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.

The demand of bike is less in spring when compared with other seasons
• The bike rentals increased in the year 2019 when compared with year 2018.
• Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
• Bike demand is less in holidays in comparison to not being holiday.
• The demand of bike is almost similar throughout the weekdays.
• There is no significant change in bike demand with working day and non-working day.
• The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall.

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

Ans.

When we have categorical variables, we can convert them to binary using the get_dummies(). Dropping the first columns as (p-1) dummies can explain p features.

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

Ans.

temp and ***atemp*** have 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.

The model has a high adjusted R squared value and the dependable variables show a p value less than 0.05. This helps in understanding the linear relationship between X and Y variables
● Plotting the Error terms by calculating the difference between ground truth Y value and predicted Y value on train dataset. The distribution plotted shows a normal distribution which in turn favours the assumptions on linear regression
● The mean of error terms are also checked to be 0

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

Ans.

Top 3 features:
1. Temp: Positive Correlation and highest correlation value among all the features
2.Year (yr): Positive Correlation and also higher as compare to other features.
3. WeatherSit (Light Snow and Mist) : Negative Correlation, as it decreases bike sharing demand increases

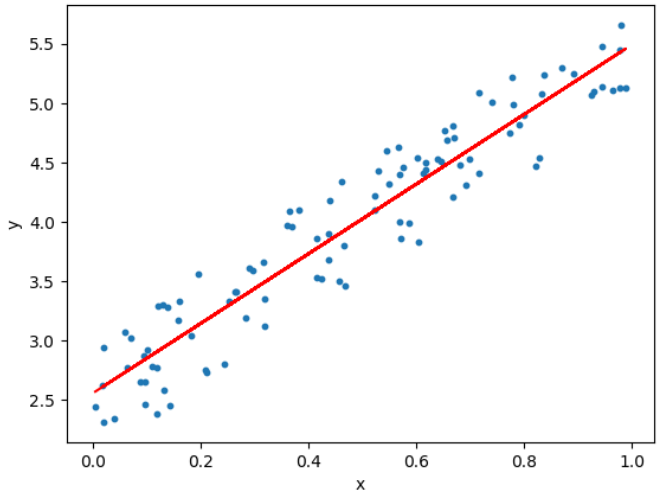
General Subjective Questions

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

Ans.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It shows how one dependent/target variable is linearly related to other independent/predictor variables. Linear regression helps us to predict about continuous/real or numeric variables such as profit, price of any product, sales count etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables.



Mathematically, we can represent a linear regression as:
y= a0 + a1x + ε

Here,
y = Dependent variable (Target Variable)
X = Independent variable (Predictor Variable)
a0 = Intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value)
ε = Random error

The motive of the linear regression algorithm is to find the best values for a0 and a1. The same concept is followed for multiple regressions when you have multiple variables.

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error. The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so to calculate this we use cost function.

Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing. We can use the cost function to find the accuracy of the mapping function. This mapping function is also known as Hypothesis function.
For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

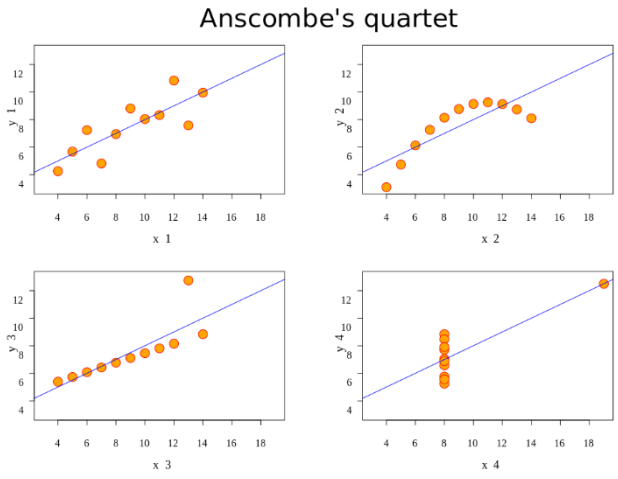
Gradient Descent: It is used to minimize the MSE by calculating the gradient of the cost function. In gradient descent, to update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to a0 and a1 is. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Once we converge on the minima, we can plot the line for the coefficients and see that it fits the data

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

Ans.

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization that signify both the importance of plotting data before analysing it with statistical properties.



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

Data sets for the 4 XY plots

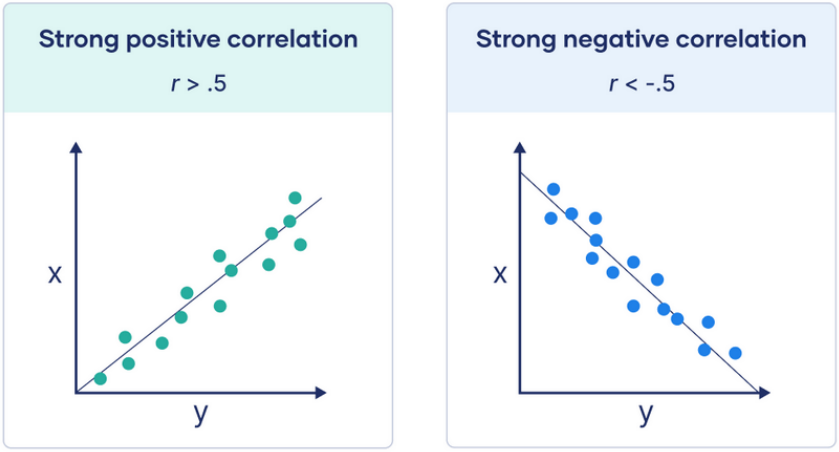
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- Data-set I – consists of set of (x,y) which represents linear relationship with little variance.
- Data-set II – shows non-linear relationship between (x,y)
- Data-set III – looks like a tight linear relationship for (x,y), except one large outlier
- Data-set IV – shows that value of x remains constants, except for one outlier

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

Ans.

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



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

Ans.

Scaling is a pre-processing step which converts all the independent features into a normalize/standardize/comparable scale. This technique is performed so that during model building the values of independent variables are brought to the same range and the coefficients of the model would be more comparable and not extreme due to the different sized values of the variables. Keeping the data scaled also helps in cleaner visualization and analysis.

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

(3 marks)

Ans.

Variance Inflation Factor calculates how well one independent variable is explained by all other independent variables combined. It measures the multicollinearity among the independent variables while doing Multiple Linear Regression.
When there is perfect correlation between two independent features then we get VIF = infinity. In perfect correlation we have r-squared = 1
So, by the formula:

VIF = 1/(1-r2)

= infinite (1/0 = infinite)

An infinite VIF value denotes that the corresponding feature may be expressed exactly by linear combination of other features which shows an infinite VIF as well.

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

(3 marks)

Ans.

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. A quantile is fraction where certain values fall below that quantile.
For example, median is a quantile where 50% data lie above this and 50% data fall below this quantile. The purpose of Q-Q plot is to determine whether two datasets come from same distribution. Whenever we interpreting a Q-Q plot, we shall concentrate on the ‘y = x’ line. We also call it the 45-degree line is statistics.