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?

Ans: From the dataset given there were numerical values and we had to create categorical and dummy values. Later, after creation of model we found out that these variables are influencing the dependent variable.

Season

- fall season has the most number of bookings done, followed by summer & winter
- spring has the lowest no of bookings

Month

- most number of bookings were done in the month from august to October
- least number of bookings were done in the month from november to February
- this shows that month can be a good parameter for booking

Weather

- clear days have the most bookings
- light rain has the least
- there is no data of heavy rain

Weekday

- most bookings are done in wednesday, saturday and least on Tuesday
- though there is not much clear pattern of this on the count variable

Holiday

- More bookings were done on holidays

Working Day

- Slightly more bookings were done on non-working day than working day

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

Ans: drop_first=True creates one column lesser than the number of categorical values. This reduces the correlation among dummy variables also. For e.g. We have 3 values for fruits (A,B,C) -> This doesn't require 3 dummy variables, it can be inferred from two columns also. Like if A =0 &B=0 then obviously its C=1. This is achieved by drop_first=True.

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

Ans: Both temp & atemp has the highest correlation with target variable.

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

Ans:

i. Linear Relationship : There is a linear relationship between dependent & independent variable.
ii. No multicollinearity : P values and VIF are in acceptable ranges
iii. Normal distribution of error term.
iv. Homoscedasticity - no visible pattern in residual values.

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

Ans:

i. Temp: temperature has a significant impact on bookings
ii. Light_rain: snow & rain are negatively impacting the bookings.
iii. Yr: the bookings of bike have increased over year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression algorithm is a method to find the best relationship between a dependent variable and independent variable. Technically this tells us about the coefficients of the independent variables and using that we can predict things like sales, cost, number of bookings, footfall etc. It uses the R squared values to find that. Mathematically the relationship can be represented with by equation:

$Y = mX + c$

- 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.

Linear regression is of two types:
- Simple Linear Regression
- Multiple Linear Regression

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.

• Auto-correlation – There is very little or no autocorrelation in the data.

• Relationship between variables – There exists linear relationship between dependent and independent variables.

• Normality of error terms – Error terms should be normally distributed

• Homoscedasticity – There should be no visible pattern in residual values.

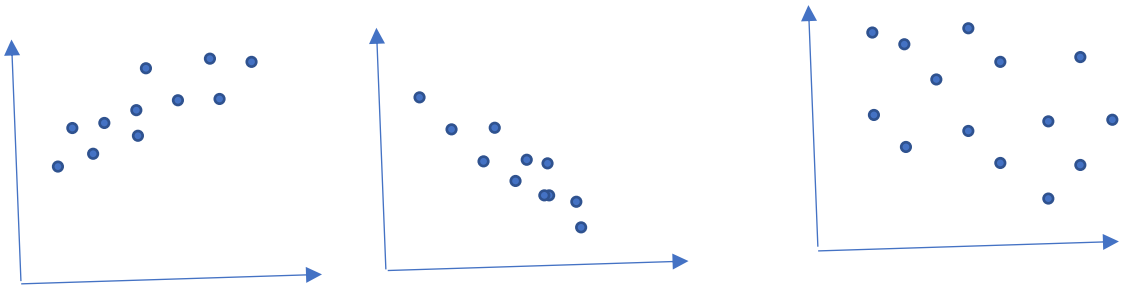
2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four dataset, with similar properties in terms of means, correlations, R-Squared, VIF and linear regression lines but on plotting them on a scatter graph it gives different representation. It is used to know the importance of EDA (exploratory data analysis) and what is the drawback of depending only on summary statistics. It also emphasizes on using data visualization to find out outliers, trends and other details which we might not get from summary statistics.

3. What is Pearson's R?

Ans: Pearson's R is a coefficient which measures how two variables are correlated. The values lie between 1 and -1.

- -1 means a strong negative relationship.
- 1 means strong positive relationship.
- 0 means no relationship.



Positive Correlation | Negative Correlation | No Correlation

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

Ans: Scaling is a technique to standardize independent features of the dataset. It actually brings different units of values on the same range.

If we are not performing scaling, then the model may take some higher values as more significant one and it will not give the perfect linear regression expression. Scaling brings the values on same range.

In normalized scaling, lowest value of the variable is mapped to 0 and the highest to 1. Hence, the values are in the range (0,1). In standardized scaling, the data is not mapped into any range rather it is transformed to have a mean of 0 and a standard deviation of 1. Also, normalized is affected by outliers while standardized is not that affected by outliers.

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

Ans: VIF stands for Variance Inflation Factor. It is used to check the level of multicollinearity in the dataset. It is calculated by:

$$VIF = 1/ (1-R^2)$$

High VIF means there is a correlation between the variables.

When the value of VIF is infinite it shows a perfect correlation between two independent variables, and 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.

Ans: The plot obtained when the quantiles of two variables are plotted against each other, then it is called as quantile – quantile plot or Q-Q plot.

Use: Since there is a comparison among the data points of two variables, it should fall on the reference line. More the deviation from the reference line, we can say that there exists different distribution in the dataset.

Importance: This helps to understand a) if there is an overlap on the common distribution b) if there is difference then why is it so