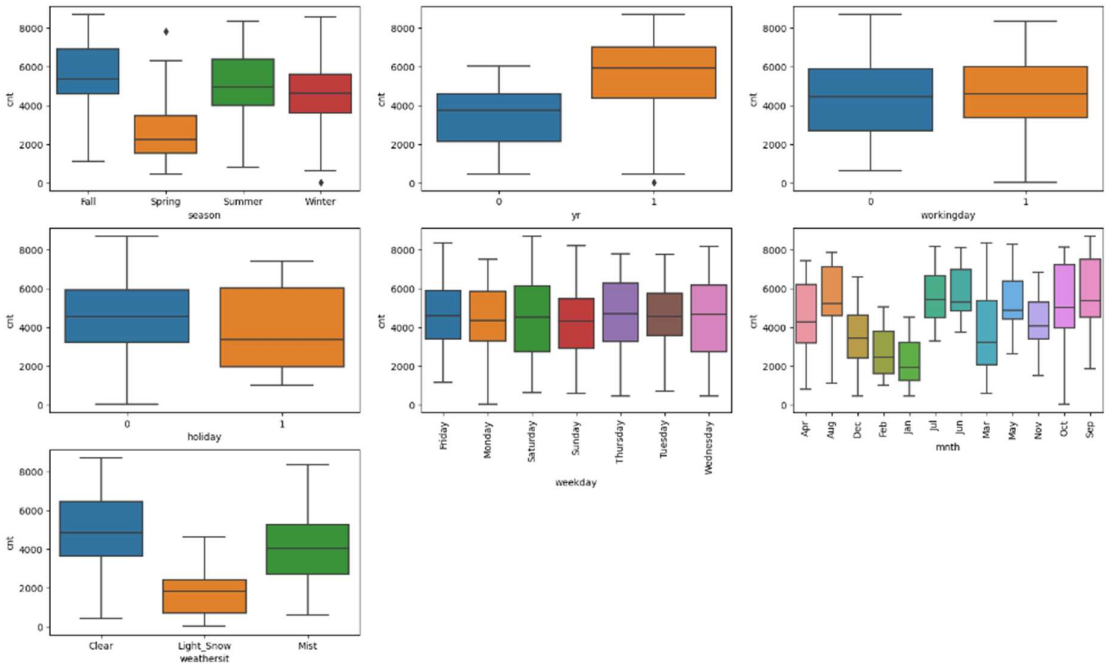
Assignment-based Subjective Questions:

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

Ans: Categorical variable analysis is done using box plots. Below are few inferences based on Categorical variable analysis.

- The cnt for the year 2019 is more than the year 2018. So this indicates there is increase in demand from year 2018 to 2019.
- cnt for the months (May, Jun, Jul, Aug, Sep, Oct) is more compared to other months. So mnth column can be possible predictor of cnt
- Fall has higher booking, and there is no much significant difference between summer, winter. Spring has lowest cnt. So possible season can be one of the predictors.
- The cnt are more on a workingday (holiday = 0). So workingday might be a possible predictor.
- As per the below plots there is very small / negligible difference on the cnt variable for all weekdays. So, weekday, from the initial observation doesn't look to be a very good predictor of the bookings (cnt).
- When the weather is Clear (weathersit), the cnt is more as compared to other weather conditions. So this is indicating weather has significance in terms of predicting the model.

Below are box plots created for categorical variable



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[for ex: "Correct", "Incorrect" and "Partially correct"] and we want to create dummy variable for that column. If one variable is not "Correct" and "Incorrect", then It is obvious that it is "Partially correct" . So we do not need 3rd variable to identify the "Partially correct".

Correct	1	0
Incorrect	0	1
Partially correct	0	0

- If categorical variable with n-levels is present, then we need to use n-1 columns to represent the same.

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

Ans: Looking at the pair plot among the numerical variables **"temp"** and **"atemp"** has highest correlation with target variable.

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

Ans: The assumptions of Linear Regression model are evaluated based on below aspects:

- Multi Collinearity:
  - The multicollinearity between variables should be insignificant (VIF value of < 5)
- Homoscedasticity:
  - No visible patterns on residuals
- Normal Distribution of Error Terms:
  - Error Terms are normally distributed.
- Linear relationship between variables:
  - Per the pair plot there was linear relationship observed with some of the variables for target variable.

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

Ans: top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- **temp**: Temperature variable has highest coefficient value of `0.5682`. Which indicates a unit of increase in temperature increase the demand by 0.5682 units.(**Positive correlation)**)
- **Light_Snow**: (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) : This has a coefficient value of `-0.2535` indicates that, a unit increase in Light_Snow variable decreases the demand by 0.2535 units.(**Negative Correlation**)
- **yr** : The coefficient of yr (year) is 0.2334 . Which indicates a increase of unit in year (2018 to 2019) increase the demand for bikes by 0.2334(**Positive correlation)**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features., is called Linear Regression.

- **Simple Linear Regression:** When the number of the independent feature, is 1 then it is known as Simple Linear regression

  Mathematical Equation: Y = β0+β1*X

  Where:

  β0 = intercept,

  β1 = coefficient

  X = independent (Predictive) variable

  Y = dependent (target) variable

- **Multiple Linear Regression:** When we try to find the dependency between one dependent variable(Y), and more than one independent (predictive) variables

  Mathematical Equation: Y = β0+β1*X1+ β2*X2+ ……. + βnXn

  Where:

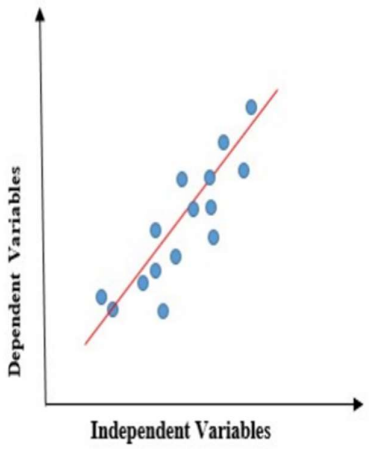
  β0 = intercept

  β1 β2 … βn = coefficients

  X1 X2 …. Xn= independent (Predictive) variables

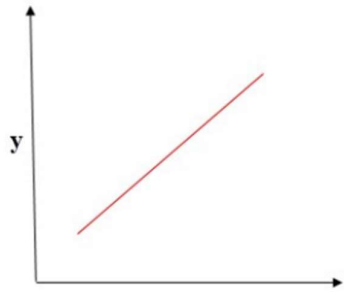
  Y = dependent (target) variable

Below fig shows linear relation between Dependent and Independent variables.

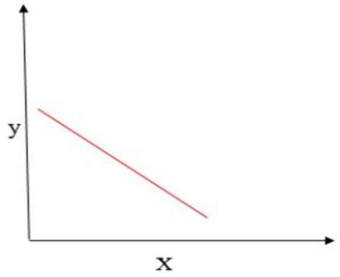


This Linear relation can be Positive or Negative relation.

Positive: When the independent variable value increases, the dependent variable value also increases.



Negative: When the independent variable value increases, the dependent variable value decreases.



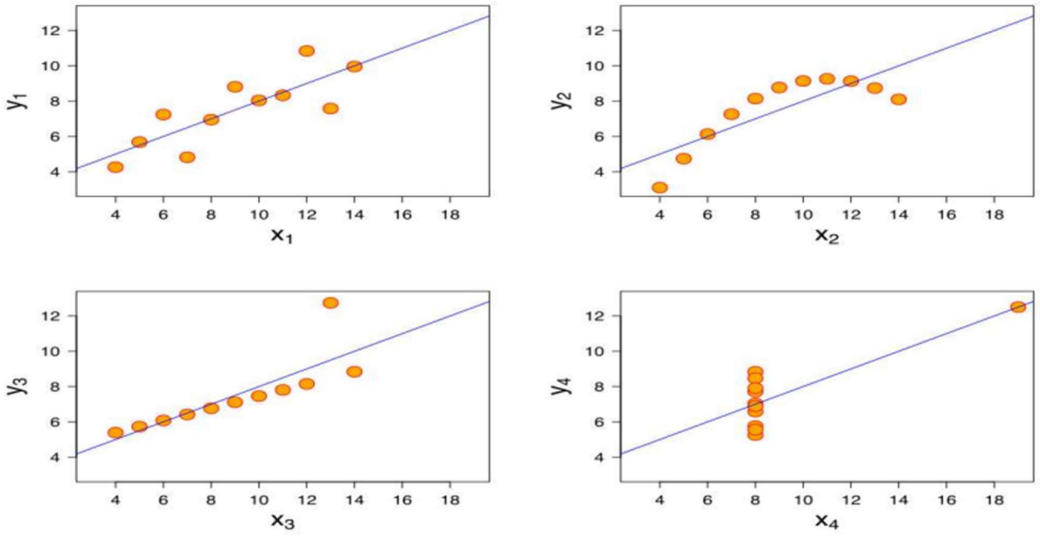
The goal of linear regression is to find the best values of β0,β1,β2, … βn to provide the best fit line for given datapoints. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. In linear Regression **MSE** (Mean Squared Error) cost function is used, which is the average of squared error.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. Which shows the importance of both graphical and numerical analysis of data.

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
				Summary Statistics						
N	11	11		11	11		11	11	11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94	3.16	1.94
r	0.82			0.82			0.82		0.82	

When the above data is plotted using a scatter plot, all datasets generate different kind of plot



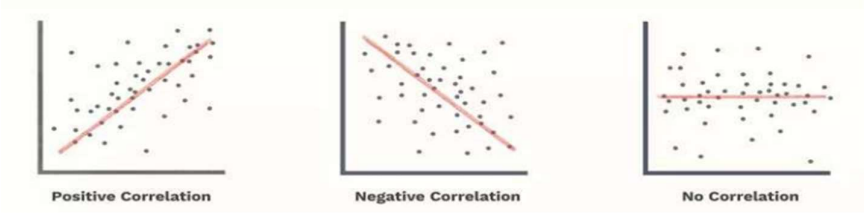
1. Dataset 1: this fits the linear regression model.
2. Dataset 2: this could not fit linear regression model.
3. Dataset 3: Though the model is linear, there is an outlier (one data point) which exerts enough influence to lower correlation, involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

   Summary: It's recommended to visualize important features in data set before implementation.

3. What is Pearson's R?

Ans: Pearson Correlation Coefficient (PCC), which is also known as Pearson's R or Correlation coefficient, is a measure of linear relationship between 2 sets of data. It is the ratio between covariance (joint variability of 2 variables), and product of standard deviations. To explain in simple terms, f 2 variables tend to increase or decrease together then it shows positive correlation. If the variables tend to increase or decrease in opposite direction, then it shows negative correlation. The value of correlation coefficient or R lies between -1 and +1.

- A value of 0 indicates no correlation between variables.
- Value > 0 indicates positive correlation
- Value < 0 indicates negative correlation



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

Ans: Scaling is a technique used to standardize or normalize the range. This is performed during data pre-processing. As the data set might vary widely, if we take the exact or face value of data, then it might lead to algorithm considering the data with large values as higher, and small values as lower, irrespective of the unit of given data.

For example: Without feature scaling, algorithm, it might be possible that 120 lbs might be considered as larger than 60 kgs, which leads to wrong predictions.

Difference between normalized and standardised scaling.

S.No	Normalized Scaling (Min-Max Scaling)	Standardizes Scaling
1	Min, Max values of features are used for scaling	Mean and Standard Deviation are used for scaling
2	it is used when features are of different scales.	This is used when we want zero mean, and standard deviation of 1 unit.
3	Scales values between [0, 1] or [- 1, 1].	It is not bounded to a certain range.
4	Affected by outliers	Not affected by outliers
5	Scikit-Learn provides a transformer (MinMaxScaler) for normalization	Scikit-Learn provides a transformer(StandardScaler) for standardisation
6	This is useful when we don't know the distribution.	This is useful, when the distribution is normal

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

Ans: A VIF value of infinite indicates that features have extremely strong relationship. In general, the larger value of VIF indicates that, there is correlation between variables. When there is an extremely strong or prefect correlation, then R-Squared value becomes 1, which leads **VIF : [1/(1-R2)]** to become infinity. In a way, this will cause multicollinearity, where variable with strong association will get dropped to bring down VIF values.

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

Ans: Q-Q (Quantile-Quantile) plot is a probability plot, a graphical technique, for determining if two data sets come from populations with a common distribution.

Uses:

The Quantile-Quantile plot is used for the following Uses:

- Determine whether two samples are from the same population.
- Whether two samples have the same distribution shape
- Whether two samples have the same tail
- Whether two samples have common location behaviour

Importance:

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 between the data samples as compared to other analytical methods.