compared to Autumn

1. **All else being equal, what is the predicted change in ice cream sales if the temperature goes up 5 degrees and price goes up by £0.25 at the same time?**

With all being equal

Temperature goes up by 5 and price goes up by 0.25

**Explanation:**

Ice Cream Sales-Initial =

Constant + Price x (-152.3) + Temperature x (7.1) 🡺 1

Ice Cream Sales -Later =

Constant + (Price+0.25) x (-152.3) + (Temperature+5) x (7.1) 🡺 2

Change in Ice Cream Sales = equation 2 – equation 1 = (7.1 x 5) + (0.25 x -152.3) = -2.75

Ice Cream Sales will decrease by 2.75 with temperature goes up by 5 degrees and Prices goes up by £0.25 at same time

1. **What percentage of the variance is described by the model?**

Adjusted – R Squared explains the percentage of variance described by the model

Adjusted – R Squared obtained after training the model = 0.4388

43.88% (**Please refer Line – 102 R code**)

1. **Is this model statistically significant at a 10% significance level?**

F-test of the overall significance is a specific form of the F-test. It compares a model with no predictors to the model that you specify. A regression model that contains no predictors is also known as an intercept-only model.

Hypotheses for the F-test of the overall significance are as follows:

**Null hypothesis:** No evidence that model is significant

**Alternative hypothesis:** Enough evidence model is statistically significant

We have significance p-value = 0.1%

If p-value for the specific F-value obtained after training regression model is less than 0.1% then we reject null hypothesis (**Please refer Liner – 102 R code**)

P-value obtained for the model < 2.2e-16 (summary of model), Which is very much less than significance benchmark (10%)

**We conclude that model is statistically significant at 10%**

1. **What are the confidence intervals of coefficients on explanatory variables at a 95% confidence level? Explain what they mean.**

Text

Description automatically generated

**Season:**

**Summer**

The 95 % confidence interval for coefficient of Season Summer is (501.3,689.9). We are confident that the actual coefficient of Season Summer in the population lies in between 501.3 and 689.9. With presence of Summer Season, the ice cream sales increase by between 501.3 and 689.9.

**Spring**

The 95 % confidence interval for coefficient of Season Spring is (258.1,412). We are confident that the actual coefficient of Season Spring in the population lies in between 258.1 and 412. With presence of Spring Season, the ice cream sales increase by between 258.1 and 412.

**Price:**

The 95 % confidence interval for coefficient of Price is (-194.5, -110.1). We are confident that the actual coefficient of Price in the population lies in between -194.5 and -110.1. With each unit of extra “Price” that employee earns, the ice cream sales decrease by between 110 and 194.

**Temperature:**

The 95 % confidence interval for coefficient of Temperature is (0.0081, 0.014). We are confident that the actual coefficient of Temperature in the population lies in between 0.0081 and 0.014. With each unit of extra “Temperature” that increases, the ice cream sales increase by between 0.0081 and 0.014.

1. **Explain if your data meets the regression conditions.**

**Residuals Vs Fitted Plot**

Chart, scatter chart

Description automatically generated

From the above viz. a scatter plot between Residuals Vs Fitted Values. We can say there is not much variance observed in the graph and no funnel like shape that indicates no Heteroscedasticity, it follows homoscedasticity which is one of the conditions of linear/multi linear regression models. Heteroscedasticity is state where, we see trend in errors but this time the trend is larger or smaller (as opposed to the errors clearly influencing each other)

**Normal Q-Q Plot**

Chart, line chart

Description automatically generated

From the above viz. all data points are very close around a straight dashed line and few going out of course (outliers). We say errors are following normal distribution which is one of the conditions for linear/multi linear regression models

**D. Prediction: (10 points) What is the predicted value of ice cream sales in a location in Country B where the average income of residents is £30,000 when the temperature is 3 degrees in Winter, and average price per serving of ice cream is £5? Also, quantify the uncertainty around this prediction using an appropriate interval at 90% confidence interval. (75 words max).**

**For Calculation Please refer – R Code (116 to 137 lines)**

Country – B

Average income of residents = £30,000

Temperature = 3

Season = Winter

Average Price per Serving of Ice Cream = £5

As per the details, Ice Cream Sales = 126.15

With confidence interval of 90%, Ice Cream Sales lies in between 45.65078 (lower range value) and 206.6427 (upper range value)

**For Hypothesis Testing Please refer the R code with the below heading (Lines – 142 onwards)**

**######################### Hypothesis Test Starts here ###################################**

**It comes at ending part of R code**

**Hypothesis Test Results:**

**B. Hypothesis testing: (10 points) Construct your hypotheses for testing the average temperature in**

**location in Country A relative to the average temperature in a location in a location in Country B.**

**Test the hypothesis and explain the result. (100 words max)**

Before we are going to check existence of differences across their temperature means,

First ensure that data should pass a test of homoscedasticity - are temperature variances same

We do this in R using Fisher's F-test

H0: No difference in their temperature in variance between 2 countries.

H1: Difference exists in variance of temperatures between 2 countries.

**We got p-value > 0.05, so we reject Alternate hypothesis and accept null hypothesis**

H0: No difference in average of temperatures between 2 countries.

H1: Difference exists in average of temperatures between 2 countries.

**We got p-value < 0.05, we reject Null Hypothesis and accept alternate hypothesis which says there exists some difference in average temperatures across countries A & B**