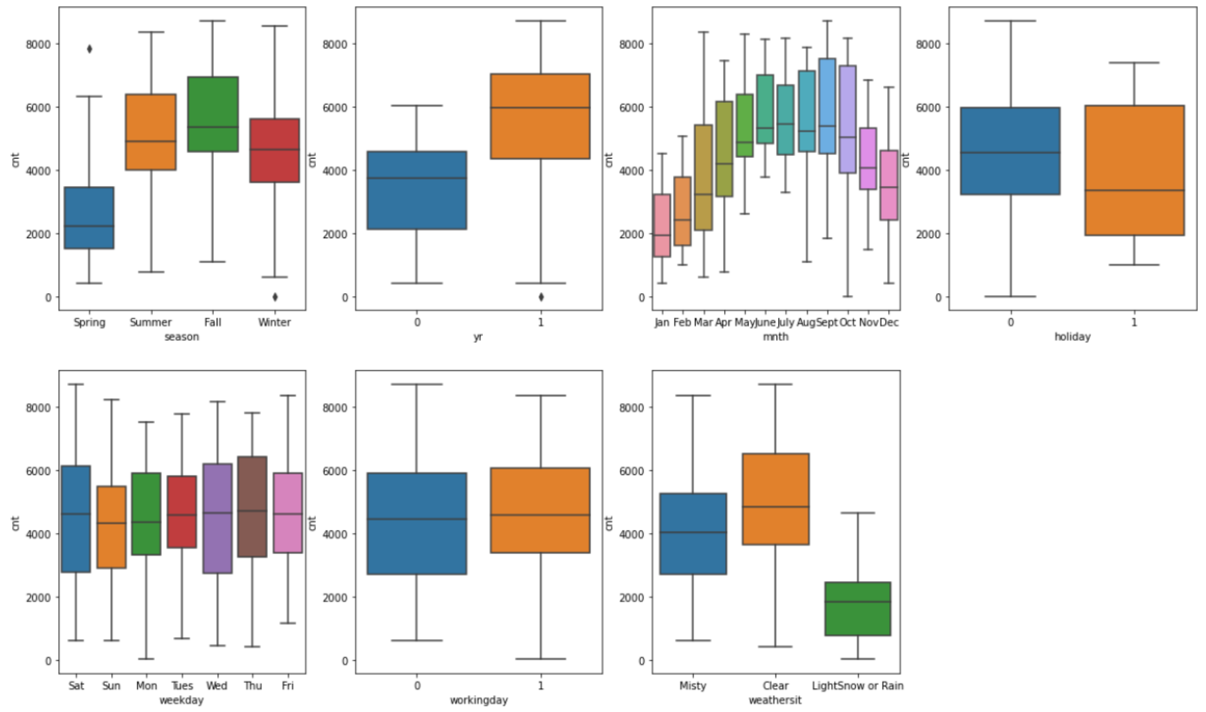
Linear Regression Assignment

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? (3 marks)

From the given dataset, the variables except ‘dtetday’ are numerical types. ML algorithms sense these data as numerical ones which may lead to wrong interpretations. (For example, it may interpret 6 > 0, for weekdays). Therefore, we convert these variables to categorical by mapping the numeric values with associated labels so that the dataset now contains meaningful categories of the variables.



From above boxplots for all the categorical variables, we can derive following:
1. There is less demand for bikes in spring and winter, high in summer and fall.
2. There is less demand in the year 2018 and high in 2019.
3. Demand is increasing continuously from Jan till June and after September the demand is decreasing. The maximum demand is in the month of September.
4. Less demand in LightSnow or Rain weather situation, and high in Clear weather situation.

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

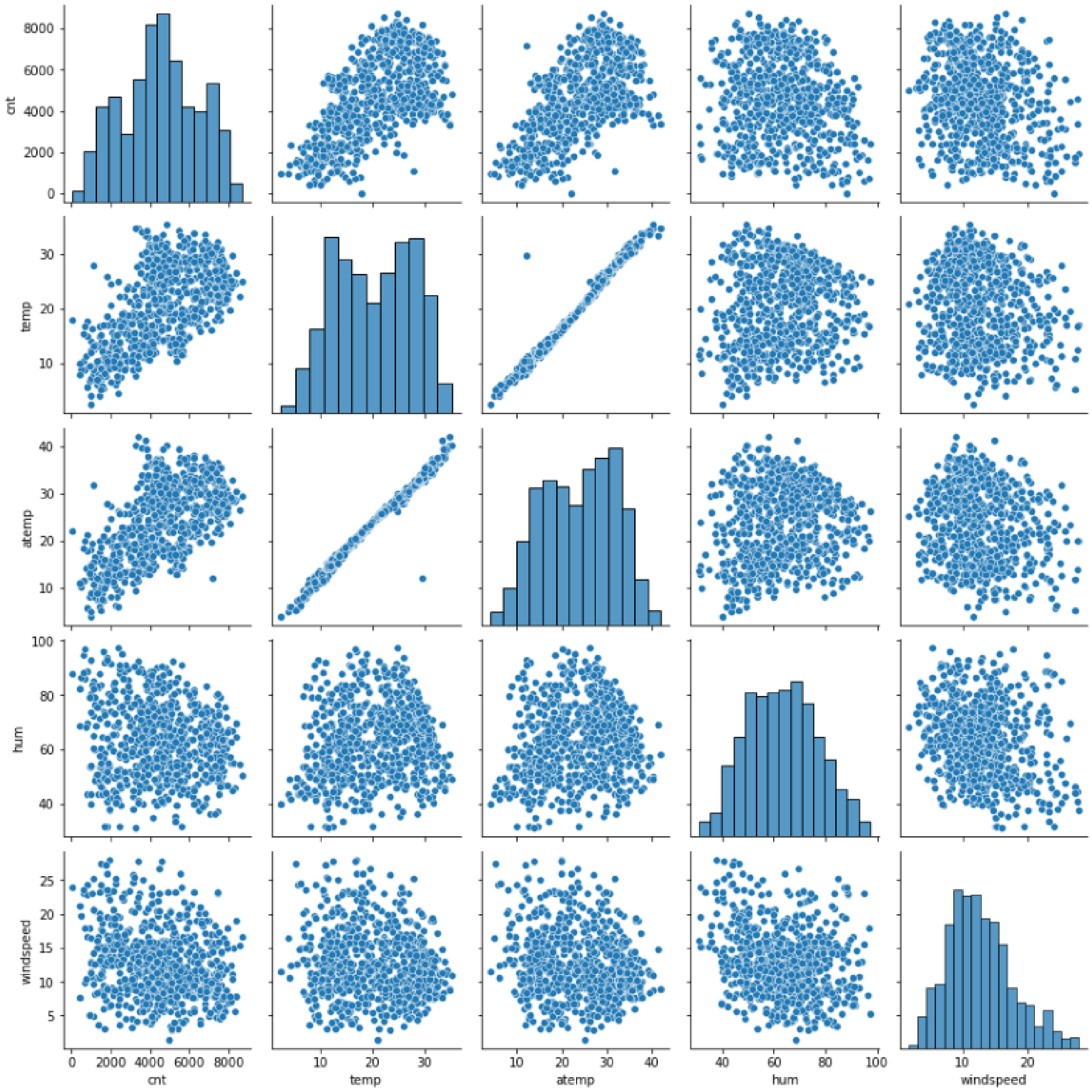
Dummy variables are created to convert categorical values into numerical values.

- While we convert ‘n’ categories into dummy variables it creates ‘n’ variables.
- Instead of creating ‘n’ variables to represent ‘n’ categories, we can drop one variable after dummy creation. Because ‘n-1’ variables can represent ‘n’ categories, this is the reason why we drop one variable using drop_first = True.

- It also reduces correlation among dummy variables, which as a result resolves multicollinearity issue when we give these dummy variables as inputs to the model building.

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

By looking at the below pairplot, we can say that ‘temp’ & ‘atemp’ has the highest correlation with the target variable ‘cnt’.



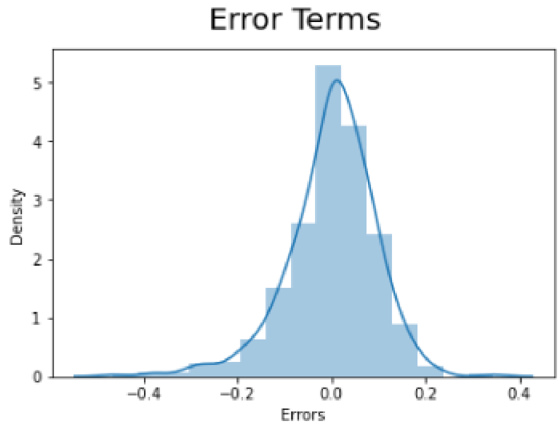
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

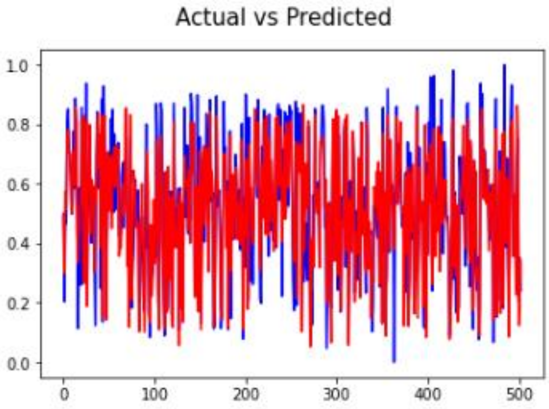
Assumptions made before model building are:

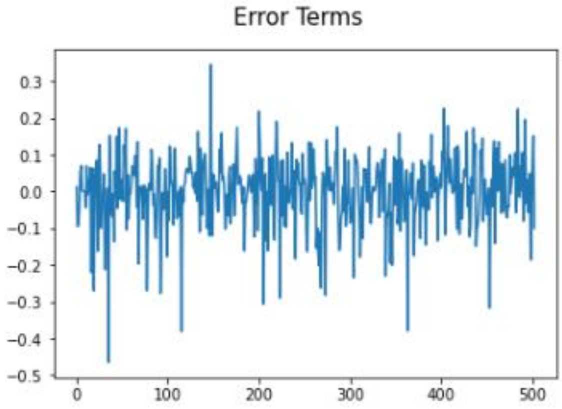
1. Linear relationship exists between X and Y
2. Error terms are normally distributed
3. Error terms are independent to each other

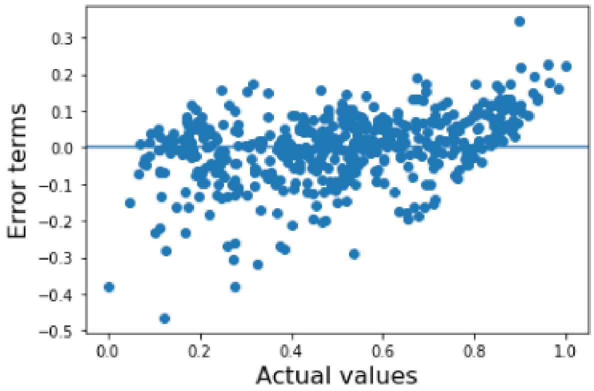
4. Error terms have constant variance (homoscedasticity)

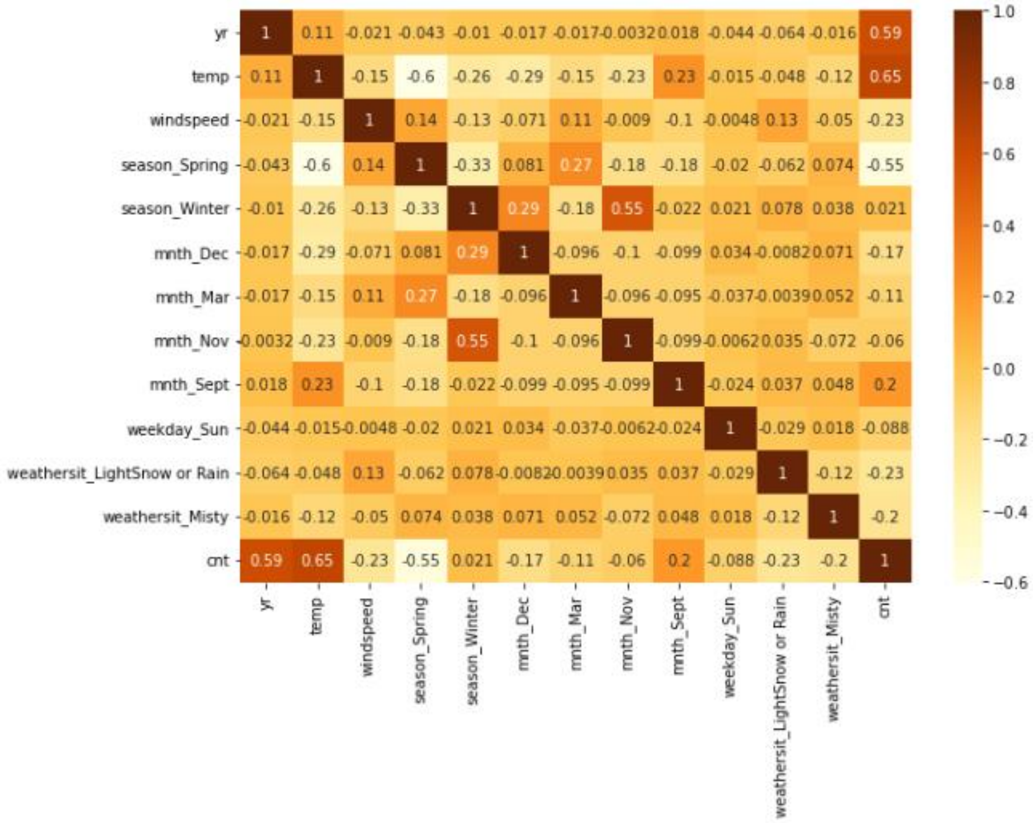
These assumptions are validated through scatter plots, heatmap and histograms after model building.











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

According to the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are

- Temperature
- Year
- Light Snow or Rain weather situation

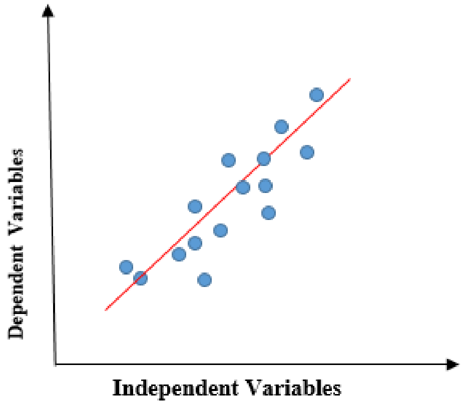
General Assignment-based Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variables and the dependent variable, consequently called linear regression.

- If there is a single input variable, such linear regression is called simple linear regression.
- If there is more than one input variable, such linear regression is called multiple linear regression.

The linear regression model gives a sloped straight line describing the relationship within the variables.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \Longrightarrow y = a_0+a_1x$$

y = Dependent Variable
x = Independent Variable
a0 = intercept of the line
a1 = Linear regression coefficient.

Steps:

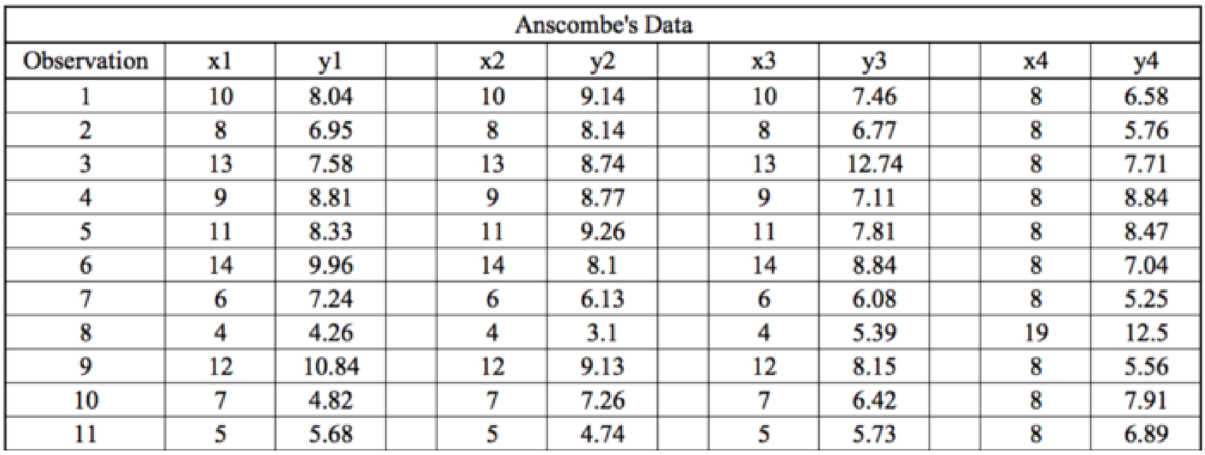
1. Reading and understanding the data
2. Visualizing the data
3. Data preparation
    a. In-order to fit a regression line, we would need numerical values and not string. Need to convert data accordingly
    b. Use dummy variables
4. Splitting data into training and testing dataset
    a. Rescale the features
    b. Dividing into X (input variables) and y (output variable) sets for model building
5. Building a linear model

    a. Identify key features using RFE first
    b. Build a linear model with identified key features using statsmodels and evaluate VIF values also
    c. Based on p-values and VIF values, drop few features and trigger step ‘b’ again
    d. Once p-values and VIF values are within acceptable limits and reasonable R2, Adjusted R2 values – we can consider that model as a good fit.

6. Model Evaluation
    a. Evaluate assumptions
    b. Compare results – actual test values vs predicted test values

2. Explain the Anscombe’s quartet in detail.

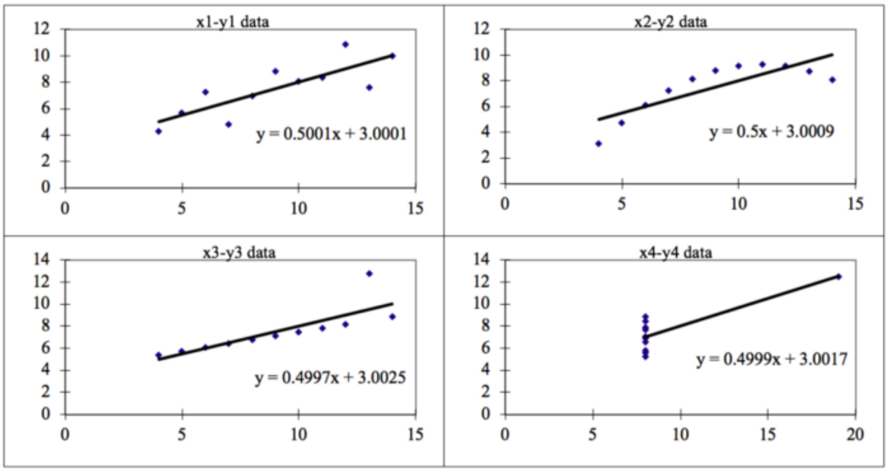
Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
Example: Four dataset are:



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

Plotted as below:



The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Therefore, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is referred to as **Pearson's R**, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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 different 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?

Scaling:
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range which also helps in speeding up the calculations in an algorithm.

Why is scaling performed:
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Difference between normalized scaling and standardized scaling:
1. In case of normalization minimum and maximum value of features are used whereas for standardisation mean and standard deviation is used for scaling.
2. Normalization ranges between 0 to 1 or -1 to 1 and standardisation is not bounded to a certain range.

3. MinMaxScaler helps to implement normalization and sklearn.preprocessing.scale helps to implement standardization in python.

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

VIF = infinite 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 resolve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q plots are also known as Quantile-Quantile plots. It is a graphical tool to help us assess if a set of data follows any particular type of probability distribution like normal, uniform, exponential.
QQ plots is very useful to determine

i. If two populations are of the same distribution
ii. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
iii. Skewness of distribution

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.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis