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)

Answer: Analysis of categorical variable:

- There are 4 seasons (1:spring, 2:summer, 3:fall, 4:winter) and if you see the boxplot summer and fall have many users that took bikes on rent.

- There are 2 years ((0: 2018, 1:2019) and according to chart, there are more users in 2019.

- Highest demand is September and lowest in January month.

- No holiday (more demand), holiday (less demand).

- Bike demand is similar every day for weekdays.

- No significant change in bike demand with working day and non-working day.

- Bike demand in seasons is:

  - `clear and Few clouds` - high
  - `Lightsnow and light rainfall` -less
  - `Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog` - No data

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

Answer: 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.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

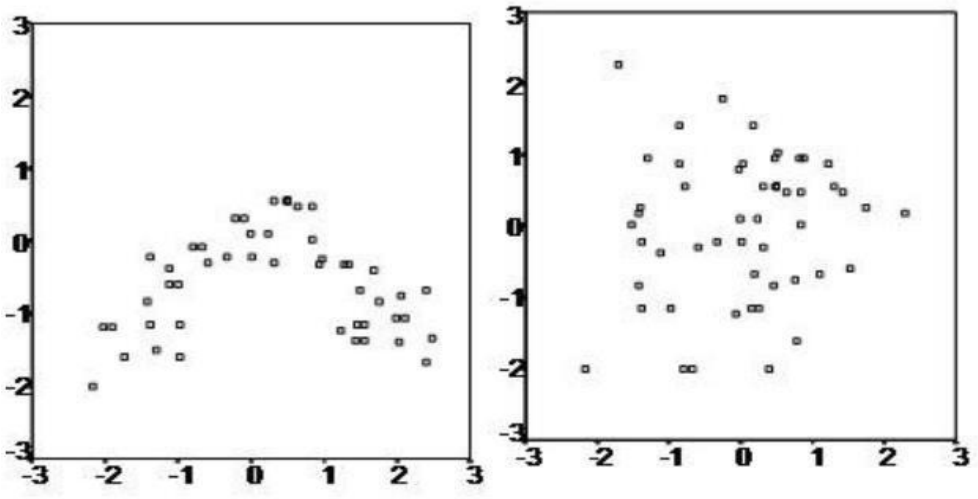
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?** (1 mark)

Answer: From the above pair-plot we could observe that, `temp` has highest positive correlation with target variable `cnt`.
We could infer following observation:
> - A positive correlation observed between `cnt` and `temp` (0.65)
> - A Negative correlation observed for `cnt` with `hum` and `windspeed` (-0.059 and -0.25)

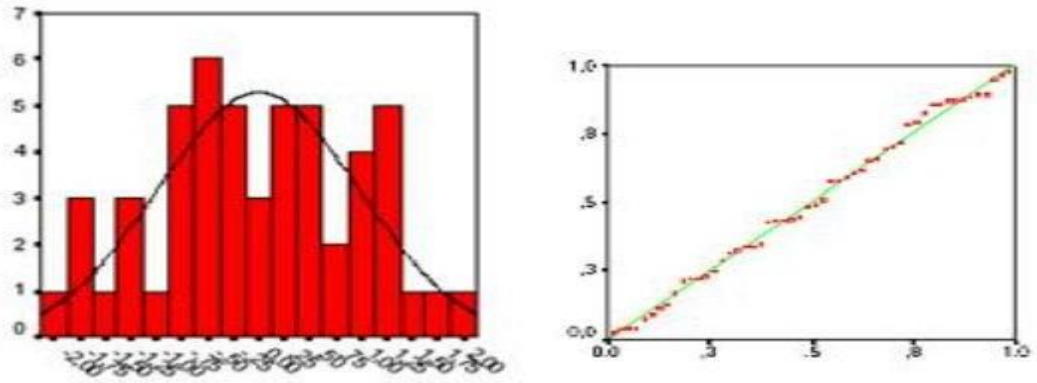
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

Answer:

1. Linear Regression needs the **relationship between independent and dependent variable to be linear**. It is also important to **check for outliers** since the linear regression is sensitive to outlier effects. The linearity assumptions can **best be tested with scatter plots**.



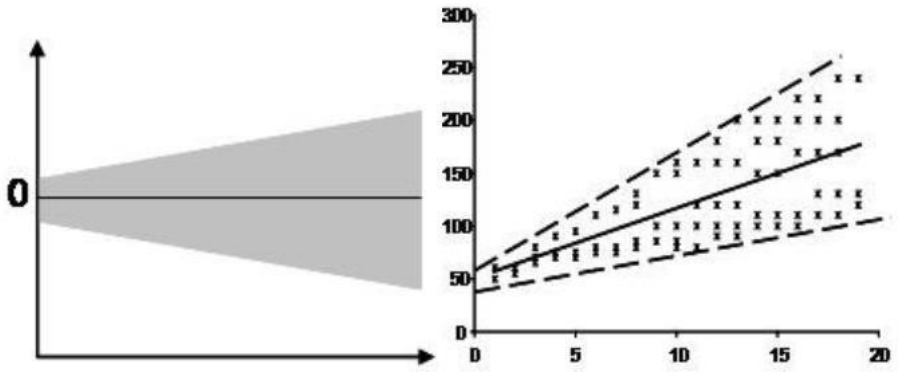
2. The linear regression analysis requires **all variable to be multivariate normal**. This assumption can best be checked with a **histogram or Q-Q plot**. Normally can be checked with the **goodness of fit test** e.g., Kolmogorov-Smirnov test. when data is normally distributed a non-linear transformation (e.g., log- transformation) might fix this issue.



3. The linear regression assumes that there is **little or no multicollinearity in data**. Multicollinearity occurs when the independent variables are too highly correlated with each other.
   Multicollinearity can be tested with 3 criteria's:
   - Correlation matrix
   - Tolerance
   - Variance Inflation Factor (VIF)

4. Linear regression analysis requires that **there is little or no autocorrelation in the data**. Autocorrelation occurs when the residuals are not independent from each other.

5. **Homoscedasticity:** The scatter plot is good way to check whether that data is homoscedastic (**residuals are equal across the regression line**).
following is the example of data that are not homoscedastic



---

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

(2 marks)

Answer: Based on final model top three features contributing significantly towards explaining the demand are:

1. Temperature (0.548)
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.283)
3. year (0.233)

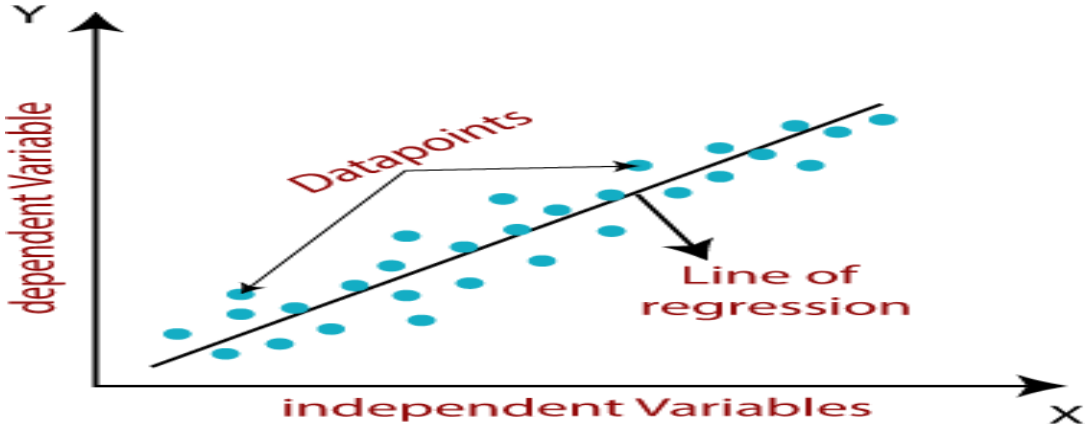
General Subjective Questions

1. **Explain the linear regression algorithm in detail.** (4 marks)

Answer: Linear regression is one of the easiest and most popular **Machine Learning algorithms**. It is a **statistical method** that is **used for predictive analysis.** Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm **shows a linear relationship between a dependent (y) and one or more independent (y) variables**, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.



Mathematically, we can represent a linear regression as:

$$y= a_0+a_1x+ \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

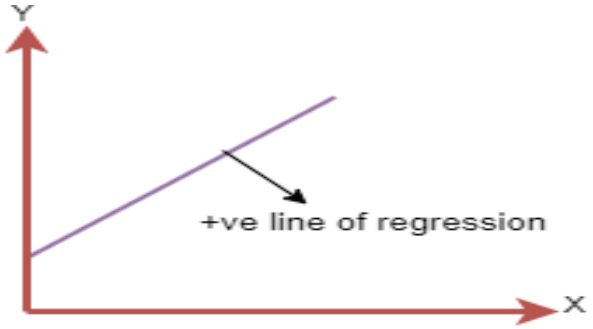
Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

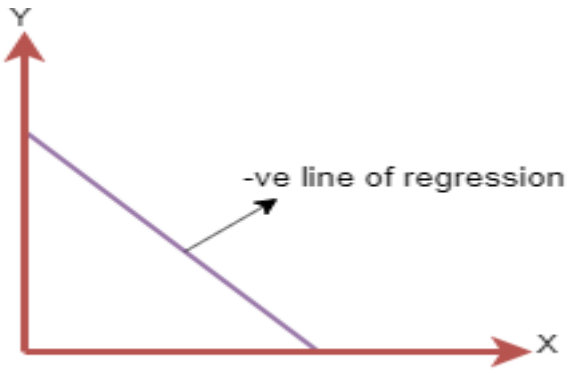
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

Finding the best fit line:

When working with linear regression, our **main goal** is to find the **best fit line** that means the **error between predicted values and actual values should be minimized**. The best fit line will have the **least error**.

The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

$$MSE = 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where,

N=Total number of observations
Yi = Actual value
$(a1x_i+a_0)$= Predicted value.

Residuals: The **distance between the actual value and predicted values** is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

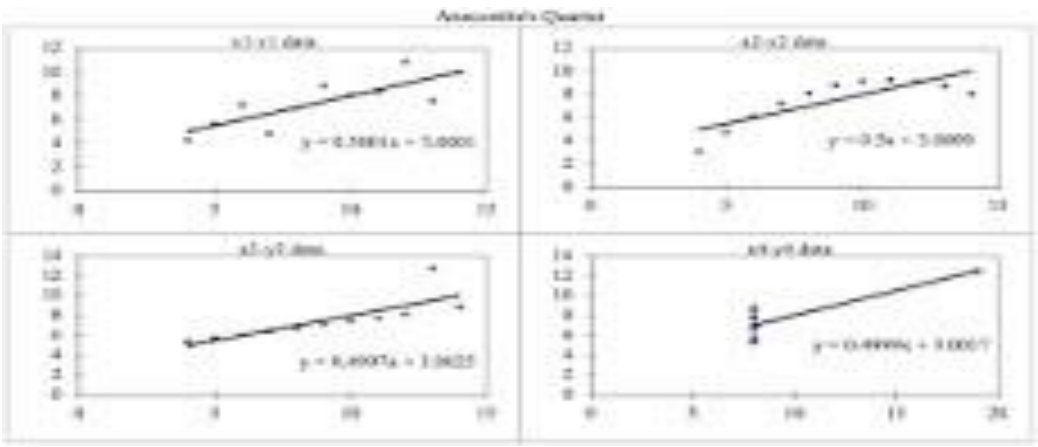
- **Linear relationship between the features and target:**
  Linear regression assumes the linear relationship between the dependent and independent variables.
- **Small or no multicollinearity between the features:**
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
- **Homoscedasticity Assumption:**
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

- **Normal distribution of error terms:**
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.
- **No autocorrelations:**
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

---

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: 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.



to illustrate the **importance** of **plotting the graphs** before analysing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x, y points in all four datasets.

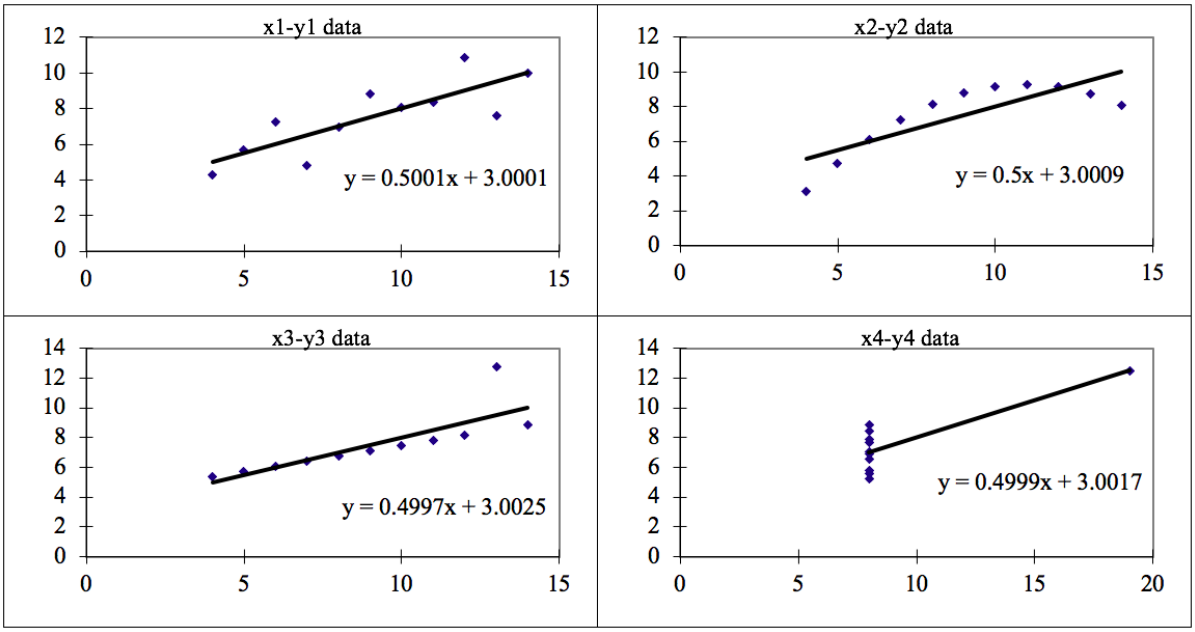
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined 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

The statistical information for all these four datasets is approximately similar and can be 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	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



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.

Conclusion: We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, 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**.

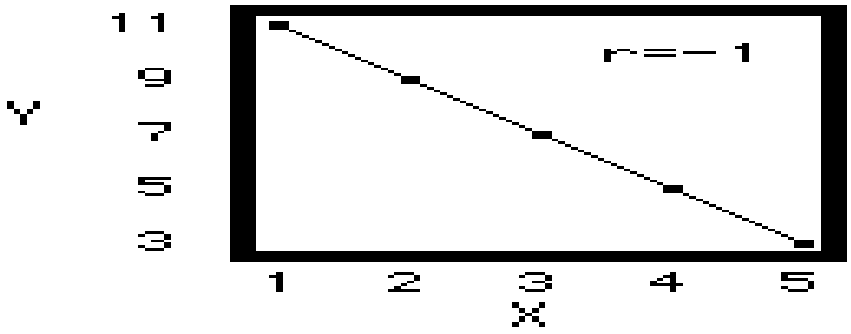
3. **What is Pearson's R?** (3 marks)

Answer: The **correlation** between two variables reflects the degree to which the variables are related. When computed in a sample, it is designated by the letter "**r**" and is sometimes called "**Pearson's r**." **Pearson's correlation** reflects the degree of linear relationship between two variables.

The correlation between 2 variables reflects the degree to which the variables are related. The most common measure of correlation is the **Pearson Product Moment Correlation** (called Pearson's correlation for short). Denoted by **Greek letter rho (ρ)** or computed in a **sample by the letter "r"** and is sometimes called "Pearson's r."

Pearson's correlation reflects the **degree of linear relationship between two variables**. It **ranges from +1 to -1**. A correlation of:

- +1 means that there is a perfect positive linear relationship between variables.
- -1 means that there is a perfect negative linear relationship between variables.
- 0 means there is no linear relationship between the two variables.



Correlations are rarely if ever 0, 1, or -1. Some real data showing a moderately.
The scatterplot below shows arm strength as a function of grip strength for 147 people working in physically-demanding jobs.

---

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** (3 marks)

Answer:

Scaling: Feature **Scaling** is a technique to standardize **the** independent features present in **the** data in a fixed range ..... If feature **scaling** is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as **the** lower values, regardless of **the** unit of **the** values.

It is **performed** during **the** data pre-processing to handle highly varying magnitudes or values or units. If feature **scaling** is not **done**, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as **the** lower values, regardless of **the** unit of **the** values.

Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$X_{changed}=\frac{X-X_{min}}{X_{max}-X_{min}}$$

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

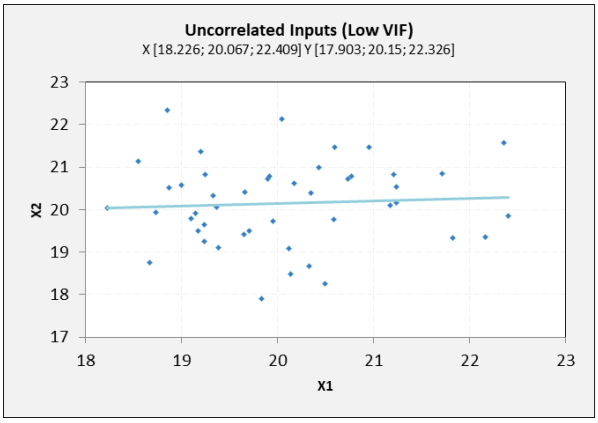
$$X_{changed}=\frac{X-\mu}{\sigma}$$

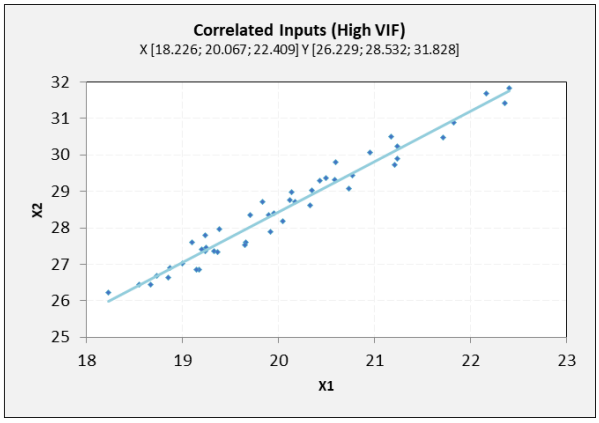
For most applications' standardization is recommended.

---

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?** (3 marks)

Answer: If there is perfect correlation, then **VIF** = **infinity**. A large **value of VIF** indicates that there is a correlation between the variables. If the **VIF** is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.



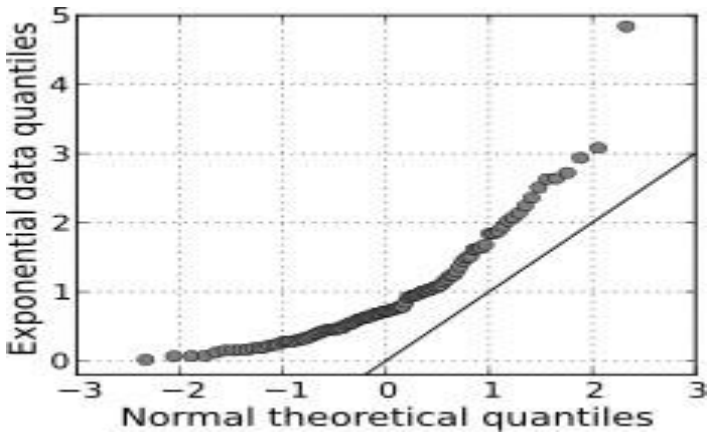


An **infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well).

---

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.** (3 marks)

USE: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the medium is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



A Q-Q plot showing the 45-degree reference line.

The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

Advantages:

1. It can be used with sample sizes also.
2. 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.

It is used to check following scenarios:

1. Come from populations with a common distribution.
2. Have a common location and scale.
3. Have similar distributional shapes.
4. Have similar tail behavior.

Why are QQ plots important?
The purpose of **Q-Q plots** is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the **Q-Q plot**; if the two data sets come from a common distribution, the points will fall on that reference line.