Assignment-based Subjective Questions

Question 1:

From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable?

Answer 1:

From the analysis of the categorical variables (season and weathersit), we can infer their impact on the demand for shared bikes (cnt). The season variable, after being converted to categories (spring, summer, fall, winter), shows how bike rentals vary with seasons. Generally, we would expect higher bike rentals during warmer seasons like summer and fall, and lower rentals in winter. The regression model's coefficients for these dummy variables indicate the extent of this variation, with positive coefficients for summer and fall suggesting higher demand in these seasons compared to the reference category, spring.

Similarly, the weathersit variable, which categorizes weather conditions (Clear, Mist, Light, Heavy), helps us understand how different weather scenarios affect bike rentals. Poorer weather conditions, such as mist, light rain, and heavy rain, are likely to reduce bike rentals. Negative coefficients for the dummy variables representing mist, light rain, and heavy rain in the regression model would confirm this expectation, indicating that bike demand decreases in these less favorable weather conditions compared to clear weather.

These observations help BoomBikes anticipate changes in bike demand based on seasonal and weather variations, allowing them to tailor their business strategies accordingly. For example, they might increase bike availability and marketing efforts during high-demand seasons or improve services to attract users even during adverse weather conditions.

Question 2:

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

Answer 2:

Using drop_first=True during dummy variable creation in one-hot encoding is crucial for avoiding multicollinearity in regression models. Multicollinearity occurs when predictor variables are highly correlated, which can make the model coefficients unstable and difficult to interpret. By dropping the first category, we eliminate the redundancy that arises when all categories sum to one, leading to perfect collinearity with the intercept term.

Additionally, dropping the first category establishes a reference category, simplifying the interpretation of the model's coefficients. Each remaining dummy variable's coefficient represents the change in the dependent variable (e.g., bike rentals) relative to this reference category. This approach not only stabilizes the model but also makes it easier to understand the relative impact of each category on the target variable.

Question 3:

Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

Answer 3:

To determine which numerical variable has the highest correlation with the target variable (cnt), we can visualize the relationships using a pair plot. A pair plot displays scatter plots for each pair of numerical variables in the dataset, allowing us to observe their relationships and identify potential correlations.

Upon examining the pair plot among the numerical variables, we often find that variables such as temp (temperature), atemp (feels-like temperature), hum (humidity), and windspeed might show different degrees of correlation with the target variable (cnt). Typically, temp and atemp tend to have the highest positive correlation with bike rentals because favorable weather conditions encourage more people to rent bikes. In contrast, hum and windspeed might show negative or lower correlations, as extreme humidity and high winds can deter bike usage.

To quantify these relationships, we can calculate the correlation coefficients between each numerical variable and cnt. The variable with the highest positive correlation coefficient indicates the strongest direct relationship with bike demand. For instance, if temp has the highest correlation coefficient (e.g., 0.8), it implies that as temperature increases, the number of bike rentals also increases significantly, highlighting temperature as a key factor influencing bike demand. This insight helps BoomBikes understand and anticipate bike rental patterns based on weather conditions, enabling better resource planning and service optimization.

Question 4:

How did you validate the assumptions of Linear Regression after building the model on the

training set?

Answer 4:

After building the linear regression model on the training set, validating its assumptions ensures the model's reliability and accuracy. Key assumptions include linearity, homoscedasticity, normality of residuals, and independence of errors. To check these, we performed residual analysis by plotting the residuals (the differences between the actual and predicted values).

First, we checked for linearity and homoscedasticity by plotting the residuals against the predicted values. Ideally, the residuals should be randomly scattered without any distinct patterns. A random scatter confirms that the relationship between the independent and dependent variables is linear and that the residuals have constant variance (homoscedasticity). Patterns or funnels in the residual plot would indicate heteroscedasticity, suggesting the variance of the errors is not constant.

Next, we assessed the normality of residuals using a histogram and a KDE plot. The residuals should form a bell-shaped curve if they are normally distributed. Additionally, we could use a Q - Q plot, where the residuals should lie along the 45-degree line if they follow a normal distribution. Finally, independence of errors was checked, ensuring that the residuals are not correlated with each other. For time series data, a Durbin-Watson test can be used to detect autocorrelation in the residuals. By validating these assumptions, we can confirm that the linear regression model is appropriate for predicting bike demand and that its predictions are reliable.

Question 5:

Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Answer 5:

Based on the final linear regression model, the top three features significantly contributing to the demand for shared bikes are temperature (temp), year (yr), and humidity (hum).

Temperature (temp) is the most influential feature. Warmer temperatures generally encourage outdoor activities, including biking. The model likely shows a strong positive coefficient for temp, indicating that as the temperature rises, the number of bike rentals increases. This relationship highlights the importance of favorable weather conditions in driving the demand for bike-sharing services.

Year (yr), indicating whether the data is from 2018 (0) or 2019 (1), also significantly impacts bike demand. The positive coefficient for yr suggests that bike rentals were higher in 2019 compared to 2018. This trend reflects the growing popularity and acceptance of bike-sharing services over time, indicating an upward trend in demand as the service becomes more established and familiar to users.

Humidity (hum) is another crucial factor affecting bike rentals. Typically, higher humidity levels can make biking uncomfortable, leading to a decrease in rentals. The model likely shows a negative coefficient for hum, indicating that as humidity increases, the number of bike rentals decreases. Understanding this relationship helps in predicting demand under varying weather conditions, allowing for better resource planning and service optimization.

General Subjective Questions

Question 1:

Explain the linear regression algorithm in detail.

Answer 1:

Linear regression is a simple way to predict one thing based on one or more other things. For example, you might want to predict how many bikes will be rented (the thing you want to predict) based on the temperature and humidity (the things you use to make the prediction).

In linear regression, you try to draw a straight line that best fits our data. If you have just one thing to predict from (like temperature), this line is on a graph with temperature on one axis and bike rentals on the other. The line shows how bike rentals go up or down as temperature changes. The goal is to find the line where the differences between the actual bike rentals and the rentals predicted by the line are as small as possible.

When you have more than one thing to predict from (like temperature and humidity), you are finding the best-fitting line in a more complex way, but the idea is the same. You use math to find the best line that shows how all the things you’re predicting from work together to affect the thing you want to predict. Once you have this line, you can use it to make predictions about bike rentals based on new temperatures and humidity levels.

Question 2:

Explain the Anscombe’s quartet in detail.

Answer 2:

Anscombe’s quartet is a set of four different sets of numbers. They have the same basic statistics, like the average, but look very different when you make a graph.

- **Same numbers**: All four sets have the same average, the same spread of numbers, and the same relationship between the numbers if you just look at the math.
- **Different graphs**: When you draw a picture of these numbers, each set looks very different. One might be a straight line, another a curve, one has an odd number far from the others, and the last is a cluster with one outlier.

Important lesson to learn here is, drawing graphs helps us understand the data better. So to explain in one line, Anscombe’s quartet teaches us that graphs are very important for really seeing what’s going on with data.

Question 3:

What is Pearson’s R?

Answer 3:

Pearson’s R is a way to measure how two things are related.

Number range: It goes from -1 to 1.

- +1: They move up together perfectly.
- -1: When one goes up, the other goes down perfectly.
- 0: They are not related at all.

Which basically translates to:

- Positive number: When one thing goes up, the other also goes up.
- Negative number: When one thing goes up, the other goes down.
- Closer to 1 or -1: Strong relationship.
- Closer to 0: Weak or no relationship.

In very simple terms, Pearson’s R is a number that tells us if two things are related and how strong that relationship is.

Question 4:

What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling?

Answer 4:

Scaling makes our data easier to use in machine learning. Scaling is a way to make numbers in our data easier to work with. It changes the numbers so they are more similar in size.

Purpose of scaling:

- Helps algorithms: Some computer programs that learn from data work better when numbers are on a similar scale.
- Faster learning: Programs can learn faster when numbers are scaled.
- Equal importance: Makes sure all parts of the data are considered equally.

Types of Scaling:

1. Normalized scaling
    - What it does: Changes numbers to be between 0 and 1.
    - How to do it: Subtract the smallest number and divide by the difference between the biggest and smallest numbers.
    - Eg. If numbers are between 50 and 100, change them to be between 0 and 1.
2. Standardized scaling
    - What it does: Changes numbers so the average is 0 and the spread is 1.
    - How to do it: Subtract the average and divide by how much the numbers is different.
    - Eg. If the average score is 50, change the scores so the average is 0.

So to answer in short:

- Normalization: Changes numbers to be between 0 and 1. Keeps the distances the same.
- Standardization: Changes numbers so the average is 0 and the spread is 1. Good for different units or extreme values.

Question 5:

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

Answer 5:

VIF (Variance Inflation Factor) can become infinite because of a situation called perfect multicollinearity.

- Why Infinite VIF happens:
    - Perfect Multicollinearity: This means one variable is exactly related to other variables. For example, if one column in our data is always double another column, they have a perfect relationship.
    - Math Issue: VIF is calculated in a way that involves dividing by a number. When variables are perfectly related, this number becomes zero. Dividing by zero gives an infinite result.
- Which basically can be translated to:
    1. If one column can be exactly calculated from other columns, it means there is no unique information in that column.
    2. The math formula used to calculate VIF involves dividing by something. With perfect multicollinearity, you end up dividing by zero, which leads to infinity.

So, VIF becomes infinite when some columns in our data are too perfectly related to each other, causing a problem in the calculation.

Question 6:

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

Answer 6:

A Q - Q plot (Quantile-Quantile plot) is a simple graph used to check if our data follows a specific distribution, usually the normal (bell-shaped) distribution.

- Axes of graph:
  - X-axis: Shows the expected values if our data were perfectly normal.
  - Y-axis: Shows the actual values from our data.
- Use in Linear Regression:
  - Checking Normality: Linear regression assumes that the errors (differences between actual and predicted values) are normally distributed. A Q-Q plot helps us see if this is true.
  - Straight Line: If the points on the Q-Q plot form a straight line, it means our data is normal. If not, our data is not normal.
- Importance:
  - Model Accuracy: Ensures the assumption that errors are normally distributed is met, which is important for the accuracy of the linear regression model.
  - Finding Problems: Helps to spot issues like skewness (data leaning to one side) or heavy tails (extreme values), which can affect the model’s performance.

So Q - Q plot is an easy way to visually check if our data or the errors from our regression model follow a normal distribution. It is important because it helps confirm that our linear regression model is working properly.