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?**

- Fall season has highest demand for renting bike, before building the model.
- After building the model spring season has the highest demand for shared bikes.
- The demand of bike sharing growing each month till June. September has the highest demand. After September we can observe the decreasing pattern on demand on the year end. This may be due to the weather condition.
- There is no informative insight from weekday.
- There is high demand when weathersit is good.i.e,Clear, Few clouds, Partly cloudy, Partly cloudy.

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

- It facilitates in lowering the more column created throughout dummy variable creation. Hence it reduces the correlations created amongst dummy variables.
- For example, lets consider 3 variables furnished, semi-furnished and unfurnished. When we create dummy variables and consider if it is not furnished and not semi-furnished as unfurnished. We can drop unfurnished column.

Example before dropping column dummy variable

unfurnished	Furnished	Semi- Furnished
0	1	0
0	0	1
1	0	0

Example After dropping column dummy variable

Furnished	Semi- Furnished
1	0
0	1
0	0

We can consider 00 as Unfurnished.

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

- temp and atemp the predictor variables showing linear relationship with the dependent variable cnt has the highest correlation with value 0.63.
- temp and atemp are highly co-related with each other with correlation value of 0.99.

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

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X. An additive relationship suggests that the effect of X on Y is independent of other variables.
2. The error terms must be normally distributed.
3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
5. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.

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

1. Temp- Temperature
2. Season_spring – On spring season
3. Yr- Year

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression Algorithm may be a machine learning algorithm supported supervised learning.

Here we are getting to specialise in rectilinear regression. Linear regression may be a part of multivariate analysis. Multivariate analysis may be a technique of predictive modelling that helps you to seek out the connection between Input and therefore the target variable.

Linear regression is one among the very basic sorts of machine learning where we train a model to predict the behaviour of your data supported some variables. within the case of rectilinear regression as you'll see the name suggests linear meaning the 2 variables which are on the x-axis and y-axis should be linearly correlated.

For example let's say you're running a advertisement and expecting a particular number of count of consumers to be increased now what you'll do is you'll look the previous promotions and plot if over on the chart once you run it then attempt to see whether there's an increment into the amount of consumers whenever you rate the promotions and with the assistance of the previous historical data you are trying to work it out otherwise you attempt to estimate what is going to be the count or what is going to be the estimated count for my current promotion this may offer you a thought to try the design during a far better way about what percentage numbers of stalls maybe you would like or what percentage increase number of employees you would like to serve the customer. Here the thought is to estimate the longer-term value supported the historical data by learning the behaviour or patterns from the historical data.

Linear regression is employed to predict a quantitative response Y from the variable X.

Linear regression mathematically- $y = mx + c$

Where,

m - Slope of the line

c – y-intercept of the line

x – independent/predictive variable

y – Dependent/target variable

2. **Explain the Anscombe’s quartet in detail.**

Anscombe's quartet consists of four data sets with virtually equal basic descriptive statistics but vastly distinct distributions and graphing appearances.

On other words, Anscombe's Quartet is a collection of four data sets that are essentially equal in terms of simple descriptive statistics, but have certain quirks in the dataset that trick the regression model if formed. When displayed on scatter plots, they have extremely distinct distributions and appear differently.

The 4 datasets mentioned are:

1. Dataset 1: this fits the linear regression model pretty well.

2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Before applying any machine learning method to the dataset, all of the relevant characteristics must be shown so that a good fit model can be created.

3. **What is Pearson’s R?**

The Pearson correlation coefficient is a measure of linear correlation between two sets of data in statistics. It is also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or simply the correlation coefficient. It is the product of two variables' covariances and their standard deviations; consequently, it is simply a normalised measurement of covariance, with the result always falling between 1 and 1. The measure, like covariance, may only describe linear correlations of variables and ignores numerous other forms of relationships or correlations. As an example, the age and height of a high school sample would be expected to have a Pearson correlation coefficient considerably more than 0, but less than 1. (as 1 would represent an unrealistically perfect correlation).

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

It is a data Pre-Processing step that is used to independent variables in order to normalise the data within a specific range. It also aids in the speeding up of algorithm calculations.

Most of the time, the acquired data set comprises characteristics with widely disparate magnitudes, units, and ranges. If scaling is not performed, the method simply considers magnitude rather than units, resulting in erroneous modelling. To address this problem, we must scale all of the variables to the same magnitude level.

Normalized scaling: It gathers all of the data between 0 and 1. sklearn.preprocessing MinMaxScaler aids in the implementation of normalisation in Python. Its is also known as MinMax scaling.

MinMax scaling= (x-min(x)) /(max(x)-min(x))

Standardization Scaling: Values are replaced by their Z scores after standardisation. It transforms the data into a conventional normal distribution with a mean of 0 and a standard deviation of 1.

Standardization= (x-mean(x)) / sd(x)

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

VIF = infinite, if there is perfect correlation. This demonstrates that there is a perfect correlation between two independent variables. In the event of perfect correlation, R2 = 1, resulting in 1/(1-R2) infinite. To resolve this issue, we must remove one of the variables from the dataset that is producing the perfect multicollinearity.

A VIF value of infinity implies that the associated variable may be stated perfectly 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.**

Q-Q Plots (Quantile-Quantile Plots) are comparisons of two quantiles. A quantile may be a fraction of values that fall below that quantile. for instance, the median may be a quantile where 50% of the info falls below it and 50% fall above it. The goal of Q-Q plots is to work out whether two sets of knowledge are from an equivalent distribution. On the Q-Q plot, a 45-degree angle is drawn; if the 2 data sets are from an equivalent distribution, the dots will fall thereon reference line.

If the two distributions being compared are comparable, the points in the Q–Q plot will roughly correspond to the route y = x. If the distributions are linearly connected, the points in the Q–Q plot will be close to, but not necessarily on, the path y = x. Q–Q plots may also be used to estimate parameters in a location-scale family of distributions graphically.

A Q–Q plot is used to match the shapes of distributions, offering a graphical representation of how features such as location, scale, and skewness differ between the two distributions.