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?**

Following are the analysis on the categorical variable, taking count as the dependent variable:

* People usually rent bikes more on non-holidays compared to holidays. The reasons for this can be people not commuting to work on weekends and using their personal vehicles as traffic not being the issue.
* In the year 2019, the median bike rents have increased. This maybe due to the increasing popularity of the bikes as well as increase in fuel prices
* The clarity of weather plays a major part when it comes to renting the bikes. Clearer the weather, more feasible the bikes are to travel and use
* In the month of spring, the renting of bikes is comparatively very less to other seasons
* Fall has the highest median, as the weather conditions are the most optimal to ride the bike
* September is the month where the use of bikes is the highest.
* Overall spread in the month plot is reflection of season plot as fall months have higher median.

1. **Why is it important to use drop\_first=True during dummy variable creation?**

Drop\_first=True is very important to use as mainly it helps to reduce the extra column created during dummy variable creation. A variable with n levels can be represented by n-1 dummy variables. Even after removing the first column, the data can be represented

Also, reducing the extra column during the variable creation helps in reducing the correlation between the dummy variables

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

From the above pair plot we could observe that, **temp** has highest positive correlation with target variable ‘cnt’.

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

* One way to validate the assumption is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.
* Checked if the independent are not correlated. Used a scatter plot to visualize the correlation. Checked the VIF, if It was VIF>=10. As it was less, there was no multicollinearity.
* Created a scatter plot that shows residual vs fitted value to check Homoscedasticity. The data points spread across equally, it means the residuals have constant variance (homoscedasticity). Otherwise, it means the residuals are not distributed equally and depicts a non-constant variance
* Looking at the plot plotted above, we can see that the residual errors don’t always have a mean value of 0.
* Used a Q-Q plot to see if the error terms have normal distribution

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

The top three features contribution significantly towards explaining the demand of shared bikes are:

* Temperature
* Year -2019
* Weather

General Subjective Questions

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

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called **simple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.

Linear regressions can be used in business **to evaluate trends and make estimates or forecasts**. For example, if a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months

To calculate best-fit line linear regression uses a traditional slope-intercept form, we use *:***y= mx + c 🡪 y=ao+a1x**

where

* y is the Dependent Variable
* x is the Independent Variable
* ao is the intercept of the line
* a1 is the Linear Regression Coefficient
* c is the constant

Linear Regression can be performed using various programs and environments. Those are:

* R linear regression
* MATLAB linear regression
* Sklearn linear regression
* Linear regression Python
* Excel linear regression

Using the above, we build a linear regression model which helps in predictive analysis to evaluate trends and make estimates or forecasts.

Basics Steps for building a Linear Regression Model are:

1] Data Understanding and Data Loading

2] Pre-Processing an EDA

3] Splitting into Train/Test

4] Missing value imputation and Scaling

5] Feature Selection

6] Modelling

7] Evaluation

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

Anscombe’s quartet are four data sets which have similar summary statistics. But upon visualization the data, we can see that the plotted graphs are not similar at all. Anscombe’s quartet is a prime example as to why visualizing and plotting the data is necessary and relying on the summary statistics is not always accurate/advisable

**It also 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.

Tools like matplotlib in python and ggplot in R are great way to visualize the data. They help in picking visual patterns, trends and outliers while analyzing the data

Let’s say there are different data sets which might be well modeled with a line, and the regression parameters will be meaningful. The same is true of this data set, but the outlier throws off the slope and intercept. After doing EDA, we can look for the

patterns and outliers

1. **What is Pearson’s R?**

Persons Correlation Coefficient or Pearson’s R or bivariate correlation is a type of correlation that is used to measure the relationship between two continuous variables

Correlation measures the relationship between two variables in the finance and the investment industries. It shows the strength of the relationship between the two variables as well as the direction and is represented numerically by the correlation coefficient. The values lie between **-1.0 and +1.0**

**Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Calculating the correlation coefficient R :

We can calculate a linear relationship b/w two given variables. Now, for us to calculate the same, we have certain requirements

* The association b/w the variables should be linear
* Scale of measurement should be interval or ratio
* There should be no outliers in the data
* Variables should be equally distributed

The formulae for the coefficient is:

R = N Σxy-( Σx)( Σy)

[N Σx2 – (Σx)2][N Σy2 – (Σy)2]

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

**Steps to calculate the correlation coefficient:**

1 ) Create a Pearson correlation coefficient table. Make a data chart, including both the variables. Label these variables ‘x’ and ‘y.’ Add three additional columns – (xy), (x^2), and (y^2).

2) Use multiplication to complete the tables and Add the columns from top to bottom

3) Use the correlation formula to plug in the values.

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

Scaling can be defined as a pre-processing step which is used on independent variables to normalize the data with a given range. We can also interpret scaling of data as transforming the given data so that it fits with a specific scale.

In a given data sets, a lot of data points can be far from each other, also varying features such as range and units. We scale variables as to reduce the distance. We generalize the variables in a sense to make the points closer to each other. Scaling just affects the coefficients and not the other values such as p-values, R2 values, etc.

The difference b/w normalized and standardized scaling is :

* In Normalized scaling we rescale the values into a range of [0,1].

In Standardized scaling we rescale the values such that the data has a mean of 0 and a standard variance of 1

* Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean and a standard deviation one.

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

If all the independent variables are statistically independent to each other, then VIF = 1.0. If there is perfect correlation, then the VIF usually goes to infinity. Perfect correlation is a rare occurrence, thus the value of VIF being infinity is not a common occurrence

An infinite VIF values shows that the corresponding variable can be expressed by a linear combination of variables that have a infinite VIF value too.

A solution to this is dropping any of the variable that is causing the perfect correlation or the ones that are causing multicollinearity

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

When we plot two quantiles plots against each other, we get a Q-Q plot. Quantile can be defined as fraction of values that fall below the quantile. For example, the median is a quantile where 50% of the data falls below that point and 50% lie above it.

We usually use Q-Q plot to find out if two sets of data come from the same distribution. It also helps us to assess if a set of data plausibly came from some theoretical distribution such as normal, exponential or uniform distribution.

A normal probability plot, or more specifically a quantile-quantile (Q-Q) plot, shows the distribution of the data against the expected normal distribution. For normally distributed data, observations should lie approximately on a straight line

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.