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: There are few insights that could be taken from the relation of categorical variables with target variable.

I. **Weathersit**:

- Demand is less on days with Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
- The bike sharing is highest on clear or on partially cloudy weather
- During cloudy or Misty weather the booking is moderate.

II. **Season**:

- Demand is less on Spring season.
- The demand is good on winter and summer seasons.
- The bookings are highest in the fall season.

III. **Month**:

- On months of Jan, Feb, Nov and Dec.
- Demand start decreasing after September.

IV. **Holiday**:

- Demand is less on all Holidays except on April mid and July first week so we can expect less on these two holidays.

V. **Year**:

- Demand increases every year.
- The demand Is higher in 2019 compared to 2018

VI. **Weekday**: It doesn't provided much relation to the dependent variable.

VII. **Workingday**: The demand is twice as high on workingdays.

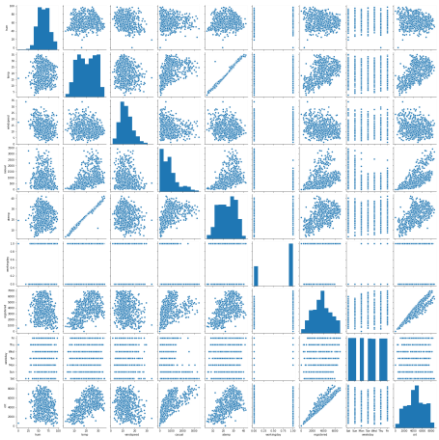
2. Why is it important to use drop_first=True during dummy variable creation?

Ans: It is important to drop_first to avoid redundant correlated feature among the created dummy variables. It also reduces a column count on the created dummy variables. The redundant variable can be any one of the dummy features where the one removed would be the explicit feature that represents the it. For example: Consider a categorical feature as gender consisting of three levels Male, female and Transgender. Since it is one-hot encoded it can be shows as

Male	Female	Transgender
1	0	0
0	1	0
0	0	1

Consider dropping first columns from this so we would have female and transgender. If female and Transgender columns has 0 and 0 means that the gender is Male, like wise if it is 1 and 0 its female… So if we don't drop Male that it would be a redundant column given us same information that 1 0 0 is Male. This shows us that one column can be ignored from the dummy variable creation using drop_first=true.

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



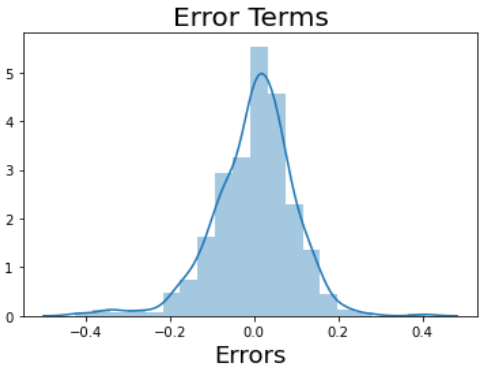
Ans:

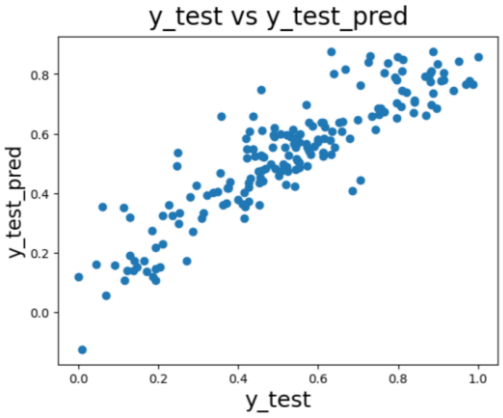
From the pair plot we can conclude that **temp** and **atemp** has the highest correlation

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

Ans:

1. The assumption about the form of the model:
   - we could say that it is a **linear assumption** ie. The dependent and the independent variables of the model is in a linear relationship.





2. The assumptions of simple linear regression were:
   - Zero mean assumption: The residuals have zero mean
   - Normality assumption: Error terms are normally distributed
   - Independent error assumption: Error terms are independent of each other
   - Constant variance assumption: Error terms have constant variance (homoscedasticity)
3. Assumptions about the estimators:
   - The independent variables are measured without errors
   - The independent variables are linearly independent of each other without multicollinearity

Assumption explenation:

In case if the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.

If the mean of residuals is zero, the mean of the target variable and the model become the same, which is one of the targets of the model.

The residuals (also known as error terms) should be independent. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify.

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

Ans: The contributing features can be determined by their coefficient. In the final model

I. **Actual Temperature(0.388)** contributes a lot towards the model output like we saw in the visualisation plot for temperature were the demand was more on high temperature and less on low temperature.
II. **Light Snow/Light Rain(-0.294)** helps in model to predict when will the bike share be less as the this is a negative coefficient.
III. **Year(0.234)** features contributed the third high towards demand of the shared bikes as it tells us the demand increases every year.

General Subjective Questions

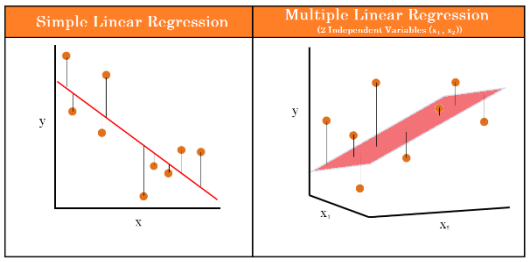
1. Explain the linear regression algorithm in detail ?

Ans: **Linear regression** is a supervised machine learning algorithm which attempts to model the relationship between input variable(s) and output variable by fitting a linear equation to observed data for prediction of future results. Input variables are considered to be an explanatory variable or predictors, and the other is considered to be a dependent variable or target variable.

The best fit line is mostly derieved by cost function/gradient descent. The linear Regression model output is mostly evaluated by the Sum of Squared Residuals Method.

The Linear Regression is of two types classified based on the number of input variables as :

- Simple Linear Regression:
  - The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line.
  - It is interpreted by the formula : $y = mx + c$ or $y = \beta_1 x_1 + \beta_0$
- Multiple Linear Regression:
  - Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. Every value of the independent variable x is associated with a value of the dependent variable y.
  - It can be interpreted by the formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n$

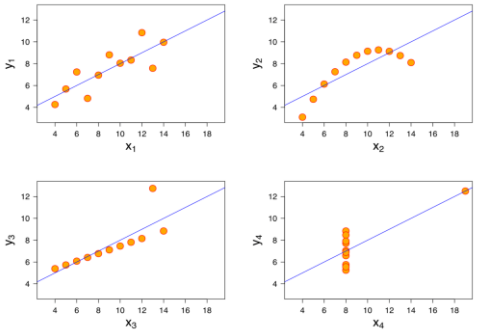


Simple linear regression has quite a few shortcomings:

- It is sensitive to outliers
- It models the linear relationships only
- A few assumptions are required to make the inference

2. Explain the Anscombe's quartet in detail ?

Ans : **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.



It demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties

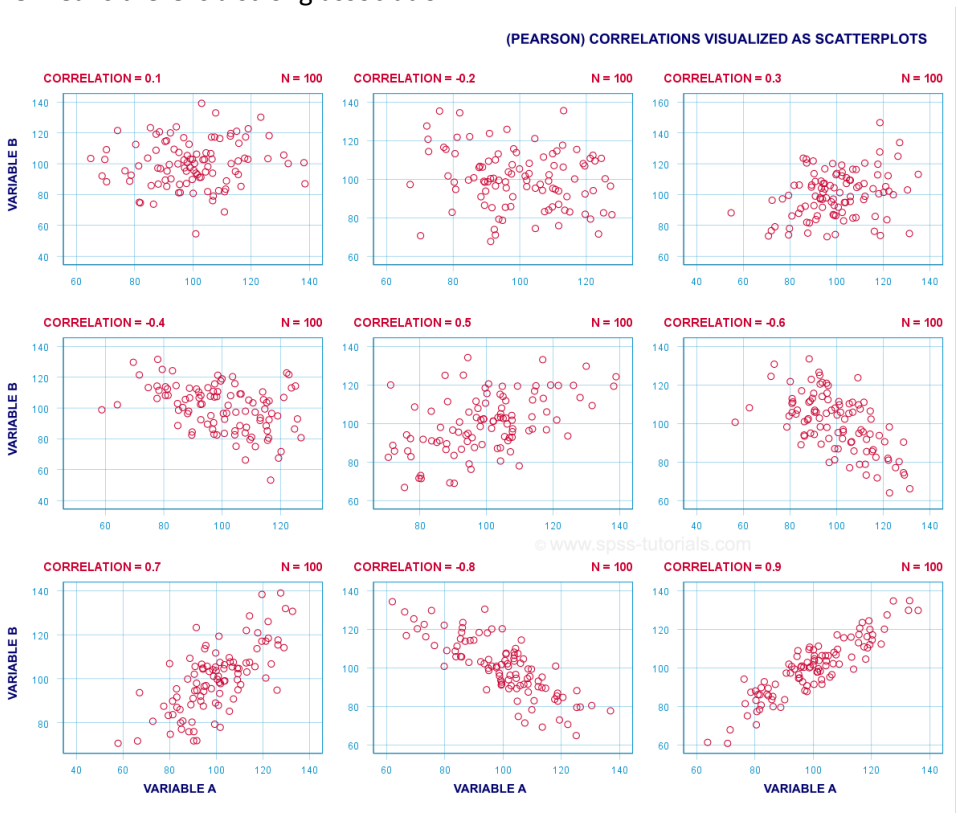
- The first scatter plot (top left) appears to be a **simple linear relationship**, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is **not linear**, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated **regression is offset by the one outlier** which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) 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.

3. What is Pearson's R?

Ans : **Pearson's correlation** (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression. **Pearson's R** are used to measure how strong a relationship is between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. 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 opposite 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



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

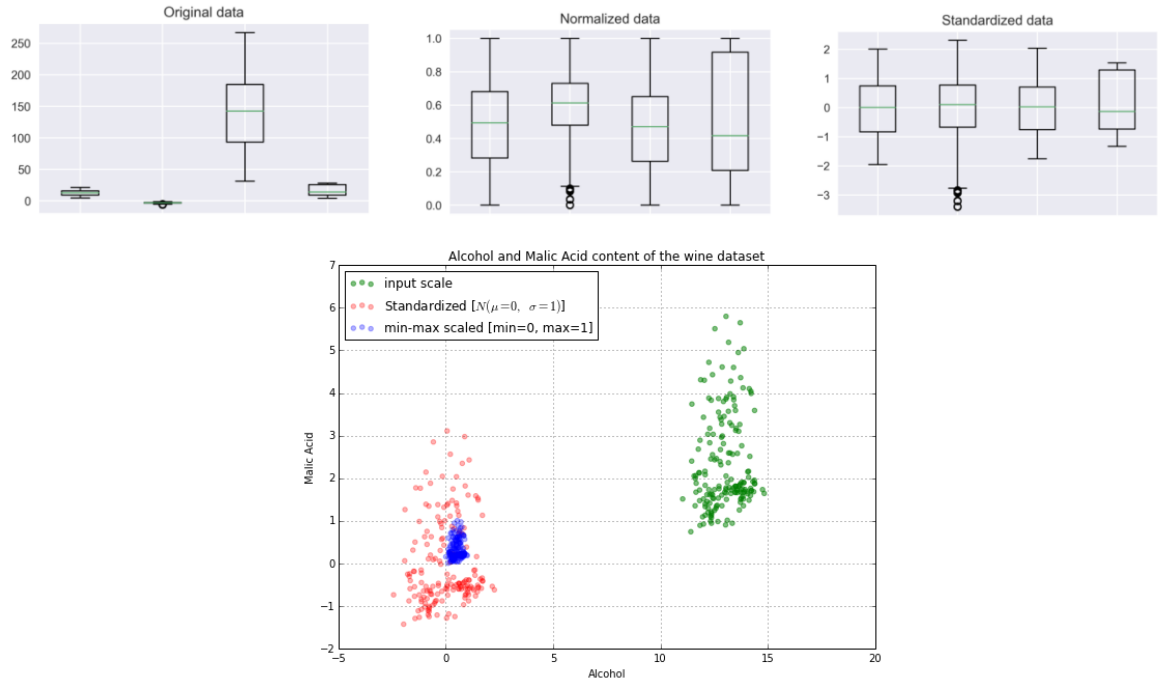
Ans: **Feature scaling** is a method used to **normalize the range** of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. So having features on a similar scale can help the gradient descent converge more quickly towards the minima ie train algorithms faster. It also helps is handling outliers in the dataset and the features that are in different scale that results in a weird coefficient. 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.

Normalisation and Standardisation are the most popular feature scaling techniques.

Normalization	Standardisation
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.	Standardization is another scaling technique where the values are centred 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.
$x = \frac{x - x_{min}}{x_{max} - x_{min}}$	$\mathrm{x} = \frac{\mathrm{x} - \mathrm{mean(x)}}{\mathrm{sd(x)}}$

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.	Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.



We can see that the normalized values lies between 0 and 1 but it looses information about outliers.

But standarsized scale has distributed the data in 1 standard deviations with median as 0 which preserves information of distribution.

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

Ans: VIF basically helps explaining the relationship of one independent variable with all the other independent variables. VIF is given by $VIF_i = \frac{1}{1-R_i^2}$. So when there is a perfect correlation R^2 becomes 1 as a result

$$\text{VIF} = \left(\frac{1}{1-R^2}\right)$$
$$\text{VIF} = \frac{1}{1-1}$$
$$\text{VIF} = \text{infinite}$$

.ie, 1/0 is statistically would be infinite limit. Thus, when we get VIF values in infinite it means that there is perfectly correlation or we could also say that the variable can be fully explained by all other variables excluding the target variable.

Example:

When we have a dataset have Celsius and Fahrenheit features both would represent a same value on transformation in this case the VIF would go infinite. or different currencies feature could also be a good example.

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 scatterplot created by plotting two sets of quantiles against one another 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 ie. a plot that quantiles the first data set against the quantiles of the second data set.

In linear regression Q-Q plot can be used when we have training and test data set received separately and then we can confirm that both the data sets are from populations with same distributions. 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.