Assignment-based Subjective

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Observations for pair plots is explained below:

1. Count is high in Year (Yr) 1 which is 2019.
2. Temp and aTemp - Resemble a linear line for cnt.
3. Casual + Registered = Cnt so we will consider Cnt in our model and drop Casual and Registered
4. temp and atemp follow a very similar distribution so we will use atemp since it would affect the customers choice more than the actual temperature
5. There is some linearity seen above between other variables and cnt so a linear regression model can be used
6. aTemp and Temp are highly correlated with Cnt

Observations for categorical variables is mentioned below:

1. Significant increase in cnt of bikers where weather is 1 (Clear, Few clouds, Partly cloudy, Partly cloudy)
2. Mean cnt of bikers is not significantly different during Working days or Weekdays
3. Cnt of bikers is highest during the months of 7, 8, 9 and 10th months (July – October).
4. In 2019 there is significant increase in the number of bikers (Cnt is significantly higher)
5. Season 3(Fall) the average bikers count(cnt) is significantly higher than other seasons

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

A. Dummy variable values are assigned to categorical variables with adequate values. Drop first true is used to reduce redundancy in data.

B. **For Example**:

```
In [64]: status = ['male','fmale','others']

In [65]: print(status)
         ['male', 'fmale', 'others']

In [67]: status = pd.get_dummies(status) #, drop_first = 'true')
         print(status)
            fmale  male  others
         0      0     1       0
         1      1     0       0
         2      0     0       1
```

C. In the above code, you will observe that we can condense Male, Fmale column into a single column with values (Male = 1 and Fmale =0). Thus optimizing the dummy values assigned to a variable.

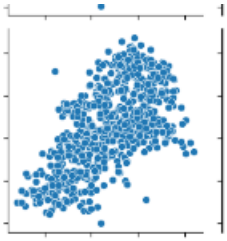
D. Observe the below code snippet

```
In [70]: status = ['male','fmale','others']

In [71]: print(status)
         ['male', 'fmale', 'others']

In [72]: status = pd.get_dummies(status, drop_first = 'true')
         print(status)
            male  others
         0     1       0
         1     0       0
         2     0       1
```

E. We can observe the redundant column Fmale is removed. This can be derived from column Male (if 1 then male, if 0 then female). Other column 1/0 will remain.
F. Dummy variables are used against few columns for this case study:
df1 =
pd.get_dummies(df,columns=["season","mnth","weekday","weathersit"],drop_first=True)

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

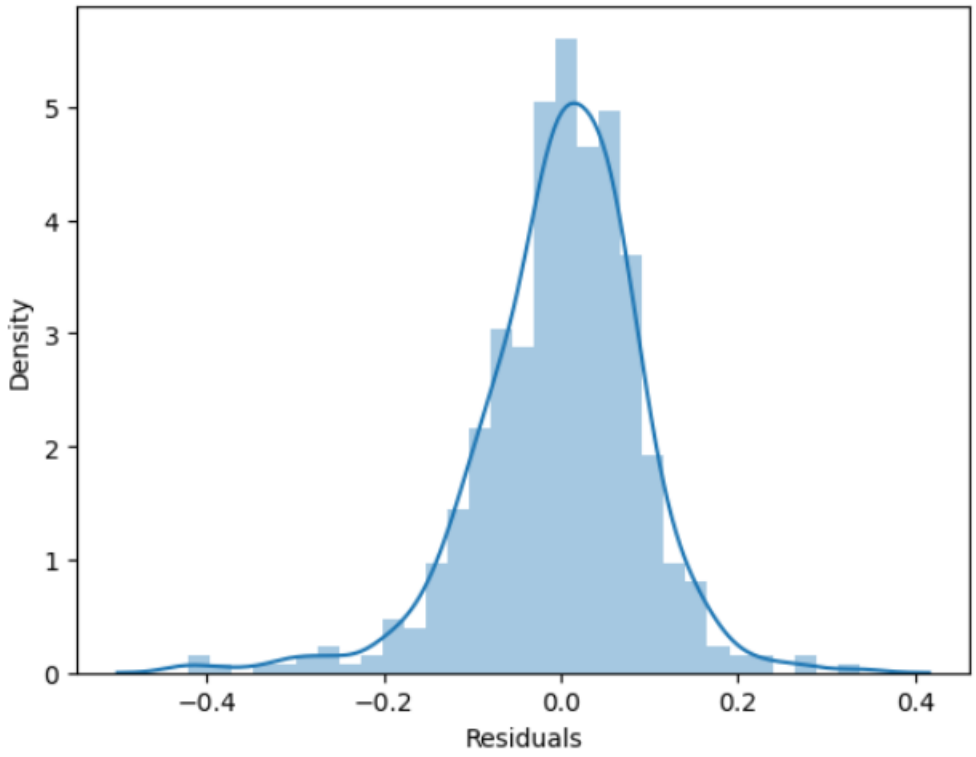
Below is the snapshot from the pair plot. Cnt and atemp have correlation



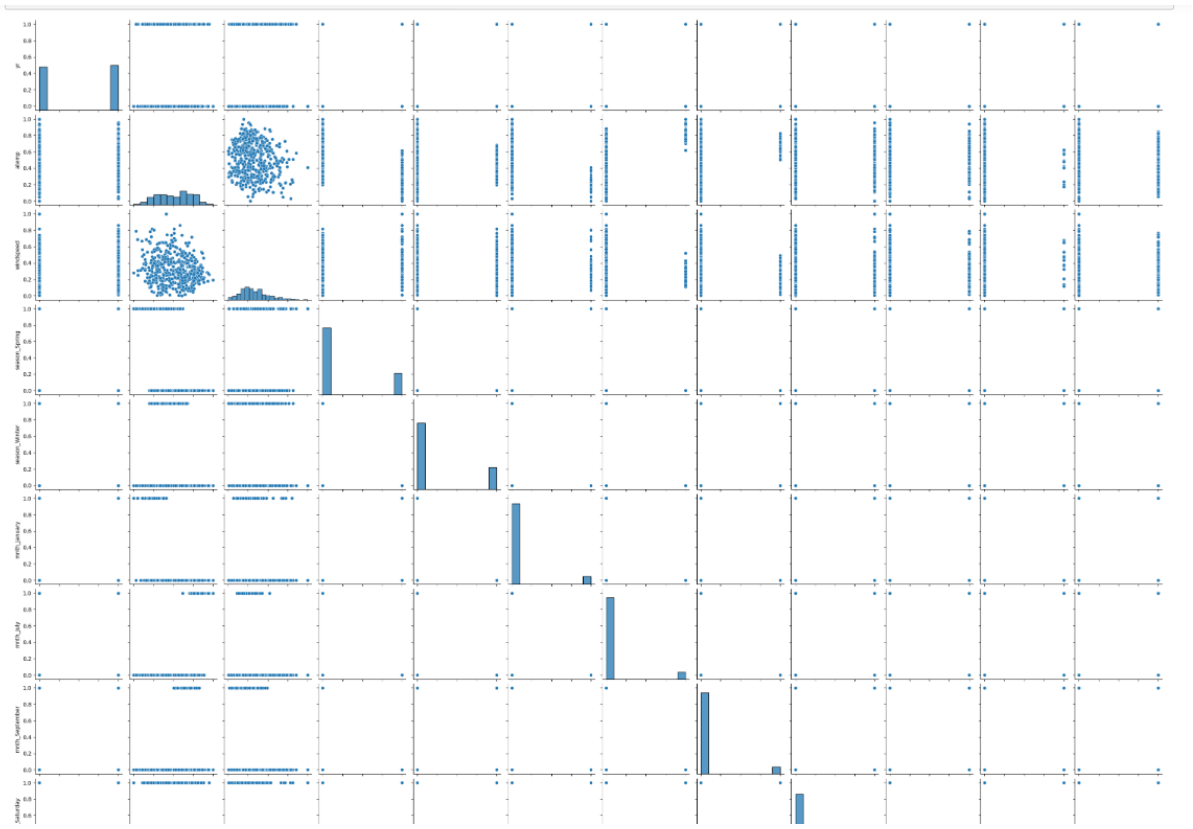
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Following are the points

1. Pair plot was used to check the relationship and linearity between Dependent (cnt) and independent variables.
2. Mean of Residuals should be 0. The residual graph of our case study is given below:



3. independent variables correlated with each other. This has been eliminated during model building. Pair plot snapshot is given below:



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

aTemp

Windspeed

Season - Spring

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

1. Linear regression algorithm is a supervised machine learning algorithm.
2. Linear regression computes the relationship between a dependant variable and single or multiple independent variable.
3. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables.
4. The equation provides a straight line that represents the relationship between the dependent and independent variables
5. Assumption for Linear Regression Model

    a. Linearity: The independent and dependent variables have a linear relationship with one another.
    b. Independence: The value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
    c. Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. Basically the amount of the independent variable(s) has no impact on the variance of the errors.
    d. Normality: The errors in the model are normally distributed.
    e. No multicollinearity: There is no high correlation between the independent variables.
6. Building the regression model
    a. We need to build a model that fits the equation y = b0 + b1x1 +b2x2….. +bnxn (Depending on the number of independent variables)
    b. Feature scaling – Independent variables at different scales need to be scaled using Min-Max or Standardization methods.
    c. Feature selection is an important step to avoid overfitting and multicollinearity. For this we first observe P value, f statistic, VIF to evaluate the relevance of variables in the model.
    d. During model building we need to remove/drop unwanted variables. This is done manually by observing the model summary (p value, f statistic, aic/bic, VIF etc). Automatically by using RFE
7. Once the model is built, we evaluate/validate the assumptions of linear regression model and build the formula.

2. Explain the Anscombe's quartet in detail. (3 marks)

1. Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.
2. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
3. Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
4. Application:

   The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)

A. Pearson's R, also known as Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.
B. The Pearson correlation coefficient is represented by the symbol "r" and takes values between -1 and +1.
C. The sign of "r" indicates the direction of the relationship: positive (r > 0) means that as one variable increases, the other tends to increase as well, and negative (r < 0) means that as one variable increases, the other tends to decrease.
D. A value of r = 0 means that there is no linear relationship between the two variables.
E. The magnitude of "r" measures the strength of the relationship. A value of r = 1 or r = -1 indicates a perfect linear relationship, while values closer to 0 indicate a weaker linear relationship.
F. Pearson's correlation coefficient is commonly used to examine the relationship between two continuous variables and to provide insights into their interactions.
G. However, correlation does not imply causation, and additional analysis is often required to draw meaningful conclusions.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

1. This is needed when we have multiple independent variables in a model, a lot of them might be on very different scales which will be difficult to interpret.
2. So we need to scale features because of two reasons:
    1. Ease of interpretation
    2. Faster convergence for gradient descent methods
3. You can scale the features using two very popular method:
    1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

    2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. 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-square, etc

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

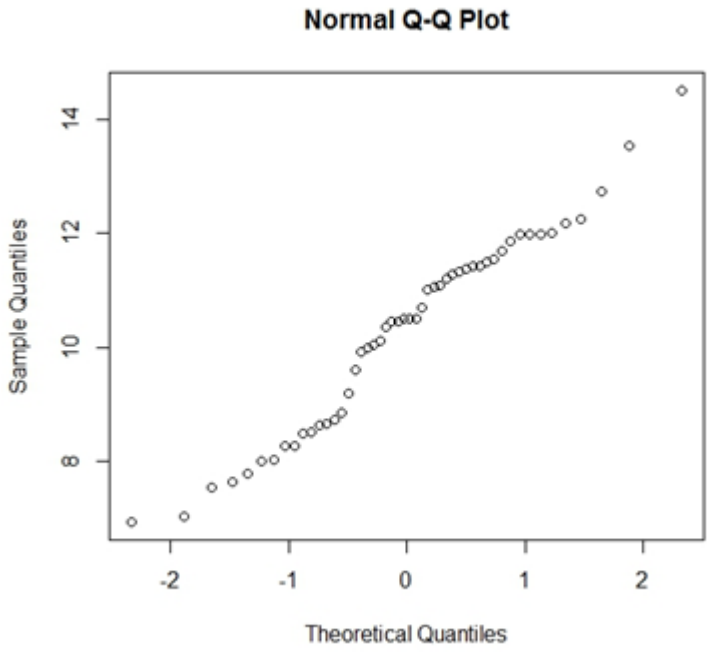
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Typically VIF should be <5, else it indicates that the independent variables are highly corelated to each other.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.