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?

From the analysis of the categorical variables in the dataset, we can infer the following:

1. **Year (yr):** The year variable has a positive effect on the dependent variable. This indicates that the demand for
   bikes has increased from 2018 to 2019.
2. **Holiday (holiday):** The holiday variable has a negative effect on the dependent variable. This suggests that the
   demand for bikes is lower on holidays compared to regular days. People may be more likely to stay at home or use
   other modes of transportation on holidays.
3. **Season (season):** The Spring season has a negative effect on the dependent variable, indicating that the demand
   for bikes is lower during Spring compared to other seasons. This could be due to weather conditions or other seasonal
   factors influencing people's preferences for bike-sharing.
4. **Weather Situation (weathersit):** The Light rain_Light snow_Thunderstorm and Mist_Cloudy categories both have a
   negative effect on the dependent variable. This indicates that adverse weather conditions, such as rain, snow,
   thunderstorms, or mist, reduce the demand for bike-sharing. People are less likely to use bikes when the weather is
   not favorable for outdoor activities.
5. **Month (mnth) and Weekday (weekday):** The demand for bikes is not uniform across months and weekdays. Some months
   and weekdays exhibit a positive effect on the dependent variable, while others do not. This suggests that the demand
   for bikes may be influenced by specific events, promotions, or trends happening during those periods.

In summary, the categorical variables provide insights into how various factors such as year, holiday, season, weather
conditions, and time-related variables affect the demand for shared bikes.

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

Using `drop_first=True` while creating dummy variables is important because it helps to avoid the "dummy variable trap"
or multicollinearity issue. The dummy variable trap occurs when one or more of the independent variables in a regression
model are linearly related, which can cause problems when interpreting the model and its coefficients.

When creating dummy variables for a categorical variable with `n` categories, you only need `n-1` dummy variables to
fully represent the information. The reason for this is that the value of the dropped category can be inferred from the
values of the remaining dummy variables.

By setting `drop_first=True`, you're essentially removing one of the dummy variables, which serves as a reference
category. The coefficients of the remaining dummy variables will then be interpreted relative to this reference
category. This approach helps prevent multicollinearity and makes the model easier to interpret.

For example, if you have a categorical variable 'season' with four categories (Winter, Spring, Summer, and Fall),
creating dummy variables without dropping the first category would result in four dummy variables. However, you can
infer the information about the dropped category (e.g., Winter) when all the other dummy variables (Spring, Summer, and
Fall) have a value of 0. By dropping one category, you eliminate the potential multicollinearity issue and make the
model more interpretable.

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

Based on your observation that both 'temp' and 'atemp' are highly correlated with the target variable 'cnt', it seems
that these two variables have the strongest linear relationship with the bike demand.

However, since 'temp' and 'atemp' are themselves highly correlated, it is important to consider multicollinearity when
building a regression model. Including both variables in the model may lead to unstable estimates and make it difficult
to interpret the individual contributions of each variable to the target. To address this issue, you can choose to
include only one of the two variables in your model.

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

After building the linear regression model on the training set, I validated the assumptions of linear regression to
ensure the reliability and interpretability of the model. The key assumptions I checked are:

1. **Linearity:** I verified that the relationship between the independent variables and the dependent variable was
linear by plotting residuals against fitted values (predicted values). I observed that the points were randomly
dispersed and showed no specific pattern, indicating that the linearity assumption held.
2. **Independence:** I ensured that the observations were independent of each other. This is often related to the study
design and data collection process.
3. **Homoscedasticity:** I checked that the variance of the residuals was constant across all levels of the independent
variables by plotting residuals against fitted values. The points were randomly dispersed and showed a constant
spread, indicating that the assumption of homoscedasticity was satisfied. If the plot showed a funnel-shaped pattern,
I would consider applying a transformation to the dependent variable or using weighted least squares regression.

4. **Normality of residuals:** I ensured that the residuals were normally distributed by plotting a histogram of the residuals.

By validating these assumptions after building the linear regression model on the training set, I ensured that the model was appropriate for the data and provided reliable predictions and interpretable coefficients.

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

The equation represents the best-fitted line for the multiple linear regression model trained on the bike-sharing dataset, with 'cnt' as the target variable, representing the total number of bike rentals on a given day. The equation can be interpreted as follows:

cnt = 0.247 * yr - 0.0754 * holiday - 0.198 * Spring - 0.3154 * Light rain_Light snow_Thunderstorm - 0.088 * Mist_Cloudy + 0.066 * 3 + 0.123 * 5 + 0.148 * 6 + 0.156 * 8 + 0.195 * 9 + 0.125 * 7 + 0.113 * 10

Each coefficient in the equation corresponds to the impact of a specific variable on the total number of bike rentals ('cnt'). The coefficients represent the change in 'cnt' for a one-unit increase in the corresponding variable, holding all other variables constant. Here's a brief explanation of the coefficients:

1. yr: A positive coefficient (0.247) indicates that the total number of bike rentals increases by 0.247 units for each additional year, suggesting an increase in demand for shared bikes over time.
2. holiday: A negative coefficient (-0.0754) implies that the total number of bike rentals decreases by 0.0754 units during holidays, indicating lower demand on these days.
3. Spring: A negative coefficient (-0.198) suggests that the total number of bike rentals decreases by 0.198 units in the spring season compared to the reference season (winter).
4. Light rain_Light snow_Thunderstorm: A negative coefficient (-0.3154) indicates that the total number of bike rentals

decreases by 0.3154 units during days with light rain, light snow, or thunderstorms, showing a lower demand in
unfavorable weather conditions.
5. Mist_Cloudy: A negative coefficient (-0.088) implies that the total number of bike rentals decreases by 0.088 units
on misty or cloudy days, indicating a slightly lower demand on such days.
6. The remaining coefficients (0.066, 0.123, 0.148, 0.156, 0.195, 0.125, and 0.113) correspond to the different months,
indicating the change in bike rentals for each month compared to the reference month (January). The positive
coefficients show an increase in bike rentals during those months compared to January.

Overall, the equation helps us understand the factors affecting the demand for shared bikes and their respective
impacts. This information can be used by the bike-sharing company to tailor their strategies and better meet customer
expectations, ultimately leading to increased revenues and customer satisfaction.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a simple yet powerful algorithm used for predicting a continuous target variable based on one or
more input features. The main idea behind linear regression is to model the relationship between the target variable and
the input features as a linear combination of the features.

In detail, the linear regression algorithm consists of the following steps:

1. **Define the model:** The linear regression model assumes that the target variable (y) can be represented as a linear
combination of the input features (X1, X2, ..., Xn), plus an error term (e). Mathematically, this can be expressed
as:
$y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + e$
where β0 is the intercept, β1, β2, ..., βn are the coefficients (or weights) for the input features, and e is the

error term, which accounts for the difference between the predicted value and the true value of the target variable.

2. **Estimate the coefficients:** The goal of linear regression is to find the coefficients ($\beta 0$, $\beta 1$, ..., βn) that
minimize the sum of squared errors (SSE) between the predicted values and the true values of the target variable.
This can be achieved using various methods, such as the Ordinary Least Squares (OLS) method or gradient descent.
In the OLS method, the coefficients are calculated using the following formula:
$\beta = (X^T X)^{(-1)} X^T y$
where X is the input feature matrix (with an added column of ones for the intercept), y is the target variable, and β
is the vector of coefficients.
3. **Make predictions:** Once the coefficients are estimated, the linear regression model can be used to make
predictions for new data points. For each new input feature vector (x_new), the predicted target value (y_pred) is
calculated as:
y_pred = $\beta 0$ + $\beta 1$X1_new + $\beta 2$X2_new + ... + βnXn_new
4. **Evaluate the model:** After fitting the model and making predictions, it is essential to evaluate the performance
of the linear regression model. This can be done using various metrics, such as the Mean Squared Error (MSE), Root
Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared (coefficient of determination).
5. **Validate the assumptions:** As linear regression relies on specific assumptions (linearity, independence,
homoscedasticity, and normality of residuals), it is crucial to validate these assumptions after building the model.
If any of the assumptions are not met, the model may need to be adjusted or an alternative modeling technique should
be considered.

In summary, linear regression is a straightforward algorithm for predicting a continuous target variable based on one or
more input features. The algorithm involves defining the model, estimating the coefficients, making predictions,
evaluating the model, and validating the assumptions. Linear regression is widely used in various

fields due to its
simplicity, interpretability, and ease of implementation.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. Each dataset consists
of 11 (x, y) points. The quartet was designed to illustrate the importance of visualizing data and not solely relying on
summary statistics when analyzing datasets.

The four datasets in Anscombe's quartet have nearly identical summary statistics, such as the mean, variance, and
correlation coefficient. Additionally, they share the same linear regression line, with very similar coefficients and
R-squared values. However, when plotted, the datasets exhibit very different patterns and trends, demonstrating the
limitations of summary statistics in capturing the underlying structure of the data.

Here is a brief description of each dataset in Anscombe's quartet:

1. **Dataset I:** This dataset follows a simple linear relationship between the x and y variables. The data points are
scattered around the regression line, making it an ideal case for linear regression analysis.
2. **Dataset II:** The data points in this dataset follow a clear non-linear, quadratic relationship. Although the
summary statistics are similar to Dataset I, linear regression would not be suitable for modeling this relationship,
and a quadratic or other non-linear model would be more appropriate.
3. **Dataset III:** In this dataset, the data points form a perfect linear relationship, except for one outlier. The
outlier has a significant influence on the regression line, highlighting the impact of outliers on model fitting and
the importance of identifying and addressing them during data analysis.
4. **Dataset IV:** This dataset consists of ten data points with the same x-value, forming a vertical line, and one
outlier with a different x-value. The correlation coefficient and linear regression line are heavily influenced by
the outlier, again demonstrating the importance of visualizing data and addressing outliers.

In conclusion, Anscombe's quartet highlights the limitations of summary statistics and the need for visualizing data to
reveal patterns, trends, and potential issues. The quartet emphasizes the importance of combining both numerical
summaries and graphical representations during data analysis to avoid misleading conclusions based solely on summary
statistics.

3. What is Pearson's R?

Pearson's R, also known as Pearson's correlation coefficient, is a statistic that measures the linear relationship
between two continuous variables. Pearson's R ranges from -1 to 1, where -1 indicates a perfect negative linear
relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship between the two
variables.

The formula for Pearson's R is:

$r = \Sigma((x\_i - \bar{x})(y\_i - \bar{y})) / (\sqrt{}\Sigma(x\_i - \bar{x})^2 \sqrt{}\Sigma(y\_i - \bar{y})^2)$

where:

- r is the Pearson's correlation coefficient
- x_i and y_i are the individual data points for the two variables
- $\bar{x}$ and $\bar{y}$ are the mean values of the respective variables
- Σ denotes the summation over all data points

In other words, Pearson's R is the covariance of the two variables divided by the product of their standard deviations.
This normalization ensures that the correlation coefficient lies within the range of -1 to 1.

Pearson's R is widely used in various fields to measure the strength and direction of a linear relationship between two
variables. A high positive or negative value indicates a strong linear relationship, while a value close to zero
suggests a weak or non-linear relationship. However, it is important to note that Pearson's R only captures linear
relationships and may not accurately describe non-linear relationships between variables. In such

cases, other
correlation measures like Spearman's rank correlation or Kendall's rank correlation may be more suitable.

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

Scaling is the process of transforming data values to a common scale, often between 0 and 1 or with a mean of 0 and a
standard deviation of 1. It is performed to ensure that all features in a dataset are treated equally when applying
certain machine learning algorithms. In other words, it helps to eliminate the effect of different units, ranges, and
magnitudes of features that can bias the model's performance.

Normalization scaling, also known as Min-Max scaling, transforms the data values to the range of 0 to 1, where the
minimum value is scaled to 0, and the maximum value is scaled to 1. This type of scaling is useful when the data has a
limited range, and there are no significant outliers.

Standardized scaling transforms the data values to have a mean of 0 and a standard deviation of 1. This scaling method
is also known as Z-score scaling. It is useful when the data has significant outliers and is not normally distributed.
Standardized scaling ensures that the transformed data has a standard normal distribution, with a mean of 0 and a
standard deviation of 1.

The main difference between normalized scaling and standardized scaling is the range of the transformed data.
Normalization scales the data values to a specific range, while standardization scales the data values to a standard
normal distribution with a mean of 0 and a standard deviation of 1. In addition, normalization is more suitable for data
with a limited range and no significant outliers, while standardization is more suitable for data with significant
outliers and non-normal distributions.

In conclusion, scaling is an essential step in data preprocessing, particularly when applying machine learning
algorithms that are sensitive to feature scales. Normalization and standardization are two common scaling methods, and
the choice of method depends on the data distribution and the presence of outliers.

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

When calculating the VIF value for a particular independent variable, the coefficient of determination (R-squared) for
that variable is calculated based on a regression model that includes all the other independent variables in the
dataset. If the independent variable is perfectly correlated with one or more of the other independent variables in the
model, then the R-squared value for that variable will be equal to 1, and the VIF value will become infinite.

In other words, when an independent variable is perfectly correlated with one or more of the other independent variables
in the model, there is no unique solution for the regression model, and the VIF value cannot be calculated. This
situation is also known as perfect multicollinearity, where one or more independent variables can be perfectly predicted
by a linear combination of the other independent variables in the model.

Perfect multicollinearity can occur due to various reasons, such as errors in data collection or processing, measurement
errors, or including redundant variables in the model. To resolve this issue, we need to identify the redundant
variables and remove them from the model. We can also consider using alternative modeling techniques, such as ridge
regression or principal component analysis, to deal with multicollinearity in the data.

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

A Q-Q (quantile-quantile) plot is a graphical technique used to assess the normality of a distribution by comparing the

observed data distribution to an expected normal distribution. In a Q-Q plot, the values of the observations are plotted
on the y-axis, and the corresponding quantiles of a standard normal distribution are plotted on the x-axis. If the
observed data follow a normal distribution, then the points in the Q-Q plot will fall approximately on a straight line.

In linear regression, a Q-Q plot is commonly used to check the assumption of normality of the residuals, which are the
differences between the observed values and the predicted values from the regression model. A Q-Q plot of the residuals
can help to determine whether the residuals follow a normal distribution, which is a key assumption of linear
regression. A deviation from normality can indicate that the model is not a good fit for the data, and the predictions
from the model may be inaccurate.

The importance of a Q-Q plot in linear regression is that it provides a visual check of the normality assumption, which
is important for the validity and reliability of the regression model. If the Q-Q plot indicates non-normality of the
residuals, then we may need to consider transforming the data or using non-parametric regression techniques. On the
other hand, if the Q-Q plot shows that the residuals are approximately normally distributed, then we can have more
confidence in the accuracy and reliability of the regression model.