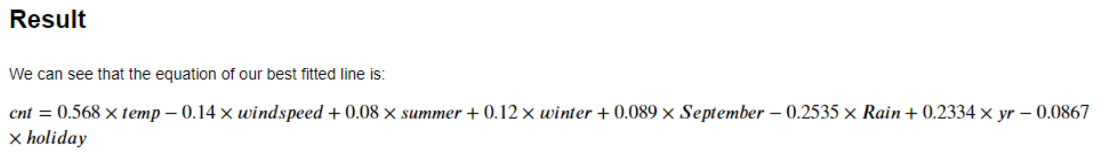
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:-

Below are the categorical variables

categorical variables:- Season, Mnth, Weekday & weathersit



Result

We can see that the equation of our best fitted line is:

$$cnt = 0.568 \times temp - 0.14 \times windspeed + 0.08 \times summer + 0.12 \times winter + 0.089 \times September - 0.2535 \times Rain + 0.2334 \times yr - 0.0867 \times holiday$$

From the Model equation we can able to infer that

Season = During summer the count is may increase

Month = September month plays vital role in increasing the count

Weekday = Saturday plays vital role in increasing the count

Weathersit = during Rain season the get downfall for counts.

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

Ans:-

Drop_first=True will drop the first column for that variable.

Example for Season column

You will have 4 values in the season column. But you will be getting only 3 variables when you specify Drop_first=True

You can drop the `fall` column, as the type of fall can be identified with just the last 3 columns where —

- `100` will correspond to `spring`
- `010` will correspond to `summer`
- `001` will correspond to `winter`
- `000` will correspond to `fall`

By this we can reduce the no of columns, avoid complexity, avoid overfit model etc.

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

Ans:- Temp has the highest correlation with the Target variable

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

Ans:-

- Residual →following the normal distribution and the mean value is Zero.
- R2 score→ From the modal the R2 value for train is 80% and for test is 77%.
- co-efficient value→From the modal,temp has the having high coefficient (0. 0.568) and which affect the count.

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

Ans:- From my model equation TEMP, YEAR, Winter play vital role in the explaining the demand of the shared bikes.

TEMP Co-efficient = 0.5854

YEAR Co-efficient = 0.2329

Month (Sep) Co-efficient = 0.0822

Weather(summer) Co-efficient =0.0750

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:- Linear regression is a type of statistical modeling that allows you to investigate the relation between the dependent and independent variable

Simple Linear Regression:- It will be used when you have one dependent variable and the relation to the independent variable has be found.

y = b0 + b1 * x1

y is the dependent variable.

x is the independent variable. We assume that this is causing the dependent variable to change in some way. This might not be a direct cause, but it implies an association between the variables

b1 is what’s known as the coefficient for the independent variable.

b0 is the constant term, or point where your trendline crosses the horizontal axis.

For example : predicting Sale value by using market spend.

Y = Sale values which is dependent Value

X = Market spend values which is independent value

How much value getting spend on the market , that much values can be expected in the sale, So sale may get increase or decrease it purely depending on how much getting spend in market .Therefore sale is dependent variable and Market spend is independent variable.

Once we find the best bo and b1 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

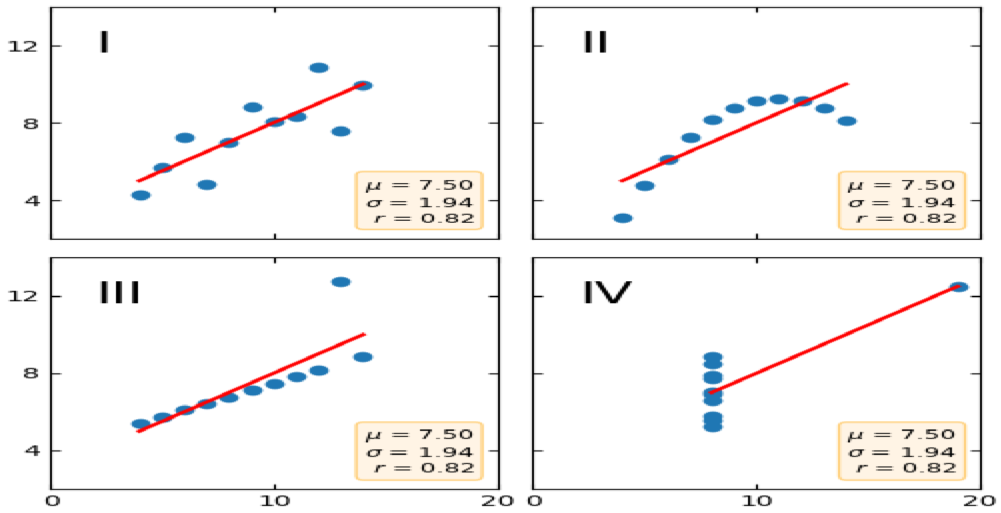
Assumption needs to be considered:

- Linear relationship between X & Y.
- Error terms are normally distributed (not x & y).
- Error terms are independent of each other.
- Error terms have constant variance.
- No Assumption on the distribution of X or Y.

2) Explain the Anscombe's quartet in detail.

Ans:- Anscombe's Quartet consists of four data sets, that when examined have nearly the identical statistical properties, yet when graphed the datasets tell a very different story. Each dataset consists of eleven (*x*,*y*) points.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



What is Pearson’s R?

The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables.

It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.

- r = -1 == > data lie on a perfect straight line with a negative slope
- r = 0 == > no linear relationship between the variables
- r = +1 == > data lie on a perfect straight line with a positive slope
- 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

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

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Why its performed :

So it is extremely important to rescale the variables so that they have a comparable scale.

If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation.

Easy of interpretation of coefficient, not p-value or model accuracy etc.,

Faster convergence for Gradient descent methods

So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

Min-Max scaling (normalization) : value between 0 & 1 (max =1 & min = 0)

(x - xmin) /(xmax- xmin)

Standardization (mean-0, sigma-1)

Std = x- mean / Sigma(sd)

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

Ans :

VIF is detection of multicollinearities by means of the variance inflation factor (VIF).

VIF values should be below < 5 == > moderate and considerable

VIF > 5 ==> severe correlation and exclude those variables from the model

The VIF can be conceived as related to the R-squared of a particular predictor variable regressed on all other includes predictor variables.:

VIF of X1 = 1/(1 - R-squared of X1 on all other Xs).

If you only have 1 X or that X is orthogonal with all the other Xs; then
VIF = 1/(1-0) = 1 - so no variance inflation.

If two Xs are perfectly correlated
VIF = 1/(1-1)= 1/0 = infinity that is the estimate is as imprecise as it can be.

The VIF is efficiently calculated (not by running a series of regressions) but as the diagonal element of the inverse of the correlation matrix of the predictors.

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

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.

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.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?