Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer1:

Analysis was done on categorical columns using the boxplot and barplot. Below are the few inference points from the visualization –

- Fall season seems to have more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings in month: may, june, july, aug, sep and oct.Trend increased starting of the year tillmid of the year and then it started decreasing at the end of year. Booking count for each month seems to have increased from 2018 to 2019.
- Clear weather have more booking. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.
- Thu, Fir, Sat and Sun have more number of bookings as compared to start of the week.
- When it's not holiday, booking seems to be less in number.
- Booking seemed to be almost equal either on working day or non-working day. But, the count increased from 2018 to 2019.
- 2019 have more number of booking from the previous year, indicating growth in business.

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

Answer2:

drop_first = True 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.

Assuming we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not X and Y, then It is obviously Z. Hence, we do not need 3rd variable to identify the Z.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer3:

'temp' variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer4:

I have validated the assumption of Linear Regression Model based on below 5 assumptions:

1. Normality of error terms
   - Error terms should be normally distributed
2. Multicollinearity check
   - There should be insignificant multicollinearity among variables.
3. Linear relationship validation
   - Linearity should be visible among variables
4. Homoscedasticity
   - There should be no visible pattern in residual values.
5. Independence of residuals
   - No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer5:

- temp
- winter

- sep

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer1:

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 −

- 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 −

Linear regression is of the following two types −

- Simple Linear Regression

- Multiple Linear Regression

Assumptions -

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

- Multi-collinearity –

 -- Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity 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 auto-correlation 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. (3 marks)

Answer2:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but vary significantly when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and not relying solely on summary statistics. Let's delve into the details of each dataset in Anscombe's quartet:

1. **Dataset I:**
 - **Summary Statistics:**

- Mean of x: 9.0
- Variance of x: 11.0
- Mean of y: 7.50
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression equation: \(y = 3 + 0.5x\)

- **Graphical Representation:**
- Consists of a clear linear relationship between x and y.
- The linear regression line represents the data well.

2. **Dataset II:**
- **Summary Statistics:**
- Mean of x: 9.0
- Variance of x: 11.0
- Mean of y: 7.50
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression equation: \(y = 3 + 0.5x\)

- **Graphical Representation:**
- Similar to Dataset I.
- Demonstrates a strong linear relationship between x and y.
- Scatter plot and regression line closely resemble Dataset I.

3. **Dataset III:**
- **Summary Statistics:**
- Mean of x: 9.0
- Variance of x: 11.0
- Mean of y: 7.50
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression equation: \(y = 3 + 0.5x\)

- **Graphical Representation:**
- Appears to have a non-linear relationship.
- A single outlier greatly influences the linear regression line.
- Highlights the impact of outliers on regression analysis.

4. **Dataset IV:**
- **Summary Statistics:**
- Mean of x: 9.0
- Variance of x: 11.0
- Mean of y: 7.50
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression equation: \(y = 3 + 0.5x\)

- **Graphical Representation:**
- Essentially a vertical line with one outlier.
- The linear regression line is heavily influenced by the outlier.
- Illustrates the sensitivity of regression analysis to influential points.

3. What is Pearson’s R? (3 marks)

Answer3:

Pearson's correlation coefficient, denoted as "Pearson's R," is a statistical metric indicating the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to

1, where 1 signifies a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 suggests no linear relationship. Calculated by dividing the covariance of the two variables by the product of their standard deviations, Pearson's R is a widely-used measure across disciplines like statistics, economics, and psychology. It's essential to recognize that while effective for linear relationships, Pearson's R may not capture non-linear associations between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer4:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.
Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Answer6:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this for the conclusion that the two data sets have come from populations with different distributions.
Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.