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?

The categorical variables that were present in the dataset are Month, Season, weather situations, Holidays, weekday, year and working day.
Effects:

- 1.Rentals were more in the year 2019 compared to 2018.
- 2.Holiday and working days don't vary much but it shows 3/4th the rental is higher for the non-holiday days.
- 3.Rentals were more during the Fall season and than in summer.
- 4.Rentals were more in partly cloudy weather.
- 5. Rentals were increasing from the mid week to end week.

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

During dummy value creation (dummy encoding) it is advisable to use drop_first=True, because to avoid formation of redundant features i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding.
eg. For a variable, say, 'Relationship' with three levels, namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

Relationship Status	Single	In a Relationship	Married
Single	1	0	0
In a Relationship	0	1	0
Married	0	0	1

We see here there is no need to define three different levels. If you drop a level, say, 'Single', you will still be able to explain the three levels.

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

The numerical variable ‘registered’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable ‘atemp’ and ’temp’ have the highest correlation with the target variable ‘cnt’.

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

Checking for Linear relationship between X as in target variable and Y as in predictors.
By plotting the histogram for Error terms and observing whether it is normally distributed or not.
By checking Error terms are independent of each other or not
By checking Error terms have constant variance (homoscedasticity).

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

Temperature, Year and winter shows high positive coefficients and they are the top features who affects positively towards the demand in bike rental, so the company should try to promote or make available of bikes when the temp are high and with the upcoming years and for the season of winter.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning.
It is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).
There are two types of linear regression:

- Simple linear regression
- Multiple linear regression

Mathematically, we can write a linear regression equation as:

y = a + bx

Here, x and y are two variables on the regression line.
b = Slope of the line.
a = y-intercept of the line.
x = Independent variable from dataset
y = Dependent variable from dataset

We implement the model by performing various tasks such as data exploration, manipulation and analysis on the dataset then we use statsmodel to estimate the model coefficients.

We estimate the model coefficients for Linear Regression by using single feature to predict quantitative response. To calculate coefficients, we use the least square criterion, which means we will find a line that will decrease the sum of squared errors.
we plot the least square line by creating a data frame with the minimum and maximum.
we plot the observed data graph and the least square line using preds value and new x value.

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum.

Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it's only limited to linear relationships. Linear regression produces the best predictive accuracy for linear relationships whereas it's little sensitive to outliers and only looks at the mean of the dependent variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

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.

Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Explanation of this output:

- 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.
- 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.
- 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.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

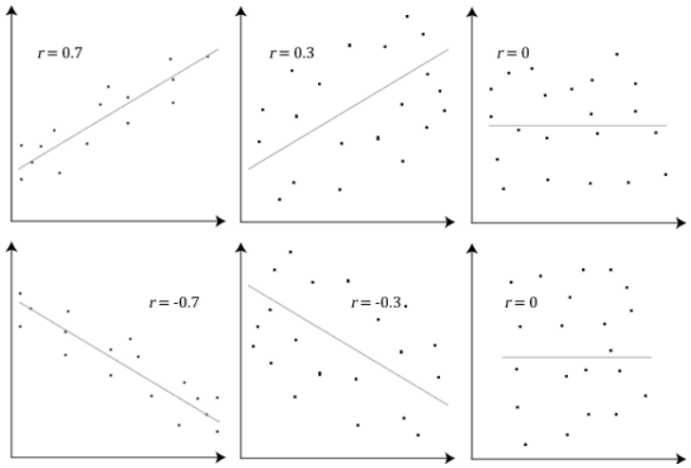
Application:
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.
- Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit.



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Potential problems with Pearson correlation.

- The PPMC is not able to tell the difference between dependent variables and independent variables
- he PPMC will not give you any information about the slope of the line; it only tells you whether there is a relationship.

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

Feature Scaling is one of the most important steps of Data Preprocessing. It is applied to independent variables or features of data. The data sometimes contains features with varying magnitudes and if we do not treat them, the algorithms only take in the magnitude of these features, neglecting the units. It helps to normalize the data in a particular range and sometimes also helps in speeding up the calculations in an algorithm.

What is Normalization?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

What is Standardization?

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Normalization vs. standardization

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

An **infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well).

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Interpretation:

q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

- **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.

X-values < Y-values: *If x-quantiles are lower than the y-quantiles.*

Different distribution: *If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*