Assignment-based Subjective Questions

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

Ans : We can infer below observation:

- Over 5K booking is happening on the Season3, Season2 and Season 4 whereas we observe
- that less than 3.5k booking is happening on season 1
- Over 4k boking happening on the range between 4 to 10.
- Over 4k booking is only happening in weathers it 1.
- most of the Bike booking happening when there is a working day
- Weekday is independent, All the days have marginally same count.

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

Ans : drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

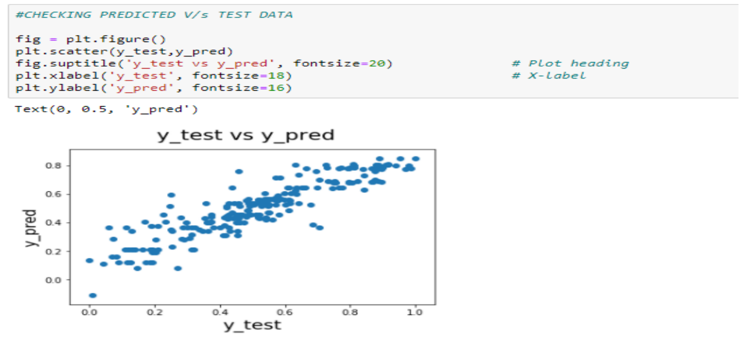
Question 3 : Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans : Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable Answer: atemp has heighest positive correlation.

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

Ans :

- Residuals are normally distributed
- There is No Multicollinearity between the predictor variables
- There is a linear relationship between temp, atemp and cnt



```
#CHECKING PREDICTED V/s TEST DATA

fig = plt.figure()
plt.scatter(y_test,y_pred)
fig.suptitle('y_test vs y_pred', fontsize=20)              # Plot heading
plt.xlabel('y_test', fontsize=18)                          # X-Label
plt.ylabel('y_pred', fontsize=16)
```

```
Text(0, 0.5, 'y_pred')
```

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

Ans : Light rain_Light snow_Thunderstorm, yr, Spring

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

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).

Mathematically the relationship can be represented with the help of following equation −

Y = mX + b

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b.

Question 2: Explain the Anscombe's quartet in detail.

Ans : Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Question 3: What is Pearson's R?

Ans : In statistics, the Pearson correlation coefficient — also known as Pearson's r, the Pearson product-moment correlation coefficient, the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data.

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

Ans: what? Scaling is data transformation technique, it simply means putting every feature value on a common scale value.

why? Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

difference? :

Normalization: It brings all of the data in the range of 0 and 1.

Standardization: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Ans : If there is perfect correlation, then VIF = infinity. 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. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.