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?**

   Here are some of the inferences on the analysis of the categorical variables and their effect on the dependent variable.
   1. The season of Fall has the highest median followed by summer as they have the best weather conditions.
   2. The median bike rentals have increased in the year 2019 compared to the year 2018. This may be due to the people getting conscious about the environment.
   3. The bike rentals are more on non-holiday days compared to holiday. This indicates that people prefer to spend time at home during the holidays.
   4. The months of Fall - June to October have a higher median value.
   5. The overall median for the weekdays and working-days are the same.
   6. The Clear weather situation has the highest median while the weather situation of Light snow has the least. The count of bike sharing is Zero for the weather situation - 4 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog'.

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

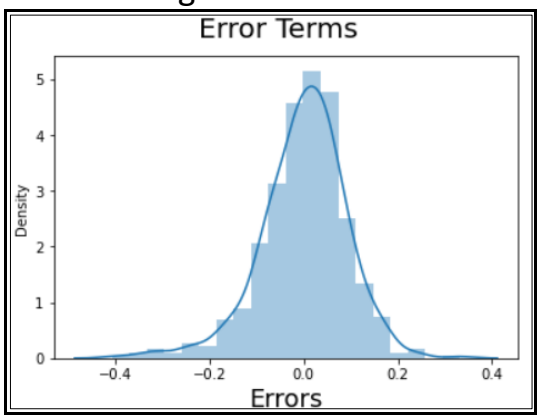
   It is important to use drop_first=True as it helps in reducing the extra column created during dummy variable creation. It helps to reduce the correlations created among dummy variables.
   Example: Let's say we have 3 types of values in a categorical column and we want to create dummy variable for that column. If one variable is not furnished and not semi-furnished, then it is obvious that it is unfurnished. So, we do not need a 3rd variable to identify the category of 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?**

   The numerical variable 'atemp' has the highest correlation with the target variable 'cnt' with a value of '0.65' followed by 'temp' with a value of '0.64'.

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

We validate the assumptions of the Linear Regression by plotting a distplot of the residuals and analysing it to see if it is a normal distribution or not and if it has a mean = 0. The diagram below shows that it is normally distributed with mean = 0.



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

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

1. temp(Temperature) - A coefficient value of '0.5480' indicates that a unit increase in temp variable, increases the bike hire numbers by 0.5480 units.
2. Light Snow(weathersit) - A coefficient value of '-0.2838' indicates that, a unit increase of this variable, decreases the bike hire numbers by -0.2838 units.
3. Yr(Year) - A coefficient value of '0.2328' indicates that, a unit increase of this variable, increase the bike hire numbers by 0.2328 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables. Liner Regression finds how the value of the dependent variable changes according to the value of the independent variable. The linear regression model provides a sloped straight line representing the relationship between the variables. It is based on the equation “y=mx+c”

Use Cases of Linear Regression are Prediction of trends and Sales Targets, Price Prediction, Risk Management, etc.

Linear regression can be further divided into two types of the algorithm - Simple Linear Regression and Multiple Linear Regression.

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

Following steps are performed while doing linear regression:

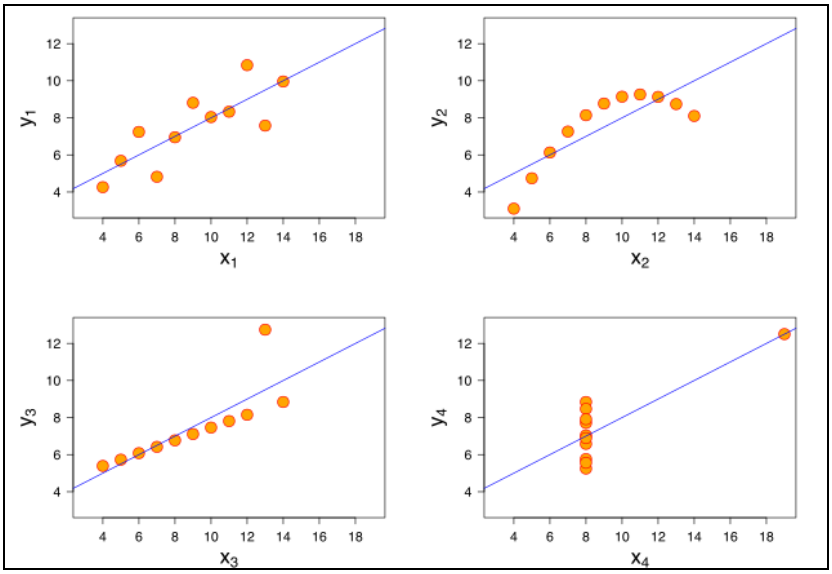
1. The dataset is divided into test and training data.
2. Train data is divided into features(independent) and target (dependent) datasets.
3. A linear model is fitted using the training dataset. Internally the api’s from python uses gradient descent algorithm to find the coefficients of the best fit line. The gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares.
4. In case of multiple features, the predicted variable is a hyperplane instead of line. The predicted variable takes the following form:

$$Y= \beta 0+\beta 1x1+\beta 2x2+\beta 3x3+\cdots+ \beta nxn$$

5. The predicted variable is than compared with test data and assumptions are checked.
6. Some of the important assumptions of Linear Regression are Linear relationship between the features and target, Small or no multicollinearity between the features, Homoscedasticity, Normal distribution of error terms.

2. Explain the Anscombe's quartet in detail.

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. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



The four datasets can be described as:

1. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
2. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
3. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
4. Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Pearson's R measures the strength of association of two variables. It is the covariance of two variables divided by the product of their standard deviation. It has a value from +1 to -1.

- A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase
- A value of 0 means no correlation
- A value of -1 means a total negative correlation. It means that if one variable increase then other will decrease

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

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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.

Normalised Scaling	**Standardized scaling**
Called min max scaling, scales the variable such that the range is 0-1.	Values are centred around mean with a unit standard deviation.
Good for non-gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers

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

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

$VIF_i = \frac{1}{1-R_i^2}$ Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator becomes 0 and the overall value become infinite. It denotes perfect correlation between the variables.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. The power of Q-Q plots lies in their ability to summarize any distribution visually.

Q-Q plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression.
- Skewness of distribution