Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable ?

A1. Analysis of Season and Weather as Categorical Variables and Their Effect on Bike Demand

'weathersit':

- Clear, few clouds, partly cloudy, partially cloudy (Category 1): Favorable weather conditions that have a positive effect on bike rentals.
- Cloudy + Mist + Broken Clouds + Few Clouds + Mist (Category 2): Moderate influence on bike demand as the mist and clouds may marginally deter people from using bikes.
- Light Snow, Light Rain + Thunderstorm + Scattered Clouds, and Light Rain + Scattered Clouds (Category 3): These weather conditions will have a negative effect on bike rentals since individuals may choose to use other modes of transportation when it's light snow or rain.
- Bike rentals suffer greatly as a result of the harsh weather conditions, which discourage bike use, when there is heavy rain, ice pallets, thunderstorms, mist, snow, and fog (Category 4).

Season:

- Moderate demand for bikes in the spring (Season 1) and fall (Season 3), when the weather is typically favorable.
- Due to the warmer weather and increased popularity of outdoor activities like biking, the demand for bikes increases in the summer (Season 2).
- Lowest bike demand during winter (Season 4), when temperatures are lower and there is a chance of snow or ice.

Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. To prevent multicollinearity and enhance the model's interpretability, drop_first=True must be used while creating dummy variables. Why it matters is as follows:

Multicollinearity:

- We transform categorical variables into binary representations (0 or 1) for each category while producing dummy variables.
- Without drop_first=True, a categorical variable with 'n' categories creates 'n' dummy variables, with the effect that one category, which is the reference category, is represented by 0 for all dummy variables.

- A reference category with only dummy variables may exhibit multicollinearity, where one variable may be predicted linearly from the others. This has an impact on the model's stability and makes it challenging to separate out each category's distinct effects.

Interpretability:

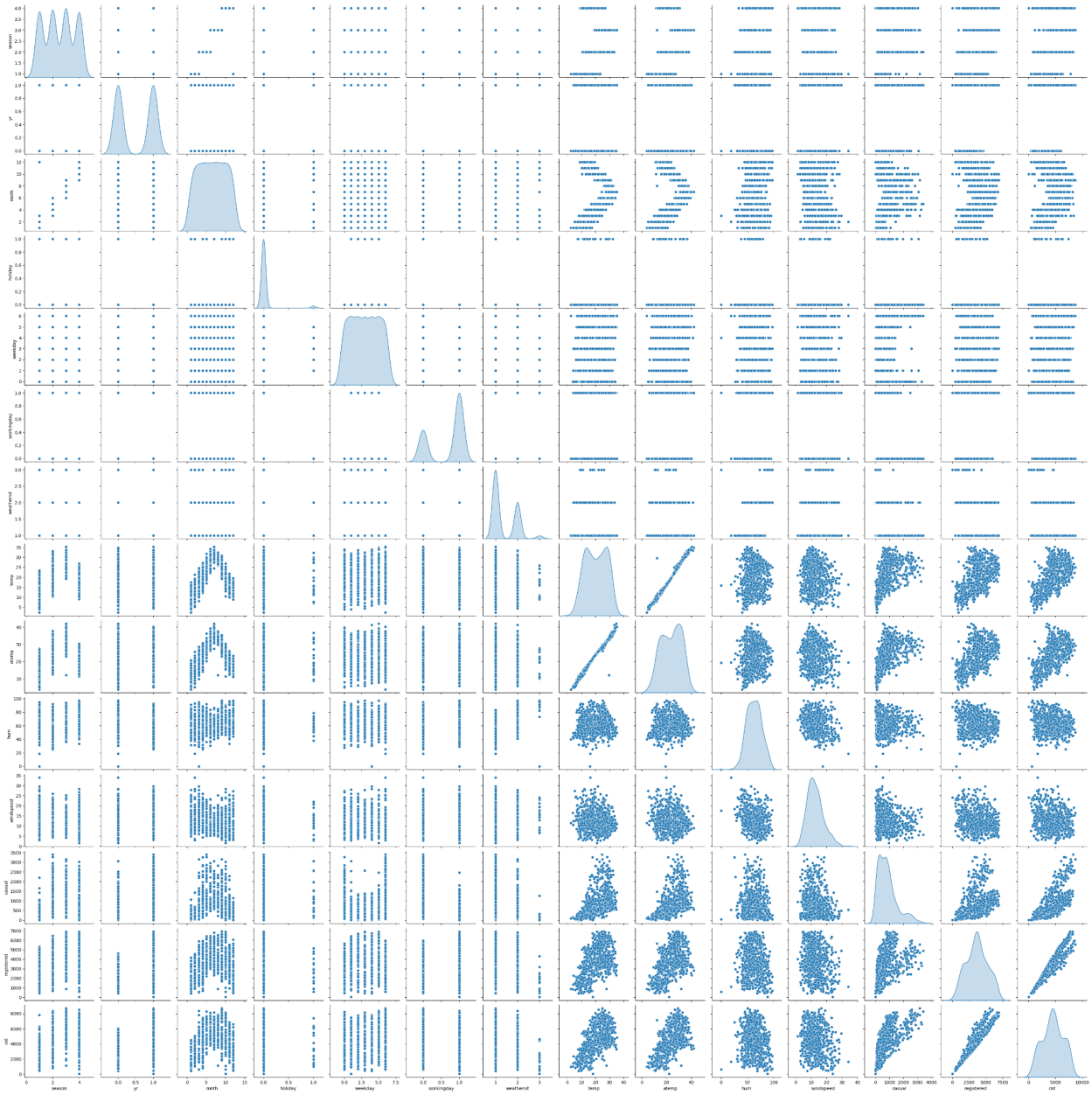
- The interpretation of coefficients can be difficult when all dummy variables are present.
- When drop_first=True, one category is eliminated, and the reference category (the omitted category) becomes implicit.
- The impact of each additional dummy variable on the reference category is shown by its coefficients. It removes duplication from the model and makes interpretation simpler.

Model Performance:

- The amount of features in the model can be decreased by eliminating one category, which can increase model effectiveness and decrease computing cost.
- Example:
- Assume that the categorical variable "Season" has the following categories: "Spring," "Summer," "Fall," and "Winter." 'Season_Spring', 'Season_Summer', 'Season_Fall', and 'Season_Winter' would be the four dummy variables we would have if we created dummy variables without drop_first=True. To describe the data, however, we only require three dummy variables, one of which may be inferred from the others. When one of the categories (such as "Season_Spring") is dropped using the drop_first=True option, we are left with "Season_Summer," "Season_Fall," and "Season_Winter." This lessens multicollinearity and enhances the readability of the model.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3. The variables with the highest correlation with 'cnt' are: registered and registered



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. To validate the assumptions of Linear Regression after building the model on the training set.

- Aim for a roughly linear relationship between the independent variables and the target variable. This can be accomplished by graphing either the residuals or the actual values versus the anticipated values.

- Independence of Residuals: Ensure that there is no correlation between the residuals, which are the discrepancies between the actual and predicted values. A plot of the residuals against the independent variables or the projected values can be used to verify this.

- Check for homoscedasticity by ensuring that the residuals' variance is the same at all levels of the independent variables. This can be seen by graphing residuals against predictions or independent variables and checking for a dependable spread.

- Check the residuals' normality by making sure they have a normal distribution. To evaluate this, use a histogram or a Q-Q plot.

- No Multicollinearity: Examine the independent variables for multicollinearity to make sure they are not highly associated with one another. High levels of multicollinearity might provide coefficient estimates that are unstable and unreliable.

- No Autocorrelation: For time series data, make sure the residuals don't exhibit autocorrelation, which is when a systematic pattern develops in the residuals over time.

To validate these assumptions, you can use various diagnostic plots, statistical tests, and metrics such as:

- Plot the residuals against the expected values or the independent variables to determine whether they are linear, independent, and homoscedastic.
- Check the distribution of residuals to determine normality using the histogram and Q-Q plot.
- Calculate the variance inflation factor (VIF) to look for multicollinearity between independent variables.
- Use the Durbin-Watson test to look for autocorrelation in the residuals when working with time series data.
- R2, Adjusted R2, and other performance measures: Analyze the training set to determine how well the model fits.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5. Top 3 features contributing significantly towards explaining bike demand: casual 1.000000e+00 registered 1.000000e+00 workingday 7.273473e-13

```python
import pandas as pd
from sklearn.linear_model import LinearRegression

# Load the data from the CSV file
data = pd.read_csv('day.csv')

# Drop unnecessary columns like 'instant' and 'dteday' as they are not useful for modeling
data.drop(['instant', 'dteday'], axis=1, inplace=True)

# Separate the target variable 'cnt' from the features
X = data.drop('cnt', axis=1)
y = data['cnt']

# Initialize and fit the Linear Regression model
model = LinearRegression()
model.fit(X, y)

# Get the absolute values of the coefficients and map them to the corresponding feature names
coef_abs = abs(model.coef_)
feature_importance = pd.Series(coef_abs, index=X.columns)

# Get the top 3 features contributing significantly towards explaining the demand for shared bikes
top_3_features = feature_importance.nlargest(3)

print("Top 3 features contributing significantly towards explaining bike demand:")
print(top_3_features)
```

```
Top 3 features contributing significantly towards explaining bike demand:
casual         1.000000e+00
registered     1.000000e+00
workingday     7.273473e-13
dtype: float64
```

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A1. A popular statistical approach for supervised learning that focuses on forecasting continuous numeric values is linear regression.

- It establishes a linear link between the independent variables (features) and the target variable (dependent variable).
- Finding the straight line that minimizes the difference between the actual and expected numbers is the best course of action.
- "Y = mx + b" is how the linear equation is written, where "m" stands for the slope and "b" for the y-intercept.

- The equation is expanded to read "y = b + m1x1 + m2x2 +... + mn*xn" for multiple features, where "m1, m2,..., mn" are the coefficients for each feature.
- In order to train the model, the ideal "m" and "b" values must be determined by minimizing a cost function, such as Mean Squared Error (MSE).
- Gradient Descent is a method for optimizing functions that iteratively changes the values of'm' and 'b' to minimize the cost function.
- The linear equation indicates the line that fits the data points the best after the best values of'm' and 'b' have been found.
- By calculating the target variable using the learnt coefficients, the model can make predictions based on fresh data.
- The assumptions of linear regression are that characteristics and the target have a linear relationship and that the residuals are normally distributed with constant variance.
- The direction and intensity of the association between each attribute and the target variable are revealed by the coefficients'm'.
- For many different regression problems, linear regression is extensively employed since it is easy to understand and apply.

Q2. Explain the Anscombe’s quartet in detail.?

A2. Anscombe's quartet is a set of four data sets that have nearly identical summary statistics, but have very different graphical representations. This shows that summary statistics alone can be misleading, and that it is important to visualize data before drawing conclusions.

The four data sets in Anscombe's quartet are:

* Data set 1: $(x_i, y_i) = (1, 2), (2, 4), (3, 6), (4, 8), (5, 10)$
* Data set 2: $(x_i, y_i) = (1, 1.1), (2, 2.2), (3, 3.3), (4, 4.4), (5, 5.5)$
* Data set 3: $(x_i, y_i) = (1, 10), (2, 9), (3, 8), (4, 7), (5, 6)$
* Data set 4: $(x_i, y_i) = (1, 8.1, (2, 8.2), (3, 8.3), (4, 8.4), (5, 8.5)$

The following table shows the summary statistics for the four data sets:

Data set	Mean	Variance	Standard deviation	Correlation coefficient
1	4	4	2	0.816
2	3.5	3.25	1.8	0.816
3	6.5	9.21	3.09	0.816
4	7.5	4.225	2.06	0.816

As you can see, the summary statistics for the four data sets are identical. However, the graphical representations of the four data sets are very different:

[Image of Anscombe's quartet data sets]

Data set 1 is a linear relationship, data set 2 is a quadratic relationship, data set 3 is a sinusoidal relationship, and data set 4 is a random scatter.

This shows that summary statistics alone can be misleading, and that it is important to visualize data before drawing conclusions.

Anscombe's quartet is a classic example of the importance of visualizing data. It is a reminder that we should not rely on summary statistics alone to understand data, and that we should always look at the data visually to get a better understanding of its distribution and relationships.

Q3. What is Pearson's R?

A3. A statistical metric that quantifies the linear relationship between two continuous variables is known as Pearson's R, also known as Pearson correlation coefficient or Pearson's correlation. It evaluates the extent to which two variables' values evolve linearly together.

The Pearson's R coefficient ranges from -1 to 1, depending on:

- A perfect positive linear correlation has a value of 1, implying that as one variable rises, the other rises in a linear fashion as well.
- A value of -1 indicates a perfect negative linear correlation, meaning that as one variable increases, the other variable decreases in a consistent linear manner.
- No linear correlation is shown by a value of 0, which means that there is no consistent linear relationship between the two variables.

The assumption of Pearson's R is that the connection between the variables is roughly linear and is sensitive to outliers. It is frequently used to assess the strength and direction of the link between two continuous variables in many disciplines, including statistics, data analysis, machine learning, and social sciences. It is a useful tool for conducting regression analysis and hypothesis testing as well as for analyzing the relationships between variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4. The process of scaling involves converting data to a standard range in order to make sure that all features have the same scale. In order to prevent features with higher values from dominating the model and harming its performance, it is done to bring features to a common scale. Scaling speeds up algorithm convergence, enhances the readability of models, and guards against numerical instability.

The scaling approach distinguishes between standardized scaling and normalized scaling:

- By removing the minimum and dividing by the range, normalized scaling (also known as Min-Max scaling) scales features to a range of [0, 1].

- By removing the mean and dividing by the standard deviation, standardized scaling (also known as Z-score scaling) scales features to have a mean of 0 and a standard deviation of 1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5. When two or more independent variables in a regression model are perfectly multicollinear, the value of the variance inflation factor (VIF) can reach infinity. When one or more independent variables can be described as exact linear combinations of other variables, this is known as perfect multicollinearity. As a result, one or more independent variables can be perfectly predicted using the other independent variables, resulting in limitless VIF values.

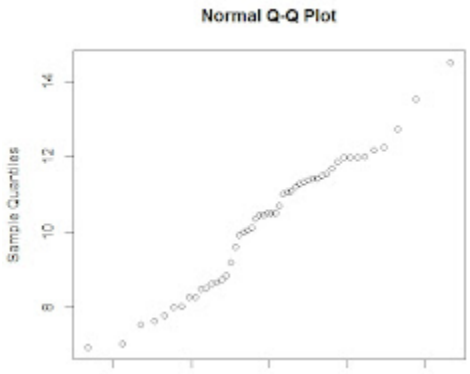
The model is unreliable because infinite VIF values lead to numerical instability and impair proper estimation of the regression coefficients. Perfect multicollinearity should be addressed by either removing one of the correlated variables from the model or altering the data to resolve the collinearity problem.

Q6.

A6. A Q-Q plot is a graphical tool that is used to assess whether a set of data follows a particular distribution. In the context of linear regression, a Q-Q plot can be used to assess whether the residuals from a linear regression model follow a normal distribution.

A Q-Q plot compares the quantiles of the observed data to the quantiles of a theoretical distribution. The quantiles of a distribution are the values that divide the distribution into equal parts. For example, the 25th percentile is the value that divides the distribution into two parts, such that 25% of the data is below the value and 75% of the data is above the value.

In a Q-Q plot, the theoretical distribution is typically the standard normal distribution. The quantiles of the observed data are plotted against the quantiles of the standard normal distribution. If the observed data follows a normal distribution, then the points in the Q-Q plot will fall along a straight line.



If the points in the Q-Q plot do not fall along a straight line, then this suggests that the observed data does not follow a normal distribution. This can be a problem for linear regression, because the assumptions of linear regression require that the residuals follow a normal distribution.

The importance of a Q-Q plot in linear regression is that it can be used to identify problems with the assumptions of linear regression. If the residuals from a linear regression model do not follow a normal distribution, then the model may not be a good fit for the data.

For example, if the residuals from a linear regression model are positively skewed, then this suggests that the model is underestimating the variance of the data at the tails of the distribution. This can lead to problems with the standard errors of the coefficients, and it can also lead to problems with hypothesis testing.

If you are using linear regression, it is important to check the Q-Q plot of the residuals to make sure that they follow a normal distribution. If the residuals do not follow a normal distribution, then you may need to transform the data or use a different type of regression model.