Linear Regression Subjective Questions

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

Answer - Categorical Variable I have amazed with boxplot and here are the analysis against each categorical variable:

- Season -> Fall have highest count booking wherein booking in spring season have low
- Year -> Compare to previous year 2019 bikes demand are high
- Month -> Sep and Oct have highest booking compare to other month
- Holiday -> Considering 0 is non holiday booking bike booking are high on non holiday
- Weekday -> As per the box plot almost booking is same for all day
- Weathersit -> Booking are high on clear and mist weather compare to Light Snow

After Analyzing all the variable we booking is high into the mist and fall season which is from May to Oct month

Q.2- Why is it important to use drop_first = True du as its difficult to analyze all variable for the model and this can give us the ring dummy variable creation?

Answer – drop_first are used when we create dummy variable for categorical columns which is having more than two features. It is important because its reduce the extra columns as its difficult to analyze all variable for the model and this can give us the wrong estimation

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

Answer – temp and atemp have higher correlation with the target variable

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

Answer – By doing the residual analysis we have identified the fitted model and checked it should normally distributed and scatter point should be approx. 0.

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

Answer – After building the model we have analyzed the given top 3 features have significant demands of the shared bikes

1. temp
2. mnth_sep
3. Winter season

General Subjective Questions

Q.1- Explain the linear regression algorithm in detail?

Answer – Linear regression algorithm is a technique that help to predict the value based on the dependent variable. Its help to build the statistical model between dependent variable

There are 2 types of linear regression model.

1. Simple Linear regression model – it is used for one dependent and one independent variable (**y = mx + c**)
2. Multiple linear regression model – it is used for various dependent variable (**y = m1x1 + m2x2 + m3x3 + …..+ mnxn +c**)

Following are the step used to build the liner regression model:

Step 1- Data Loading/Reading, Understanding and visualizing the Data(EDA)

Step 2- Prepare Data for Modeling (train-test split and rescaling)

Step 3- Building the model

Step 4- Residual Analysis

Step 5- Prediction and evaluation on the test set

Q.2- Explain the Anscombe's quartet in detail.

Answer Anscombe's are designed to highlight the importance of visualizing data and the dangers of solely relying on summary statistics. the datasets have very different underlying patterns, which demonstrates the limitations of only using basic statistical measures to understand data. It help in detecting outliers, nonlinear relationships, and other nuances that affect model selection and interpretation.

It contains 4 different datasets with different patterns

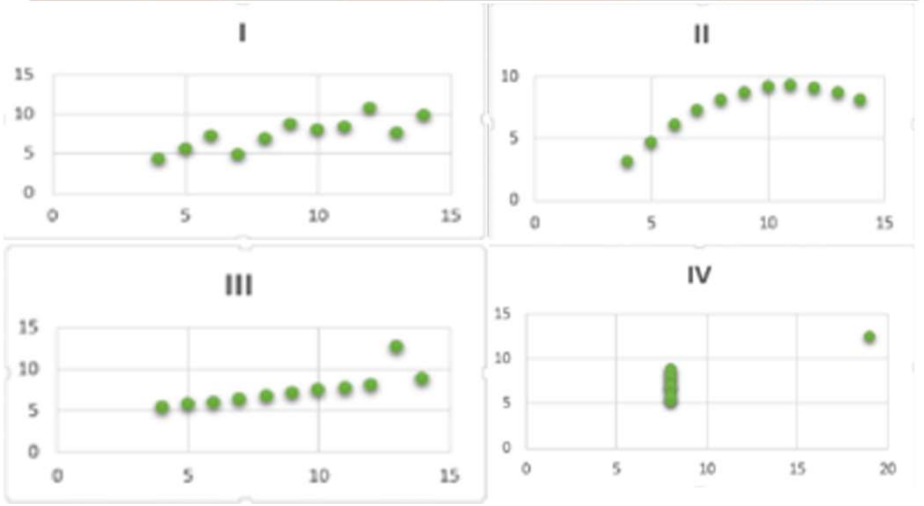
1. **(I) Linear Relationship –** This data set shows clear relation with X and Y where y increases X increases

2 . **(II) Non-Linear Relationship –** This data set have non-linear relationship with X and Y

3. **(III) Outlier Influence –** This is mostly follow the linear relationship but it contain one outlier that impact the linear regression line

4. **(IV) Almost Identical Statistics, Non-linear Relationship –** This dataset contain 3 group of data points that exhibit the non-linear relationship

	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
Mean	9	7.500909	9	7.500909	9	7.5	9	7.500909
Variance	11	4.127269	11	4.127629	11	4.1226	11	4.123249
Correlation	0.816421		0.816237		0.816287		0.816521	



Q.3- What is Pearson's R?.

Answer Pearson's R is is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges between -1 and 1.

- A correlation coefficient of +1 indicates a perfect positive linear relationship, means if the one variable increases, the other variable also increases.
- A correlation coefficient of -1 indicates a perfect negative linear relationship, means if the one variable increases, the other variable decreases.
- A correlation coefficient close to 0 indicates little to no linear relationship between the variables.

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

Answer Scaling is performed to to standardize the range of features. It involves the transforming the data so that it fits within a specific range. Scaling is particularly important when dealing with algorithms that are sensitive to the magnitude of the input variables, such as distance-based algorithms or those that use gradient descent for optimization.

MinMax Scaling (Normalize Scaling) - It scale the features between range 0 and 1. its effected the outliers.

Standardized Scaling - it transforms features to have a mean of 0 and a standard deviation of 1. its is less effected the

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

Answer This happened when one of the predictor variables is perfectly collinear with a combination of other predictor variables in the model. One predictor can be exactly predicted by a linear combination of the others, making it impossible to compute the VIF. It is indication that there is a severe multicollinearity problem in model that needs to be addressed before proceeding with further analysis.

If we create dummy variables from categorical variables and include all the dummy variables without dropping one reference category, it can be leading to perfect collinearity and infinite VIF.

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

Answer Quantile to Quantile plot is the powerful tools for assessing the distributional similarity of data to a theoretical distribution, especially the normal distribution. In linear regression, Q-Q plots are used to check the assumption of normally distributed residuals, which is crucial for the validity and interpretation of regression models.

It helps to visually determine whether the data follows the expected distribution or deviates from it.