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?

Answer. I've used the boxplot and scatter plot to analyses categorical columns. The following are a few things we may conclude from the visualization.

- The fall season appears to have drawn in more bookings. And, from 2018 to 2019, the number of bookings has risen dramatically in each season.
- The majority of reservations were made in the months of May, June, August, September, and October. As the year progressed, the trend grew until the middle of the year, when it began to decline as we approached the conclusion of the year.
- Bookings are higher on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
- When it is not a holiday, the number of bookings appears to be lower, which seems sensible because during holidays, the number of bookings is higher.
- Bookings appeared to be nearly equal on working and non-working days.
- 2019 saw a higher amount of bookings than the previous year, indicating positive business growth.

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

Answer. It's important to use drop first = True since it reduces the extra column formed during dummy variable construction. As a result, the correlations between dummy variables are reduced.

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

Answer. The variable 'temp' 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. On the basis of the following five assumptions, we have validated the assumption of the Linear Regression Model.

- Normality of error terms
- The distribution of error terms should be normal.
- Multicollinearity check
- Multicollinearity between variables should be insignificant.
- Linearity between variables should be visible.
- In residual values, there should be no discernible pattern.

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

Answer. 1. temp
2. winter
3. sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer. The statistical model that analyses the linear connection between a dependent variable and a set of independent variables is known as linear regression. The term "linear connection" refers to the fact that when the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes as well (increase or decrease).

Assumptions –

1. Multi-collinearity

The linear regression model implies the data has minimal or no multi-collinearity. Multi-collinearity is defined as when the independent variables or features are reliant on one another.

2. Auto-correlation

Another hypothesis The linear regression model assumes that the data has little to no auto-correlation. Auto-correlation arises when residual errors are dependent on one another.

3. Relationship between variable

The linear regression model implies a linear relationship between the response and the feature variables.

4. Normality of error terms

The distribution of error terms should be normal.

5. Homoscedasticity

In residual values, there should be no discernible pattern.

The following equation can be used to illustrate the relationship mathematically-

$Y=mx+c$

The dependent variable, Y, is the one we're trying to predict. We're making predictions based on the independent variable X. The slope of the regression line that shows the impact of X on Y is m. The Y-intercept is a constant called c. If $X = 0$, Y then $Y = c$.

Furthermore, as described below, the linear relationship might be positive or negative in nature.

1. Positive Linear Relationship:

If both the independent and dependent variables rise, the connection is said to be positive.

2. Negative Linear relationship:

If the independent variable rises while the dependent variable falls, the connection is said to be positive**.**

3. Linear regression has the 2 types:

1. Simple linear regression
2. Multiple linear regression

2 Explain the Anscombe's quartet in detail

Answer. Anscombe's quartet consists of four datasets with virtually similar elementary statistical features that, when graphed, appear radically different. There are eleven (x,y) points in each dataset. They were created by statistician Francis Anscombe in 1973 to show the significance of charting data before studying it, as well as the impact of outliers on statistical features.

In his vision, Francis John "Frank" Anscombe, a well-known statistician, discovered four sets of 11 data points and asked the council to plot them as his dying wish. Below are the four sets of 11 data points.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The council then calculated the mean, standard deviation, and correlation between x and y using solely descriptive statistics.

3. What is Pearson's R?

Answer. The Pearson correlation coefficient (PCC) is a measure of linear correlation between two sets of data in statistics. It is also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation. It is the covariance of two variables divided by the product of their standard deviations; thus, it is effectively a normalised measurement of covariance with a value between 1 and 1.

The Pearson's correlation coefficient ranges from -1 to +1, depending on the following factors:

- The data is fully linear with a positive slope if r = 1. (i.e., both variables tend to change in the same direction)
- The data is fully linear with a negative slope if r = -1. (i.e., both variables tend to change in different directions)
- There is no linear relationship if r = 0.
- r > 0 <5 indicates a poor relationship.
- r > 5 <8 denotes a moderate relationship.

- r > 8 denotes a strong correlation.

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

Answer. Normalization is the process of rescaling values into a range of [0,1].
Typically, standardisation entails rescaling data to a mean of 0 and a standard deviation of 1. (unit variance).

Difference between normalized scaling and standardized scaling

Normalized Scaling:
1. Scaling is done using the minimum and maximum value of features.
2. It's utilised when there are multiple scales of characteristics.
3. Values are scaled between [0, 1] and [-1, 1].
4. The n-dimensional data is squished into an n-dimensional unit hypercube using this technique.
5. When we don't know the distribution, it's useful.
6. It's also known as Scaling Normalization.

Standardized Scaling:
1. Scaling is done using the mean and standard deviation.
2. When we wish to secure a zero mean and unit standard deviation, we employ it.
3. It is not restricted to a specific range.
4. Outliers have a considerably less impact on it.
5. When the feature distribution is Normal or Gaussian, it is beneficial.
6. It's also known as Z-Score Normalization.

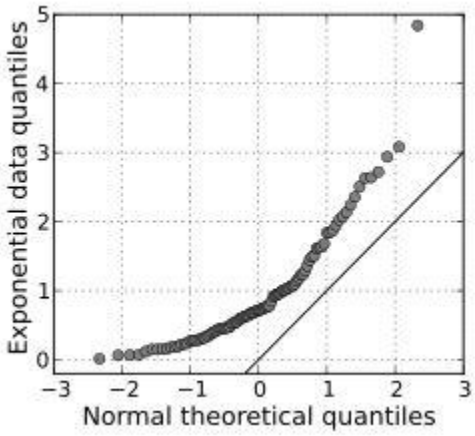
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer. VIF = infinity if there is perfect correlation. This demonstrates that two independent variables have a perfect correlation. We get R2 = 1 in the event of perfect correlation, which leads to 1/(1-R2) infinity. To overcome this issue, we must remove one of the variables that is producing the perfect multicollinearity from the dataset. An infinite VIF value suggests that a linear combination of other variables can exactly express the related variable (which show an infinite VIF as well).

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

Answer. Q-Q charts (Quantile-Quantile plots) are plots that compare two quantiles. A quantile is a percentage of the population in which specific values fall below it. The median, for example, is a quantile where 50% of the data falls below it and 50% of the data falls above it. Q Q plots are used to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45 degree angle is drawn; if the two data sets are from the same distribution, the points will fall on that reference line.

The 45 degree reference line on A Q Q plot:



The points in the Q–Q plot will roughly lie on the line y = x if the two distributions being compared are similar. The points in the Q–Q plot will roughly lie on a line if the distributions are linearly connected, but not necessarily on the line y = x. Q–Q plots can also be used to estimate parameters in a location-scale family of distributions graphically.

A Q–Q plot is used to compare the morphologies of two distributions, offering a graphical representation of how features like location, scale, and skewness differ in the two distributions.