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?

(3 marks)

- Demand for rented bikes is fairly high from May to Oct months which belong in the Summer and Fall seasons
- Demand in 2019 was way higher than in 2018. This may be a result of more users adopting the service
- Demand is fairly high during clear and cloudy weather. There is no demand during heavy rain/snowfall
- Demand is slightly higher during working days as compared to non-working days. The inverse is applicable for holidays where demand reduces during holidays
- There's not much variation in demand per day of the week. But the highest demand is on Monday(6)

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

(2 mark)

drop_first=True is used to drop an additional variable which is created as part of dummy creation. This reduces an additional column in the feature set and also reduced multicollinearity among the dummy variables created.

Eg: Suppose there is a column that captures information on the *Mode of Transport.* This particular column can hold 3 values:

- Car
- Bike
- Boat

When we create dummy variables, these are the following scenarios:

Actual Value	Dummy_Car	Dummy_Bike	Dummy_Boat
Car	1	0	0
Bike	0	1	0
Boat	0	0	1

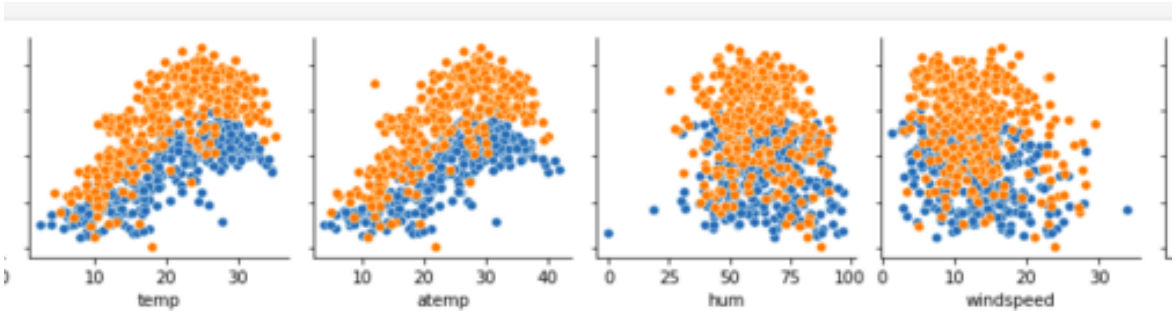
At this point we have 3 dummy variables. But we can express the same with just 2 variables as well:

Actual Value	Dummy_Car	Dummy_Bike
Car	1	0
Bike	0	1
Boat	0	0

As you can see when Car and Bike are '0', we can safely say that the mode of transport is a boat.

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

(1 mark)



temp has the highest correlation with cnt (target variable)

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

(3 marks)

- Normality of Errors:
  - Plot a histogram of error density and check if it fits a normal distribution
  - Check if the mean of the errors is 0
- Low/No Multicollinearity
  - Check VIF values for the features
  - If VIF < 10 for all variables, then multicollinearity is low
- Linear Relationship between independent and target variable
  - Plot a CCPR plot, it'll help visualise the behaviour of target var with respect to the independent var. It can be positively/negatively correlated or not correlated at all.
- Independence of Residuals
  - This can be evaluated by checking the Durbin-Watson factor. Values closer to 2 show that residuals are non-auto-correlational.
- Homoscedesticity

- This is checked by a qualitative scan of residuals vs count scatterplot. If there is no pattern found, it means the model is homoscedastic.

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

(2 marks)

Keeping yr aside, the top contributing features are as follows:

- temp **(+0.421)**
- weathersit_Light Rain **(-0.275)**
- windspeed **(-0.152)**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

(4 marks)

Linear regression is a statistical model that explains or predicts the target variable given one or more independent variables with the premise that they are related linearly. This means either the value of the target increases/decreases with an increase/decrease in value of independent var (positive correlation) or it behaves inversely and increases with a decrease in value of the independent variable (negative correlation).

This model is defined by the eq for a straight line:

$$Y = mX + C$$

where,

Y: target variable
X: Independent variable
m: the slope of line
C: Intercept of the line on y-axis

For a positive correlation, the slope will be positive and for a negative correlation, the slope will be negative.

2. Explain the Anscombe's quartet in detail.

(3 marks)

3. What is Pearson's R?

(3 marks)

It is a measure of linear correlation between two sets of data. it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationship or correlation.

A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

(3 marks)

Scaling is a technique to normalize the range of independent variables or features. It is necessary to ensure an unbiased comparison of coefficients once the model is complete. Scaling is important to bring variables that have different units to the same magnitude.

Standardized Scaling	Normalized Scaling
It is used when we want to ensure zero mean and unit standard deviation.	It is used when features are of different scales.
Not bound by any scale	Scales values between [0, 1] or [-1, 1]
Mean and the standard deviation is used for scaling	Min and Max value of features are used for scaling

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

(3 marks)

VIF is the variance inflation factor. It determines the factor by which one independent variable is correlated with another. High values of VIF signify high multicollinearity. It is calculated using:

$$\mathbf{VIF = 1/(1-R^2)}$$

Where R^2 is the R squared score.

Let's break this down further. For VIF to be Infinite,

$$(1-R^2) = 0$$

Therefore, $R^2 = 1$

And $R^2 = 1$ signifies perfect correlation which means the feature with VIF as infinite is completely defined by other independent variables.

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

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.

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.