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)

Below are the inferences drawn from the categorical variables

- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during september month.
- The weekday box plots indicates that there is no significant difference in bike sharing in each day.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.
- The year box plots indicates that more bikes are rent during 2019.

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

It helps in reducing the extra column created during dummy value creation. Hence it reduces the correlation created among the dummy variables.

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

Registered variable has the highest correlation with the target variable(0.95)

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

- Calculated the residual value between the y_train and y_predicted and verified the error is normally distributed
- Verified that there is a linear relationship between x and y
- Verified the VIF value to understand the multicollinearity between variables . Since values are within the permissible range there is no multicollinearity between variable

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

- Temp(positively correlated)- The demand of bike is likely to increase with increase in Temperature

- Yr(positively correlated)- The demand of bike seem to be more in 2019 compared to 2018. It is likely to increase each year
- Weathersit_3 (Negatively correlated) – The demand of Bike is likely to decrease in - Light Snow, light rain+ Thunderstorm + Scattered clouds, Light rain+ Scattered cloud

General Subjective Questions

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

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables .It finds how the value of the dependent variable is changing according to the value of the independent variable.

If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.

The best fit line in linear regression is calculated by using the slope – intercept formula

$$y = b0 + b1x$$

Where y is the dependent variable

X is the independent variable

b0 is the intercept of the line

b1 is the linear regression coefficient

The goal of the linear regression algorithm is to get the best values for b0 and b1 to find the best fit line and the best fit line should have the least error.

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination

2. Residual Standard Error (RSE)

R2 or Coefficient of Determination -it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

Where RSS is Residual sum of squares

TSS is sum of error of data from mean

$$RSS = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

$$\mathbf{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Residual Standard error (RSE) – It is the estimate of the standard deviation of irreducible error (the error which can't be reduced even if we knew the true regression line; hence, irreducible). In simpler words, it is the average deviation between the actual outcome and the true regression line.

The below Assumptions are taken into consideration in a linear regression model.

Linear relationship between X and Y

2. Error terms are normally distributed (not X, Y)

3. Error terms are independent of each other

4. Error terms have constant variance (homoscedasticity)

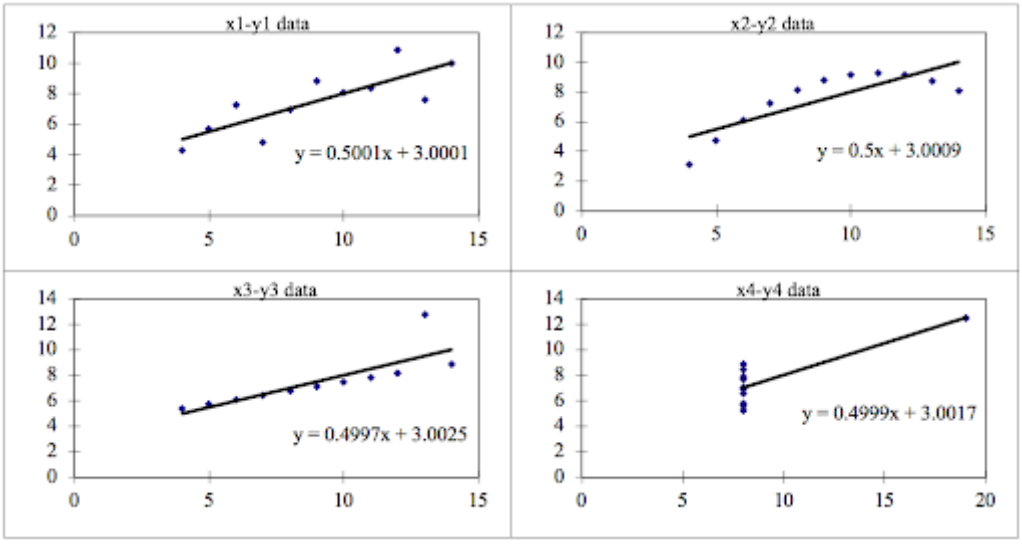
Once the a straight line on the data is fitted , hypothesis testing is required to be performed for the Beta coefficient

The model is assessed considering the R2 value , t-statistics, F-statistics values, adjusted R2 , VIF values(multiple Linear regression)

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. Below is the details

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



Data set 1: fits the linear regression model very well

Data set 2 cannot fit the linear regression model

Data set 3 : shows the outliers involved in the data set, which cannot be handled by the linear regression model

Data set 4:Doesnot fit linear regression model

Hence ,Anscombe's quartet says about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC

It measures the linear correlation between two variables. It lies between -1 to +1.

Pearson r value of −1 indicates a perfect negative linear relationship between variables,

Pearson r of 0 indicates no linear relationship between variables,

Pearson r r of 1 indicates a perfect positive linear relationship between variables

It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

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

Scaling is a step of data-pre-processing which is applied to the independent variables to normalize the data to a particular range .

Many times there is a huge difference in the data range of different independent variables. The algorithm only takes the value of the data not the units . Hence to speed up the calculation of the algorithms , scaling is performed.

Difference between Normalized Scaling and Standardized Scaling-

Normalized scaling is also called as Min-Max scaling. It brings all the values to a range of 0 and 1.

It is affected by outliers. Scikit-Learn provides a transformer called MinMaxScaler for Normalization. It is useful when we don't know about the distribution

Standardization brings all the data to a standard normal distribution which has a mean zero and Standard deviation 1. It is much less affected by outliers.. Scikit-Learn provides a transformer called StandardScaler for standardization. It is useful when the feature distribution is Normal or Gaussian.

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

(3 marks)

This happens when there is a perfect correlation between two independent variables. The corresponding variable can be expressed exactly by a linear combination of the other variables . In this case the R2=1 which implies 1/(1-R2) is infinity.

We need to drop one of the variables from the dataset which is causing this multicollinearity.

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

(3 marks)

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.