

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)

Answer –

- 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, july, aug, sep and oct. 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 more number of bookings as compared to the start of the week. When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
 - Booking seemed to be almost equal either on working day or non-working day.
 - 2019 attracted more 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? (2 mark)

Answer - using drop_first=True during dummy variable creation is important to avoid multicollinearity in the dataset, especially when using these dummy variables in regression models. When you create dummy variables for a categorical feature with n unique categories, you typically create n dummy variables. However, including all n dummy variables in the model can lead to multicollinearity because one of the dummy variables can be perfectly predicted from the others..

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

Answer - **Temperature (temp)**

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

Answer –

- Linearity: Check the linearity assumption by plotting the actual vs. predicted values. The relationship should be linear.
- Normality of Residuals: Plot a histogram or a Q-Q plot of the residuals to check if they are normally distributed. You can also use statistical tests like the Shapiro-Wilk test for normality.

- **Homoscedasticity:** Plot the residuals vs. predicted values (or fitted values). The plot should show no clear pattern (i.e., the variance of residuals should be constant across all levels of predicted values).
 - **Independence of Residuals:** Check for autocorrelation in the residuals using a Durbin-Watson test or a plot of residuals vs. time (for time series data).
 - **Multicollinearity:** Check for multicollinearity among the independent variables using variance inflation factor (VIF) values. VIF values greater than 10 indicate multicollinearity.
 - **Outliers and Leverage Points:** Identify outliers and leverage points by plotting standardized residuals vs. leverage. Points that are both outliers and have high leverage may need further investigation.
 - **Model Fit:** Evaluate the overall fit of the model using metrics like R-squared, adjusted R-squared, and F-statistic.
 - **Cross-validation:** Use cross-validation techniques (e.g., k-fold cross-validation) to assess the model's performance on unseen data.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer -

1. **Temperature**
2. **weathersit_3**
3. **Year (yr)**

General Subjective Questions

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

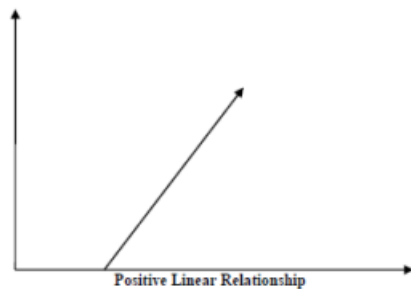
Answer- Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease),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

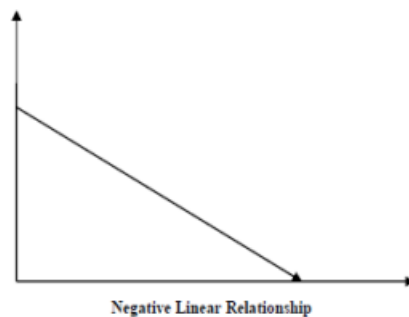
Positive Linear Relationship:

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 negative if independent increases and dependent variable decreases. It can be understood with the help of following graph –



2. Explain the Anscombe's quartet in detail. (3 marks)

Answer - Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their

similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

3. What is Pearson's R? (3 marks)

Answer - Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases

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

Answer -

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. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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

Answer –

If there is perfect correlation, then $VIF = \infty$. 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 $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

(3 marks)

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

(3 marks)

Answer -

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 reference line. The greater the departure from this reference line, the greater the evidence 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.