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)
   1. From the data we can see that:
      1. Bike rentals are more during Summer and fall compared to other seasons
      2. The count of bike rentals increased from June till October complementing the season data
      3. We can also see that clear skies contributed to more rentals compared to other weather situations
      4. Bike rides also increased during the year 2019 assuming that popularity was gaining on the second year of service
      5. More bikes are rented during Wednesday’s and Saturday’s
      6. We can also see that more bikes are rented during a non-working day
2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)
   1. drop\_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   1. The variable temp seems to have highest correlation (0.63)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   1. Residual plots: Create a residual vs fitted value scatter plot to see if the red line is flat on 0, which indicates that the residual errors have a mean value of zero
   2. Plotting y test and y pred – This will help us understand the spread to check if the graph is linear.
   3. Homoscedasticity – Check if the variance of errors are constant.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   1. Temp with coeff of 0.3531
   2. Sit1 – Clear skies with coeff of 0.2499
   3. windspeed with negative coeff of -0.1676

General Subjective Questions

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

Linear Regression is helpful in finding out how an outcome changes based on another dependent variable. For example, if you consider price of a car, we can draw a relation between how the size of the car’s engine affects the price of the car. The idea is to find a straight line that best describes the relationship between car price and engine size in our example.

A point to be noted is that, in our example, I have shown only one variable influencing the price of the car. In reality, many factors such as brand, size, number of seats, colour etc can influence the outcome(price of the car)

So, we can classify linear regression into 2 main types.

1. Simple Linear Regression – As the name suggests, in this type of linear regression, we consider only one factor we are interested in, to check how it influences the outcome. E.g, number of cylinders to price of the car. In simple terms, simple linear regression tries to draw a line through a set of data points, where the line predicts the outcome. Equation for a simple linear regression would look like:

y = c + mx

where y is the outcome we like to predict

c is where the line crosses the y axis

x is the input variable

m is the slope of the line - how much influence the input variable x have on the outcome y

1. Multiple Linear Regression – Here we consider more than one factor to check how all those influence the outcome. We choose the factors that contributes more and eliminate factors that contributes less or no change to the outcome. Equation for a simple linear regression would look like

y = c + m1x1 + m2x2 + …. + mnxn

where n is the number of factors

Regression Algorithm:

1. **Data Collection:** First, we gather data that includes what we want to predict and the factor that might influence it.
2. **Dummy values creation:** We convert the categorical variables to dummy variables as those would otherwise be difficult to include due to their non-numeric nature.
3. **Scaling:** Some of the variables would not be in range of the other variables. For example, engine capacity would be in large range where as number of doors would be typically between 2 – 5. We would need to scale these features to bring everything within the range.
4. **Splitting into Test and Train:** We then divide the data into test and train. Typically, we allocate 70% to train and 30% to test
5. **Finding the Best Line:** We run the algorithm on the train dataset. The algorithm then tries to find the best line through the data points. This best line is the one that minimizes the difference between the actual data points and the predictions the line makes. The difference between these two values is called an error or residual.
6. **Optimization:** Linear regression uses a method called Ordinary Least Squares (OLS) to minimize the total error by adjusting the slope (m) and intercept (c). The total squared difference between the real data and the predicted data should be as small as possible. We use RFE (Recursive Features Elimination) to arrive at the features that have best fit.
7. **Making Predictions:** Once the algorithm has found the best line, we can use it make predictions on the test dataset

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

Whenever we are given data to build models, it is not enough to rely only on the numerical summary statistics. It is very important to visualize the data using plots and graphs to analyse the data before we build models.

The Anscombe’s quartet is a set of 4 identical data set whose summary statistics such as mean, median, standard deviation etc are very similar. But when we plot them in a graph, we can see that all the datasets are completely different from each other.

This emphasizes that graphs are essential to a good statistical analysis.

Each dataset in Anscombe’s quartet has:

• The same mean of both the x and y variables.

• The same variance for x and y.

• The same correlation between x and y.

• The same linear regression line, represented by: y = 3 + 0.5x

To emphasize how similar the four datasets are in terms of their summary statistics, here are the shared values for all datasets:

• Mean of x-values: 9

• Mean of y-values: 7.50

• Variance of x-values: 11

• Variance of y-values: 4.12

• Correlation between x and y: 0.816

• Linear regression equation: y = 3 + 0.5x

Anscombe’s Quartet teaches us that while summary statistics are useful, they can never replace the power of visualizing the data. Plotting our data can reveal important patterns, trends, or anomalies that the numbers alone won’t show. It reminds us to always dig deeper and not just rely on averages and correlations.

3. What is Pearson’s R? (3 marks)

Pearson’s R is used to analyze relationship between two variables. The Pearson correlation coefficient is used to quantify the strength and direction of a linear relationship between two continuous variables.

Direction:

When the direction is positive, it means that when one increases, the other variable increases in a linear manner.

Similarly, when the direction is negative, it means when one variable increases, the other variable decreases in a linear manner.

When the value of the Pearson’s coefficient is 0, it means there is no linear correlation between the two variables

Strength:

The Pearson’s coefficient is denoted by the letter r. When:

r = 0.7 to 1.0 (or -0.7 to -1.0): Strong positive (or negative) correlation.

r = 0.4 to 0.7 (or -0.4 to -0.7): Moderate positive (or negative) correlation.

r = 0.1 to 0.4 (or -0.1 to -0.4): Weak positive (or negative) correlation.

r = 0: No correlation.

In python we can use pearsonr(x ,y) to calculate Pearson’s coefficient, where x & y are numeric vectors of same length.

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

1. **Scaling:** Some of the variables would not be in range of the other variables. For example, engine capacity would be in large range where as number of doors would be typically between 2 – 5. We would need to scale these features to bring everything within the range.
2. There are 2 types of scaling:
   * + 1. Min-Max Scaling – Min – Max scaling takes the lowest value as 0 and highest value as 1. It fits all the variables relative to 0 and 1. This is done using the formula below:

Xscaled = (x – xmin)/(xmax – xmin)

Where x is the original data point

xmin is the minimum value of the feature

xmax is the max value of the feature

* + - 1. Standardized scaling – Standardized scaling converts the data of a feature to have a mean of 0 and a standard deviation of 1. This ensure data is centred around 0 and has unit variance. This is done using the formula below:

Xscaled = (x -mean)/standard deviation

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

When the value of the Variance Inflation Factor (VIF) becomes infinite, it indicates a situation called perfect multicollinearity in the dataset. This occurs when one or more features (independent variables) in a regression model are exactly linearly dependent on other features.

The formula for calculating VIF is:

VIF = 1 / (1-R^2)

Where R^2 is coefficient of determination.

When there is multicollinearity, R^2 would be 1. When we substitute it to the formula for VIF, it becomes 1 / (1-1) which is infinite.

This happens because the model cannot distinguish between the linearly dependent variables, and this leads to the inability to compute a unique estimate for the regression coefficients. In simple terms, the model is confused because two or more variables carry exactly the same information.

In our bike sharing example, if we retain the variables registered and casual, it would be an example for multicollinearity as the cnt variable is sum of registered and casual.

In order to eliminate infinite value for VIF, we do the following:

1. Retain only one variable if 2 or more variables are perfectly collinear
2. We can also combine the variable like summing up or average to have only one variable

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

(3 marks)

A Q-Q plot is an important diagnostic tool for validating the normality assumption of residuals in linear regression. By comparing the residuals of the model with a theoretical normal distribution, the Q-Q plot helps identify issues such as:

• Violations of the normality assumption.

• The presence of outliers or heavy tails.

• Potential model mis-specifications or the need for transformations.

Using a Q-Q plot, we can assess whether the linear regression model meets necessary assumptions for reliable predictions and statistical inference.

If the points in the Q-Q plot lie close to a straight line, it suggests that the residuals are normally distributed, confirming the assumption of normality.

If the points deviate significantly from the line, especially in the tails, it suggests that the residuals are not normally distributed, and the model may need adjustments (e.g., by transforming the dependent variable).