Assignment-based Subjective Questions

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

Answer: Here are some of the inferences:

- The number bookings increased from 2018 to 2019.
- After "Oct" month, it shows decreasing bookings.
- Once we approach the weekend, the number of bookings are comparatively high.
- Clear weather attracts more bookings.
- Also, people book less during holidays since, they might want to spend to do rest during weekends at homes.

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

Answer: It helps in reducing an extra unwanted column during the dummy variablecreation, with the removal of that extra column, it does reduce our effort to manually remove independent correlated variables.

Q. 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 'cnt'.

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

Answer: Based on Normality of Error Terms, Multi-collinearity Check, Homoscedasticity, and Independence of Residuals

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

Answer:

- **Temp**
- **Weather having Light_snowrain**
- **Holiday and winter season**

General Subjective Questions

Q. Explain the linear regression algorithm in detail.

Answer: It's a statistical method used for modeling the relationship between a dependent variable and one or more independent variables.

It assumes that the relationship between the variables can be approximated by a linear function.

Detailed explanation of the linear regression algorithm:

1. Linear regression models the relationship between the independent variables (xxx) and the dependent variable (yyy) as:

$$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\varepsilon y$$

- x1,x2,…,xn are the independent variables.
- y is the dependent variable.
- β0 is the intercept.
- β1,β2,…,βn are the coefficients (slopes) that represent the change in y for a one-unit change in the corresponding x.
- ε is the error term, representing the difference between the observed and predicted values of y.

The objective of linear regression is to find the best-fitting line that minimizes the difference between the predicted values and the actual values of the dependent variable.

2. Cost Function:

The most common method to train a linear regression model is by minimizing the sum of squared errors (SSE) or mean squared error (MSE).

3. Gradient descent:

Its an iterative optimization algorithm used to minimize the cost function in the direction that reduces the cost.

4. Assumptions:

Linear regression assumes:

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: The observations are independent of each other.
- Homoscedasticity: The variance of the errors is constant across all values of the independent variables.
- Normality: The errors follow a normal distribution.

Q. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet comprises four datasets that have identical simple statistical properties,yet appear very different when graphed. Each of the 4 datasets consist of eleven data points (x, y).

The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

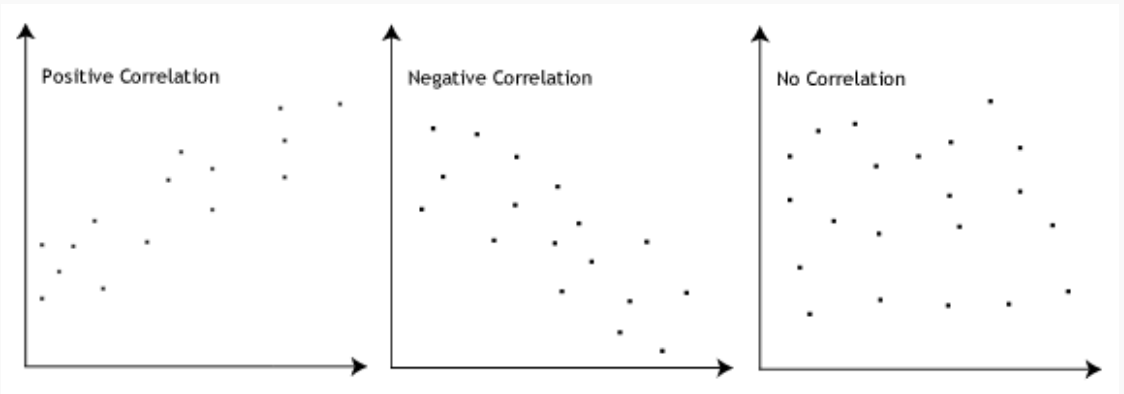
In each panel, the Pearson correlation between the x and y values is the same, r = 0.82 (approx.).In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.

Q. What is Pearson's R?

Answer: Pearson correlation coefficient (PCC), also referred to as **Pearson's R** is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations. Thus, it is a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- R = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- R = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- R = 0 means there is no linear association
- R > 0 < 5 means there is a weak association
- R > 5 < 8 means there is a moderate association
- R > 8 means there is a strong association



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

Answer: Its a step of data pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
The collected dataset contains features that are highly varying in magnitudes, units and ranges. In case the scaling step is skipped, the algorithm only takes magnitude in account and not units which therefore leads to incorrect modelling. To solve this issue, we need to do scaling in order to bring all the variables to the same level of magnitude.

Difference between normalized scaling and standardized scaling:

- Normalized scaling (also known as MinMax scaling) rescales the values into a range of [0, 1] whereas, Standardized scaling rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
- Standardised scaling will affect the values of dummy variables but MinMax scaling will not.

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

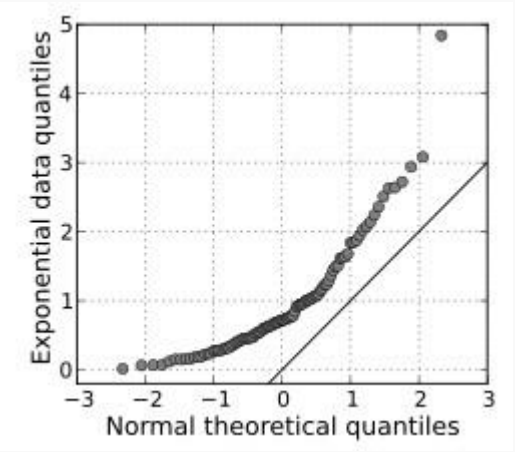
Answer: If there is perfect correlation, then VIF value turns out to be “infinity”. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/(1-R2) which gives the value “infinity”.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which shows an infinite VIF as well).

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

Answer: Its a **Quantile-Quantile** plots which are the plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. A quantile ranges from 0 to 100 percentile. The purpose of Q-Q plot is to figure out if the two datasets come from the same distribution.

Here is a Q-Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.