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: There were 6 categorical variables in the dataset.
We used Box plot (refer the fig above) to study their effect on the dependent variable ('cnt') .
The inference that We could derive were:
season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median
of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a
good predictor for the dependent variable.
weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median
of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30%
of total booking. This indicates, weathersit does show some trend towards the bike bookings
can be a good predictor for the dependent variable.
holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which
means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the
dependent variable.
weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on
all days of the week) having their independent medians between 4000 to 5000 bookings. This
variable can have some or no influence towards the predictor. I will let the model decide if this
needs to be added or not.
workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

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.
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Ans:There is linear relationship between temp and atemp. Both of the parameters cannot be used in the model due to multicollinearity. We will decide which parameters to keep based on VIF and p-value w.r.t other variables

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

Ans: VERY LOW Multicollinearity between the predictors and the p-values for all the predictors
seems to be significant. For now, we will consider this as our final model (unless the Test data metrics are not significantly close to this number).
The Coefficient values from the model of all the variables are not equal to zero which means we are able to reject Null Hypothesis
F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant the Model is.
F-statistic: 233.8
Prob (F-statistic): 3.77e-181
The F-Statistics value of 233 (which is greater than 1) and the p-value of '~0.0000' states that the overall model is significant.
The Residuals were normally distributed after plotting the histogram . Hence our assumption for Linear Regression is valid
VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5

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

Ans - As per our final Model, the top 3 predictor variables that influences the bike booking are:
Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp
variable increases the bike hire numbers by 0.5636 units.
Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070
units.
Year (yr) - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases
the bike hire numbers by 0.2308 units.
So, it's suggested to consider these variables utmost importance while planning, to achive
maximum Booking
The next best features that can also be considered are
season_4: - A coefficient value of ‘0.128744’ indicated that w.r.t season_1, a unit increase in
season_4 variable increases the bike hire numbers by 0.128744 units.
windspeed: - A coefficient value of ‘-0.155191’ indicated that, a unit increase in windspeed
variable decreases the bike hire numbers by 0.155191 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

2. Explain the Anscombe’s quartet in detail.

Ans: They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

3. What is Pearson's R?

Ans: Pearson correlation coefficient ,measures the strength and direction of the relationship between two variables.

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

Ans: Scaling is step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling: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**).

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 leads 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).

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.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.