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

ANS:

• Year wise : In year 2019 we have around 62.33 % of total Bike sharing i.e. Bike sharing is increasing with the years and the with the increase in the age of Bike Sharing Company

• Month wise : Jan , Feb and March saw a great rise in Bike Sharing which is more then 100 percent each

• Season wise : Highest no of bike sharing will be in fall season and then in summer and then in winter and spring have least no of bike sharing.

• Weather-sit : Company Many Notice increase in Bike sharing count for below weather

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

Weather-sit : Bike sharing count will decrease with snow, rainfall and thunderstorm .

• Holiday : Bike sharing is more on non holidays so it seems people spend time with family during holidays or may be using private Vehicles.

• Working and non working days are almost uniformly distributed but bike sharing is slightly more on the working day

• With the clear weather (cloudy & Misty) during Fall & Summer season between May to Oct month for the year 2019 with moderate temp , low humidity and less windspeed we can experience more bike sharing.

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

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multi collinear) and one variable predicts the value of others. When we use one-hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a dummy variable trap. So, the regression models should be designed to exclude one dummy variable that's why we us drop_first=True during dummy variable creation For : E.g. Let's consider the case of gender having two values male (0 or 1) and female (1 or 0). Including both the dummy variable can cause redundancy because if a person is not male in such case that person is a female, hence, we don't need to use both the variables in regression models. This will protect us from the dummy variable trap.

Q3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

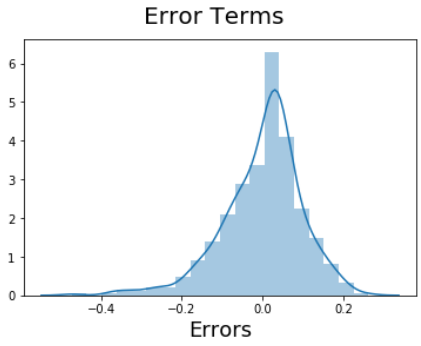
temp & atemp have highest correlations(63 Percent) with the cnt (i.e. target) variable.

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

Using Residual Analysis and checked wether the Error terms are normally distributed with mean zero using distplot.

```
In [57]: # Plot the histogram of the error terms
         fig = plt.figure()
         sns.distplot((y_train - y_train_pred), bins = 20)
         fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
         plt.xlabel('Errors', fontsize = 18)

Out[57]: Text(0.5, 0, 'Errors')
```



We will also calculate r2 in both train and test.

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

Temperature (Temp)

-A coefficient value of `0.478564` indicated that a temperature has significant impact on bike rentals

Light Rain & Snow (weathersit =3)

-A coefficient value of `-0.291333` indicated that the light snow and rain deters people from renting out bikes

Year (yr)

-A coefficient value of `0.233911` indicated that a year wise the rental numbers are increasing

- `So from 2018 to 2019 there has been 23% growth in terms of rental numbers`

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

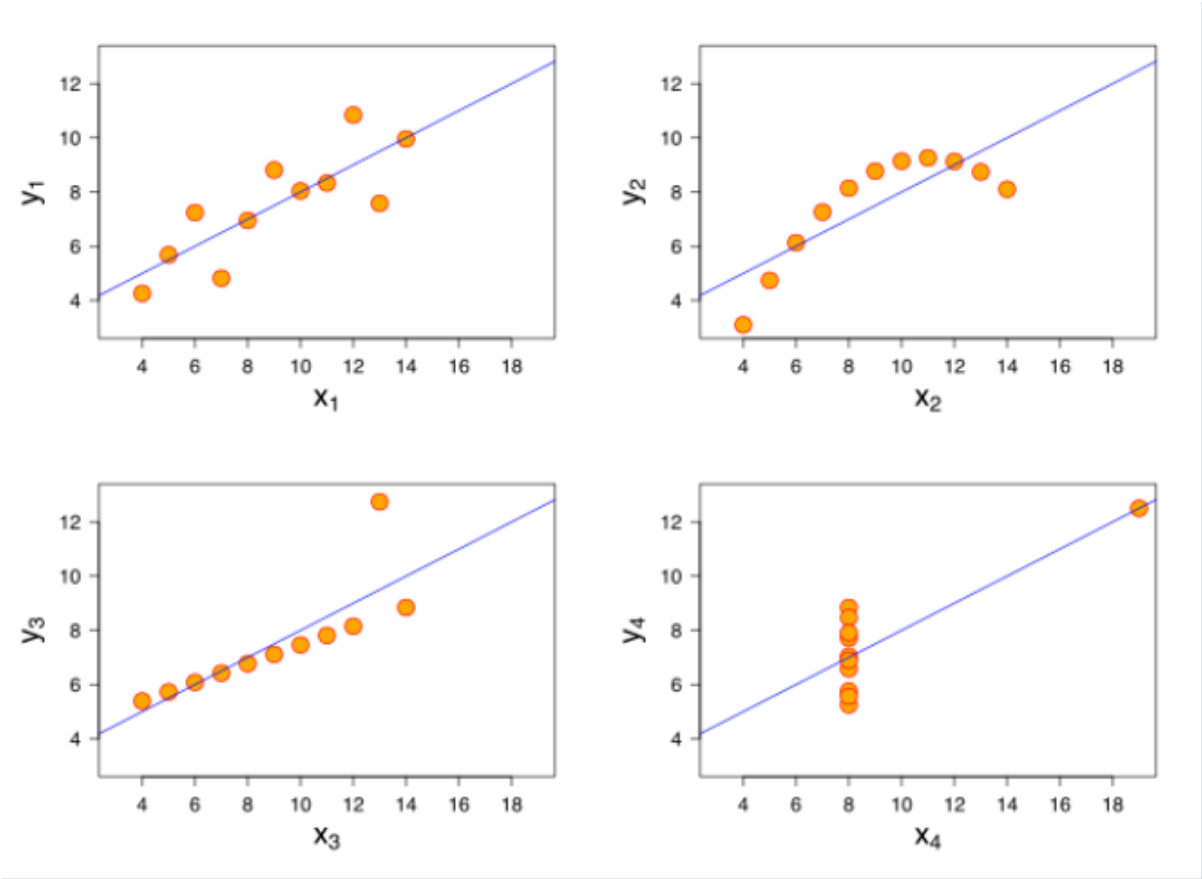
Regression analysis is nothing but a predictive modelling methodology that aims to investigate the relation that exists between independent variables or predictors and dependent variables or targets. This is done by fitting a line or curve to different data points in a way that we can minimise the difference in data point distances from the line, or the curve. Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum.

Types:

1. Simple Linear Regression
   Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.
2. Multiple Linear Regression:
   Multiple linear regression (MLR), also known simply as multiple regression, is **a statistical technique that uses several explanatory variables to predict the outcome of a response variable**. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable. Example Factors that implement rain

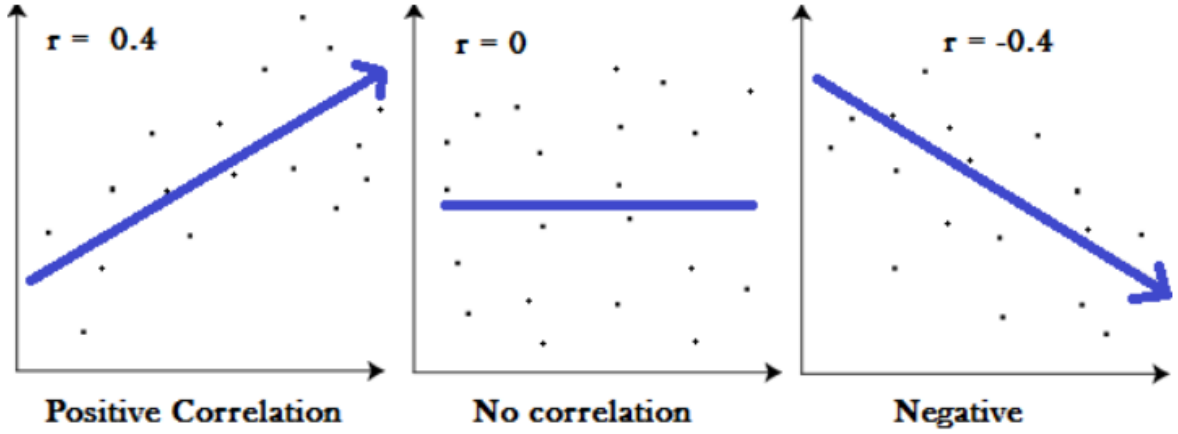
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. It has been rendered as an actual musical quartet.



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

The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.



Formula

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

Here,

- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =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? (3 marks)

Scaling is a technique of bringing down the values of all the independent features of our dataset on the same scale. Feature selection helps to do calculations in algorithms very quickly. It is the important stage of data preprocessing. If we didn't do feature scaling then the machine learning model gives higher weightage to higher values and lower weightage to lower values. Also, takes a lot of time for training the machine learning model.

Why Scaling

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

If a feature in the dataset is big in scale(Salary) compared to other(Age) then in model it gives higher weightage to higher values and lower weightage to lower values. this big scaled feature becomes dominating and needs to be normalised/standardised.

Standardising: The variables are scaled in such a way that their mean is zero and standard deviation is one.

x = x − mean(x) sd(x)

Use-case of Standardiser

- In most of the Machine Learning models.
- Anywhere, where there is no need to scale features in the range 0 to 1.
- Since, it transforms the normal data distribution to standard normal distribution, which is the ideal & expected to have, most of the time it is the best to use in machine learning models

Normalisation : The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data . Use-case of Normalisation: • Every situation where the range of features should be between 0 to 1.

For example, in Images data, there we have color pixels range from 0 to 255(256 colors in total), here Normaliser is the best one to use. • There can be multiple scenarios where this range is expected, there it is optimal to use MinMaxScaler.

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

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.

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. 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 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

- come from populations with a common distribution
- ii. have common location and scale
- iii. have similar distributional shapes
- iv. have similar tail behavior

Interpretation

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.