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?

From the analysis of the categorical variables, there seems to be a trend in the following variables:

1. Seasons - There is a higher number of bikes used in summer and fall. The same pattern shows in the months as well
2. Weather - Highest number of bikes used in clear weather
3. Year - There is a significant increase in the number of bikes used in 2019

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

It is important to keep our model as simple as possible. We do not want any information that is redundant to our model. And that means we need to carefully select only the useful variables.

When a categorical variable is encoded using dummy variables, each category becomes its own column to indicate which category an observation belongs to. However, as it turns out, we need one less column to represent the total number of categories to represent all categories. For example, if our variable had 3 categories, we need only 2 categories to represent all 3 states of that variable **(n-1, n = number of categories)**.

The **drop_first=True** command eliminates the first column since it is redundant.

This is also important since the R-Squared is sensitive to the number of variables. Hence this helps us keep our model as light as possible.

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

The Variable **'temp'** (ie, temperature) has the highest correlation with the target variable

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

The model that was chosen was validated on the basis of statistical inferences such as

1. The goodness of the fit - R^2
2. The goodness of the fit - Adjusted R^2
3. Mean squared error
4. Predictive accuracy of the model
5. Analysis of the residual
6. P-values of variables
7. F - statistic

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

Based on the model, the top 3 features were

1. 'Temp' (temperature)
2. Light_rainsnow (weather)
3. 2010 (year)

General Subjective Questions

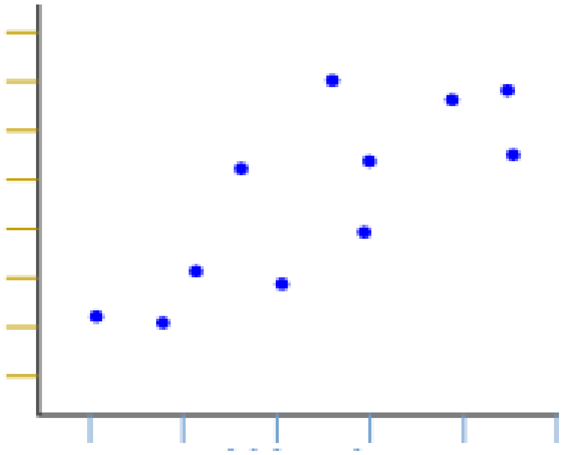
1. Explain the linear regression algorithm in detail

The linear regression algorithm is one of the most basic and widely used algorithms in machine learning. The algorithm is used for forecasting and making predictions based on the data it has studied. The algorithm identifies linear relationships (assuming they are linearly related) in order to make predictions.

The Linear regression makes use of a well-known formula in math/geometry ie, the equation of a line. $y = mx + b$, where y is what we are predicting (y-axis), m is the slope or the coefficient, x (x-axis) is our input value, and b is the y-intercept or the constant

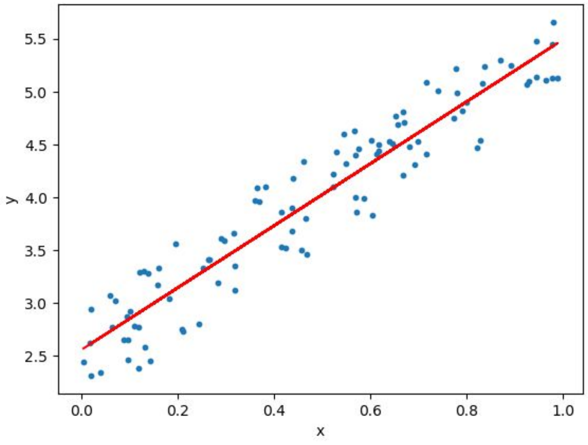
For simplicity sake, let's assume our data set has only 1 predictor and 1 target variable (the same principle applies to multiple variables)

Assume, that we have a dataset which comprises of prices of houses and their size in the area. Now when we plot this data, we see a trend like this. The x-axis is the size (predictor variable) and the y-axis is the price of the house (target variable)



We notice a trend. We see that as the area of the house increases, the prices increase as well. The regression algorithm studies this relationship, and now tries to predict the price of a house for an area that was not in the data set.

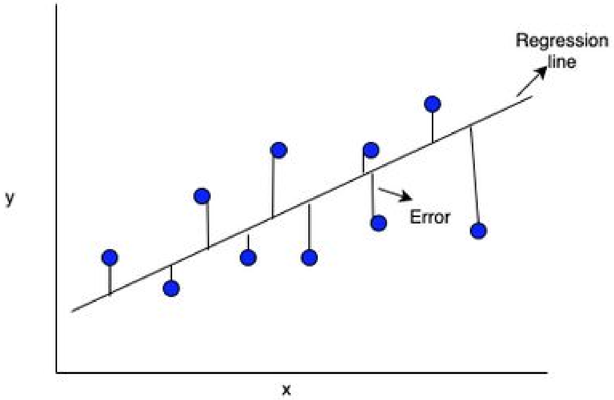
It does so by *fitting a line in the data.*



So for every point on the x axis, the lindicates indicates the predicted value on the y axis.

Finding the line of best fit

The way we find the line of best fit line is to minimize the **cost function** by minimizing the error value
The error value is the difference between the predicted value and the actual value. The algorithm fits a line in a manner that the square of the residual is at its minimum for that data set. This is also called the mean squared error (MSE).



$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

This regression line is expressed in the form of y = mx+b. However, in real life, any outcome is usually a result of various factors. We can add multiple variables to this equation to find relations between various variables.

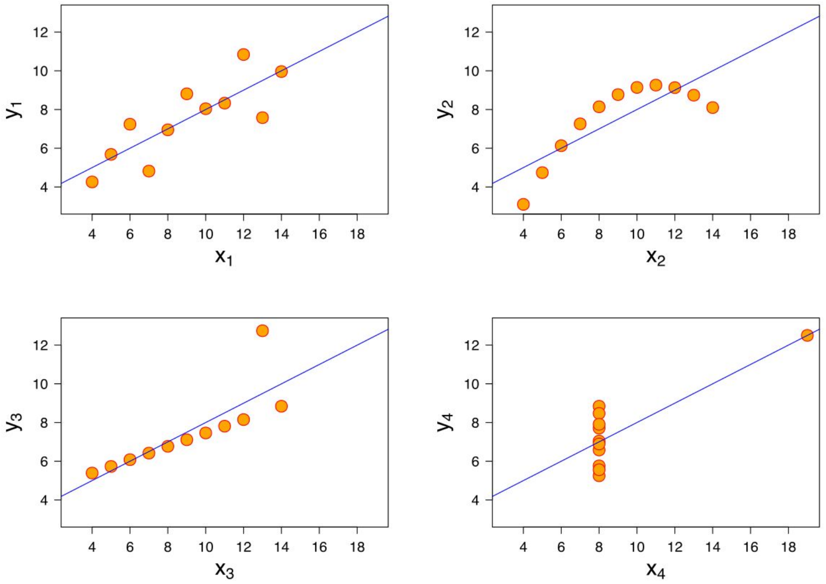
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

2. Explain Anscombe’s quartet in detail.

Anscombe’s quartet (developed by Francis Anscombe) consists of 4 data sets that demonstrate how very different data can have the same statistical information.

Each of the datasets consists of 1 predictor and one target variable.

1. The first dataset consists of data that has a linear relationship.
2. The second dataset is no linear that takes the shape of a parabola
3. The third dataset has a perfectly linear relationship and one extreme outlier
4. The fourth data shows an example when one high-leverage point is enough to produce a high correlation coefficient.



The mean, variance, standard deviation, correlation coefficient, R-squared, the regression line is nearly identical in all 4 datasets, even though the data sets are very different.

This shows the importance of plotting our data, and not just going by statistical information.

3. What is Pearson's R?

The correlation coefficient is a measure of how strong 2 variables are correlated. The Pearson's R is a widely used method to calculate the correlation coefficient between 2 variables.

The Pearson's R gives a value between -1 and 1, where -1 indicates a perfect negative correlation and 1 indicates a perfect positive correlation and 0 indicates no correlation between the variables. For example, the higher the work experience, the higher the salary is likely to be. Hence there is a strong positive correlation.

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

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

In most cases, when we are finding relations between various predictor variables and the target variable, the variables will be on vastly different scales. For example, our target variable could be in ‘Kilometer units’ in the range of 0 to 10 and our predictor could be in Millions of dollars, and another variable could be in thousands of dollars.

Even though our regression can find patterns and make predictions, it will be difficult to interpret since there are so different units in vastly different scales.

Scaling is performed to bring all of these variables in one scale, making interpretation easy to comprehend.

The 2 popular types of scaling are normalization and standardization.
Normalization - This technique is to re-scales features with a distribution value between 0 and 1. For every feature, the minimum value of that feature gets transformed into 0, and the maximum value gets transformed into 1.

Standardization - This technique re-scales the features of the distribution such that the mean of the data is 0 and the standard deviation is 1.

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

A high VIF (Variance inflation factor) value indicted a high multi-collinearity between its variables. In most cases, we want to get rid of these variables since its effect it is already being expressed by other variables.

However, when there is a perfect correlation between the variables, the VIF becomes infinity. We can look at the formulae to see why

VIF = $1/1 - R^2$

If our data is perfectly correlated, the R^2 will be 0. And when the R^2 is 0, the denominator becomes 0 and the fraction becomes infinity.

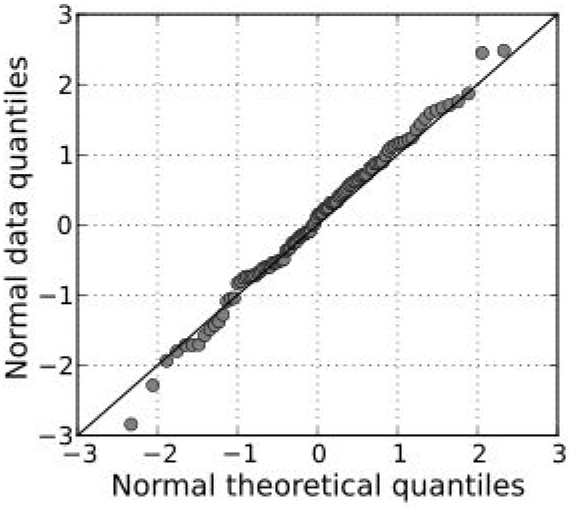
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot (quantile-quantile plot) is a plot that tells you of the datasets follow a normal distribution.
The data set is first converted into which quantile it falls into (sample quantile). It is then plotted against the quantile of a normal distribution (theoretical quantile).

If the sample quantile and the theoretical are similar, we will get a straight line from the graph. This means our data set is (approximately) normally distributed.

The Q-Q plot is important in linear regression. For our linear model to work, we make an assumption that our data is normally distributed, in order to get an accurate p-value. We also make an assumption that the residuals are normally distributed.

Hence it is important to use the Q-Q plot to graphically validate our assumptions.


Normal data quantiles
Normal theoretical quantiles