Assignment Based Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

ANSWER: I have plotted the categorical variables with the target variables on box plot has inferred following effect on target:

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of May, June, Jul, Aug and Sept. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted a greater number of booking from the previous year, which shows good progress in terms of business

2. **Why is it important to use drop_first=True during dummy variable creation?**

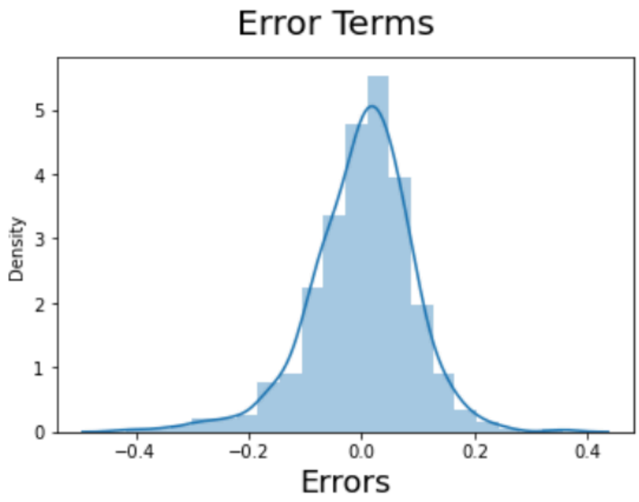
ANSWER: drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

ANSWER: ‘temp’ variable has the highest correlation with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

ANSWER: - Residuals distribution should follow normal distribution and centred around 0(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

ANSWER: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- spring
- sept

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

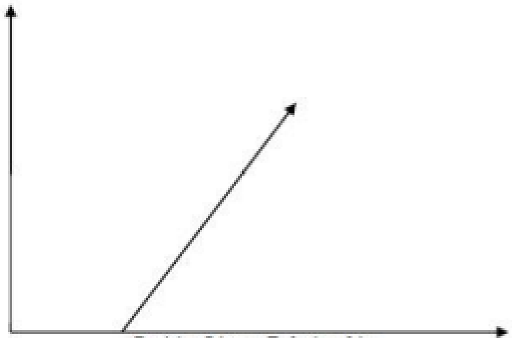
ANSWER: Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −
$Y = mX + c$
Here,
Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship:

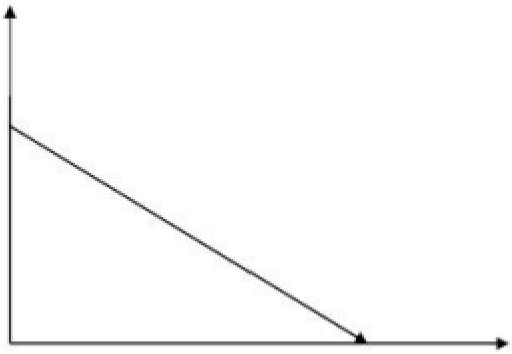
▪ A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

- Negative Linear relationship:

▪ A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Linear regression is of the following two types −
1. Simple Linear Regression
2. Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model −

• Multicollinearity – Linear regression model assumes that there is very little or no multicollinearity in the data. Basically, multicollinearity occurs when the independent variables or features have dependency in them.
• Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
• Relationship between variables – Linear Regression model assumes that the relationship between response and feature variables must be linear.
• Normality of error terms – Error terms should be normally distributed
• Homoscedasticity –There should be no visible pattern in residual values.

2. **Explain the Anscombe's quartet in detail.**

ANSWER: Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasise both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties

- The first scatter plot (top left) appears to be a simple linear relationship.
- the second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. **What is Pearson's R?**

ANSWER: Pearson's R is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data? R = 1 means the data is perfectly linear with a positive slope R = -1 means the data is perfectly linear with a negative slope R = 0 means there is no linear association.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

ANSWER: Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

ANSWER: VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.
VIF = 1 / (1 – Ri 2)

If there is perfect correlation, then VIF = infinity where Ri is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity".

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

ANSWER: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?