Assignment - based subjective Questions

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

Ans:
The demand of bikes is less in the month of Spring when compared to the other seasons

- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- The demand of bike is almost similar throughout the weekdays.
- The demand bike increased in the year 2019 when compared with year 2018.
- Bike demand is less in holidays in comparison to not being holiday.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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 us in reducing the extra column created during dummy variable creation. It reduce correlations created among dummy variables.
A variable.

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

Ans:
From the correlation map, temp, atemp and days_old seems to be highly correlated.
We can see that cnt is linearly increasing with temp indicating linear relation.

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

Ans: we can see from the scatter plots, temperature has a clear linear relationship with cnt.

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

Ans :
The top 3 features contributing significantly towards the demands of shared bikes are:
1. Weatherise_light_snow (negative correlation)
2. yr_2019(positive correlation)
3. Temp(positive correlation)

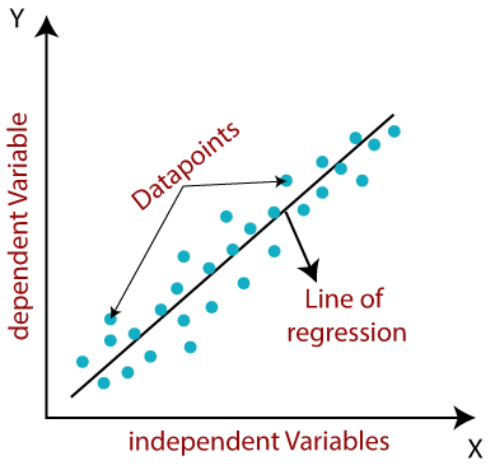
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans : linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. 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. Consider the below image:



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 divided into two types :

* Simple Linear Regression

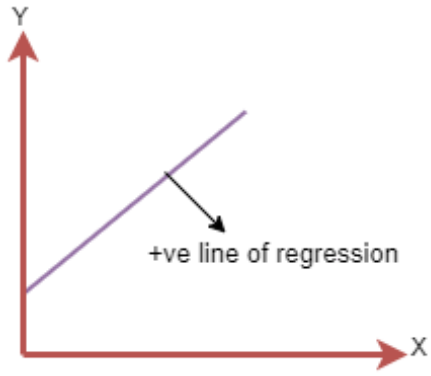
* Multiple Linear Regression

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.

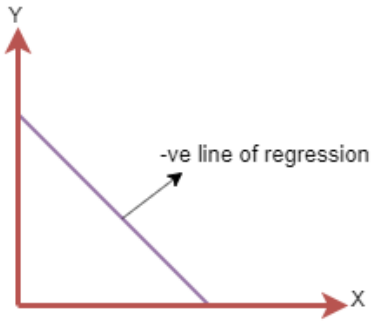
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **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: $Y= -a_0+a_1x$

Steps involved in Linear Regression Algorithm

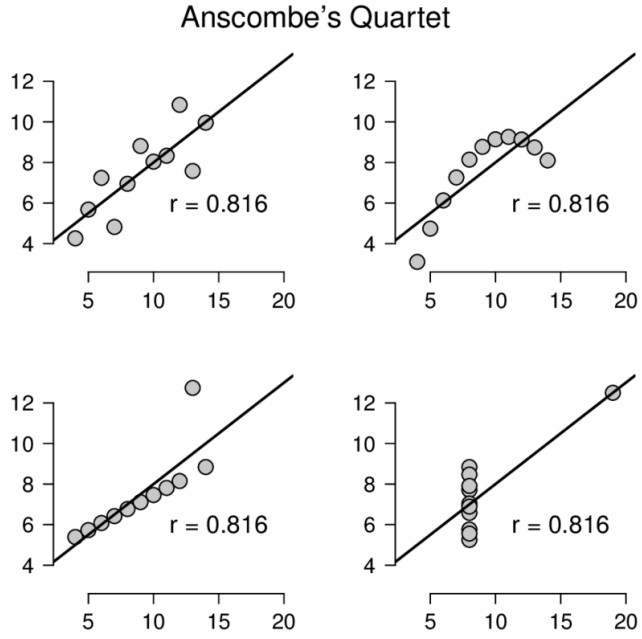
Since we have covered the basic concepts now let's look at the steps involved in the linear regression algorithm.

1. Prepare the given data. Read more from here.
2. Decide the hypothesis function (i.e. for simple linear regression, y = a + bx is the hypothesis function)
3. Initialize **a,** and **b** with some random values.
4. Update the parameters **a,** and **b** using gradient descent algorithm i.e.
    - Calculate y_predicted, **y_predictedi = a + bxi**
    - Calculate cost function,
    - $$j = \frac{1}{n}\sum_{i=1}^{n}(y\_actual_i - y\_predicted_i)^2$$
    - Compute the gradient of cost function with respect to parameters (**dJ/da**, **dj/db**)
    - Update a and b using that gradient:
        - **a = a – lr*(dJ/da)**
        - **b = b- lr*(dJ/db), lr** is learning rate.
    - Repeat from steps I to iv until the desired result is obtained (i.e. cost function is minimized)

5. Once the gradient descent is completed we will get updated values of **a,** and **b** for which the cost function is minimum. And line corresponding to those values will be the best fit line.

2. Explain the Anscombe's quartet in detail.

Ans : 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.



Anscombe's quarter highlights the importance of plotting data to confirm the Validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, r=.816. in fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.

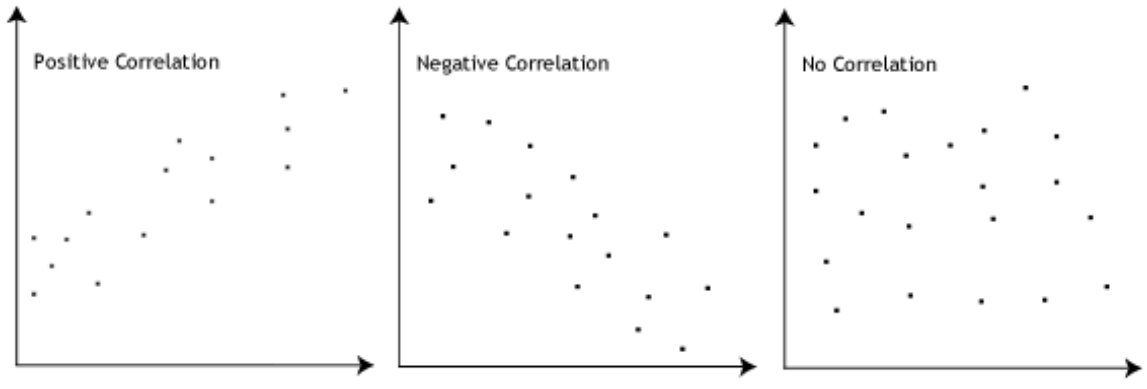
3 What is Pearson's R ?

Ans:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between –1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Ans :

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, 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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

S.NO.	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.

3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6	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.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect 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).

The value of VIF is calculated by the below formula:

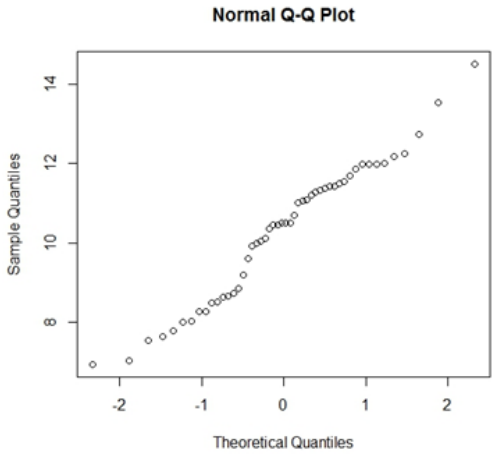
$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
Importance of Q-Q plot: Below are the points:
I. The sample sizes do not need to be equal.
II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
III. The q-q plot can provide more insight into the nature of the difference than analytical methods.