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)

Ans. From my analysis of the categorical variables, it shows that “weather” has a prominent effect on the dependent variable. Due to changes in “weather” conditions, the monthly usage/rental of bikes is also affected, and hence, even “month” & “season” seem to have an effect on the dependent variable.

The presence of warm and pleasant weather conditions during the summer season (May, June, July) and the fall season (Aug, Sept, Oct) promotes the most counts of rentals/usages of the bikes.

In the winter season and start of the spring season ie months from Nov to Feb, there is the highest dip in the number of rentals.

When it comes to weather conditions, there are 0 rentals for Heavy weather that is attributed to “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.”

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

Ans. This prevents multicollinearity, which is like when your data is too buddy-buddy, confusing your model. It also makes calculations clearer and avoids the "dummy trap" where your columns are so tight-knit that you can't separate their effects.

In short, ‘drop_first=True’ helps you dodge a data pitfall, keeps things less confusing, and makes your model's life easier by avoiding multicollinearity.

3. Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. If you drop the variables ‘casual’, and ‘registered’ then, ‘temp’ (Temperature) has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. Validation of assumptions can be done by Residual Analysis. This is what I followed to check the accuracy of assumptions :

- I started by analyzing the residuals, which are the differences between the actual and predicted values. I created a scatter plot of residuals against predicted values. If the residuals are randomly scattered around zero without any clear pattern, it indicates that the assumption of constant variance (homoscedasticity) is met.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks)

Ans. Weather conditions, months, day

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans. Imagine you're trying to draw a straight line through a cloud of points on a graph. You want this line to best represent how those points are scattered. That's what linear regression is all about.

Here's how it works:

- Line Finding: You have a bunch of data points (like dots on a graph) that show a relationship between two things. One is your "input" (let's call it X), and the other is your "output" (we'll call this Y). You want to find a line (a simple straight line) that fits these dots in the best way possible.
- Best-Fit Line: The idea is to find a line that minimizes the difference between the dots and the line itself. In other words, you're aiming to reduce the "error" as much as possible. This line is called the "best-fit line" or the "regression line."
- Equation of the Line: The equation of this line is like this: $Y = mX + b$. Here, 'm' is the slope of the line, which tells you how steep it is, and 'b' is the intercept, which tells you where the line crosses the Y-axis.
- Finding m and b: The big job is figuring out the right values for 'm' and 'b' that make your line fit the data points as closely as possible. This is where the magic happens. You want to tweak 'm' and 'b' until your line is as close to those dots as it can be.
- Minimizing Error: To figure out the best 'm' and 'b', you'll use a bit of math. The goal is to make the vertical distance between the dots and the line (the error) as small as possible. There's a method called the "least squares" method that helps you find these 'm' and 'b' values.
- Prediction Time: Once you've got your 'm' and 'b', your regression line is ready. Now, you can predict new 'Y' values for any 'X' values that were not in your original data. You just plug in the 'X' values into the equation ($Y = mX + b$), and it tells you where these new points should be on the line.
- Interpreting the Line: The slope 'm' tells you how much the 'Y' changes when 'X' changes by one unit. The intercept 'b' gives you the starting point of the line on the Y-axis.
- Assumptions: Just a heads up, linear regression has some assumptions. It assumes there's a linear relationship between your variables and that the errors (the differences between your line and the actual data points) are random and evenly spread out.

That's the essence of linear regression. It's like finding the straight story in a bunch of scattered points. It's used in all sorts of places, from predicting house prices to understanding how factors like temperature and humidity influence crop yields.

2. Explain Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's quartet is like a set of four secret twins that teach us a big lesson about data. Imagine you have four different groups of data points. They might look totally different when you plot them, but here's the mind-blowing part: they all have the same average, same variance, same correlation, and even the same linear regression line! But what's happening is that they show us how important it is to not just blindly trust numbers. Just because things look okay on the surface, it doesn't mean they're the same deep down.

- The Classic: This one's straightforward. It's a bunch of points that follow a pretty clear linear pattern. The linear regression line fits right in.
- The Outlier: Now, this is where it gets fun. The second twin has almost the same linear regression line, but there's a sneaky outlier that messes things up. This reminds us that one noisy point can mess with our nice line.
- The Parabola: The third twin is playing with a parabola. It's curving upwards, and a linear regression line just doesn't fit well. It reminds us that not all data can be squeezed into a straight line.
- The Horizontal Line: The fourth twin is like, "Who needs ups and downs?" It's just a bunch of points in a straight, horizontal line. Again, the linear regression line is a bit off.

Anscombe's quartet reminds us that even though numbers look good on paper, there's often a bigger story behind them. It's like having four identical-looking houses, but when you step inside, each one has a totally different setup.

3. What is Pearson's R? **(3 marks)**

Ans. Pearson's R is like a friendship detector for data. It's a number that tells you how two sets of numbers are buddies with each other.

Imagine you have two sets of data, like ice cream sales and temperature. Pearson's R swoops in to check if they have a relationship.

- If the temperature goes up and the ice cream sales go up too, R gets all excited and dances close to +1. It means they have a positive correlation, like two besties hanging out.

- But if the temperature goes up and ice cream sales go down, R sulks and heads toward -1. It's a negative correlation like when one friend is in a bad mood, the other is down too.
- Now, if R is close to 0, it's like they're not really friends. No correlation. They just do their own thing, like two people walking down the street without noticing each other.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans. Scaling is done to make sure all your data values have similar ranges. Imagine you're comparing age and income. Age might be between 0 and 100, while income could range from a few dollars to thousands. Without scaling, the big numbers like income might dominate the comparison, leaving age in the dust.

Now, let's talk about the two kinds of scaling:

- Normalized Scaling: This is like making everyone's values fit between 0 and 1. It's perfect when you want to keep proportions intact. Each data point gets divided by the maximum value in the dataset. So, if you had a bunch of scores, the highest score would become 1, and the rest would be relative to that.
- Standardized Scaling: Here, it's all about giving everyone the same average (0) and standard deviation (1). This is awesome when you want to compare data points on the same scale, while also considering their distribution. It's like giving each data point a makeover, so they all look similar.

In short, scaling helps data behave and play nicely together. Normalized scaling squishes values between 0 and 1, while standardized scaling gives them the same average and spread.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. The occurrence of an infinite value for the Variance Inflation Factor (VIF) typically arises from perfect multicollinearity within the dataset. Perfect multicollinearity occurs when one independent variable in a regression model is a perfect linear combination of one or more other independent variables.

Mathematically, if a variable can be expressed as a constant multiple of another variable or a combination of other variables, this leads to the generation of an infinite VIF for the variable involved in the linear combination.

In practical terms, perfect multicollinearity can render the matrix inversion process, which is a crucial step in calculating VIF, impossible. As a result, the VIF value becomes undefined or infinite.

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

Ans. Q-Q stands for "Quantile-Quantile," which sounds fancy but is pretty simple. It's a plot where you put the ordered values of one set of data on one axis and the expected values from a reference distribution on the other axis. Basically, it helps you see if your data and the reference are on the same wavelength.

In linear regression, a Q-Q plot is like a detective tool. It helps you check if the residuals (the differences between the predicted and actual values) are alright and following a normal distribution. This matters because linear regression often assumes that your data behaves normally. If your residuals fall in line with that straight reference line on the Q-Q plot, it's like a thumbs up for normality. It means your model assumptions are holding up well. But if your residuals start dancing all over the place, curving away from the line, it's like a signal that something's not quite normal.

So, it's an important tool to give your model a health check. If the Q-Q plot is happy, your linear regression has more confidence. If not, it's a sign you might need to dig deeper into your data or consider adjusting your model.