**Assignment-based Subjective Questions**

***Q1** : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*

**Ans**: Season, weather situation, holiday, month, workingday and weekday were categorical variable , box plot was use to analyse them against target variable.

These influence dependent variable as below:

**Season :** During Summer/Fall season there is considerably more distribution and median of data

**Month :** We can see there is more Bike rentals taken during months of those seasons and hence months June - October are mostly months in which demand is higher. Demand is increasing each month till June. September month has highest demand.

**Weather Situation:** During heavy rain there is less demand, however its more during clear weather

**Holiday**: Holiday has less demand

**Weekday:** Weekend have more demand then weekday

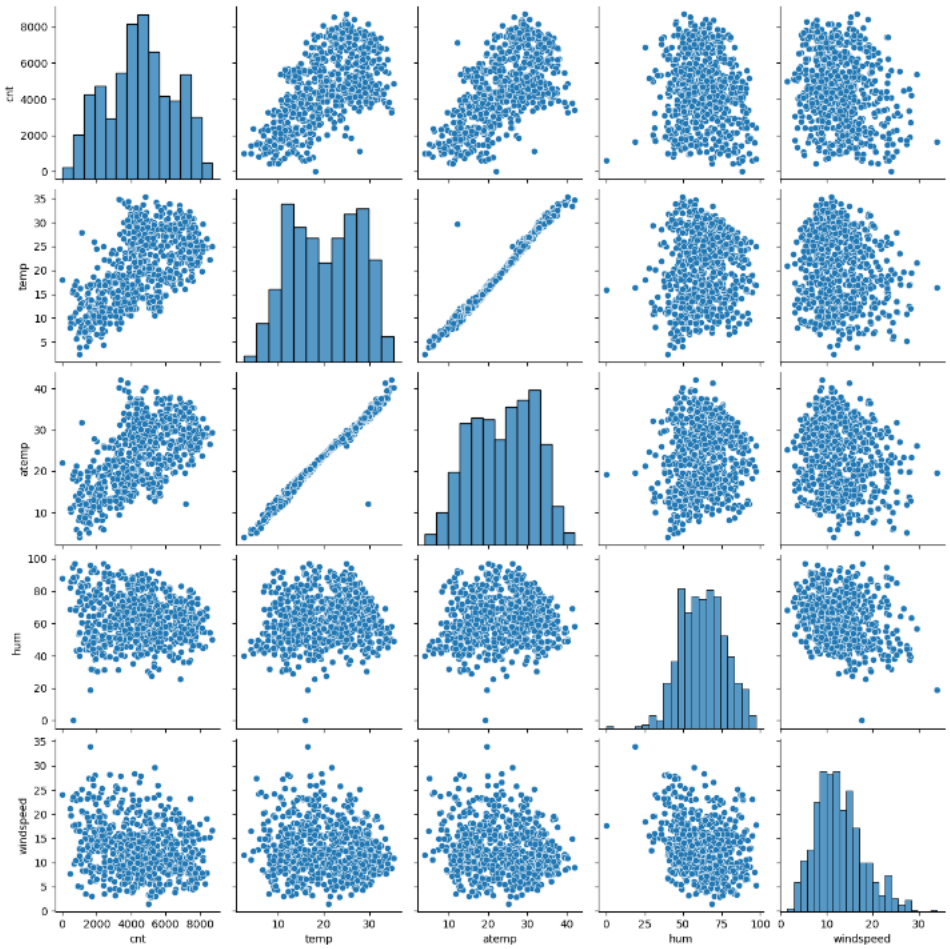
**Working day**: Little impact on dependent variable

***Qn2**: Why is it important to use drop_first=True during dummy variable creation?*

**Ans**: If we dont remove the 1 column then all dummy variable will be correlated, so we removed redundant column.Also if all dummy variable will be kept it can result in multicollinearity between them so we loose a column to have data in proper relation.

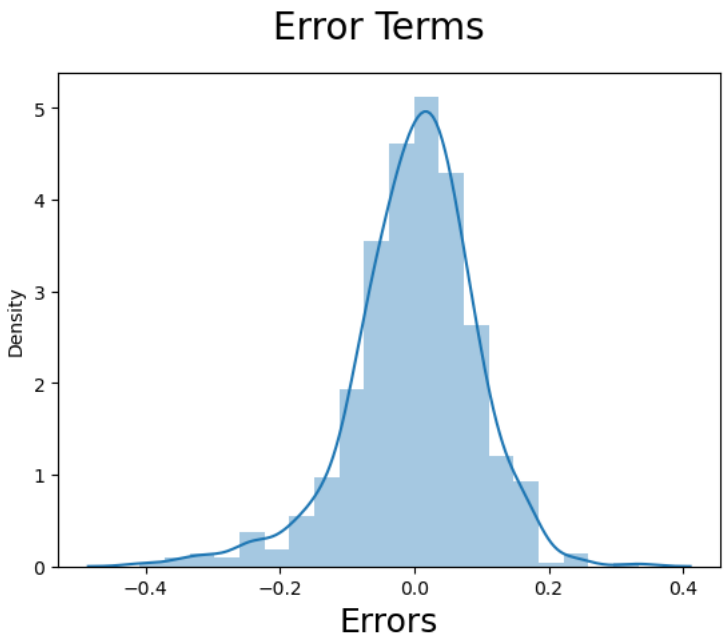
***Qn3**: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?*

**Ans**: As below parplot, Temp and aTemp has highest correlation;




***Qn4**: How did you validate the assumptions of Linear Regression after building the model on the training set?*

**Ans**: By preparing residual error diagram :




```
39]:  # residual is centeretd around 0, thickness around 0.4 on left
      # Errors are normally distribured here with mean 0. So everything seems to be fin
```

As show above most of the values are around mean 0.

***Qn5**: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?*

**Ans**: **Temperature :** With  coefficient of 0.5677 which is highest, so if temp will increase then bike rental will also increase by 0.5677 unit

**Year**: With coefficient of 0.2296 , bike rental will increase with unit of 0.2296 with year

Weather situation Light snow: With coefficient of -0.2103 , unit increase in this weather will decrease demand of bike

**General Subjective Questions**

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

**Ans**: Linear regression is a type of supervised Machine learning algorithm used for the prediction of numeric values. Regression most commonly used predictive analysis model.

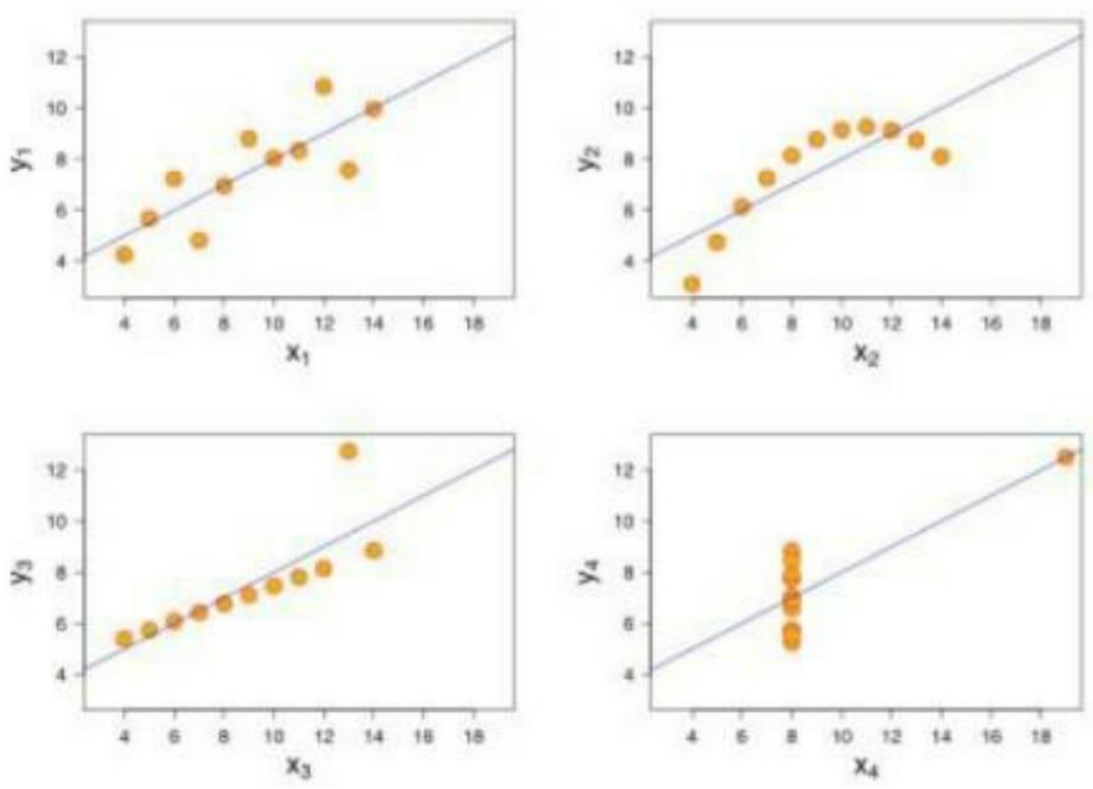
Equation of Linear regression is ‘y=mx+c’. It assumes that there is a linear relation between the dependent variable(y) and the predictors or independent variable(x). In regression, we calculate the best fit line which describes the relation between the dependent and independent variable. Regression is performed when the dependent variable is of continuous data type and independent variables could be of any data type like continuous, categorical etc. It tries to find the best fit line which shows the relationship between the dependent variable and independent variable with least error. Regression is classified into simple linear regression and multiple linear regression.

- **Simple linear Regression**: SLR is used when the dependent variable is predicted using only one independent variable.

- **Multiple linear regression**: MLR is used when the dependent variable is predicted using multiple independent variables.

***Qn2 :*** *Explain the Anscombe's quartet in detail.*

**Ans:** Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influencial observations on statistical properties.

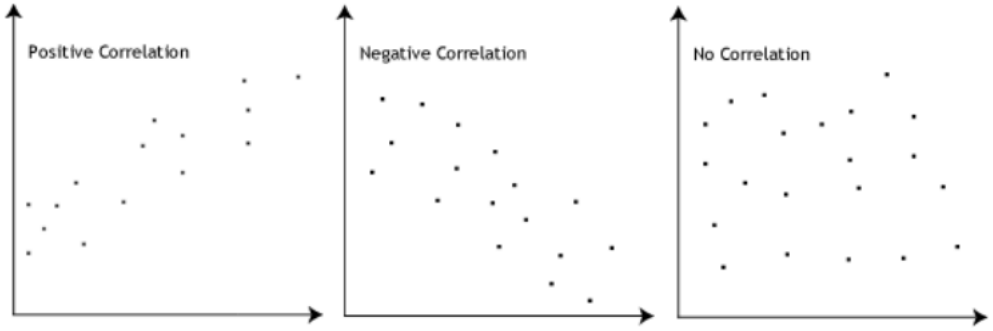



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally while there is a relation between them, it's not linear.
- In the third graph(bottom left), the distribution is linear but should have a different regression line. The calculated is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- The fourth graph, shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables

***Qn3:*** *What is Pearson's R?*

**Ans:** Pearson's r 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 variables tend 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 Pearson 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 variables. A value greater than 0

indicates a positive association; that is, as the value of one variable increases, 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 increases, the value of the other variable decreases. This is shown in the diagram below:




**Qn4:** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

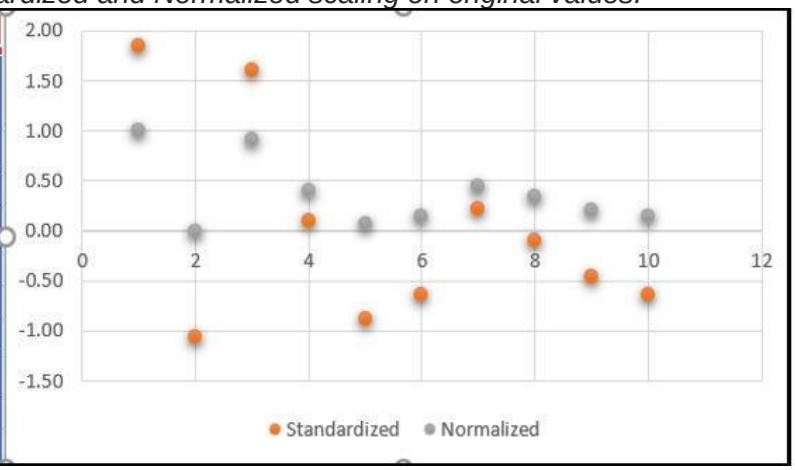
**Ans:** Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh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

**Example**: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

***Example:***

*Below shows example of Standardized and Normalized scaling on original values.*

| Original | Standardized | Normalized |
|---|---|---|
| 59 | 1.85 | 1.000 |
| 11 | -1.06 | 0.000 |
| 55 | 1.61 | 0.917 |
| 30 | 0.09 | 0.396 |
| 14 | -0.88 | 0.063 |
| 18 | -0.64 | 0.146 |
| 32 | 0.21 | 0.438 |
| 27 | -0.09 | 0.333 |
| 21 | -0.46 | 0.208 |
| 18 | -0.64 | 0.146 |



**Qn5:** You might have observed that sometimes the value of VIF is infinite. Why does this happen?
**Ans:** When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. If there is perfect correlation, then VIF = infinity. large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

**Qn6:** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

**Ans:** The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
**Use of Q-Q plot:** A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
**Importance of Q-Q plot**: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.