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 : We did the analysis based on the following variables

1. Season: People prefer to take bikes mostly when it Fall/Summer/Winter and not in spring.
2. Year: There was a growth in the number of people opting for the bikes in 2019 as compared to 2018
3. Month: Distribution of months is contrary to the season distribution. While we see winter having higher percentiles same is not replicated in the months of Dec-Feb
4. Holiday: Median of population opting for bikes on non-holidays is more than that of holidays
5. Weekday : Medians for all the days are more or less the same
6. Working Day : This variable counters our data based on holiday.
7. Weather Situation: In case of Heavy rains or light snow we see a decrease in the number of people taking bikes. People are more likely to take bikes when there is a clear sky or light mist

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

Ans: Drop first allows us to exclude 1 of the values of categorical columns. This help us reduce unnecessary columns since for a column with n unique values we only need n-1 encoded columns. A row with all n-1 columns as 0 will mean that the value is same as the droped column

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

Ans : The predictor variable 'atemp' and ‘temp’ are highly correlated with our target variable 'cnt'.

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

Ans There are four assumptions associated with a linear regression model:

1. Linearity: based on Pairplot we saw the linearity of the `numerical variables` with `cnt`
2. Mean of Residual: -3.3665e-16 which is very close to 0
3. Homoscedasticity: The variance of residuals is not dependent on the variable X_train_ml due to which we can reject heteroscedasticity
4. Normality: The residual terms are normally distributed, but not perfectly a normal curve due to lower sample size.
5. No Perfect Multicollinearity based on the heatmap for the X_train input training variables vs cnt

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. Windspeed
2. workingday_Yes
3. season_spring

General Subjective Questions

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

Ans: Linear regression may be defined as the statistical model that analyzes 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).

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

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

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

Ans : Scaling is the act of squeezing data point between a limit. It is performed in linear regression so as to squeeze columns having different ranges of values between a fixed value range. standardized scaling distributes values normally using mean and standard deviation while normalized scaling uses minima and maxima of data

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

Ans : VIF becomes infinity when there is perfect correlation. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.