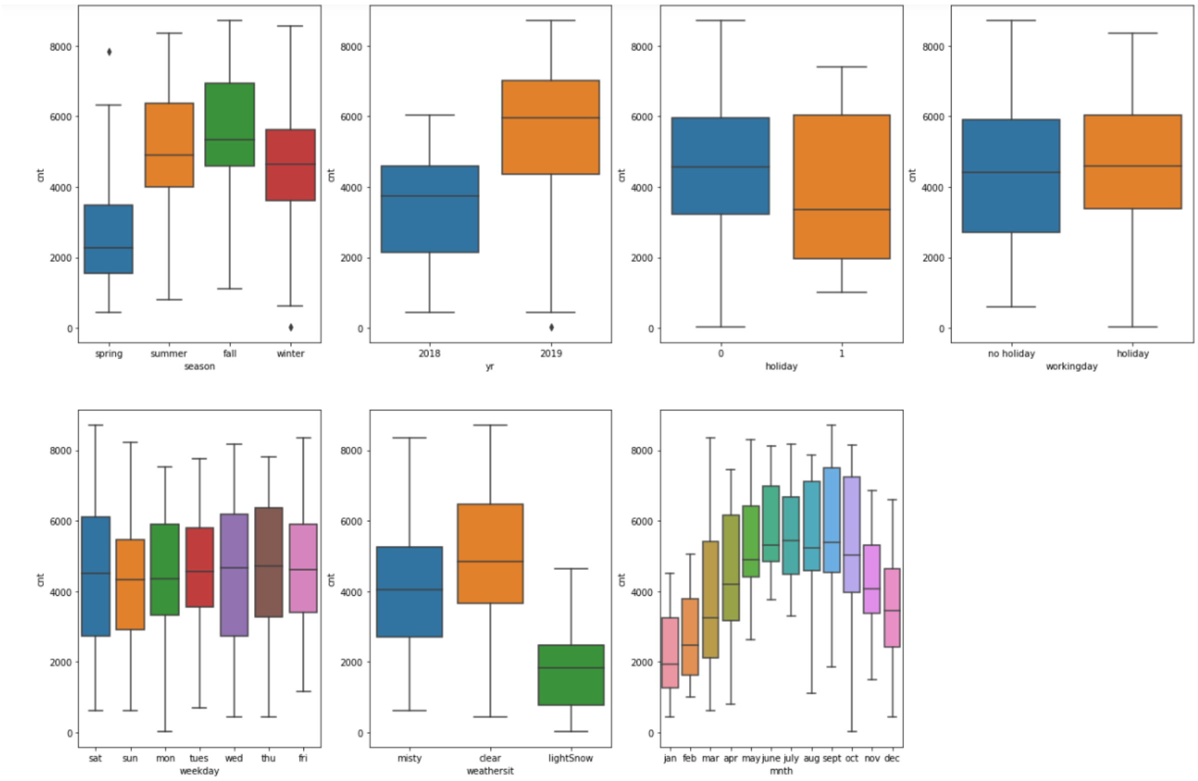
Assignment-based Subjective Questions

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

Ans: Following is the list of categorical variables and its effect on dependent variables –



1. yr: Maximum bikes were rented in year 2019.
2. season: Fall season has slightly high rented bikes and minimum bikes rented were in the spring season.
3. mnth: Demand is increasing from January till June and then there is drop in demand from September to December. Maximum demand is observed in the month of September.
4. weekday: On Saturday there is high demand of bikes.
5. weathersit: Less demand in Light Snow weather and high when the weather is Clear.

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

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Ans: As temp and atemp has high correlation with each other. So both atemp and temp are has highest correlation with the target variable.

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

Ans: Below points are considered for validating the assumptions of linear regression model on training set:
1.Co-efficient values – Non-zero co-efficient indicate that all there is relationship between independent and dependent variables.
2.P-value – P-values are less than 0.05 which indicate they are significant to model.
3.VIF – VIF is less than 5, so that there is no multicollinearity between predictor variables.
4.F-statistic and Prob(F-statistic) – F-statistic high and low Prob(F-statistic) indicates that overall model fit is significant and not just by chance or only predictor variables are significant.

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

Ans: Top three features contributing significantly towards explaining the demand of the shared bikes are as follows –
1.Temperature
2.Year
3.Season winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression – Linear regression is the supervised machine learning model which attempts to explain the relationship between dependent (output variable) and independent variable/s (predictor variable).

Types of linear regression:
1.Simple linear regression
2.Multiple linear regression

Algorithm:
Below are the steps for linear regression –
1.Import required libraries like pandas, seaborn, matplotlib, sklearn and statsmodels.
2.Read and understand dataset.
  - Check for missing values
  - Understand potential independent variables based on dataset and business requirement.

3.Prepare data for modelling

- Handle categorical and binary variables.
- Check if assumptions are met as per type of regression model.

4. Split dataset into train and test set

5.Train the model

- Check for significant variables (using train set) in case of multiple linear regression.
- Drop out insignificant variables
- Repeat step 5 until best coefficients found

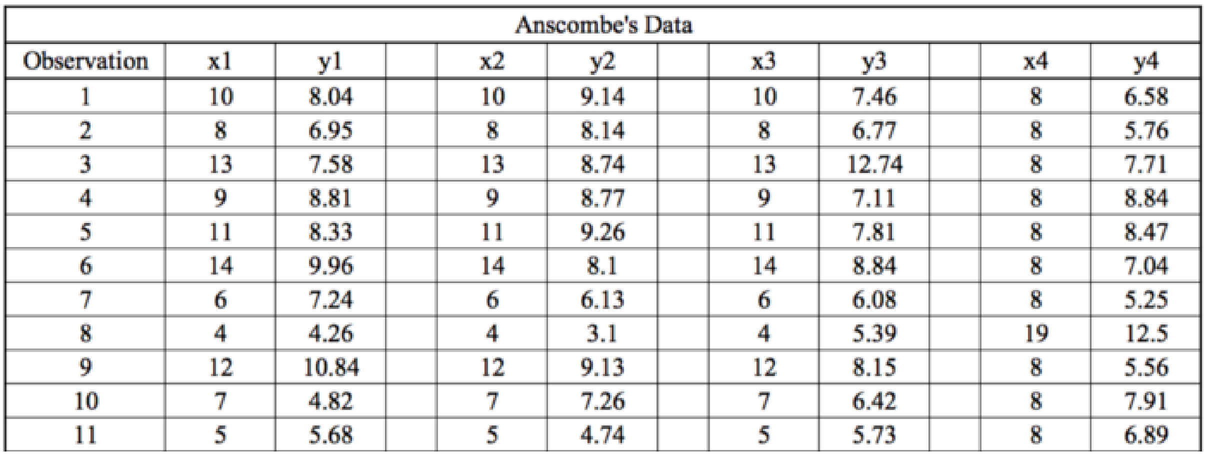
6. Predictions and model evaluation on test set.

- Predict target variable values using test set.
- Evaluate the model using cost function.

2. Explain the Anscombe’s quartet in detail.

Ans: 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.

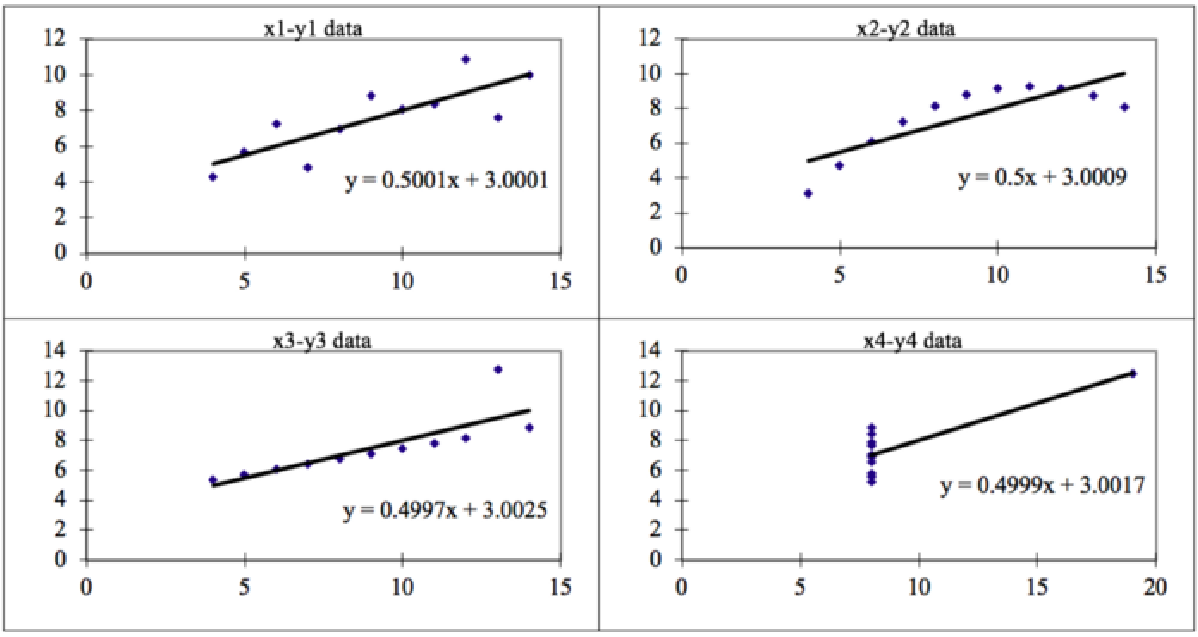
Example: Four datasets are:



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?

Ans: 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?

Ans:
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.sklearn.preprocessing.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?

Ans: 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.

Ans: 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