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?

Answer: The categorical variables does have an effect on the dependent variables. We see that some of the categories under the columns like “year”, “weathersit” and “season” does have an impact on the dependent variable. For example, we see that the bookings increase during the fall season (season column) and when the weather is clear (weathersit) and with every passing year (year column).

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

Answer: The main reason to use drop_first=True is to avoid the situation of multi-collinearity. We see that whenever we create dummy variables, we will end with one column which basically carries the same information as the other n-1 columns i.e this column is highly correlated with the rest of the columns. Removing this column is very important as retaining this column would make this column dependent on other columns, which is against our assumption for linear regression model that predictor variables are independent.

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

Answer: If we remove the registered column, then the column that is most correlated with target column is **atemp.**

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

Answer: There are 4 major assumptions. Below are those assumptions and the ways to validate them:

- Linear relationship between X and Y: Validated through the pairplot of independent variables and the target variable.
- Error terms are normally distributed: Validated by plotting the distplot of error terms. If they are normally distributed, like the one in our model case, then this assumption is validated
- Error terms are independent of each other: Validated by plotting a scatter plot of the error terms or the residual terms. If there is no obvious pattern in the error terms and are randomly scattered then they are independent.
- Error terms have constant variance (homoscedasticity): Validated by plotting a scatter plot of the error terms or the residual terms. If the error terms are equally distributed around the zero line, then homoscedasticity assumption is validated

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: The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Atemp
2. Year
3. weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is one of the simplest type of models. It tries to model the relationship between the continuous predictor variables and the continuous response variable. This model assumes that there exists a linear relationship between the predictor variable and the response variable.

The model works on the principle of obtaining a line that best fits the data i.e when we plot predictor variable and the response variable, there exists a line which passes through all the data points with zero or least possible error. Here the error refers to the gap between the predicted value (the line) and the actual value. This error is also termed as residual.

Since there is a linear relationship between the predictor and the response variable, the equation of the best-fit line takes the form: y = beta0 + beta1*x1+ beta2*x2. Where x1 and x2 are the predictor variables, y is the response variable, beta1 and beta2 are the co-efficient of x1 and x2 respectively, and beta0 is the Y-Intercept.

2. Explain the Anscombe's quartet in detail.

Answer: It refers to the 4 datasets which, though statistically resemble each other, but when plotted yield different pattern. When we say statistically similar value, we mean their summary statistics like the mean, Standard deviation etc. Typically, we assume that if the dataset has similar summary statistics then their nature is also same. But that has been disproved using the Anscombe's quartet where in though they are statistically same, but when plotted, they yield different distribution. Anscombe quartet is always a great reminder of the fact that one shouldn't just blindly look at the summary statistics but should also visualize the data before drawing the final conclusion

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

Answer: Pearson's R or Pearson's Correlation Coefficient gives us a measure of the correlation between two variables. Two variables are said to be corelated when change in one of the variables either positively or negatively influences the other variable. The pearson's R can take value between -1 and +1. If the value is -1 it means that the variables are negatively corelated i.e increase in one variable results in decrease of other variable by equal amounts. If the value is +1 it means that the variables are positively corelated i.e increase in one variable results in increase of other variable by equal amounts.

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

Answer : Scaling is a data pre-processing technique wherein we try to normalize the range of the data of either an independent or a dependent variable.

One of the main reasons for why we perform scaling is for the fact that there are some of the variables which has a higher range when compared to the rest of the other variables during training a Machine Learning model. This higher range variable might influence the model to a larger extent than the smaller range variable. For eg, in context of evaluating a car's performance, if one variable, for eg cars weight, is measured in gms and other variable is measured in kms, then the weight variable will have higher values as compared to kms. This would result in the final model having larger co-efficient for weight as compared in kms. Such large variance in co-efficient (beta values) would lean the model towards higher beta and hence more biased.

Normalized scaling is a scaling technique wherein we will compress the range of the data so that it lies between 0 and 1. Standardized scaling is one wherein we try to compress the range of the data using the mean and the standard deviation. Doing so, the range of the resultant data is not defined unlike Normalized data where it is bound between 0 and 1. Another difference is, normalization will result in losing the shape of the data but standardization will preserve the shape of the data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: VIF value can be infinite when the R-Squared value is 1 i.e whenever there is a perfect correlation between the variable of concern and the rest of the other variables. When r-squared value is 1, the denominator of VIF is 0 and hence 1/(1-r-squared) is infinite. This implies that the target variable for whom the VIF is infinite is completely redundant, and dependent on other independent variables. This would naturally break our linear regression assumption.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: Q-Q plot is a tool that enables us understand if two datasets come from populations with same underlying distributions. A Q-Q plot is a plot of quantiles of the first dataset against the quantile of the second dataset. If the points lie on the line which is at an angle of 45 degree originating from origin, then the two datasets are from the same distribution i.e either normal or exponential. If they are scattered around, above or below, then the datasets belong to 2 different distributions.

One practical application is to validate whether the given dataset has the distribution similar to that of a theoretical distribution. Its also useful to determine the residuals follow a normal distribution. Q-Q plots can also be used to determine the skewness of the plot. If the left-side of the plot deviates from the straight line then its left skewed, and if the right side of the plot deviates from the straight line, then its right skewed.