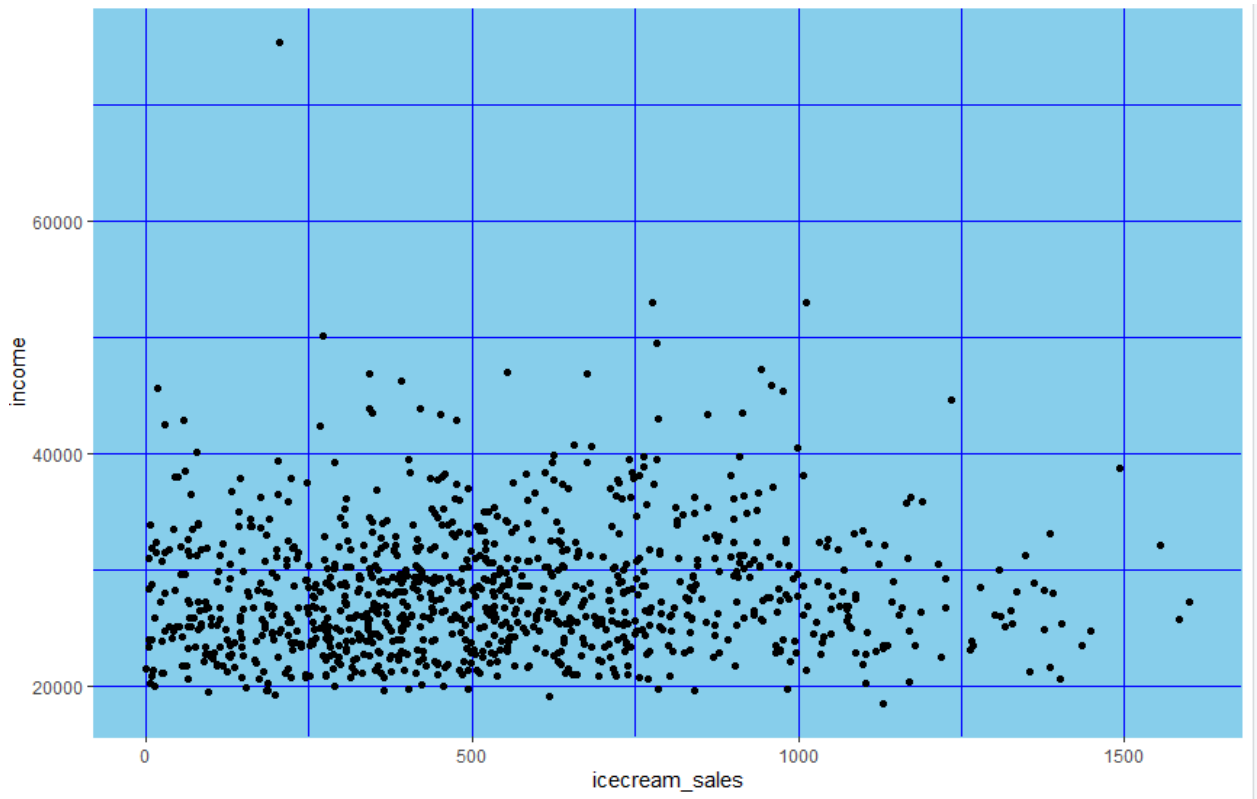
A. Exploratory data analysis (EDA): (30 points) Perform exploratory data analysis (EDA) and explain your variables numerically and graphically. Please do not replicate the same investigation in numeric and visual explorations. A brief interpretation should accompany your R output and plot. (575 words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

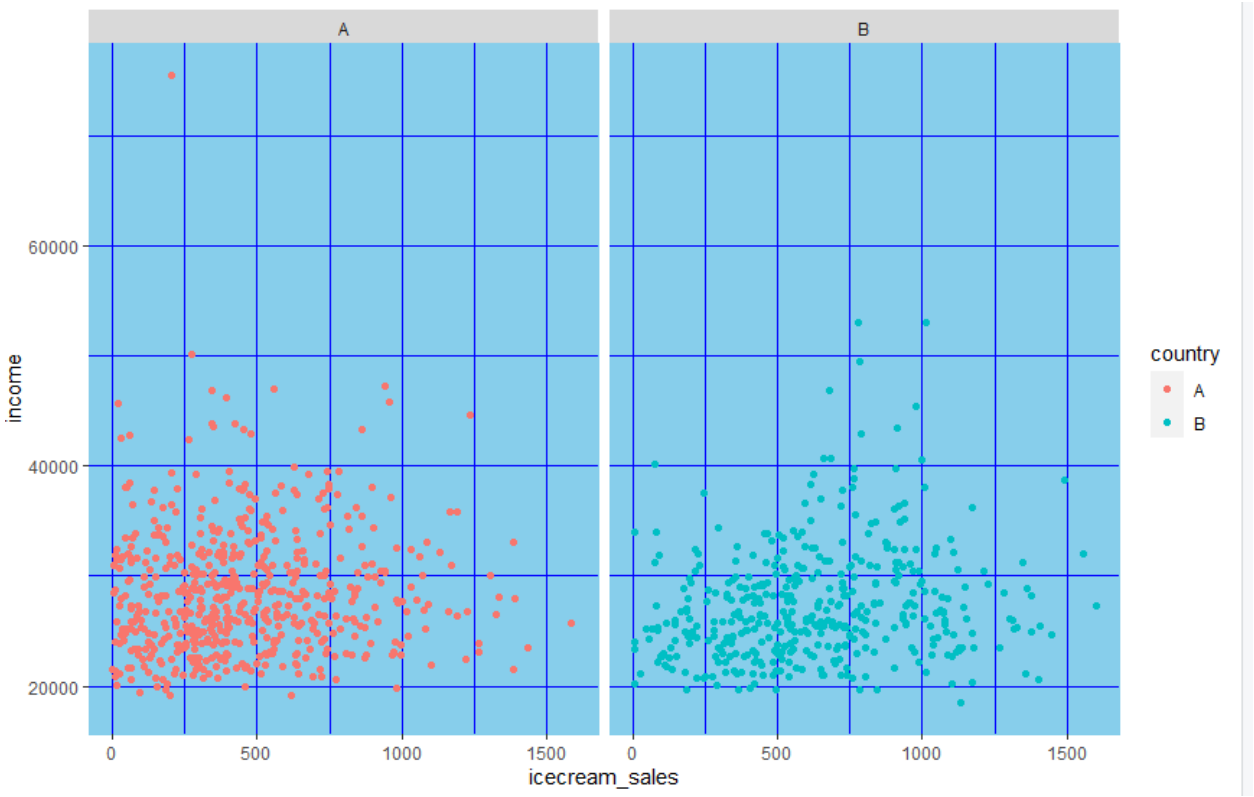
For exploratory data analysis kindly refer code lines between 24 to 96

Sales Vs Income - We are going to analyze if the income has any effect over sales



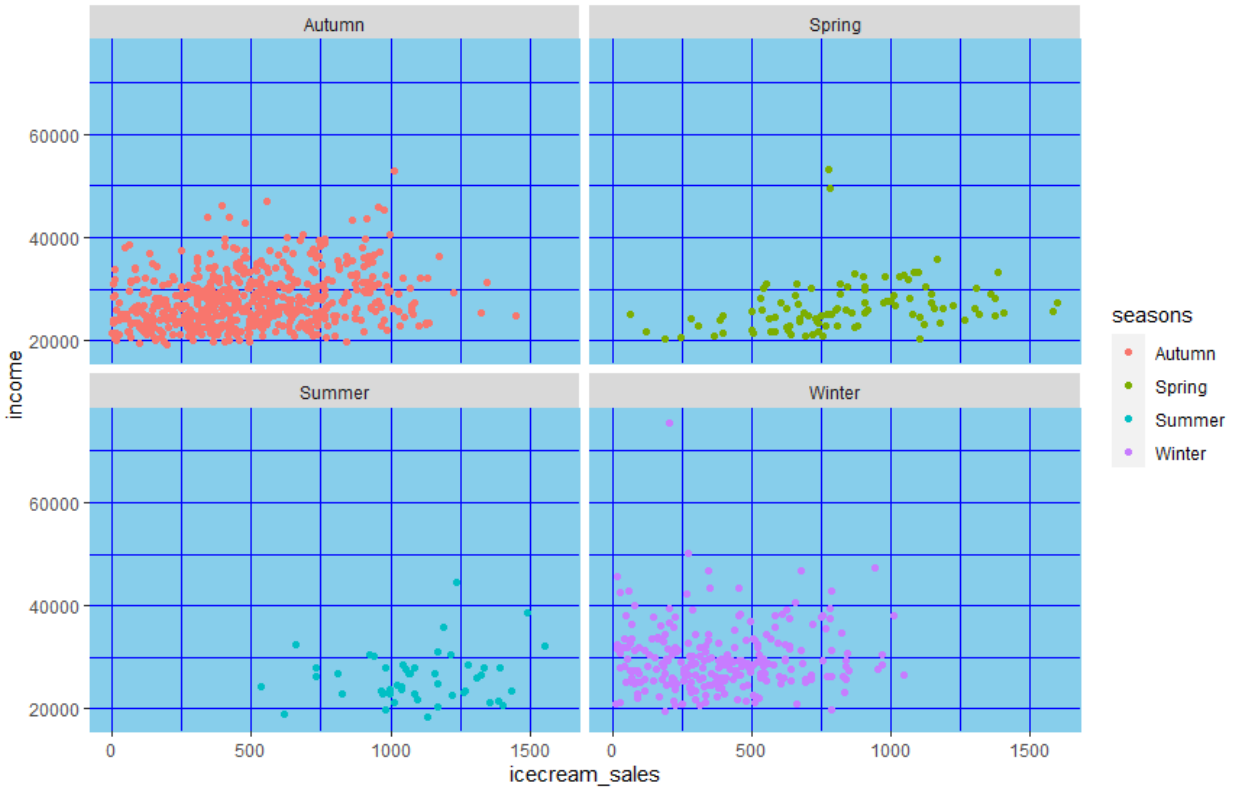
It can be said that employee salaries do not have a significant impact on ice cream sales. As you can see from the graph, even with low salaries, ice cream sales are high, and high-value employees have slightly higher ice cream sales.

Sales Vs Income over each country - We are going to analyze their distribution separately



It can be said that employee salaries do not have a significant impact on ice cream sales. As you can see from the graph, even with low salaries, ice cream sales are high, and high-value employees have slightly higher ice cream sales. For Country A, the scatter plot shows that income has little effect on ice cream sales, but for Country B, income has little effect on ice cream sales.

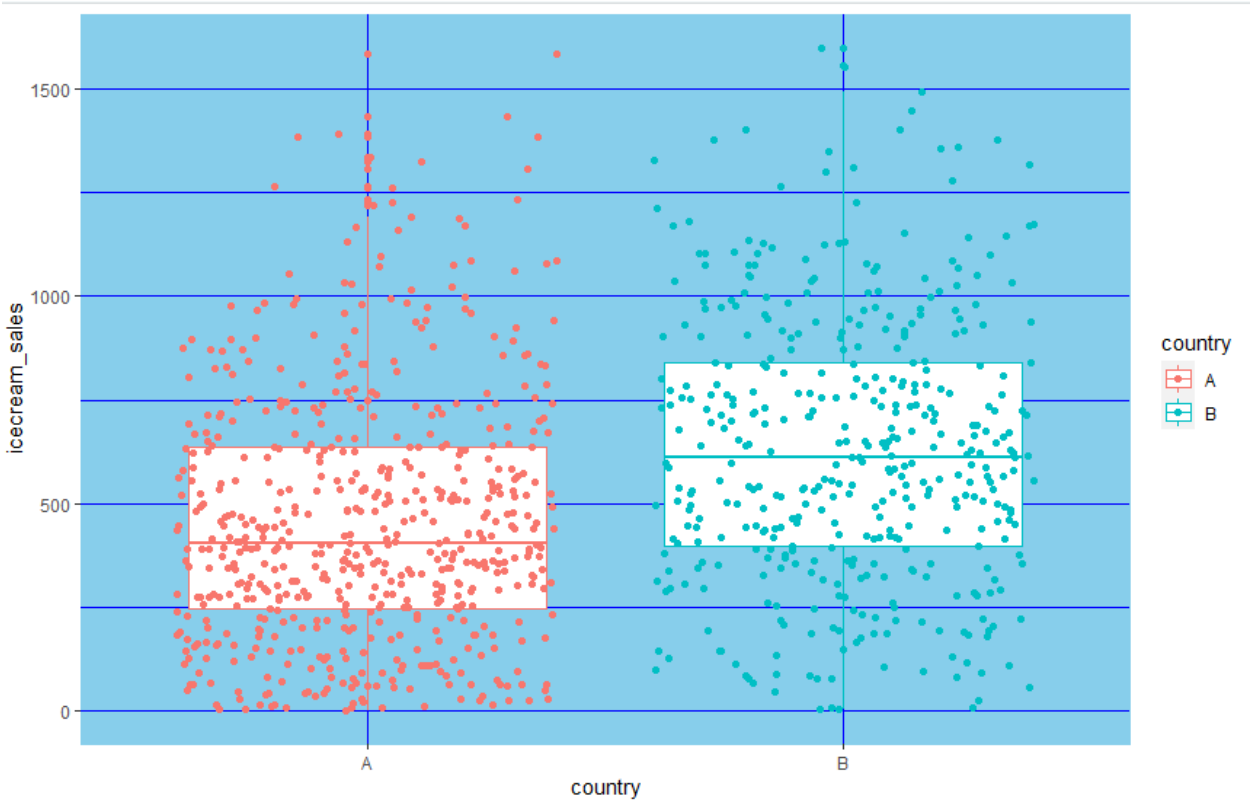
Sales Vs Income over seasons - We are going to analyze if the income has any effect over ice cream sales



It can be said that employee salaries do not have a significant impact on ice cream sales. As you can see from the graph, even with low salaries, ice cream sales are high, and high-value employees have slightly higher ice cream sales. For Country A, the scatter plot shows that income has little effect on ice cream sales, but for Country B, income has little effect on ice cream sales. All we can say is that the dates were so scattered that it happened more above Autumn> Winter> Spring> Summer. As the season progresses, income will have less impact on ice cream sales, he said.

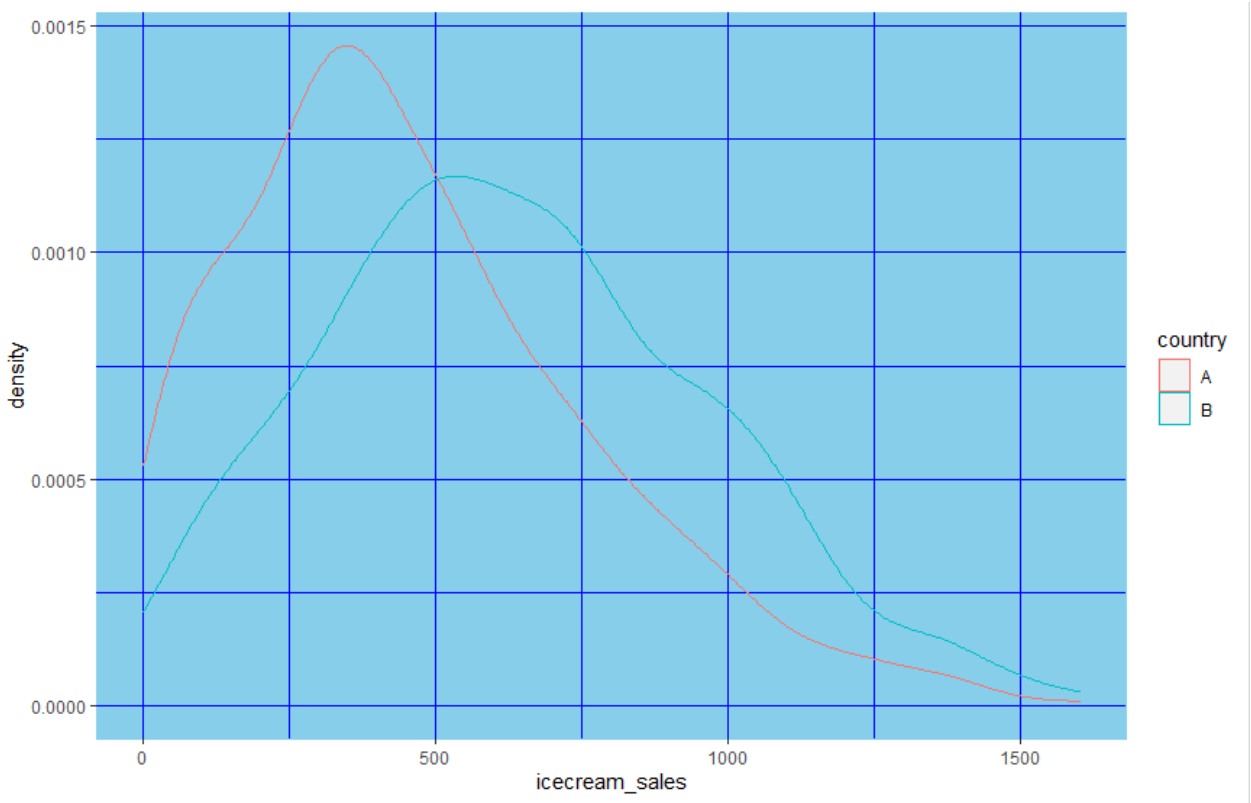
Sales:

Sales over country



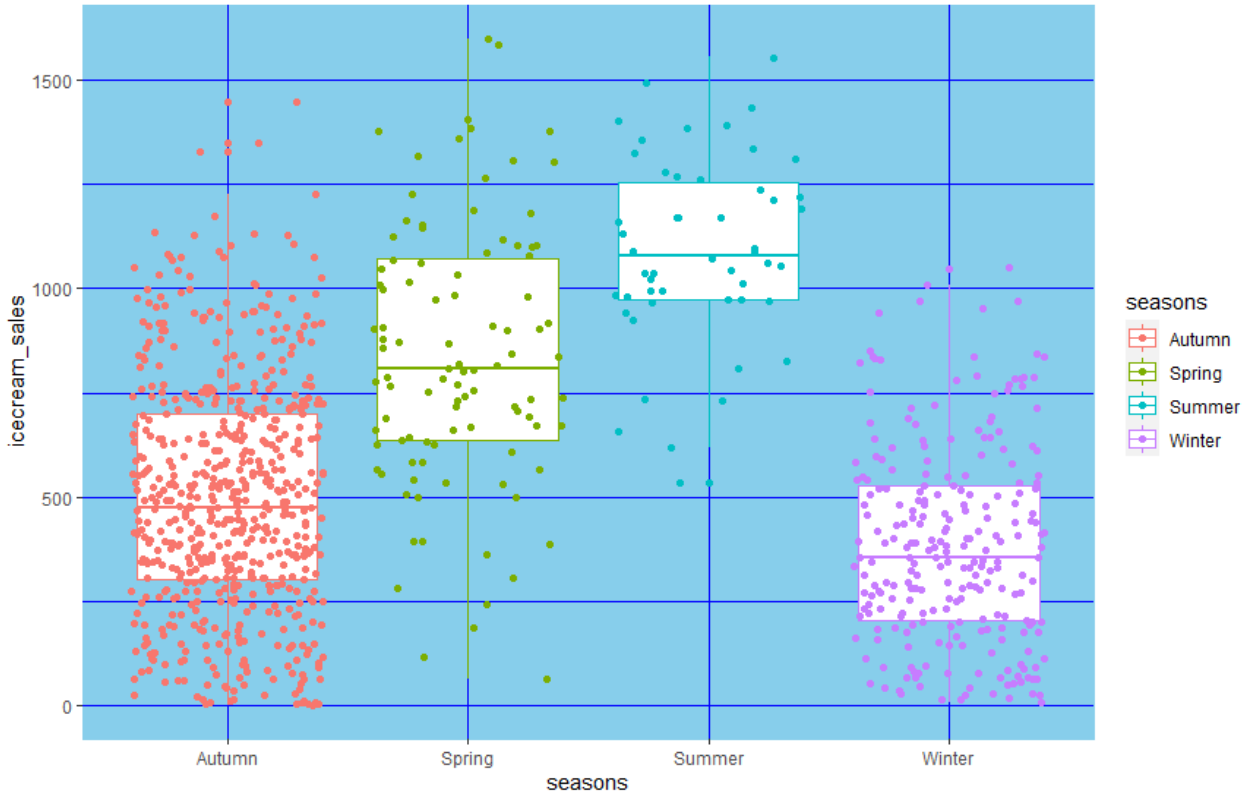
There are more records in country A than in country B. The average ice cream sales in Country B are higher than those in Country A. You can also see that most of B's ice cream sales are better than A's.

Sales over country



It can be seen that most of the ice cream sales from Country A are lower than those of Country B.

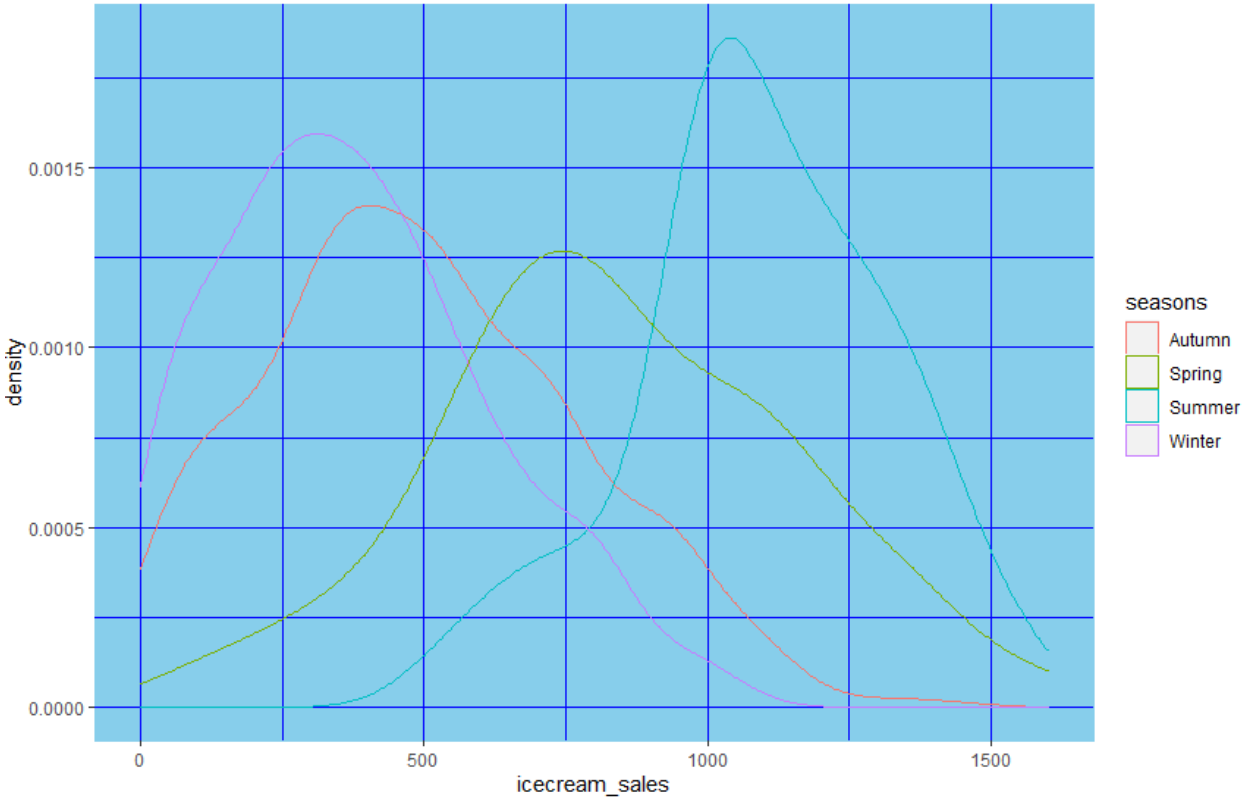
Sales over Seasons



From the jitter plot on the boxplot above, we can say two statements.

Fall> Spring> Winter> Summer data collection has done more Or sales increased in Autumn> Spring> Winter> Summer In addition, the data is evenly distributed across all seasons (less skew)

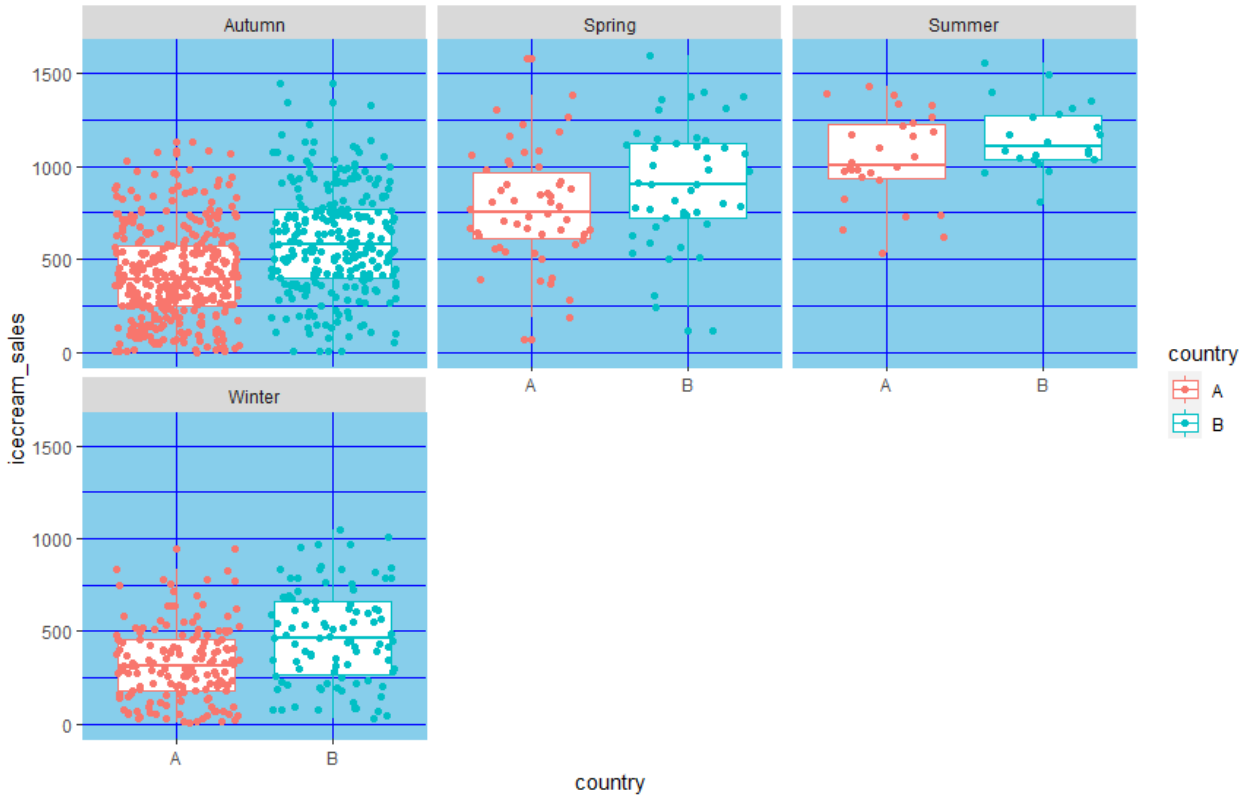
Sales Over Seasons



There is some bias in the summer season (higher sales were more common)

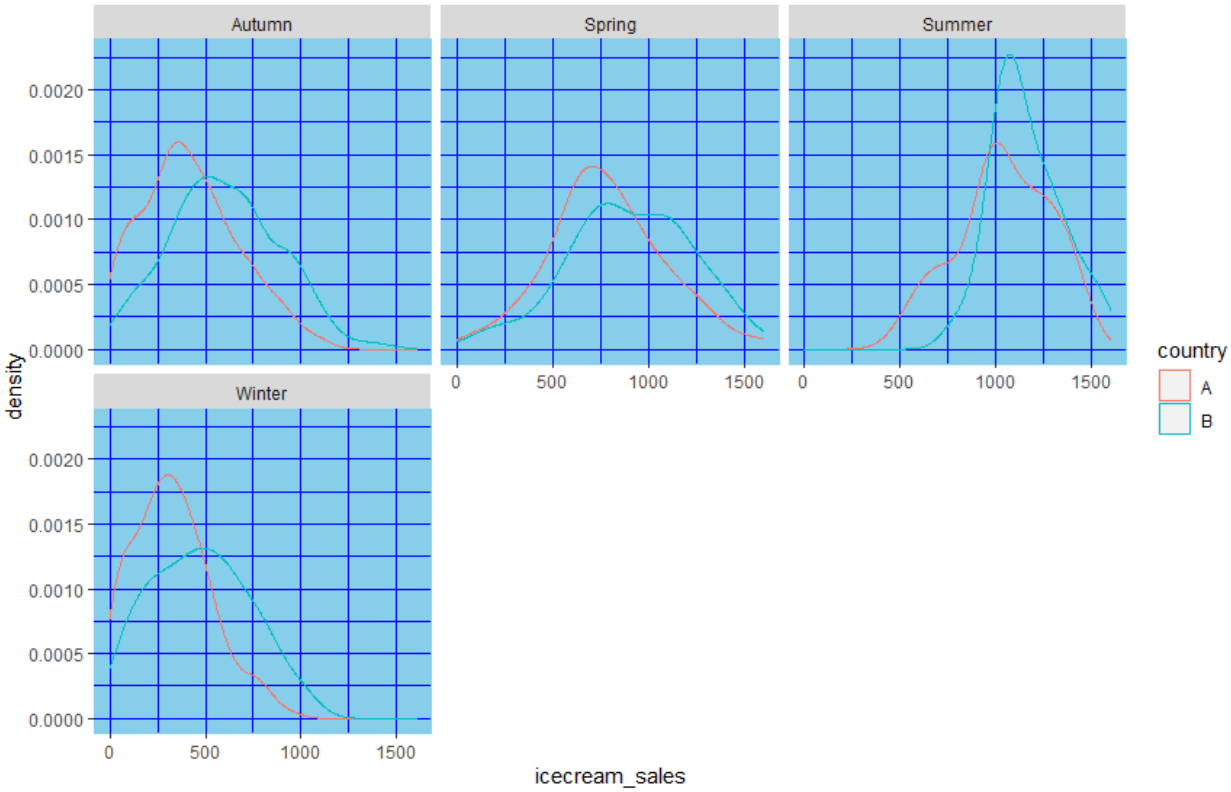
For the rest of the season (autumn, spring, winter), the dates are almost evenly distributed (ice). Sale of cream)

Sales Over Country & Seasons



That is, from the above. It can be said that there is a lot of winter data, but sales in summer> spring> winter> autumn are also high. The important thing is that Country B is the leader in ice cream sales.

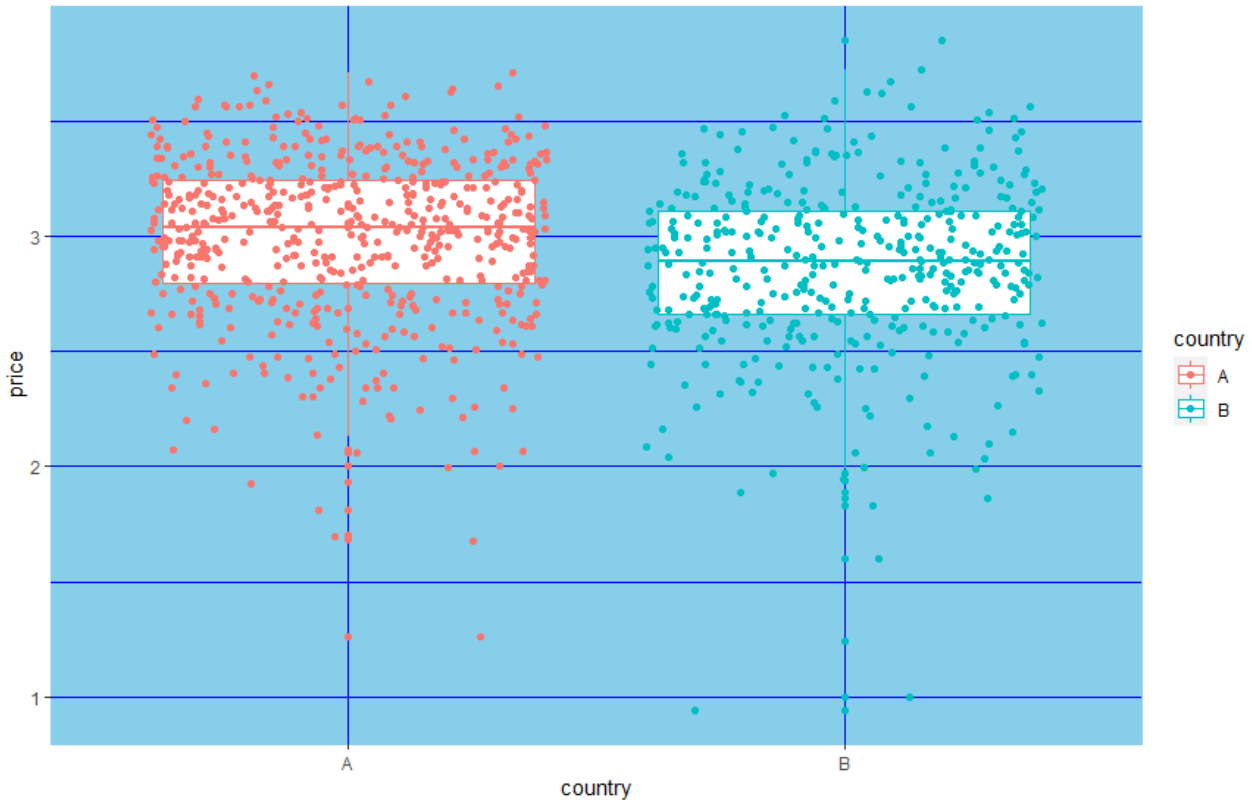
Sales Over Country & Seasons



In density visualization, all country B leads in terms of ice cream sales, but our value comes from country A (in terms of quantity).

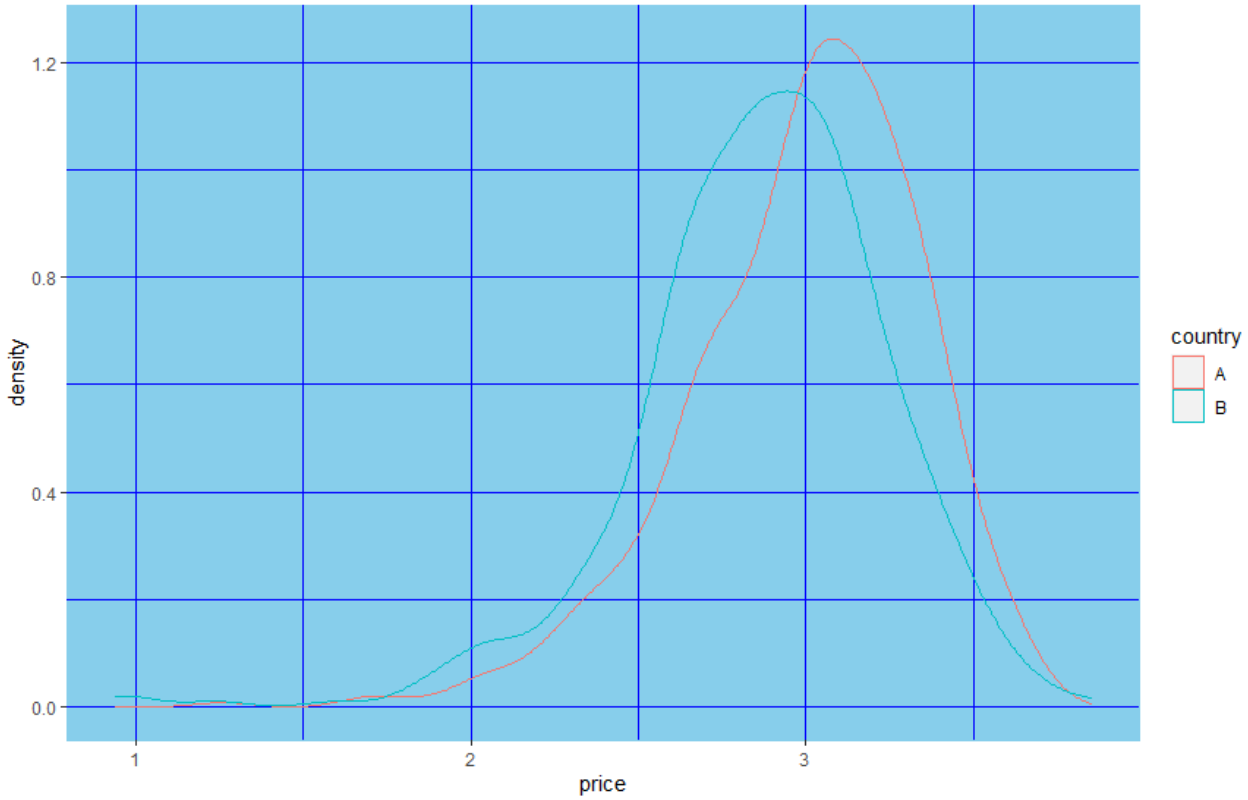
About Ice Cream Average Price

Average Price Over Country



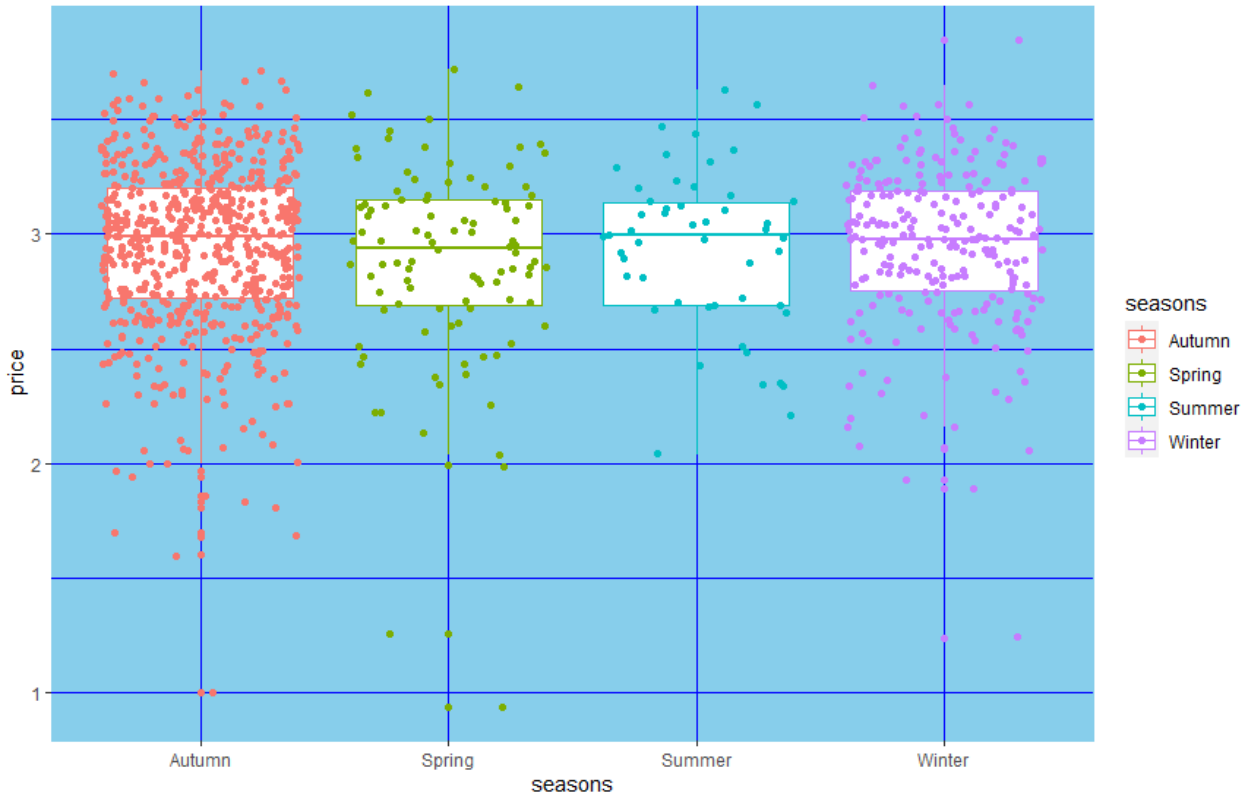
Jitter plot on a boxplot-From a point of view, the average ice cream prices are about the same (center line), and it can be said that Country B has the lowest average ice cream price in more numbers than Country A. But in A it is compared a little higher in B

Average Price Over Country



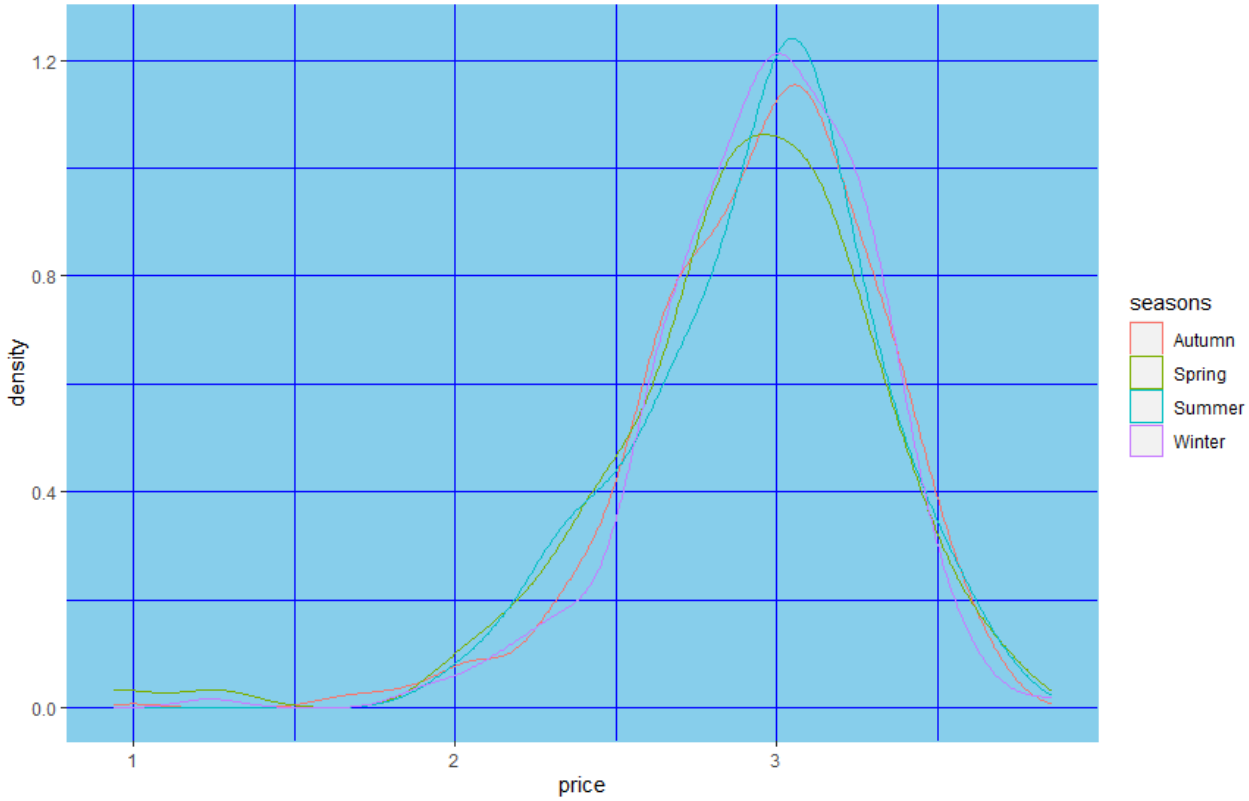
You can see that Country A sells more ice cream at a higher price than Country B (although both Country A and Country B sell more ice cream, as shown in the density graph below). We sell cream at a higher price).

Price Over Seasons



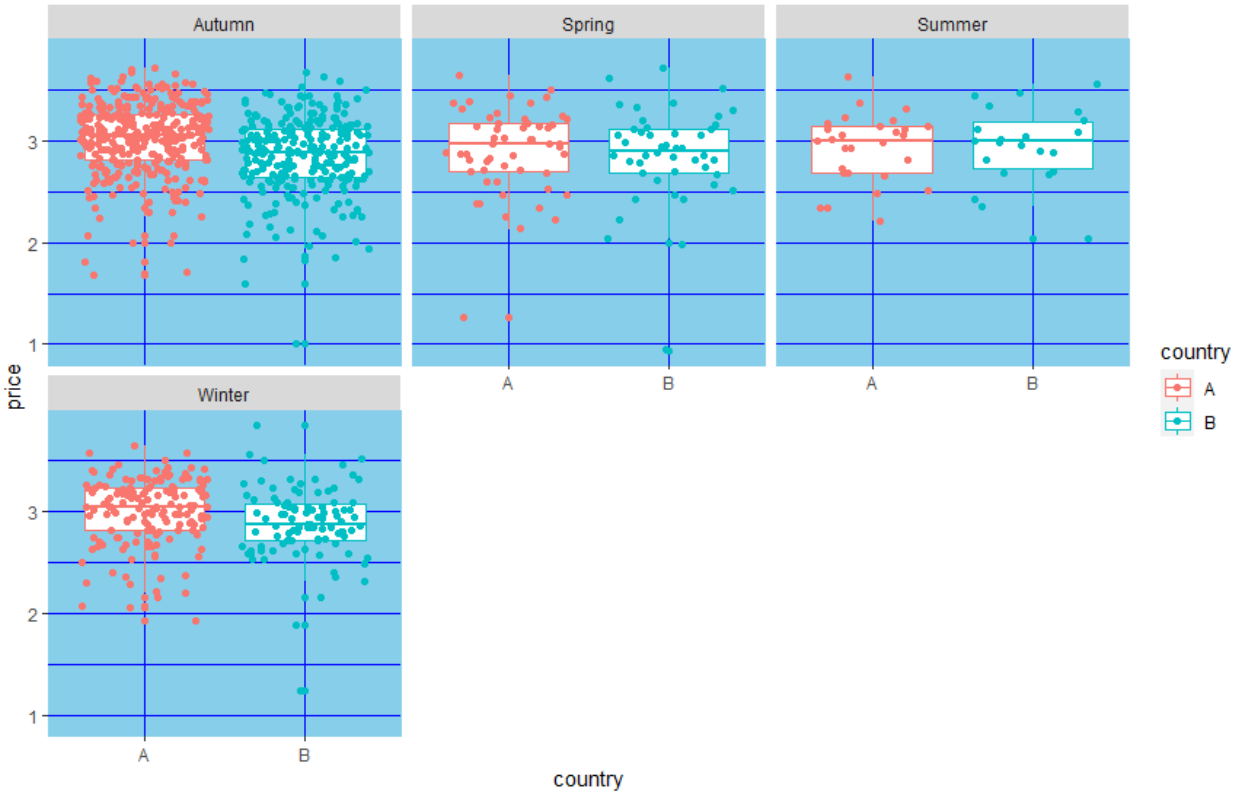
The average price of ice cream is about the same regardless of the season Also, the median price distribution for autumn and winter-most records are as above.

Average Price Over Seasons



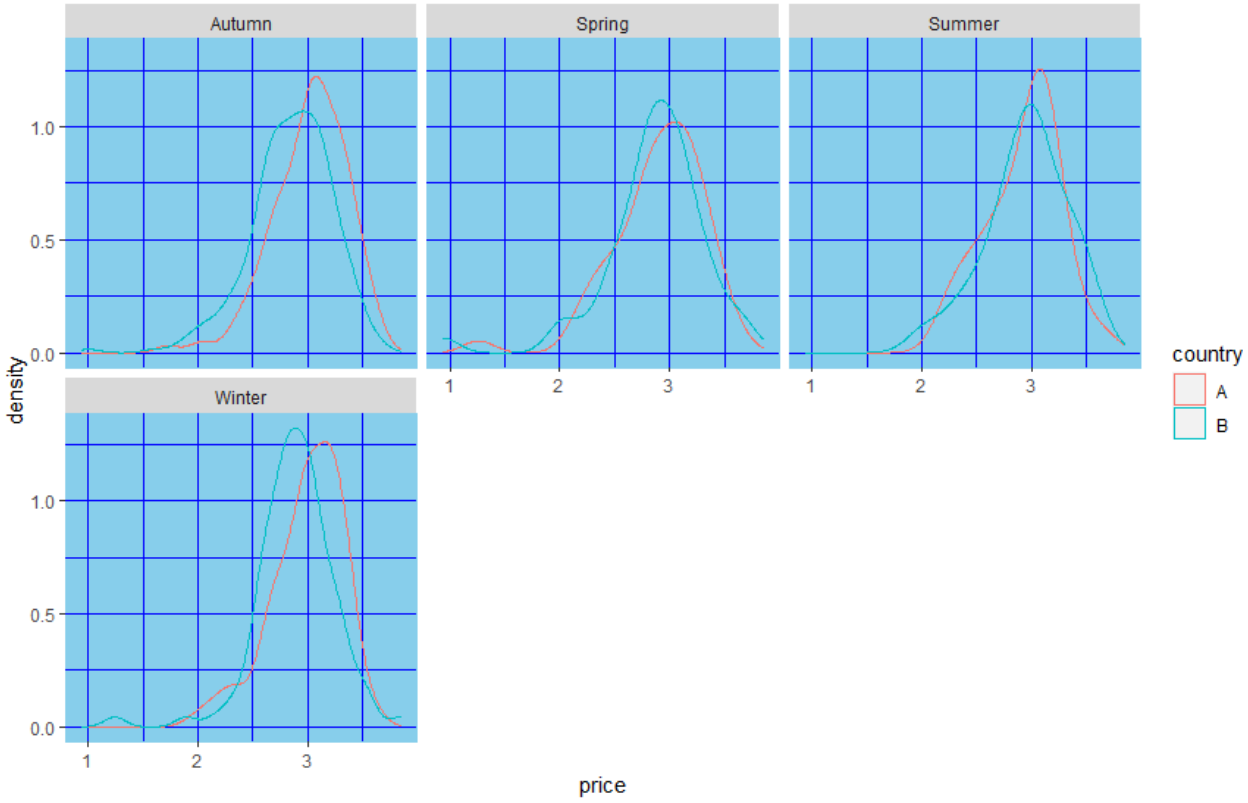
It is clear that the price distribution is about the same for all seasons and that prices tend to be slightly higher.

Average Price Over Country & Seasons



The average ice cream price in Country A is slightly higher in all seasons compared to Country B.

Average Price Over Country & Seasons



Country A is slightly above Country B at certain ice cream prices in all seasons And the data from both countries are distorted over all seasons with high prices for ice cream.

Multi Linear Regression Model Summary refer lines from 100 to 130 in 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)

Specifically, answer the following questions:

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

Regression Equation:

There are two categories of country traits, converted into one hot coding of dummies, and the same applies to the four categories of seasonal traits. When training the model with data using

multiple linear regression, You will get the coefficients for each characteristic except one category at a time. A dummy one-hot encoding function obtained from the country and season functions. Therefore, Country_B and Season_Winter have NA. Those effects are taken into account along with the intercept values 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

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

• Price, Country_A, Season_Spring, and Season_Summer have a reasonable impact on ice cream sales, based on the coefficient sizes of all characteristics obtained after the training model.
• Of these, Price & Country_A has a significant negative impact on ice cream sales.
• Season_Spring and Season_Summer have a positive effect on ice cream sales.
• Income and temperature have the least impact on ice cream sales
• Temperature has a moderate effect on ice cream sales

3. **All else being equal, what is 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?**

   All other ice creams are the same. Use the coefficients obtained above (coefficient values mentioned in
   Country A with an average income of £20,000 has value 1459
   Country B with an average income of £30,000 has value 1710

   Country B has 251 units more ice cream sales compared to country A

4. **All else being equal, what is the predicted change in ice cream sales if the price goes up by £0.50 and temperature goes up by 2 degrees at the same time?**

   With all being equal
   Price goes up by 0.50 and temperature goes up by 2
   Explanation:
   Ice Cream Sales-Initial =
   Constant + Price x (-152.3) + Temperature x (7.1) ➔ 1
   Ice Cream Sales -Later =
   Constant + (Price+0.50) x (-152.3) + (Temperature+2) x (7.1) ➔ 2
   Change in Ice Cream Sales = equation 2 – equation 1 = (7.1 x 2) + (0.50 x -152.3) = -61.95

   Ice Cream Sales will decrease by 61.95 with temperature goes up by 2 degrees and Prices goes up by £0.50 at same time

5. **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 – 125 R code**)

6. **Is this model statistically significant at a 5% significance level?**

   F test of overall significance is a special form of F test. Compares a model without a predictor to the specified model. Regression models that do not contain predictors are also called intercept-only models.

The hypothesis of the F-test for overall significance is as follows: Null hypothesis: There is no evidence that the model is significant

Alternative hypothesis: Sufficient statistic model is statistically significant There is a significance p value = 0.1% If a particular F-number obtained after training the regression model has a p-value of less than 0.1%, it rejects the null hypothesis (see Liner-125 R Code). 2.2e16 (model summary) far less than the P value & let significance benchmark (5%) corresponding to the model We conclude that the model is statistically significant at 5%.

7. **What are the confidence intervals of coefficients on explanatory variables at a 90% confidence level? Explain what they mean.**

```
> confint(linear_regression_model)
                          2.5 %         97.5 %
(Intercept)      3.680062e+02  706.79118826
income           8.097484e-03    0.01382867
price           -1.944690e+02 -110.09420537
temperature      4.140566e+00   10.03519356
country_A       -1.727197e+02 -110.00385006
country_B                  NA            NA
seasons_Autumn   1.311166e+01  104.17030721
seasons_Spring   2.584739e+02  412.69609289
seasons_Summer   5.013378e+02  689.89431149
seasons_Winter             NA            NA
> |
```

Season:

Summer

The 90% confidence interval for the summer factor is (501.3.689.9). We are confident that the actual summer factor of the population is between 501.3 and 689.9. During the summer season, ice cream sales will increase between 501.3 and 689.9.

Spring

The 90 % self-belief c program language period for coefficient of Season Spring is (258.1,412). We are assured that the real coefficient of Season Spring withinside the populace lies in among 258.1 and 412. With presence of Spring Season, the ice cream income boom with the aid of using among 258.1 and 412.

Price:

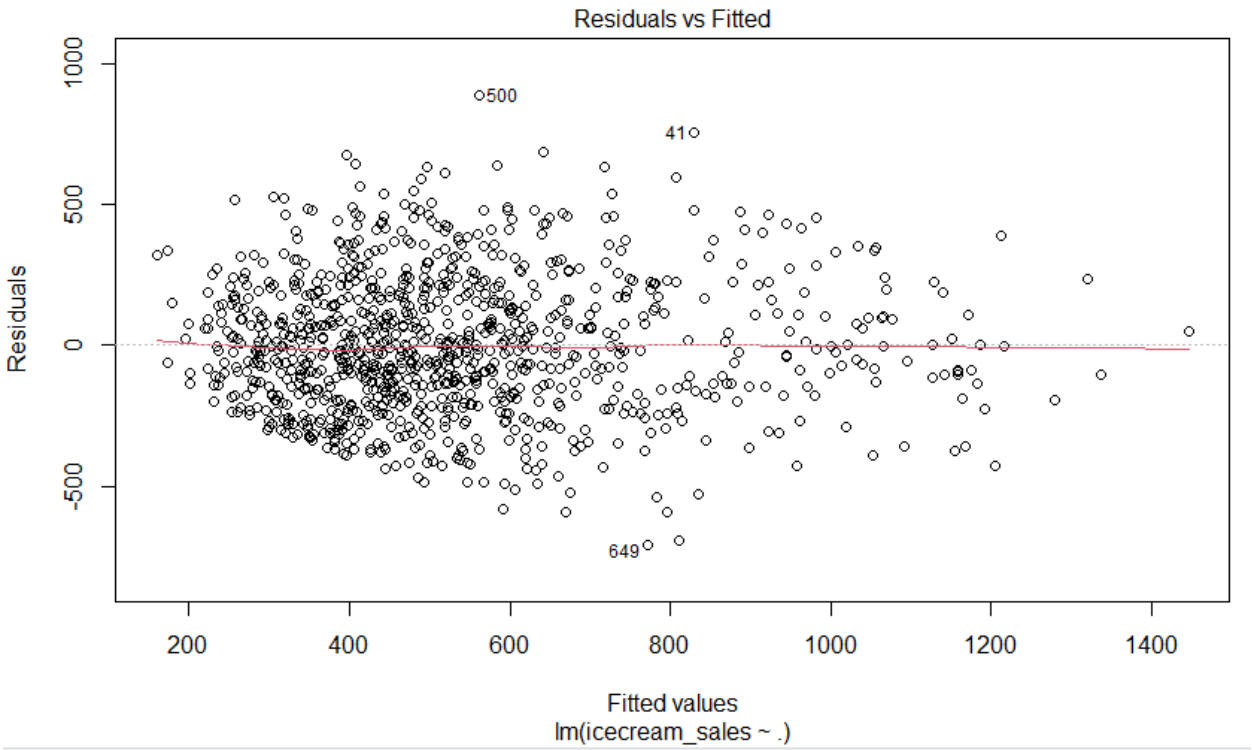
The 90% confidence interval for the price factor is (194.5, 110.1). We are confident that the real price factor of the population is between 194.5 and 110.1. As the unit of "price" earned by an employee increases, ice cream sales decrease from 110 to 194.

Temperature:

The 90% confidence interval for the temperature coefficient is (0.0081, 0.014). I am convinced that the actual temperature coefficient of the population is between 0.0081 and 0.014. With each additional "temperature" unit, ice cream sales increase from 0.0081 to 0.014.

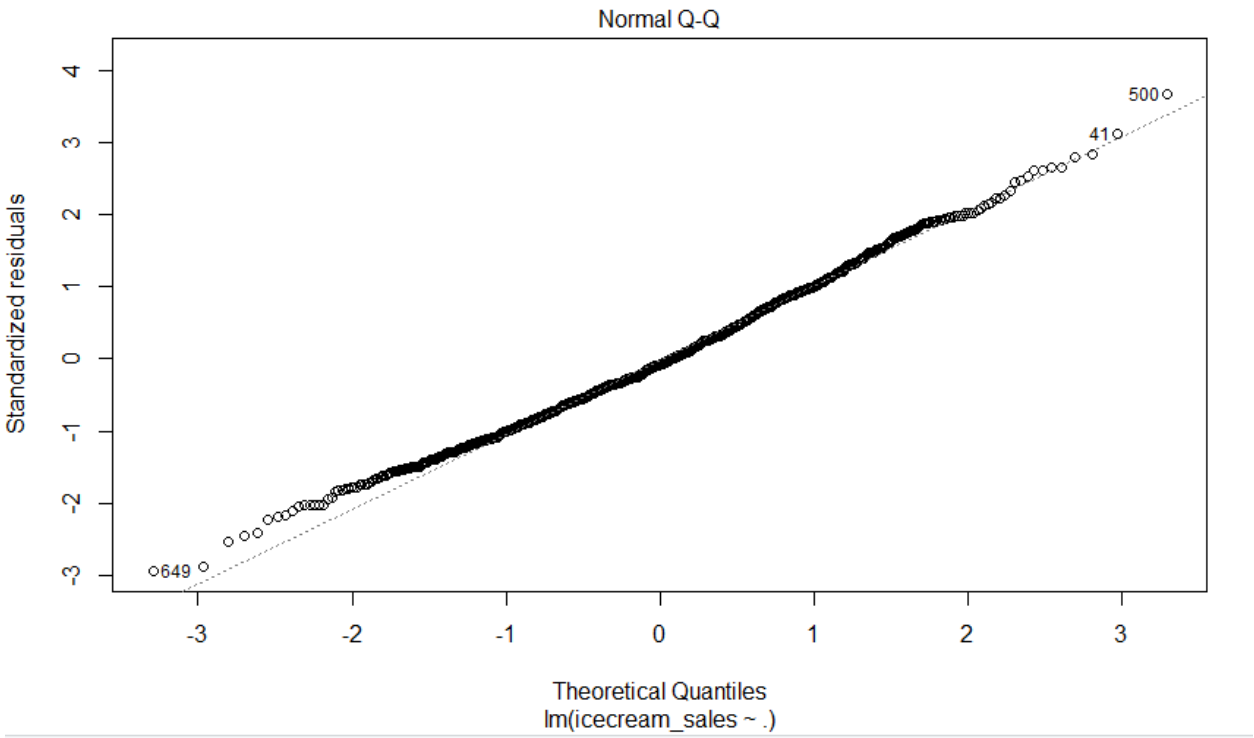
8. Explain if your data meets the regression conditions.

Residuals Vs Fitted Plot



That is, from the above. Scatter plot between residuals and approximations. The graph shows little variance and does not show non-uniform variance, but it can be said that it is not a funnel-shaped shape that follows homoscedasticity, which is one of the conditions of the linear / multiple linear regression model. Non-uniform variance is a condition in which there is a tendency for errors, but this time the tendency is higher or lower (as opposed to the clear influence of errors on each other).

Normal Q-Q Plot



That is, from the above. All data points are very close to the dashed straight line, with few data points disappearing spontaneously (outliers). The error is said to follow a normal distribution, which is one of the conditions of a linear / multiple linear regression model.

D. Prediction: (10 points) What is the predicted value of ice cream sales in a location in Country A where the average income of residents is £30,000, the temperature is 23 degrees in Spring, and average price per serving of ice cream is £3? Also, quantify the uncertainty around this prediction using an appropriate interval at 95% confidence interval. (75 words max). For Calculation Please refer – R Code (116 to 137 lines)

Country – A
Average income of residents = £30,000
Temperature = 23
Season = Spring
Average Price per Serving of Ice Cream = £3
As per the details, Ice Cream Sales = 126.15
With confidence interval of 95%, Ice Cream Sales lies in between 716.1015 (lower range value) and 817.2799 (upper range value)

For Hypothesis Testing Please refer the R code with the below heading (Lines – 142 onwards)
It comes at ending part of R code
Hypothesis Test Results:

B. Hypothesis testing: (10 points) Construct your hypothesis for testing the average ice cream sales in a location in Country A relative to the average sales in a location in Country B. Test the hypothesis and explain the result. (100 words max)

Before checking the difference in ice cream sales means, first make sure that the data must pass the homoscedasticity test, Are the sales variance is equal. This is done in R using Fisher's F-test.

H0: There is no difference in their sales in variance between 2 countries - Null Hypothesis
H1: There is difference exists in average of sales between 2 countries - Alternate Hypothesis

Since the p-value exceeds 0.05, we reject the alternative hypothesis and accept the null hypothesis.

H0: There is no difference in their sales in variance between 2 countries - Null Hypothesis
H1: There is difference exists in average of sales between 2 countries - Alternate Hypothesis

It has a p-value < 0.05, reject the null hypothesis and accept the alternative hypothesis that there is some difference in the average temperature between countries A and B.