*Assignment-based Subjective Questions 1.*

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

*A.We can observe that the dependent strong correlation with fall season, year 2019, August, September, October months, good weathersit*

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

*A. drop_first=True will delete the 1st variable of the dummy variables as it's not actually required. We can have n-1 values for a dummy variable*

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

*A. We can observe that the dependent variable have good correlation with temp, atemp variable*

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

*A.I have validated the r2 score and linearity of training and test data*

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

A.We can say temperature, wind speed, year, holiday are the top contributing features

General Subjective Questions

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

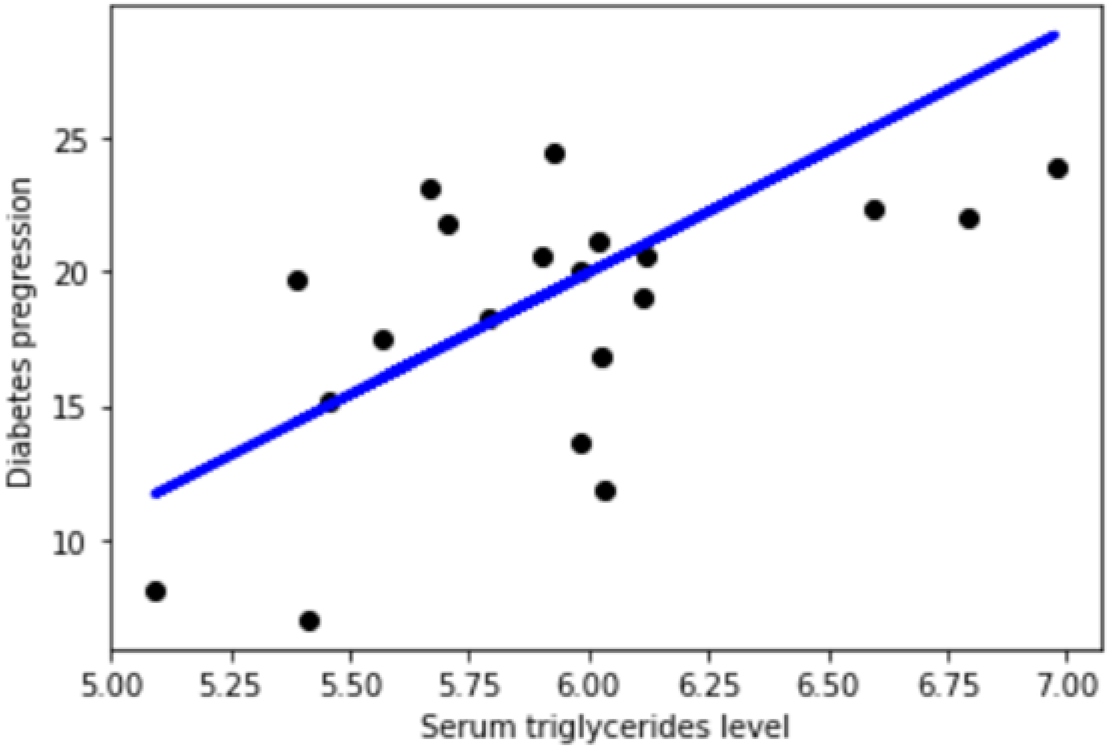
A. Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots$$

In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

In the following image, a linear regression model is described by the regression line y = 153.21 + 900.39x. The model describes the relationship between the dependent variable, Diabetes pregression, and the explanatory variable, Serum triglycerides level. A positive correlation is shown. This example demonstrates a linear regression model with two variables. Although it is not possible to visualize *models with more than three variables, practically, a model can have any number of variables.*



*A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.*

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

*A.* ***Anscombe's quartet*** *comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.*
*The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.*

***Anscombe's quartet*** *is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.*
*The four datasets of* ***Anscombe's quartet.***

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

*The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit). The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0* *indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.*

3. *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?*
   A. *It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.*

   B. *Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.*

   C. *It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.*

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

*MnMaxscalng = x-min(x)/max(x)-min(x)*

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

*Standardization=x-mean(x)/sd(x)*

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

*A. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.*

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

*A.Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.*

*This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.*

*Few advantages:*

*a) It can be used with sample sizes also* *b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.*

*It is used to check following scenarios:*

*If two data sets —*

*i. come from populations with a common distribution*

*ii. have common location and scale*

*iii. have similar distributional shapes*

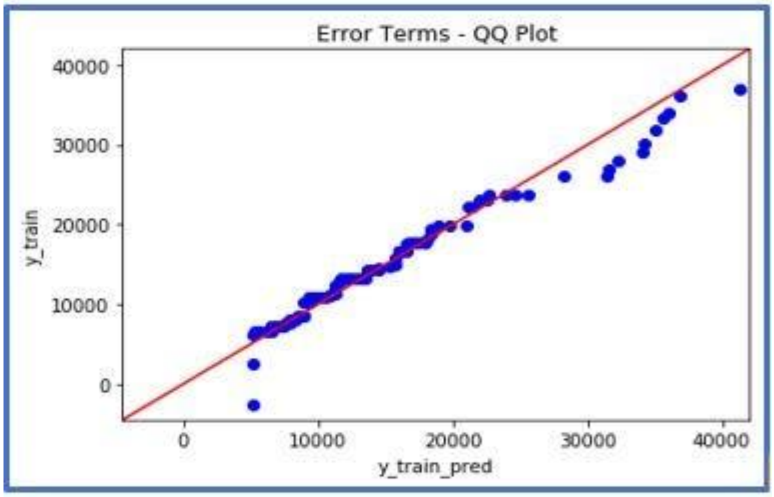
*iv. have similar tail behavior*

*Interpretation:*

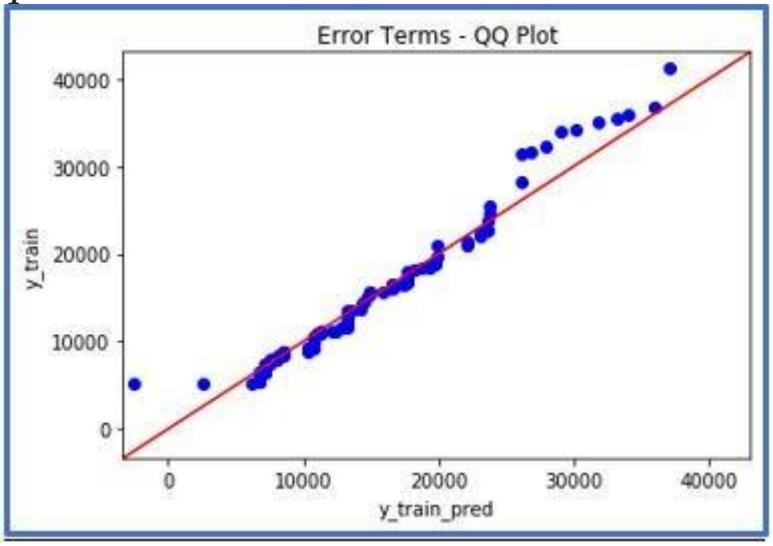
6. *A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.*

   *Below are the possible interpretations for two data sets.*

   *a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis* *b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.*



*c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.*



*d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*