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)

Answers:-

- From the analysis of the categorical variable we can clearly say that the demand of bike is very less in the month of spring as per other months.
- In the fall month bike demand is high as we can see from the graph
- When we compare 2019 with 2018 bike demand is higher in the year 2019 as per graph
- When the weather is clear the bike demand is high but when there is a rain or snow bike demand is low we can see from the graph
- In the month of May to Octobers is show high in bike demand
- There is surprise data from the graph that if the is working or not there is no change in the bike demand
- In weekday also bike demand is approximately similar as per graph.

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

Answers:-

- It is important to drop dummy variable because it reduce the collinearity between dummy variable.
- It is important to delete the dummy variable as extra columns will get deleted and some time it misleads our data for prediction.
- In dummy variable there might be some variable which does not give any meaning so we can drop that column and keep that column which contains the same information.
- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Answers:-

- Temp and atemp as highest correlation with the target variable among the all numerical variables as per heat map.

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

Answers:-

- Error terms should be normally distributed
- The Two variable should be in a linear relationship
- All the variables should be multivariate normal
- There should be no multicollinearity in the data set

- There should be no auto correlation int the data
- There should be Homoscedasticity among the data

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

Answers:-

- Temperature
- winter
- sep

General Subjective Questions

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

Answers:-

Linear regression in Machine Learning is a supervised algorithm and the most used regression algorithm. In simple words, linear regression means fitting the best fit line between independent and target variables with the least mean square error.

Before implementing linear regression, we should check whether the data is following these assumptions:

- Data should be linear
- No Multicollinearity
- No auto-correlation
- Homoskedasticity should be there

- Linear regression is one of the very basic forms of the Machine learning where we train a model to predict the behaviour of the data based on some variable .In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

  Mathematically, we can write a linear regression as:
  y=a+bx
  Where a and b given by the formula below,

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Hence, x and y are the two variables on the regression line

b= Slope of the line.

a=y-intercept of the line.

x=independent variable from dataset.

y=Dependent variable from dataset.

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

Answers:-

- Anscombe's quartet comprises four dataset that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consist of even(xy) points .They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outlier on statistical properties.

  Simple understanding:-

  Once Francis John 'Frank' Anscombe who was a statistician of great reputed found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points .Those 4 sets of 11 data-points are given below

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

  After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

- Apply the statistical formula on the above data-set,

  Average Value of x = 9

  Average Value of y = 7.50

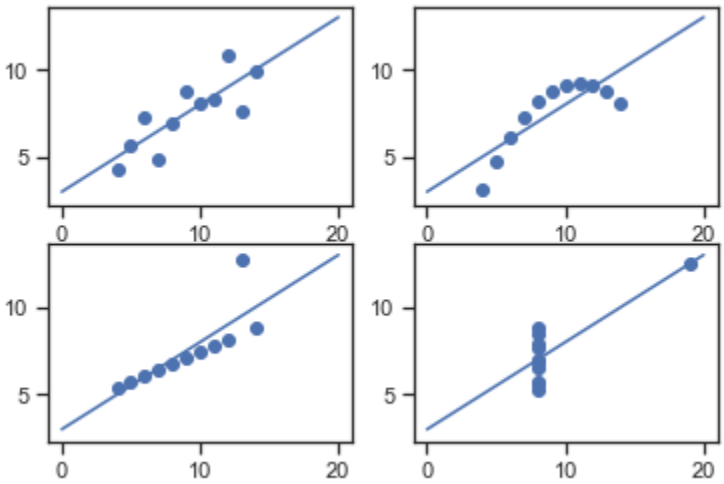
  Variance of x = 11

  Variance of y =4.12

  Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



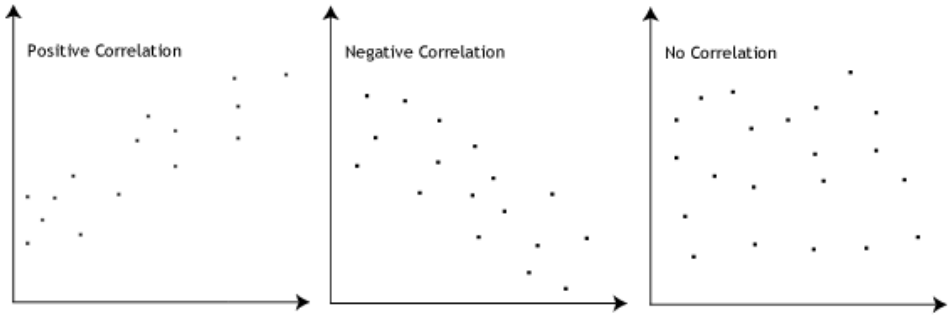
Graphical Representation of Anscombe's Quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

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

Answers:-

- Pearson's 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 variable tends 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.
  The Person 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 variable . A value greater than 0 indicates a positive association; that is, as the value of one variable increase, 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 increase ,the value of the other variable decrease .This is shown in the diagram below;



Pearson r Formula

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

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

Answers:-

- It is a step of data pre-processing which is applied to independent variable to normalize the data within a particular range. It also help in speeding up the calculation in an algorithm.
- Most of the time ,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 variable to the same level of magnitude
- It is important to note that scaling just affect the coefficient and none of the other parameters like t-statistics, F-statistics, p-values, R-squared ,etc.

Normalized scaling:-

1) Minimum and the maximum value of the feature are used for scaling
2) It is when features are of different scale
3) Scale value between [0,1] or [-1,1]
4) It is really affected by outliers
5) Scikit-learn provides a transformer called minmaxscalar for normalization.

Standardizes scaling:-

1) Mean and standard deviation is used for scaling.
2) It is used when we want to ensure zero mean and unit standard deviation
3) It is not bounded to a certain range.
4) It is much less affected by outliers.

5) Scikit-learn provide a transformation called StandardScaler for standardization

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

Answers:-

- If there is perfect correlation , then VIF = infinity. This show a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 = 1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variable from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of the other variable (which show 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

Answers:-

- The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.
- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions. This plot shows if residuals are normally distributed.
- Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.
- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.
- QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator is Gaussian either, so the standard confidence intervals and significance tests are invalid.