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?

Answer:

- Maximum bike rental in **season 3** which is **'Fall'** and specifically in month **September and October**.
- On Holiday bike rental is comparatively less than **Non-Holiday** day
- Weather suitable for maximum bike rental is "**Clear, Few clouds, Partly cloudy, Partly cloudy**". (No bike rental in "Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog".)

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

Answer: **drop_first=True** drops first column created after while dummy variables. This insures we don't have any redundant column which might increase the correlation between the created dummy variables.

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

Answer: The variable which has highest correlation with the target variable("cnt") is "**temp"** and "**atemp"**. (We are not considering the "casual" and "registered" variable because target variable is sum of the same.)

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

Answer:

Linear Relationship between the features and target: Pair plot shows linear relationship between 'cnt' and 'temp'.

Multicollinearity between the features: We can have calculated the VIF for every variables and they are under 5, which represent little to no correlation between feature variables.

Homoscedasticity: We have verified the error term is the same across all values of the independent variables

Normal distribution of error terms: We have plotted the error term and it follows normal distribution.

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

Answer: Top three feature contributing significantly towards explaining the demand of the shared bikes are:

1. **temp** (**Temperature** has co-efficient of 0.5788 meaning demand in bike rental will increase by 0.5788 if there is an increase of 1 unit of temperature).
2. **Yr** (**Year** has co-efficient of 0.2277 meaning demand in bike rental increase by 0.2277 if there is a year change, this shows increasing trend by year)
3. **weathersit_3** (weathersit_3 has co-efficient of -0.2364 meaning bike rental demand will decrease by 0.2364 in weather condition of **Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: **Linear regression** is machine learning algorithm based on supervised learning. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between input and the target variable. If the there is a single input variable, then it is **simple linear regression** and if there are more than one variable then it is **multiple linear regression**.

The relation between dependent variable(y) and independent variables(X) is represented by above graph.

The red line is referred to as the best fit straight line based on given points.

Linear relationship can be Positive or Negative linear relationship.

$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
θ1: intercept
θ2: coefficient of x

Use Cases of Linear Regression:

- Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
- Price Prediction – Using regression to predict the change in price of stock or product.
- Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

Best fit line should have the least error means the error between predicted values and actual values should be minimized. Once we have best fit line then that can be used for predicting the value of 'y' for the input value of 'x'.

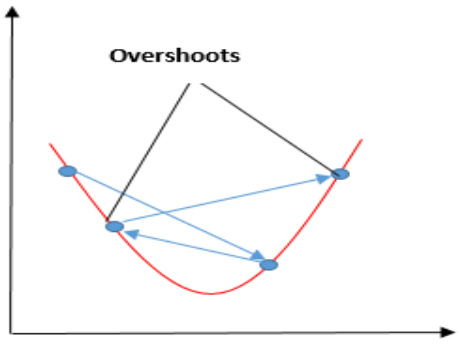
Cost Function

The cost function helps to determine the best possible values for **θ1 and θ2.**
Cost function of linear regression is the **Root Mean Squared Error(RMSE)** between predicted y value(pred) and true y value(y).

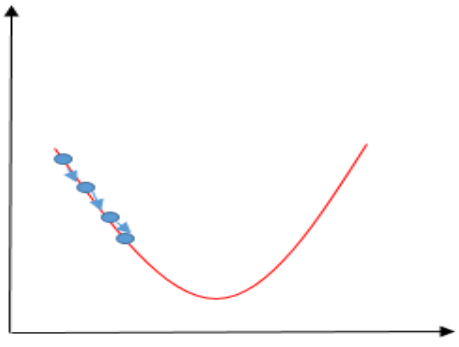
$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Gradient Descent:

Gradient descent is a method of updating **θ1** and **θ2** to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (**θ1**, **θ2**=> xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.



High learning rate

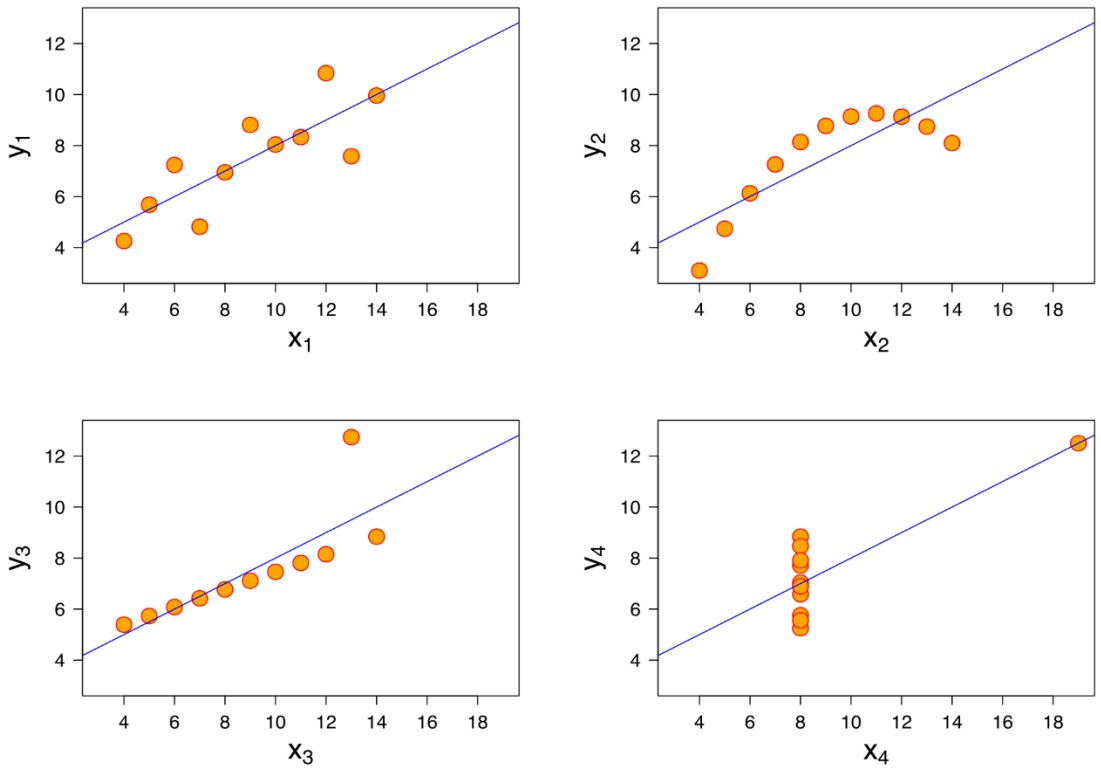


Low learning rate

2. Explain the Anscombe's quartet in detail.

Answer: **Anscombe's quartet** emphasis on importance for **data visualization** in data statistics.

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.



The above four plot have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

The **statistical information** for all four datasets are approximately **similar** and computed as follows:

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	

Out of all four datasets only first dataset fits linear regression model pretty well. This show importance of **data visualization** and how any regression algorithm can be fooled by the same. So it's better to visual the data before implementing any model.

3. What is Pearson's R?

Answer: Pearson's R also know known as **product-moment correlation coefficient** or **bivariate** correlation. As the name suggests, its statistic that measures the linear correlation between two variables. Value for Pearson correlation lies between **-1 to 1**.

Pearson's correlation coefficient cannot be used for nonlinear relationships cannot differentiate between dependent and independent variables.

Assumptions for Pearson r correlation:

1. Both variables should be **normally distributed**.
2. There should be no **significant outliers**.
3. Each Variables should be **continuous**.
4. Variable should have **linear relationship**.

5. There should not be any **blank** observations.
6. **Homoscedascity** should be present

The formula for calculating the Pearson's coefficient is:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Example:

Step 1: Make a Pearson correlation coefficient table. Make a data chart using the two variables and name them as X and Y. Add three additional columns for the values of XY, X^2, and Y^2. Refer to this table.

Step 2: Use basic multiplications to complete the table.

Person	Age (X)	Income (Y)	XY	X^2	Y^2
1	20	1500	30000	400	2250000

2	30	3000	90000	900	9000000
3	40	5000	200000	1600	25000000
4	50	7500	375000	2500	56250000

Step 3: Add up all the columns from bottom to top.

Person	Age (X)	Income (Y)	XY	X^2	Y^2
1	20	1500	30000	400	2250000
2	30	3000	90000	900	9000000
3	40	5000	200000	1600	25000000
4	50	7500	375000	2500	56250000
Total	140	17000	695000	5400	92500000

Step 4: Use these values in the formula to obtain the value of r.

r = [4 * 695000 - 140 * 17000] / √{4 * 5400 - (140)^2} {4 * 92500000 - (17000)^2}

= [2780000 - 2380000] / √{21600 - 19600} {370000000 - 289000000}

$= 400000 / \sqrt{\{2000\}\ \{81000000\}}$

$= 400000 / \sqrt{162000000000}$

$= 400000 / 402492.24$

$= \mathbf{0.99}$

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

Answer: Scaling is a step of data pre-processing for bring features/features to common or normalized scale so it can be compared other variables/features. We need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model.

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 have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1
- ***sklearn.preprocessing.MinMaxScaler*** helps to implement normalization in python.

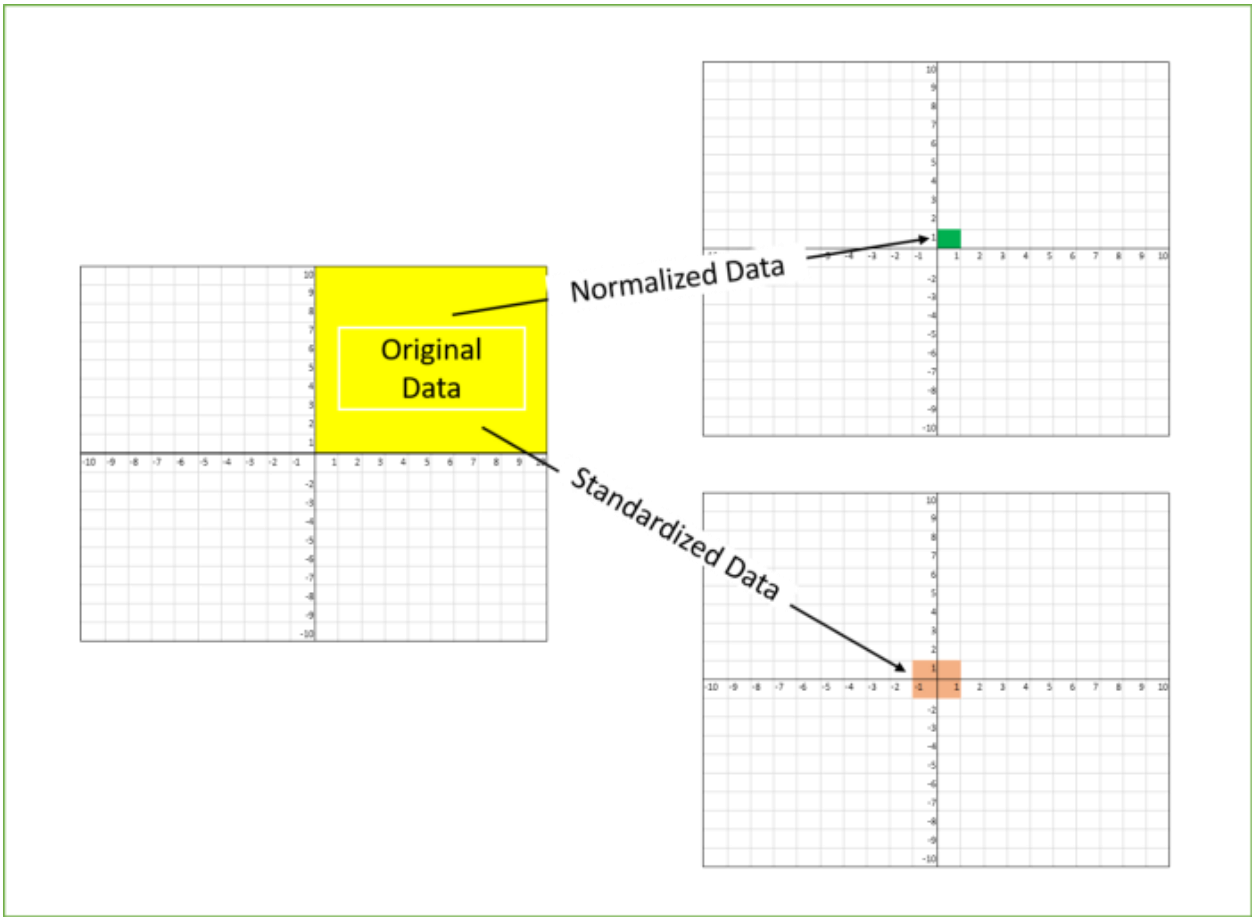
$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- ***Standardization*** *replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ***sklearn.preprocessing.scale*** *helps to implement standardization in python.*
- *One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.*



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

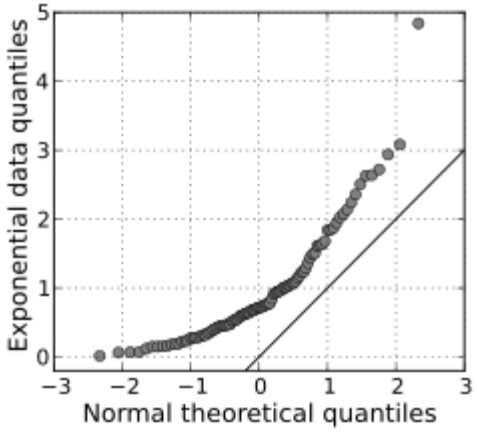
Answer: **VIF** is "**Variance Inflation Factor**" represents the correlation between two variables or features.

If value of VIF is **infinite**, it means **prefect correlation** between the variables/features.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a **linear combination** of other variables (which show an infinite VIF as well).

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

Answer: **Q-Q Plots** (Quantile-Quantile plots) are plots of **two quantiles** against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool 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.



It can be **used** for checking below scenarios on the data sets:

- Come from populations with a **common distribution**
- Have common **location** and **scale**
- Have similar **distributional shapes**
- Have similar **tail** behavior

Below are the possible interpretations for two data sets.:

- **Similar Distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degrees from x -axis
- **Y-values<X-values**: If y-quantiles are lower than the x-quantiles
- **X-values < Y-values**: If x-quantiles are lower than the y-quantiles
- **Different distribution**: If all point of quantiles lies away from the straight line at an angle of 45 degrees from x -axis