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)

A. Analyzing the box plots of the categorical variables against rentals count, following have been our observations.

a. **Year**: The average number of rentals is increasing with year which can be due to increasing popularity among people.
b. **Season**: More rentals were observed during fall season followed by summer, winter and spring.
c. **Holiday**: There were more average rentals during non-holiday days which can be due to people using rentals to commute to work/school.
d. **Weekday**: No specific behaviors were observed during any of the weekdays for rentals.
e. **Working day**: Average rental count for working and non-working days were almost same.
f. **Weather situation**: Weather with Clear, Few clouds, Partly cloudy, Partly cloudy had seen more average number of rentals than misty or rainy days.

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

A. While creating dummies by default we create number of dummy variables equal to number of discrete categorical variable in the categorical variable (n). But among these 'n' dummy variables, one of the variable will be redundant. If we do not use drop_first = True, then 'n' dummy variables will be created, and these predictors('n' dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

A. After looking at the pair plot we find that feature 'registered' has the highest correlation of 0.95 with target variable.

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

A. Following are the assumptions of Linear Regression :

a. Linear relationship between X and Y – This has been validated by plotting a line plot between Independent and Target Variables
b. Error terms are normally distributed (not X, Y) – This has been validated by plotting a histogram of the error terms.
c. Error terms have constant variance (homoscedasticity) – This has been validated by plotting a scatter plot of the error terms with predicted terms and confirmed that there is no pattern in the error terms.
d. Error terms are independent of each other – This has been validated by plotting line plot between predictions and residuals and there were not any auto-correlation between error terms.

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

A. From the final model we found that below three features were contributing significantly towards explaining the demand of the shared bikes

   a. Feeling Temperature : with coefficient `0.412`
   b. $weathersitLightRain$ [Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds]: with coefficient `0.275`
   c. Year: with coefficient `0.236`

General Subjective Questions

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

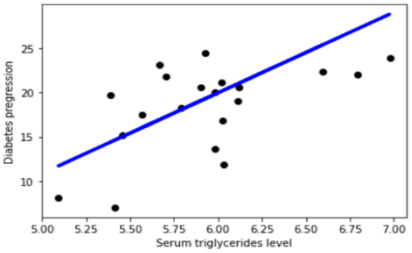
A. Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line 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:

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

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.

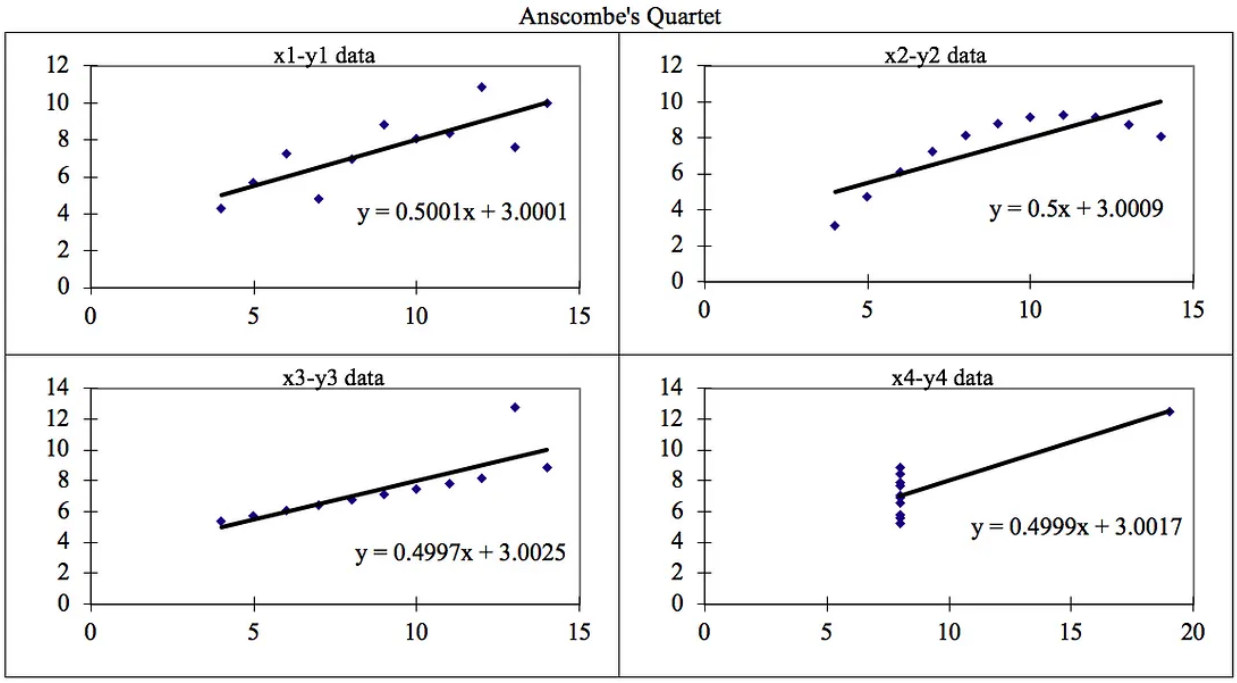
In the following image, a linear regression model is described by the regression line y = 153.21 + 900.39x. The model describes the relationship between the dependent variable, Diabetes pregression, and the explanatory variable, Serum triglycerides level. A positive correlation is shown. This example 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 can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the

Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

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

A. Pearson Correlation Coefficient (r):

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight:The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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

A. Scaling:

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.

Why it is performed:

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.

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.

Difference between normalized scaling and standardized scaling:

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.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.
Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.
This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

Formula and Calculation of VIF

The formula for VIF is:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

$R_i^2 = \text{Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones}$

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:

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

Plotting a Q-Q plot:

Below are the steps to generate a Q-Q plot for team members age to test for normality

a. Take your variable of interest (team member age in this scenario) and sort it from smallest to largest value. Let's say you have 19 team members in this scenario.

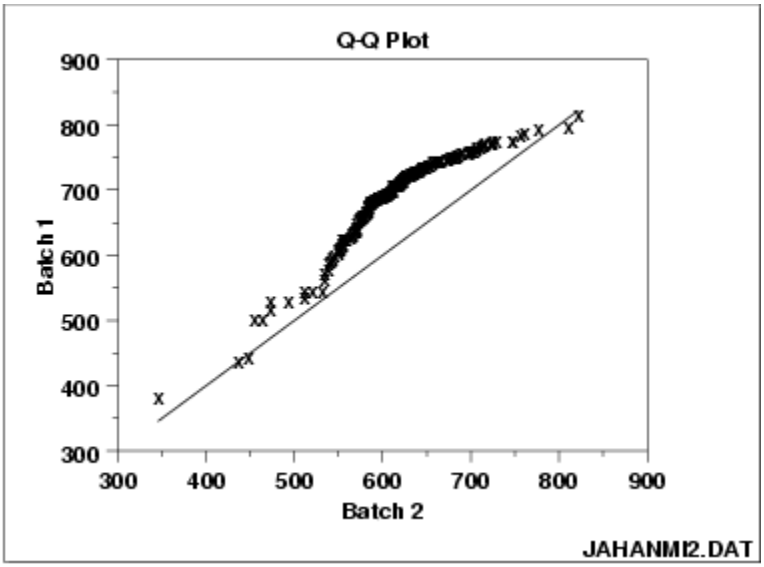
b. Take a normal curve and divide it into 20 equal segments (n+1; where n=#data points)

c. Compute z score for each of these points

d. Plot the z-score obtained against the sorted variables. Usually, the z-scores are in the x-axis (also called theoretical quantiles since we are using this as a base for comparison) and the variable quantiles are in the y-axis (also called ordered values)

e. Observe if data points align closely in a straight 45-degree line

f. If it does, the age is normally distributed. If it is not, you might want to check it against other possible distributions



Importance:

Q-Q plot can also be used to test distribution amongst 2 different datasets. it is possible to compare the distributions of two datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.

A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.