# 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?

Kindly review the following examination of categorical variables within the dataset through box plots:-

>> High bike demands were observed in June, July, August, September, and October.

>> On holidays, there was a relatively lower average demand, possibly due to fewer people going out.

>> The demand in 2019 was notably higher compared to 2018.

>> More significant demand was observed in Summer and Fall compared to Spring and Winter.

1. Why is it important to use **drop\_first=True** during dummy variable creation?

>> It aids in minimizing the creation of unnecessary variables when generating dummy

variables.

>> It assists in mitigating correlations among dummy variables.

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

>> “**temp**” and “**atemp**” has the highest correlation with the target variable

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

The following considerations were addressed for Linear Regression :-

>> Assessing the p-value for each feature, ensuring it is less than 0.05.

>> Evaluating the VIF value for each feature, ensuring it is preferably less than 5, with a

check between 5 and 10 when necessary (valid VIFs are generally less than 5).

>> Verifying that error rates exhibit a normal distribution.

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

Top 3 features contributing are:

* + temp
  + winter
  + light snow rain

# General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm that calculates the linear association between a dependent variable and one or more independent features. When there is only one independent feature, it is referred to as Univariate Linear Regression. In cases where there are multiple features, it is termed as Multivariate Linear Regression.

## Assumption for Linear Regression Model

Linear regression serves as a robust tool for comprehending and forecasting variable behavior, but its accuracy and reliability hinge on meeting specific conditions:

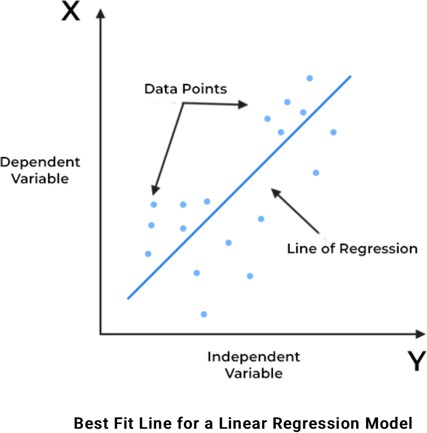
**Linearity**: The independent and dependent variables have a linear relationship with one another.

**Independence**: The observations in the dataset are independent of each other.

**Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant.

**Normality**: The residuals should be normally distributed.

**No multicollinearity**: There is no high correlation between the independent variables. A sloped straight line represents the linear regression model with equation y = mx + c.



In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the ‘best fit line.’ The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

There are two main types of linear regression:

**Simple Linear Regression**: This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable

**Multiple Linear Regression**: This involves more than one independent variable and one dependent variable.

## The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

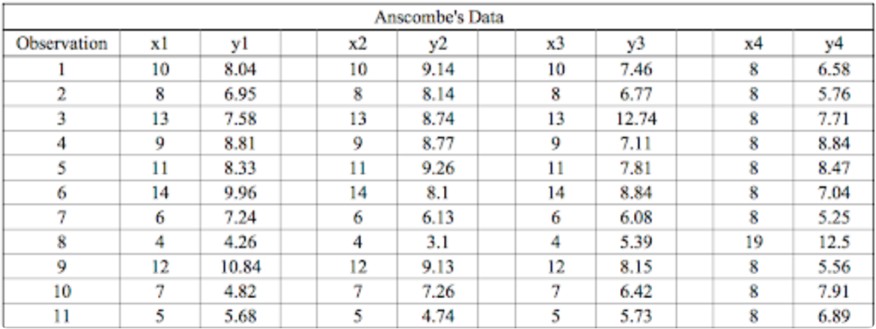
1. Explain the **Anscombe’s quartet** in detail.

Anscombe’s quartet comprises four datasets that exhibit nearly identical simple descriptive statistics. However, subtle peculiarities in each dataset can mislead regression models. Although the quartet shares similar descriptive statistics, their distributions differ significantly, leading to distinct visualizations on scatter plots.

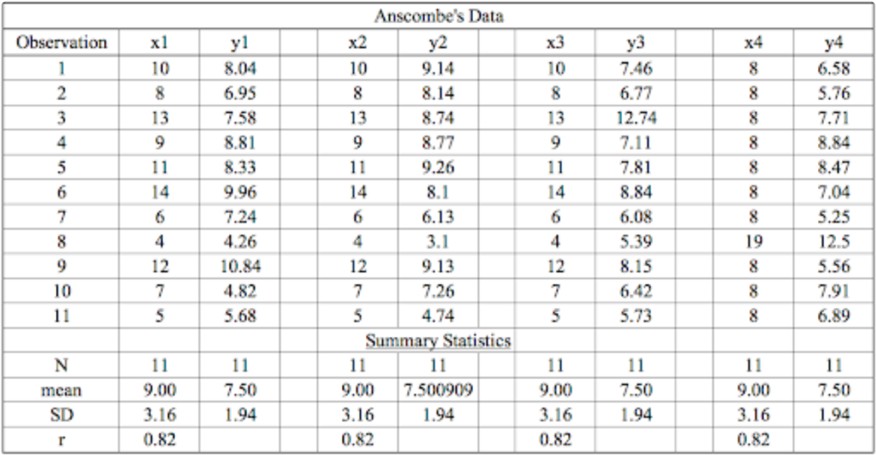
The Purpose of Anscombe’s quartet:

Anscombe’s quartet underscores the importance of visually inspecting data before applying various algorithms for model building. This underscores the necessity of plotting data features to observe sample distributions, aiding in the identification of anomalies such as outliers, data diversity, and linear separability. Moreover, it emphasizes that linear regression is suitable only for datasets exhibiting linear relationships and is ill-equipped to handle other types of datasets.

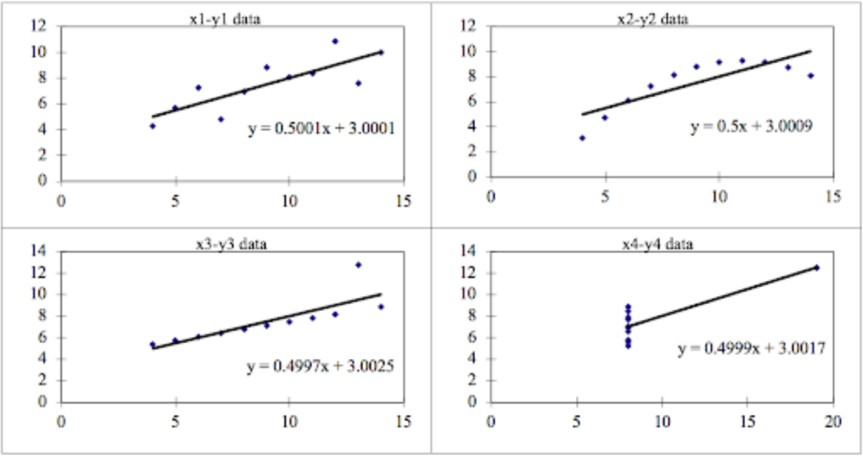
We can define these four plots as follows:



The statistical information for these four data sets are approximately similar. We can compute them as follows:



However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn’t interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

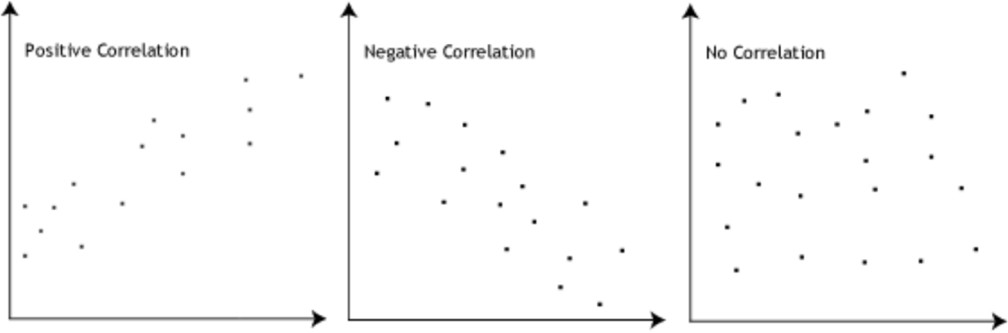
* Dataset 1 – Fits the linear regression model pretty well
* Dataset 2 – Cannot fit the linear regression model because the data is non-linear
* Dataset 3 – Shows the outliers involved in the dataset which cannot be handled by the linear regression model
* Dataset 4 - Shows the outliers involved in the dataset which also cannot be handled by the linear regression model

As you can see, Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

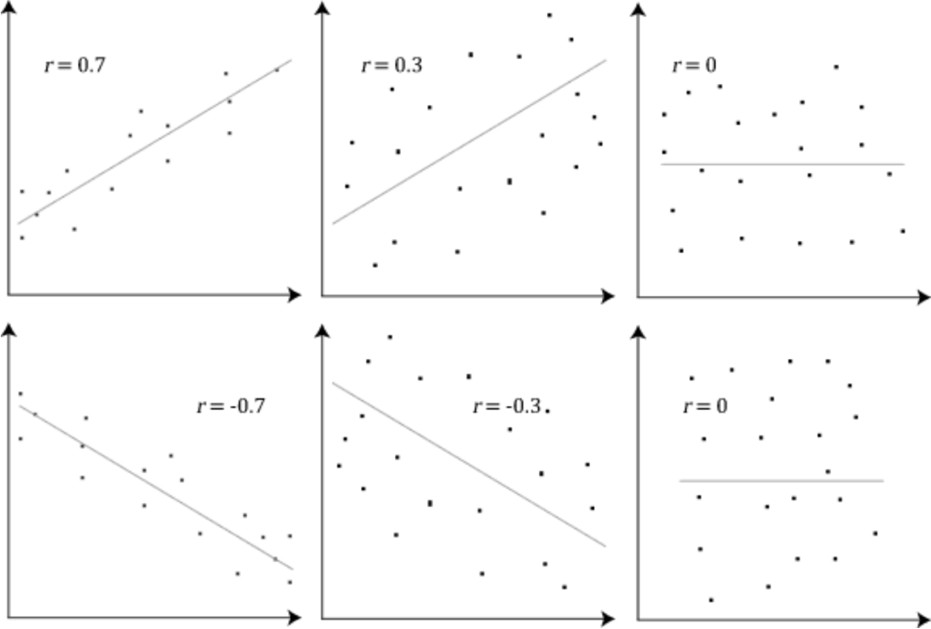
1. What is Pearson’s R?

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

**The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1**. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



**The stronger the association of the two variables, the closer the Pearson correlation coefficient**, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature

scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.

Hence, **Scaling the data can help to balance the impact of all variables on the distance calculation and can help to improve the performance of the algorithm**. In particular, several ML techniques, such as neural networks, require that the input data to be normalized for it to work well.

Normalized scaling and Standardized scaling are the 2 types of scaling techniques.

|  |  |
| --- | --- |
| Normalized Scaling | Standardized Scaling |
| Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| It is used when features are of different  scales | It is used when we want to ensure zero  mean and unit standard deviation. |
| Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| It is really affected by outliers. | It is much less affected by outliers. |
| Scikit-Learn provides a transformer called **MinMaxScaler** for Normalization. | Scikit-Learn provides a transformer  called **StandardScaler** for standardization. |
| This transformation squishes the n- dimensional data into an n-dimensional  unit hypercube | It translates the data to the mean vector of original data to the origin and squishes or  expands. |
| It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |

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

A Variance Inflation Factor (VIF) serves as a metric quantifying the extent of multicollinearity within regression analysis. Multicollinearity occurs when there is a correlation among multiple independent variables within a multiple regression model, potentially impacting the accuracy of regression results. Therefore, the VIF provides an estimate of how much the variance of a regression coefficient is inflated due to multicollinearity.

An infinite VIF indicates perfect correlation, signifying a substantial relationship between variables. A high VIF value suggests a correlation among variables; for instance, a VIF of 4 indicates that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity. This inflation results in a corresponding doubling of the standard error of the coefficient (as the square root of variance is the standard deviation).

Moreover, a perfect correlation between two independent variables is reflected in a scenario where R2 equals 1, leading to a VIF calculation of 1/(1-R2), which becomes infinite. To address this issue of perfect multicollinearity, it is necessary to remove one of the variables causing this correlation from the dataset.

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

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

## Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

## Importance of Q-Q plot:

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 than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.