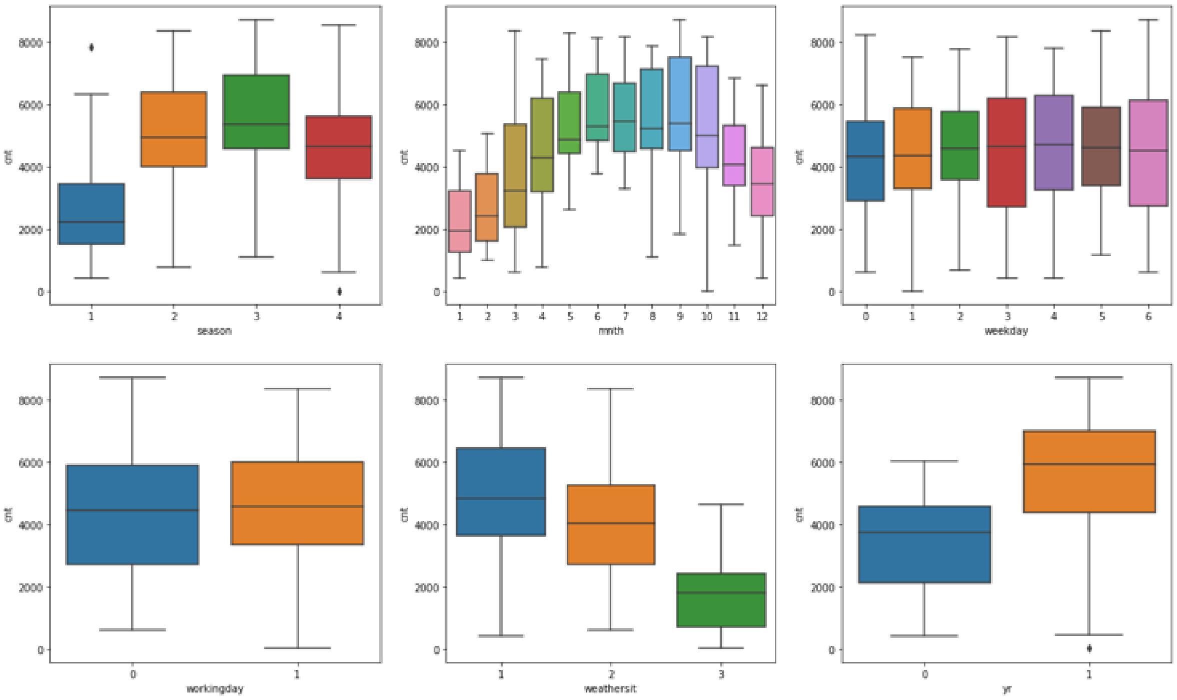
Assignment-based Subjective Questions

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

Ans:



From the box plot, we can easily conclude that:
1. Bike sharing count is more in season of summer and fall compared to winter and spring.
2. Bike sharing count is more in month of June to October than other months.
3. Bike sharing count is more when weathersit is Clear, Few clouds, partly cloudy, Partly cloudy.
4. Bike sharing count is more in year 2019
5. Bike Sharing count have no major impact of working day and weekday.

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

Ans: Using drop_first=True is more common in statistics and often referred to as "dummy encoding". When we convert the categorical variables to dummies, indirectly we are giving importance to each value in a categorical column by making each value as a column. If we don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted. For e.g. if we have a categorical column for gender for Male and Female, then after creating dummy variables , we can easily drop any one of column, as the other columns zero values will represents the deleted column values.

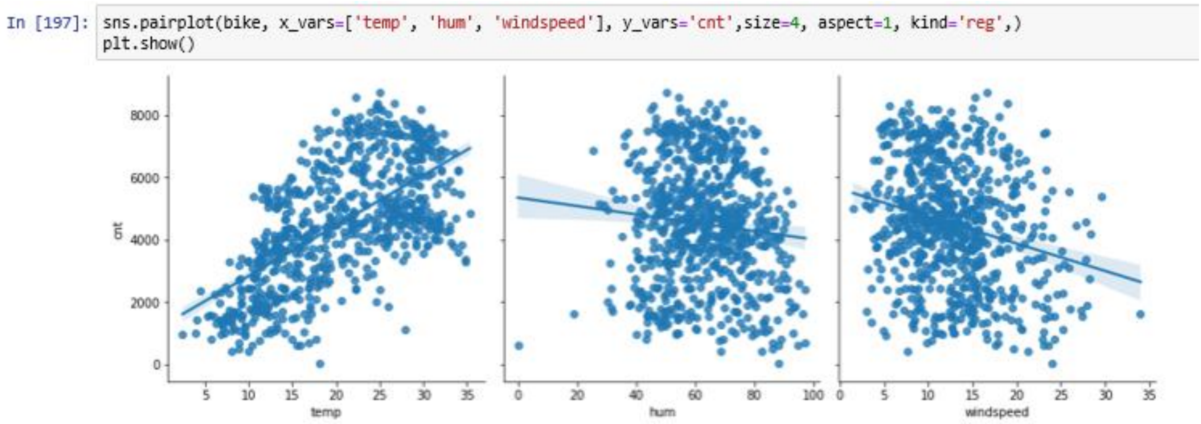
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: From pair plots, we can conclude that, Registered and Casual were highly correlated with cnt, that shows that both casual and registered are similar to cnt, hence they needs to be ignored in analysis as they directly relates to cnt.

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

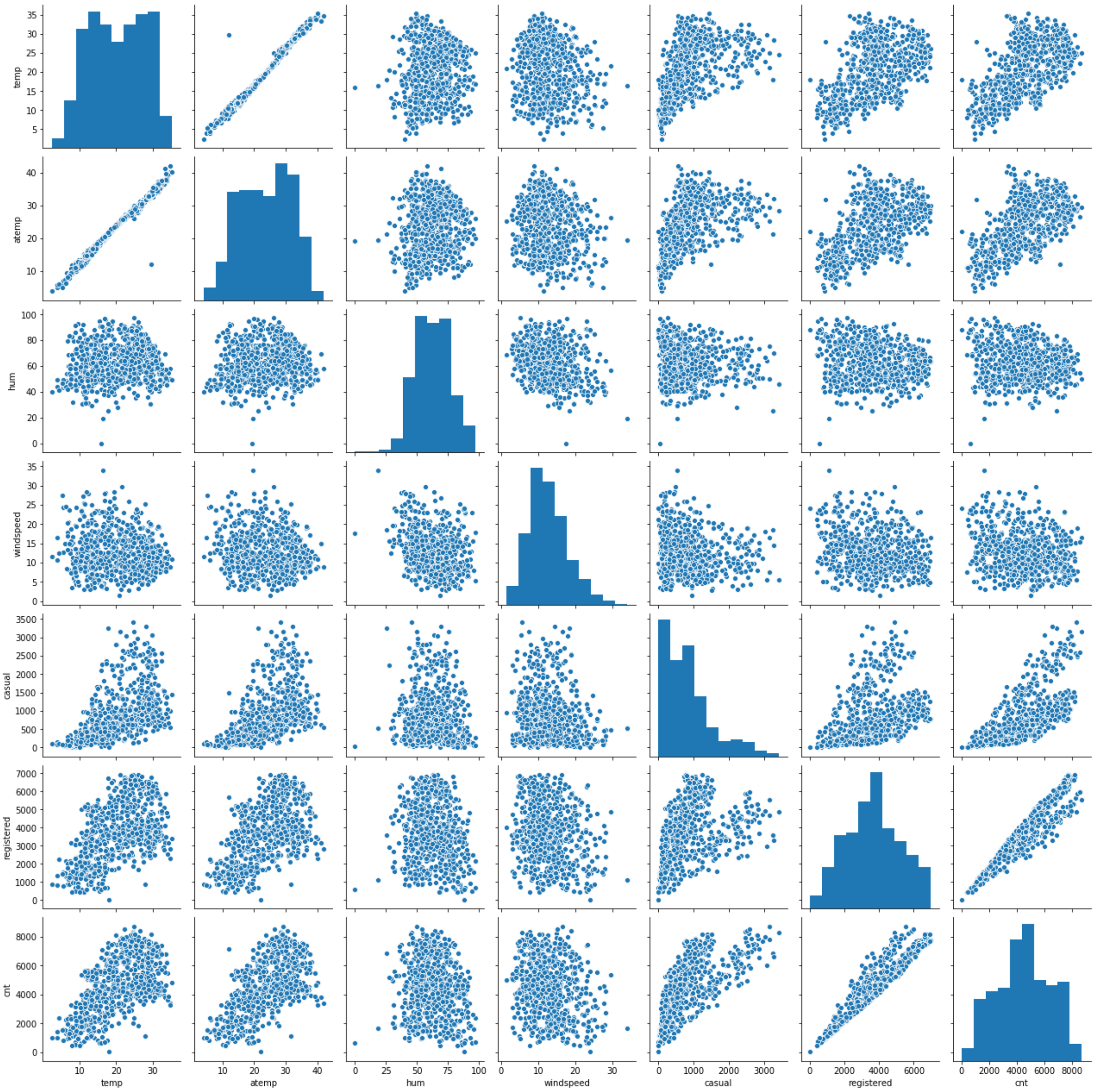
Ans: Assumptions of Linear Regression. There are 5 basic assumptions of Linear Regression Algorithm:

1. ***Linear Relationship between the features and target:** According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target. So below plot show that there is a linear relationship between 'cnt' and independent variables*



```
In [197]: sns.pairplot(bike, x_vars=['temp', 'hum', 'windspeed'], y_vars='cnt',size=4, aspect=1, kind='reg',)
          plt.show()
```

***2. Little or no Multicollinearity between the features:** Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables.It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model.Pair plots and heatmaps(correlation matrix) can be used for identifying highly correlated features.*

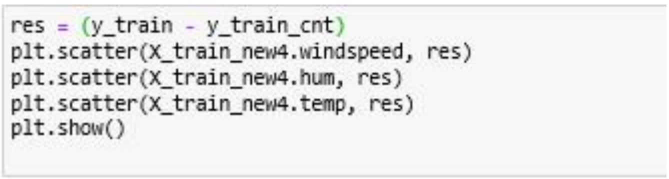


From pair plots, we can conclude that

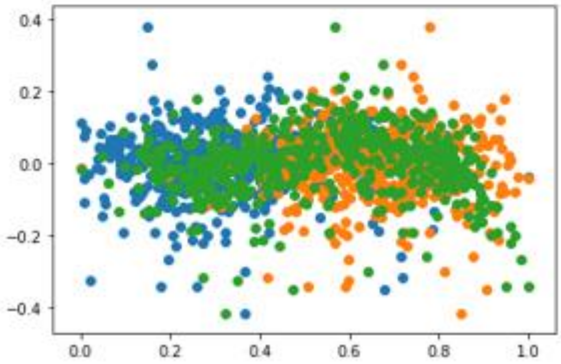
1. registered, casual and cnt are highly correlated, that shows that both casual and registered are similar to cnt, hence they needs to be ignored in analysis as they directly relates to cnt.

2. atemp and temp are highly correlated and hence we can drop one of them from the analysis as they act as redundant variables.

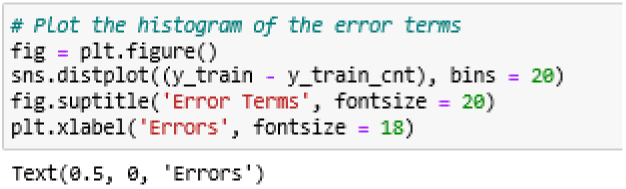
*3. **Homoscedasticity Assumption:** Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity.*



```
res = (y_train - y_train_cnt)
plt.scatter(X_train_new4.windspeed, res)
plt.scatter(X_train_new4.hum, res)
plt.scatter(X_train_new4.temp, res)
plt.show()
```

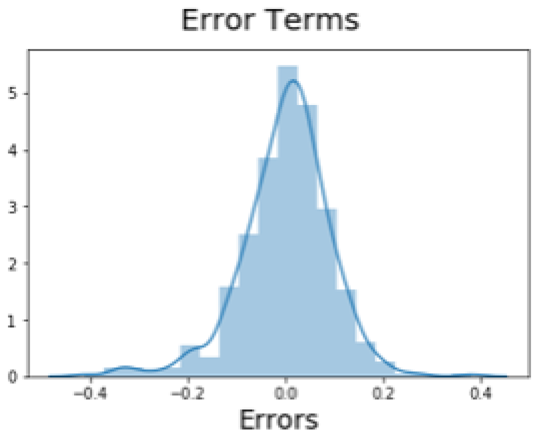


4. Normal distribution of error terms: *The fourth assumption is that the error(residuals) follow a normal distribution.*

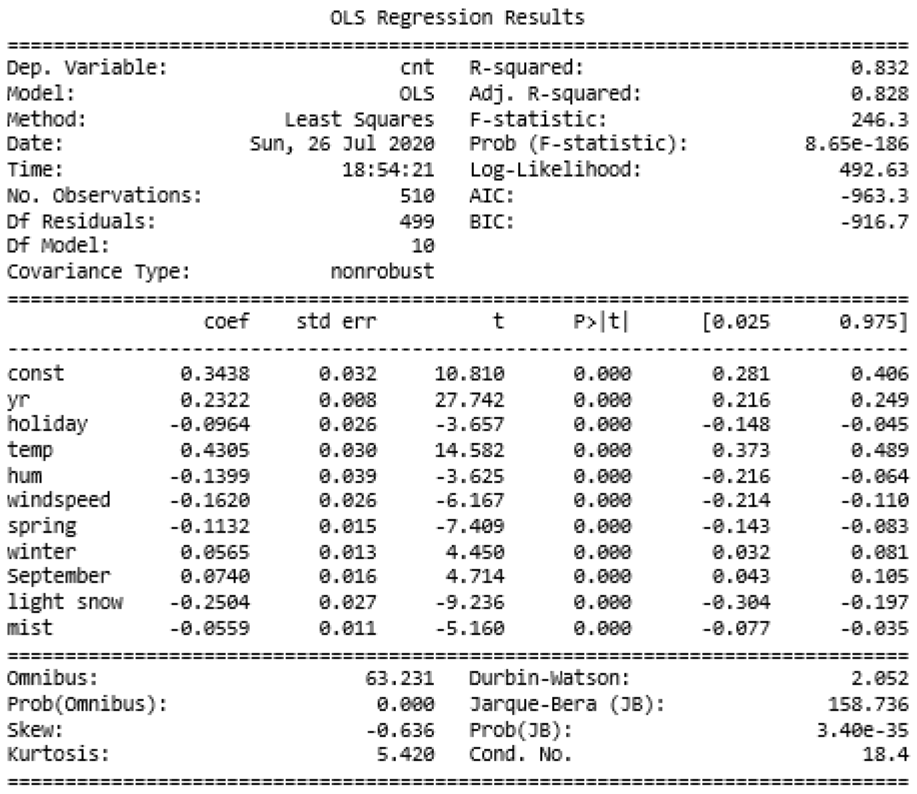


```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

```
Text(0.5, 0, 'Errors')
```



5. Little or No autocorrelation in the residuals: *Autocorrelation can be tested with the help of Durbin-Watson test. From the above summary note that the value of Durbin-Watson test is 2.052 quite close to 2 as said before when the value of Durbin-Watson is equal to 2, r takes the value 0 from the equation 2*(1-r),which in turn tells us that the residuals are not correlated.*



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.832
Model:                            OLS   Adj. R-squared:                  0.828
Method:                 Least Squares   F-statistic:                     246.3
Date:                Sun, 26 Jul 2020   Prob (F-statistic):         8.65e-186
Time:                        18:54:21   Log-Likelihood:                 492.63
No. Observations:                 510   AIC:                            -963.3
Df Residuals:                     499   BIC:                            -916.7
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3438      0.032     10.810      0.000       0.281       0.406
yr             0.2322      0.008     27.742      0.000       0.216       0.249
holiday       -0.0964      0.026     -3.657      0.000      -0.148      -0.045
temp           0.4305      0.030     14.582      0.000       0.373       0.489
hum           -0.1399      0.039     -3.625      0.000      -0.216      -0.064
windspeed     -0.1620      0.026     -6.167      0.000      -0.214      -0.110
spring        -0.1132      0.015     -7.409      0.000      -0.143      -0.083
winter         0.0565      0.013      4.450      0.000       0.032       0.081
September      0.0740      0.016      4.714      0.000       0.043       0.105
light snow    -0.2504      0.027     -9.236      0.000      -0.304      -0.197
mist          -0.0559      0.011     -5.160      0.000      -0.077      -0.035
==============================================================================
Omnibus:                       63.231   Durbin-Watson:                   2.052
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              158.736
Skew:                          -0.636   Prob(JB):                     3.40e-35
Kurtosis:                       5.420   Cond. No.                         18.4
==============================================================================
```

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: From Linear regression model, we can conclude that cnt = 0.2322 * Yr -0.0964 * holiday + 0.4305 * temp - 0.1399 * hum - 0.1620 * windspeed - 0.1132 * spring + 0.0565 * winter + 0.0740 * September - 0.2504 * light snow - 0.0559 * mist.*

Standardized coefficients signify the mean change of the dependent variable given a one standard deviation shift in an independent variable. thus coefficient of predictors can be compared to assess the importance. Thus based on high coefficient values below factor have highest influence:

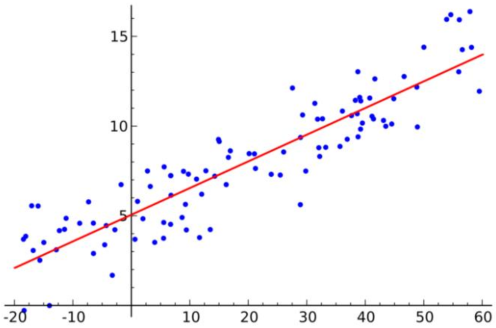
- *Temp*
- *Light Snow*
- *Yr*

General Subjective Questions

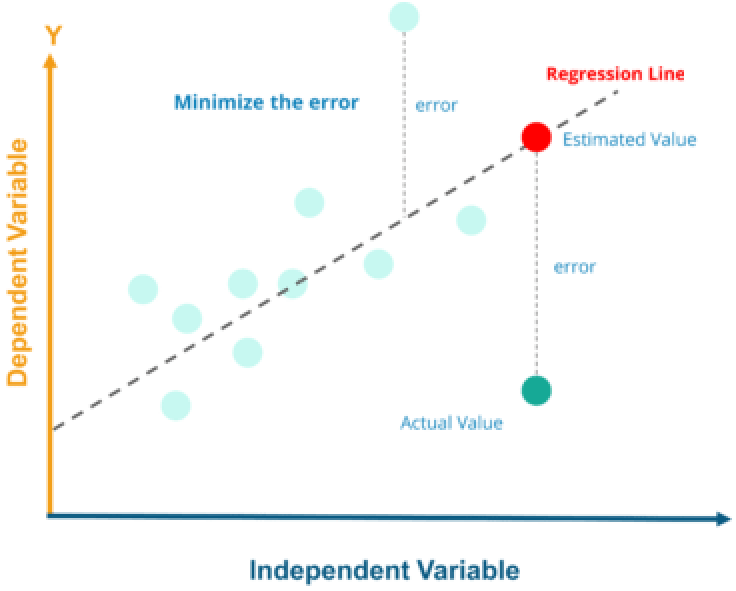
1. **Explain the linear regression algorithm in detail.**

Ans : *Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.*

Linear Equation => y = B1*X1 + B0



Finding the best fit line: *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***.**



1. ***Cost Function:*** *The cost function helps us to figure out the best possible values for B0 and B1 which would provide the best fit line for the data points. Since we want the best values for B0*

and B1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

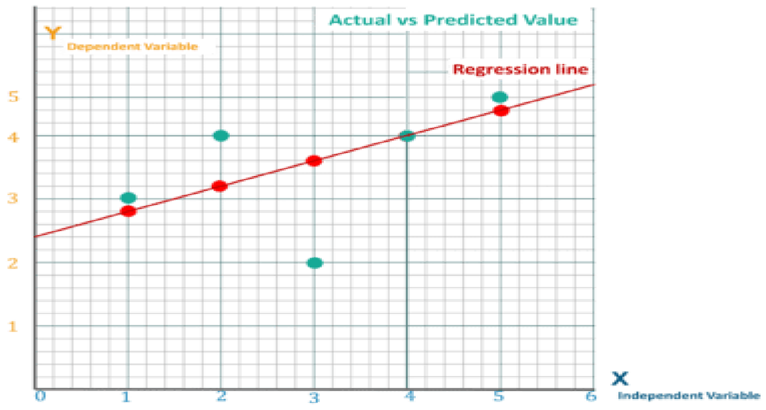
We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function.

2. *Gradient Descent:*

- *Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.*
- *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.*

3. ***Least Square Method:*** *Least squares is a statistical method used to determine the best fit line or the regression line by minimizing the sum of squares created by a mathematical function. The "square" here refers to squaring the distance between a data point and the regression line. The line with the minimum value of the sum of square is the best-fit regression line.*

4. ***R Square Method – R–squared*** *value is the statistical measure to show how close the data are to the fitted regression line*



y = actual value

y¯ = mean value of y

yp = predicted value of y

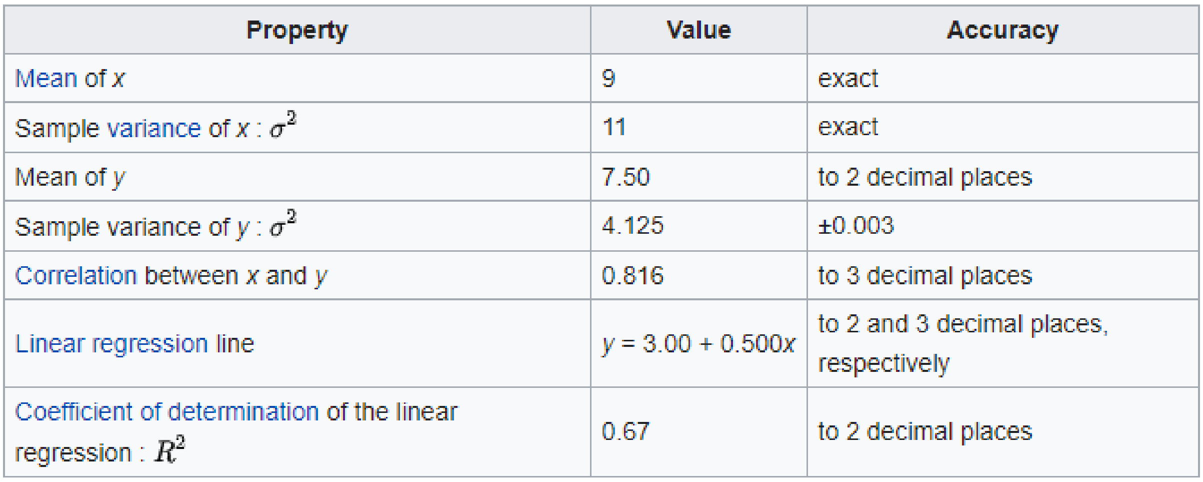
$$R^2 = 1 - \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

R-squared does not indicate whether a regression model is adequate.

2. Explain the Anscombe's quartet in detail.

Ans: *Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business. But there's a danger in relying only on summary statistics and ignoring the overall distribution. Anscombe's Quartet is one way saving us from the danger of summary statistics.*

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.The summary statistics show that the means and the variances were identical for x and y across the groups :

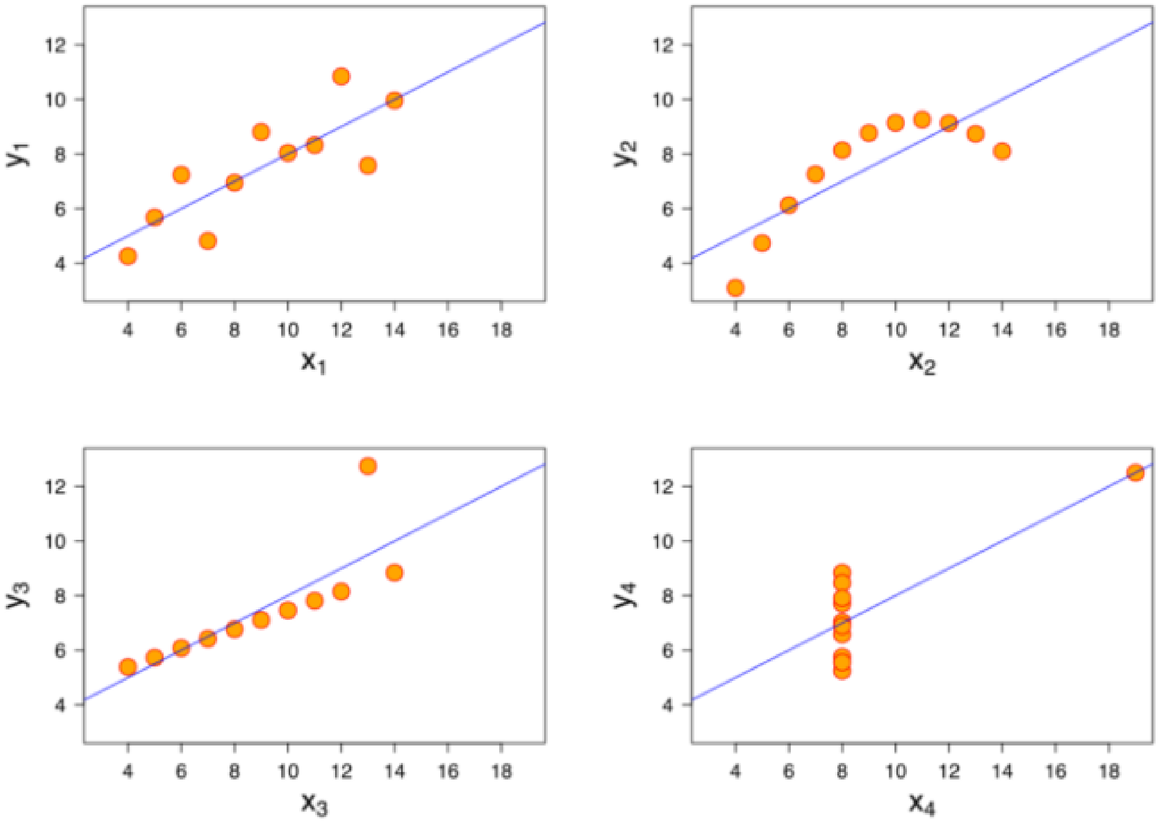


Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : σ^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : σ^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :

1. *Dataset I appears to have clean and well-fitting linear models.*

2. *Dataset II is not distributed normally.*
3. *In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.*
4. *Dataset IV shows that one outlier is enough to produce a high correlation coefficient.*



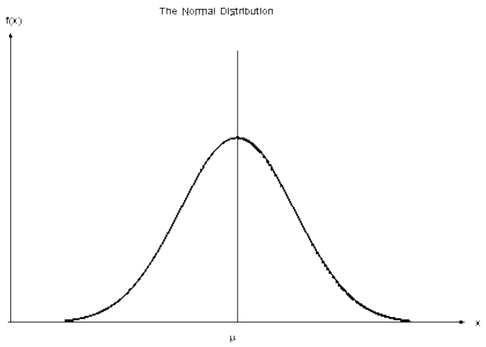
This quartet emphasizes the importance of visualization in Data Analysis.

3. What is Pearson's R?

Ans: *Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. a + sign indicates a positive relationship and a - sign indicates a negative relationship.*

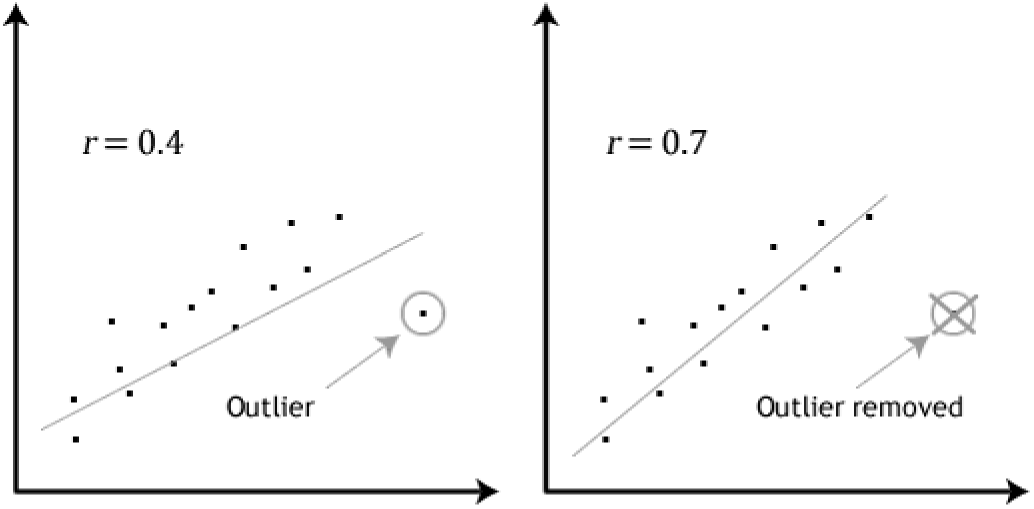
Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. You'll come across Pearson r correlation.

Assumptions



1. For the Pearson r correlation, both variables should be ***normally distributed.***

2. There should be ***no significant outliers***

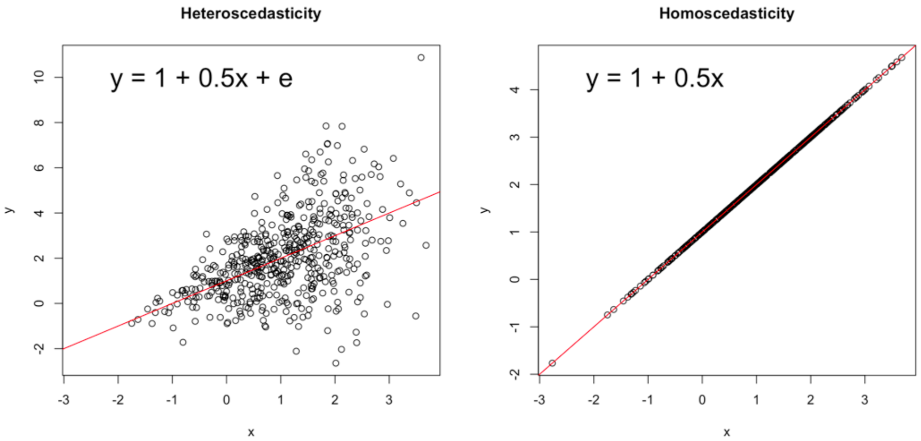


3. Each variable should be ***continuous*** *i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.*

4. The two variables have a ***linear relationship****.*

5. The observations are ***paired observations.***

6. ***Homoscedascity*** *Homoscedascity simply refers to '****equal variances****'. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic*



.

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

Ans: *Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.*

Techniques to perform Feature Scaling:

***Normalization** is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.*

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- *When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0*
- *On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1*
- *If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1*

***Standardization** is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.*

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ *is the mean of the feature values and* σ *is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.*

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

Ans: The variance inflation factor (VIF) *quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing* collinearity/multicollinearity*.*

$$VIF = \frac{1}{1 - R^2}$$

If VIF is equal to infinity then it means that residual is 1 and hence it shows that there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

You can assess multicollinearity by examining tolerance and the Variance Inflation Factor (VIF) are two collinearity diagnostic factors that can help you identify multicollinearity. Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variable's tolerance is 1-R2. A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some suggest that a tolerance value less than 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and non-significance, multicollinearity may be an issue.

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

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.
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. Residual plots and Q-Q plots are used to visually check that your data meets the homoscedasticity and normality assumptions of linear regression. Q-Q plots let us check that the data meet the assumption of normality. They compare the distribution of our data to a normal distribution by plotting the quartiles of our data against the quartiles of a normal distribution. If our data are normally distributed then they should form an approximately straight line.

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