Linear Regression Bike Sharing Assignment

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)**

Ans: As per my analysis the categorical variables in the dataset were season, yr, mnth, weathersit, holiday and working day.

I could infer about their effect on the dependent variable

i. Season

The number of bike bookings are higher in Summer, Fall and Winter. The number of bookings drop in Spring season.

ii. Year/yr

The number of bookings increases every year. we can observe that the booking is increased in 2019 compare to 2018.The number of bookings next year will even be higher. It might be due to the fact bike rentals are getting more popular as people are becoming more conscious about their health and environment.

iii. Months/mnth

The average number of the bike bookings were happening in the months of May to Sep with a median of over 4000 bookings per month. It shows that the mnth can be a good predictor for the dependent variable.

iv. Weather Situation/weathersit

- In Clear, Few clouds, Partly Cloudy weather maximum bike bookings. Median of count for Clear and Cloudy weather lies between 4000 to 6000
- Less bookings when weather is with light snow or light rain median is close to 2000
- No bike bookings when heavy rain or heavy snow is there

v. Holiday

Most of the bike bookings were happening when there is not an holiday. It means holiday can't be a good predictor for the dependent variables.

vi. working day

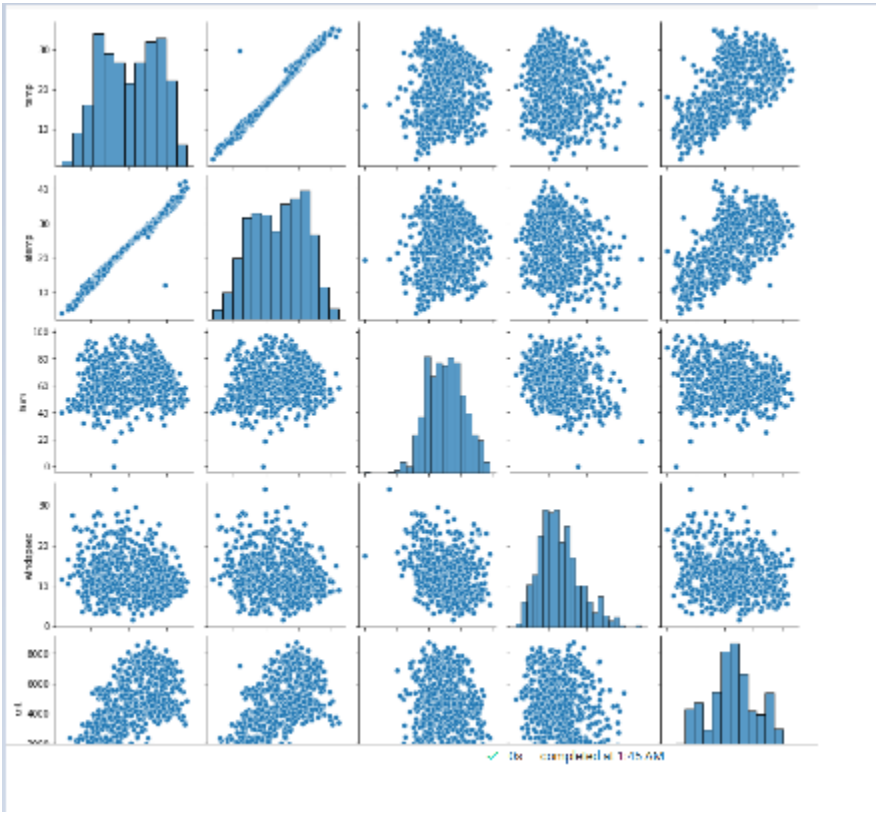
We can see that most of the bike bookings were on the higher end on days which were marked as non-working days

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

Ans: When drop_first is set to True, n-1 dummy variables are created for the categorical variable. For data analysis, a dummy variable n-1 is sufficient. So it is very important to use drop_first = True as it helps to reduce the extra columns created when the dummy variable is created. Therefore, it reduces the correlations generated during dummy variables. If one of the dummy variables generated from the categorical variable is not deleted, there is a constant variable (intercept) that creates a multicollinearity problem, so it is redundant with the data set.
Example: Iterative models can have convergence problems and skew the list of variable importance. Another reason is that multicollinearity occurs between dummy variables when all dummy variables are present. You lose one column to control this.

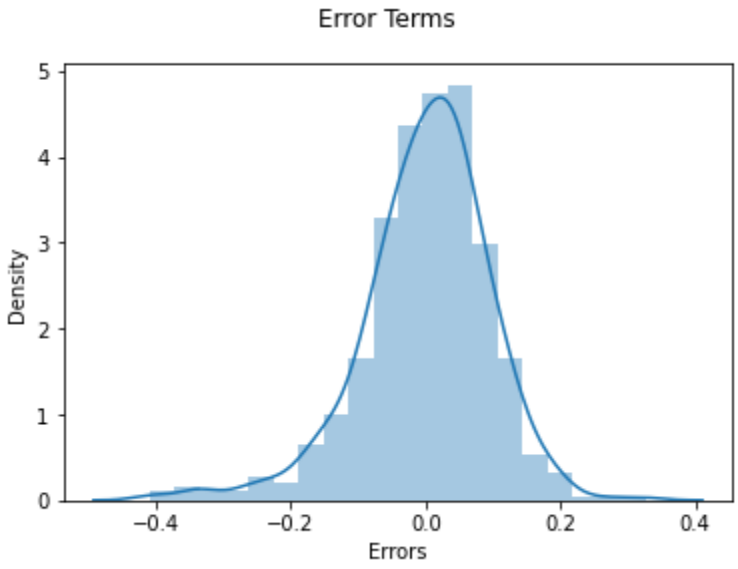
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Ans: By looking at the pair-plot among the numerical variables, the two variables "temp" and "atemp" are highly correlated with target variable (cnt).



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

Ans: After building a linear regression model on the training data set, the assumption is that the errors are normally distributed. To support this, residual analysis was performed. The remainder is the error of the difference between the actual y value and the y value predicted by the model. The residual distribution should follow a normal distribution and should be centered at 0 (mean = 0). Check this assumption about the residuals by plotting the variance of the residuals and checking whether the residuals are normally distributed. The chart above shows that the residuals are distributed around mean = 0.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Ans: Based on the final model, top 3 features contributing significantly towards explaining the demand of the shared bikes

Temperature (temp) : The demand for bike booking rises with the increase in temperature. For every unit increase in temperature, the bike booking count increases, when all other variables are kept constant.

Year(yr) : As the year variable increases, there is an increasing demand of bikes. For every increase of one year, the bike booking count is expected to increase, keeping all other variables as constant.

Light Rain & Snow : Indicates that the light snow and rain deters people from renting out bikes:

General Subjective Questions

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

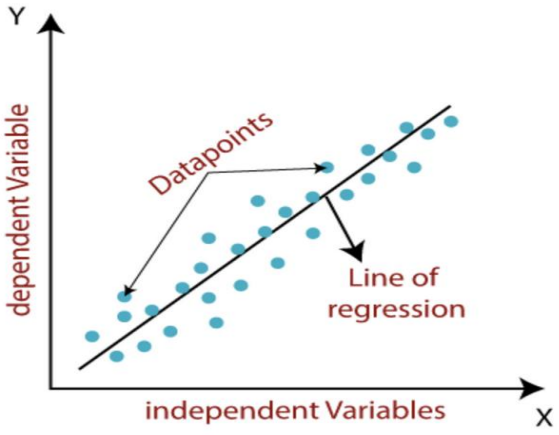
Ans: Linear regression is a machine learning algorithm based on supervised learning. It does regression. Regression models the target value of the prediction based on independent variables. It is primarily used to determine and predict relationships between variables. Different regression models depend on the type of relationship between the dependent and independent variables under consideration and the number of independent variables used.
Linear regression works by predicting the value of the dependent variable (y) based on the given independent variable (x). Therefore, this regression method finds a linear relationship between X (input) and Y (output).

A simple linear regression equation can be expressed by

$$Y = m*x + c$$

Here,
y : is the predicted variable (dependent variable),
m is slope of the line,
x is independent variable,
c is intercept(constant).

It is cost function which helps us to find the best possible value for m and c which in turn provide the best fit line for the data points



Types of Linear Regression
Linear regression is of the following two types –

- **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line.

  Formula: **Y=β0+β1X1 +ϵ**
- **Multiple Linear Regression**: It shows the relationship between one dependent variable and several independent variables.

  Formula: **Y=β0+β1X1+β2X2+…+βpXp+ϵ**

Where **β**1, **β**2, **β**3 are coefficients or slopes for the variables **X**1, **X**2 and **X**3 respectively and **β**0 is the intercept

A model is a linear when it is linear in parameters relating the input to output variables. The dependency need not be linear in terms of inputs for the models to be linear. For example, all equations below are linear regression, and they define the model that represents the relationship between model parameters.

Linear Regression Type	Mapping Relation	Equation Type
Simple Linear Regression	$y \rightarrow X; X = x_1$	$y = \beta_0 + \beta_1 x_1$
Multiple Linear Regression	$y \rightarrow X; X = [x_1, x_2]$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

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

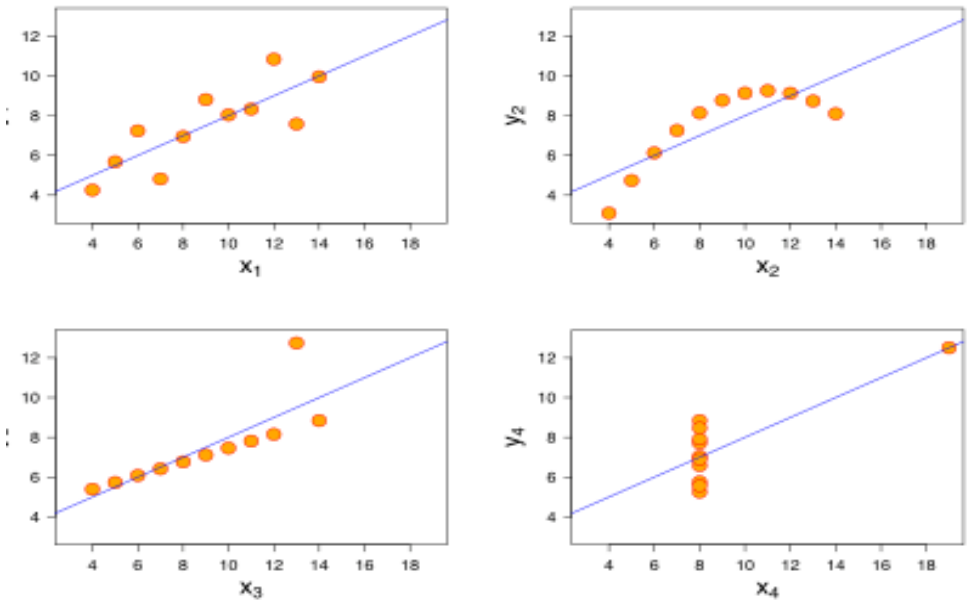
Ans: Anscombe's Quartet is a modal example of the importance of data visualization, developed by statistician Francis Anscombe in 1973 to demonstrate the importance of plotting data before analyzing it for its statistical properties. It consists of 4 data sets, each data set of 11 (x, y) points. The main thing to analyze for these data sets is that they all have the same descriptive statistics (mean, variance, standard deviation, etc), but have different graphical representations.

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

If we apply the statistical formula on the above given data-set, we get the following values

i. Average Value of x = 9
ii. Average Value of y = 7.50
iii. Variance of x = 11
iv. Variance of y =4.12
v. Correlation Coefficient = 0.816
vi. Linear Regression Equation : y = 0.5 x + 3

Four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties

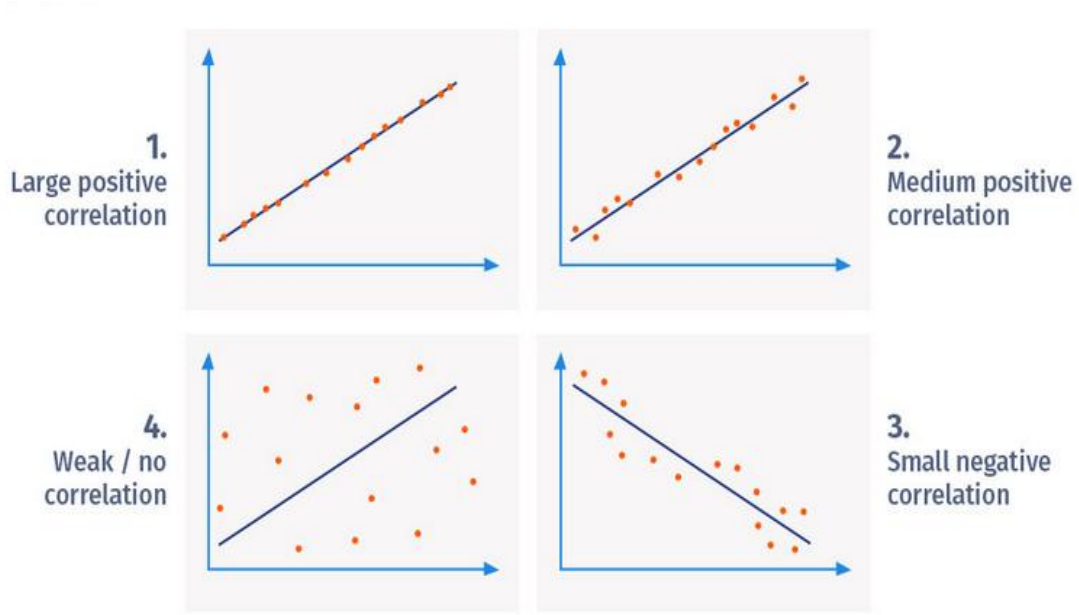


Statistical Properties

- The first scatter plot (top left) appears to be a simple linear relationship. So this fits the linear regression model pretty well
- The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line. shows the outliers involved in the dataset which cannot be handled by linear regression model
- In the last, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables

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

Ans: Pearson's r is a numerical summary of the strength of the linear association between the
variables. If the variables move up and down together, the correlation coefficient will be positive. Pearson's R measures the strength of a linear relationship between two variables. Simply put, Pearson's correlation coefficient calculates the effect of changing one variable when the other changes.



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

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

The formula for Pearson's R is

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

Here,

R = Correlation coefficient

xi=values of the x-variable in a sample

$\bar{x}$=mean of the values of the x-variable

$\bar{y}$=values of the y-variable in a sample

yi=mean of the values of the y-variable

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

Ans: Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data preprocessing to handle highly varying magnitudes or values or units. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.

Why scaling is performed because, 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 modeling. To solve this problem, 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.

Normalized scaling:

i. It is a process where the variables are scaled in the range of 0 and 1.
ii. It is also called as MinMaxScaling
iii. In python, **sklearn.preprocessing.MinMaxScaler** helps to implement normalization

MinMaxScaling: X = X – Min(x) / Max(x) – Min(x)

Standardized scaling:

i. It is a process where the variables are scaled in a way that each data has it's mean as 0 and a standard deviation of 1.

ii. In python, sklearn.preprocessing.scale helps to implement standardization

iii. One disadvantage of normalization over standardization is that some information in the data is lost, especially regarding outliers.

Standardization: X = X – Min(x) / SD(X)

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

Ans: The Variance Inflation Factor(VIF) is a measure of collinearity among predictor variables
within a multiple regression.

The formula of **VIF is 1/(1-R^2).**

- Here R^2 denotes that how much variable is co-related to other variables.
- When R^2 = 1 then VIF = Infinity.
- That means when there is a perfect co-relation then VIF will be infinity.

If there is perfect correlation, then VIF = infinity. Where R is the R-square value of that independent variable which we want to check how well the independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.
So, VIF = 1/(1-1) which gives us VIF = 1/0 which results the VIF value as infinity

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

Ans: Q-Q (Quantile-Quantile) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential or Uniform. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

It can be used to identify

- If 2 samples are similarly distributed or not based on the fit- line passing closely w.r.t to the plotted points or not
- It can explain if the distribution scale is similar or not depending on the angle or slope of the fit-line
- It can also be used to explain what kind of distribution best fits the sample data by fitting q-q plot between quantiles of the dataset and quantiles of different distribution(uniform/normal etc)

It is used to check the below scenarios:

- If two data sets come from populations with a common distribution
- If two data sets have common location and scale
- If two data sets have similar distributional shapes
- If two data sets have similar tail behavior