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-
   Following are the infer about effect on the dependent variable,

- Number of bookings are higher in fall season followed by summer.
- number of bookings stars rising from the starting of year it is at pick at middle of year & stars decreasing towards the end of year.
- More bookings are there in Saturday as compared to other days
- Number of bookings are higher in clear weather sit
- Most of the bookings has been done during the month of May, June, July, august, September and October.
- Trend increased starting of the year till mid of the year and then it started decreasing towards the end of year.
- Number of bookings for each month is increased from 2018 to 2019.
- Clear weather has more booking. Booking increased for each weather in 2019 as compare to in 2018.
- Wen, Thu, Fri, Sat have a greater number of bookings as compared to the other days of week.
- Bookings are more on working day as compare to non-working day. the count increased from 2018 to 2019.
- More numbers of booking done in year 2019 as compare to 2018.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer-
   Importance to use drop_first=True during dummy variable creation,
   drop_first=True is important to use, as 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.

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

   Answer-

   'Temp' variable 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? (3 marks)**

   Answer-

   I have validated the assumptions of Linear Regression upon following assumptions-

   1.**Normality of error terms**-It should be normally distributed.
   2.**Multicolinearity** -There should be insignificant multicollinearity among variables.
   3.**Linear relationship**- Linearity should be visible among variable
   4. **Homoscedasticity**- There is no visible pattern in residual values. (conical).
   5.**Independance of residuals**- No auto - corelation

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

   Answer-

   Following are the top 3 features contributing significantly towards explaining the demand of the shared bikes,

   1.Temp
   2.Winter
   3.Summer

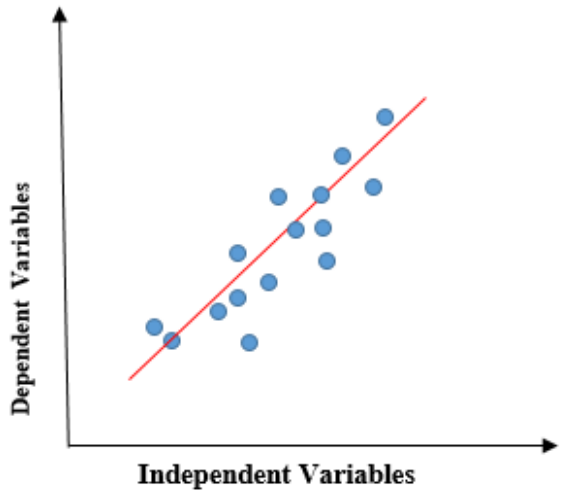
General Subjective Questions

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

   Answer-

   linear regression algorithm-

   Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line.

Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

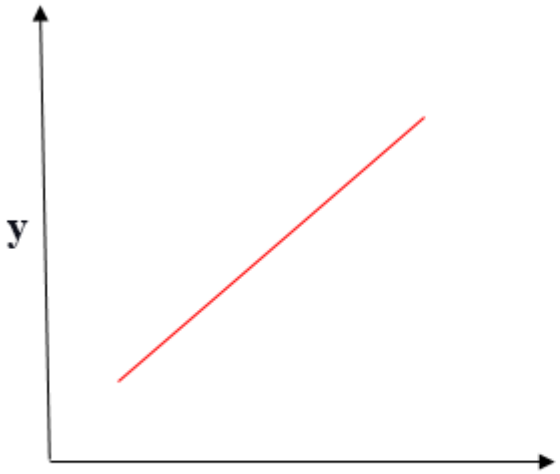
$$y = mx+b \implies y = a_0+a_1x$$

y= Dependent Variable.
x= Independent Variable.
a0= intercept of the line.
a1 = Linear regression coefficient.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

Positive Linear Relationship

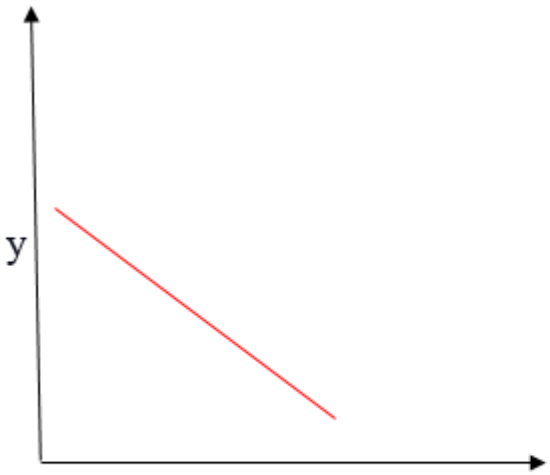
If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



$y = a_0 + a_1 x$

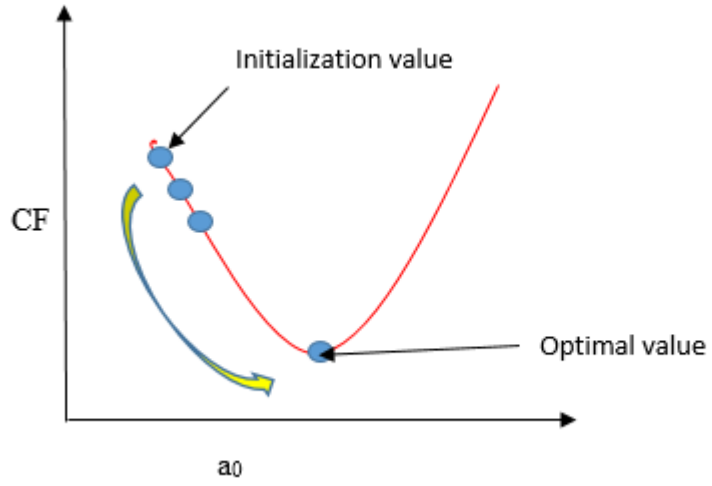
Negative Linear Relationship

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



$y = -a_0 + a_1 x$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.



To update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives for a_0 and a_1. The partial derivates are the gradients, and they are used to update the values of a_0 and a_1. Alpha is the learning rate.

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.

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:

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

2. Explain the Anscombe's quartet in detail.
Answer-
Anscombe's quartet

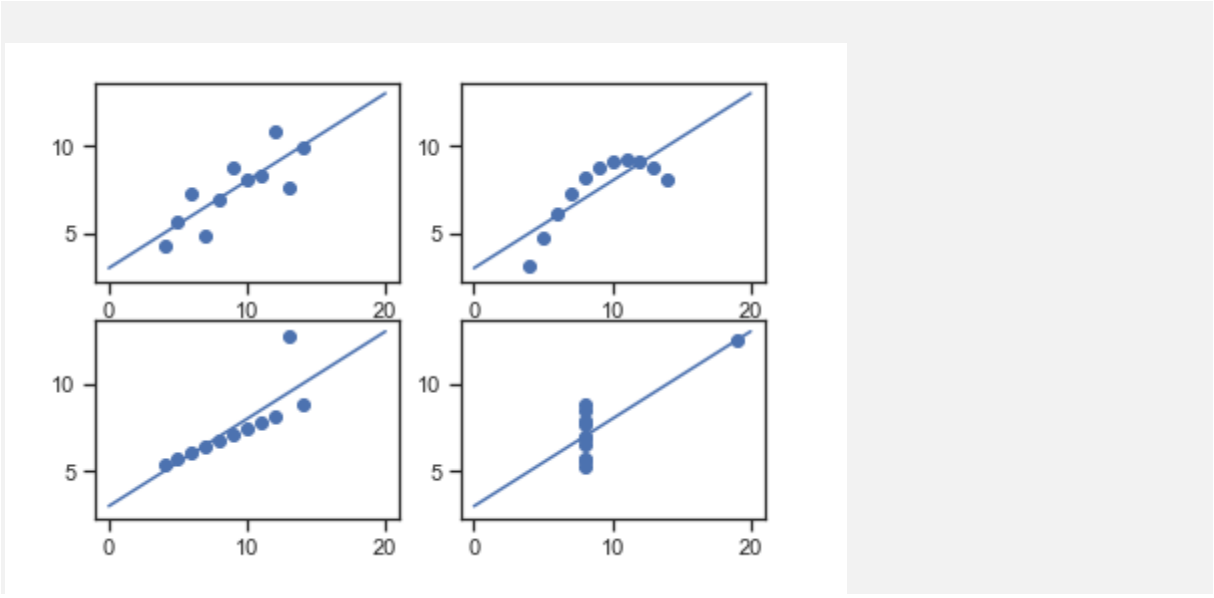
Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Four Data-sets,
Apply the statistical formula on the above data-set,
Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation: y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

- Data-set II — shows a curve shape but doesn’t show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

3. **What is Pearson’s R?**
   Answer-

   Pearson’s R-

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula given 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

Some steps are needed to be followed,

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.

Person	Age (X)	Income (Y)	XY	X^2	Y^2
1					
2					
3					
4					

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 / √{2000} {81000000}

= 400000 / √162000000000

= 400000 / 402492.24

= 0.99

The positive value of Pearson's correlation coefficient implies that if we change either of these variables, there will be a positive effect on the other. For example, if we increase the age there will be an increase in the income.

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

   Scaling-
   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.

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.

Difference between normalized scaling and standardized scaling?

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	Scikit-Learn provides a transformer called StandardScaler for

	Normalization.	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**?
   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.

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

   Q-Q plot-

   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 median is a quantile where 50% of the data fall below that point and 50% lie above it.
   The use and importance of a Q-Q plot in linear regression-

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.