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?

Ans.

- Season: There is a positive trend in number of customers in summer, fall and winter.
- Year: Very good increase in the count of customers showing in the year 1.
- Month: The trend in customer count is similar to the season.
- Working day: Higher count in weekend and holidays compared to other days. But a wider spread is there in working days
- Weathersit: Clear weather shows good a positive trend in count other than other weather situations
- Weekdays: This is showing similar tend in count ,Saturday and Wednesday is having wide spread in count.
- Holiday: on holidays, there is positive trend in bike counts.

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

Ans.

There will extra column created during the dummy variable creation. So **drop_first=True** command will remove the extra columns created and reduces the correlation created between the dummy variables.

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

Ans.

Temp

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

Ans.

- Residual analysis. Plotted the residual graph and found that residuals are normally distributed.
- Calculated the r2_score of train data, got the value as 82.15% on train data.
- Calculated the Mean Squared Error value, got this value close to zero.

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

Ans.

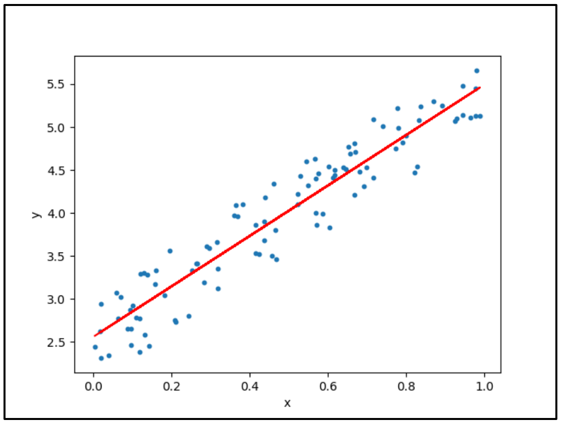
- Yr, season and month

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans.

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. 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 above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

y=mx+b ---→ y=a0+a1x

Y = Dependent Variable.

x = Independent Variable

2. Explain the Anscombe's quartet in detail

Ans.

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 mea**n** of all x,y points in all four datasets

This tells us about the importance of visualizing 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 only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

3. What is Pearson's R?

Ans.

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

- **Pearson r Formula**

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

r =correlation coefficient

xi=values of the x variable in a sample

x=mean of the values of the x variable

yi= values of the y variable in a sample

y= mean of the values of the y variable

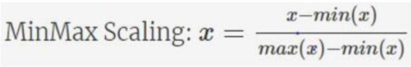
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans

Scaling 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.

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 must 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.

Normalization/Min-Max Scaling:



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

- It brings all the data in the range of 0 and 1.

Standardization Scaling:

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

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

Ans.

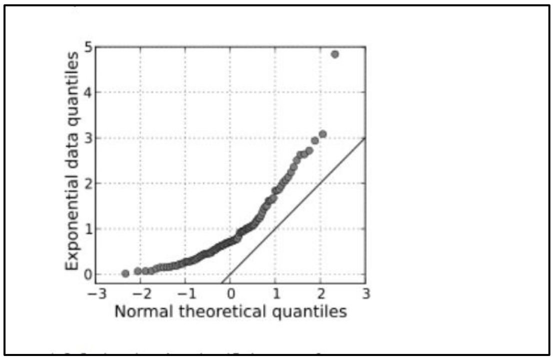
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (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.

Ans.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. 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.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a normal quantile-quantile (Q-Q) plot**.** The points are not clustered on

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

the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

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