Name :- Rajat Mandaniyan

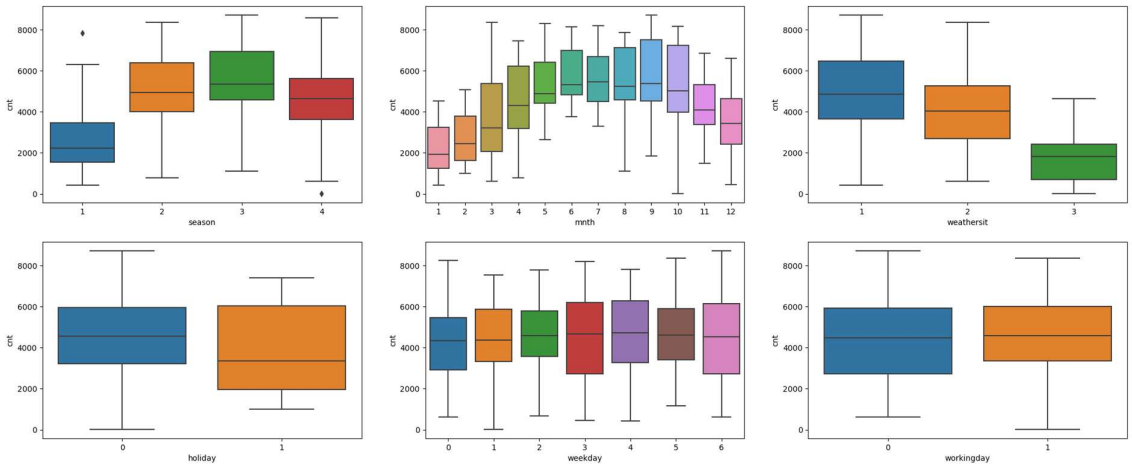
EmailD :- rajatkajob@gmail.com

Batch MLC-46 Batch (AI and ML)

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)

There are 6 categorical variables in the dataset.



We used Box plot (refer the fig above) to study their effect on the dependent variable ('cnt') .

The inference that We could derive were:

- **Season:** Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- **mnth:** Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- **weathersit:** Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent Variable
- **holiday:** Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- **weekday:** weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

- **workingday:** Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

Ans :- If there are n dummy variable values for the categorical field , then it can be represented with n-1 values or (n-1) values are sufficient to uniquely identify all the possible values for the categorical field.

For e.g :- Categorical field :- Season have four values season1 , season2 , season3, season4

Then it can be represented as

Season1 :- 100

Season2 :- 010

Season3 :- 001

Season4 :- 000

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

Ans:-

```
In [142]: 1 bike_num.corr()
```

Out[142]:

	temp	atemp	hum	windspeed	cnt
temp	1.000000	0.989610	0.173758	-0.161614	0.627185
atemp	0.989610	1.000000	0.186440	-0.190693	0.630458
hum	0.173758	0.186440	1.000000	-0.281480	-0.059645
windspeed	-0.161614	-0.190693	-0.281480	1.000000	-0.239927
cnt	0.627185	0.630458	-0.059645	-0.239927	1.000000

Outcome :- The above Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'

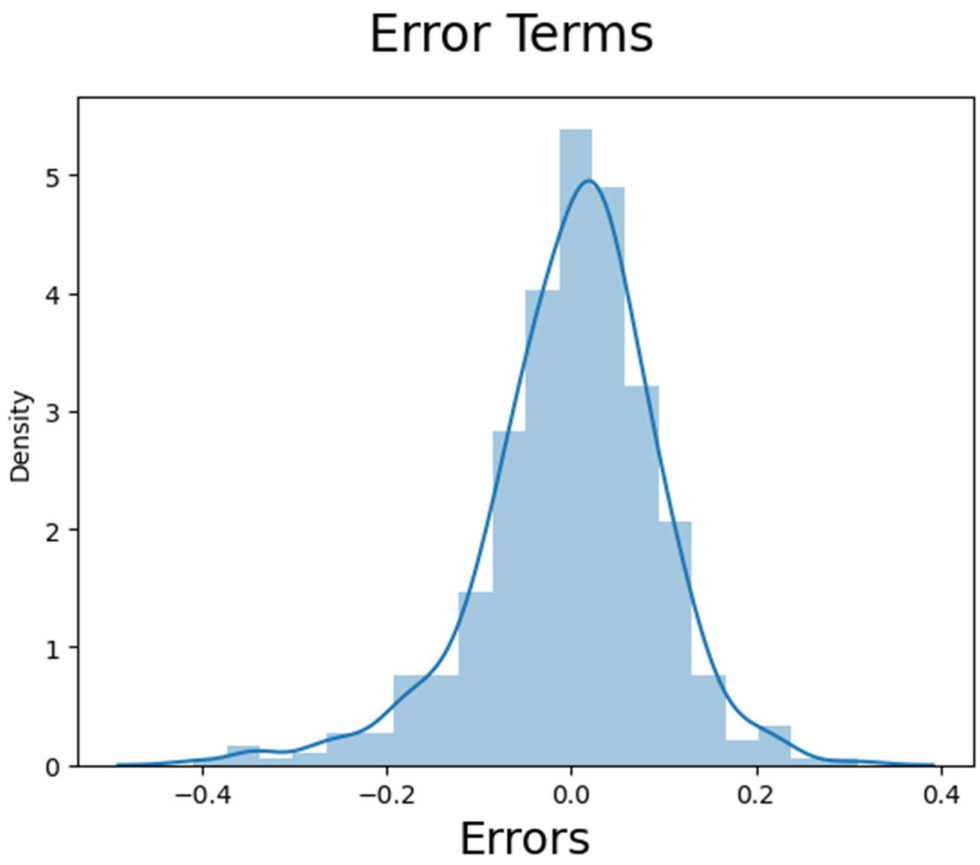
As per the above co-relation chart for numeric variables :-

atemp:- has the gighest correlation with the target variable variable with correlation value = 0.63

followed by temp as second highest with correlation value = 0.627

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

Ans :- Validation by Plotting the residuals (y_train – y_pred) as histogram (distplot) and by analyzing whether it is a normal distribution and it has mean of zero (0) or not .



From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

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:- Final Model is

The equation of best fitted surface based on model lr6:

cnt = 0.084143 + (yr × 0.230846) + (workingday × 0.043203) + (temp × 0.563615) − (windspeed × 0.155191) + (season2 × 0.082706) + (season4 ×0.128744) + (mnth9 × 0.094743) + (weekday6 ×0.056909) − (weathersit2 × 0.074807) − (weathersit3 × 0.306992)

As per our final Model, the top 3 predictor variables that influences the bike booking are:

Temperature (temp) - A coefficient value of '0.5636' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3069' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3069 units.

Year (yr) - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

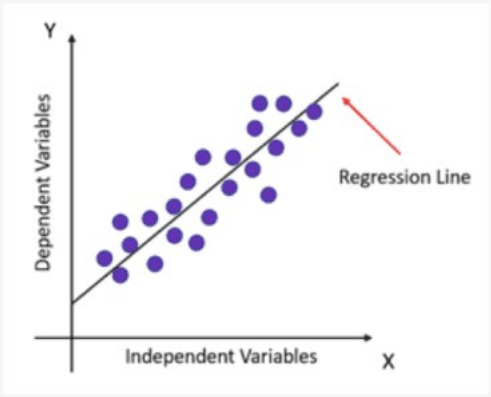
So, it's suggested to consider these variables utmost importance while planning, to achive maximum Booking The next best features that can also be considered are

General Subjective

Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans :- It is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). You learnt about these two types of linear regression under this module:
   ● Simple linear regression
   ● Multiple linear regression
   1. Simple Linear Regression The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straightline.
   The straight line is plotted on the scatter plot of these two points.

   The standard equation of the regression line is given by the following expression: $Y = \beta_o + \beta_1 X$

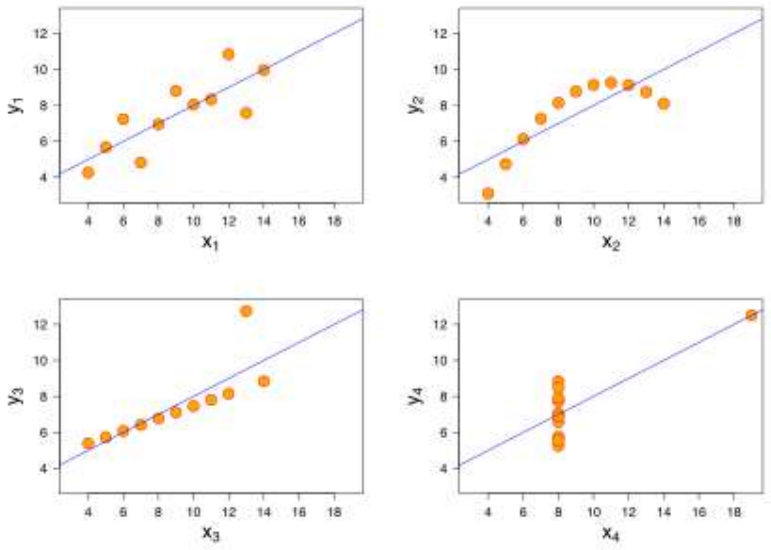


   The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

   When a number of independent variables more than one, the governing linear equation applicable to regression takes a different form like:y= c+m1x1+m2x2… mnxn where represents the coefficient responsible for impact of different independent variables x1, x2 etc. This machine learning algorithm, when applied, finds the values of coefficients m1, m2, etc., and gives the best fitting line.

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]



For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s2 x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s2 y	4.125	±0.003

Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : {\displaystyle R^{2}}	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, 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.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.[2][3][4][5][6]

The datasets are as follows. The x values are the same for the first three datasets.[1]

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76

13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is not known how Anscombe created his datasets.[7] Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.[7][8] One of these, the Datasaurus Dozen, consists of points tracing out the outline of a dinosaur, plus twelve other data sets that have the same summary statistics.[9][10][11] Datasaurus Dozen was created by Justin Matejka and George Fitzmaurice. The process is described in their paper "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing".

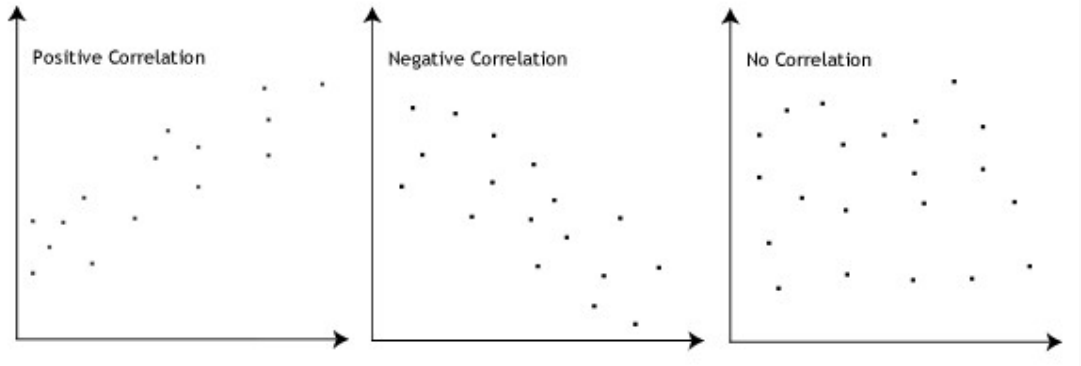
The Datasaurus Dozen, just like Anscombe's Quartet, shows why visualizing data is important, as the summary statistics can be the same, while the data distributions can be very different.

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



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

- 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 normalized scaling and standardized scaling? (3 marks)
   Ans

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.

Reason for scaling is performed :-

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.

Difference between normalized scaling and standardized scaling :-

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.

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? (3 marks)
   Ans :

The VIF of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining explanatory variables. A rough rule of thumb is that the VIFs greater than 10 give some cause for concern.

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).

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

Quantile-Quantile (Q-Q) 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 distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

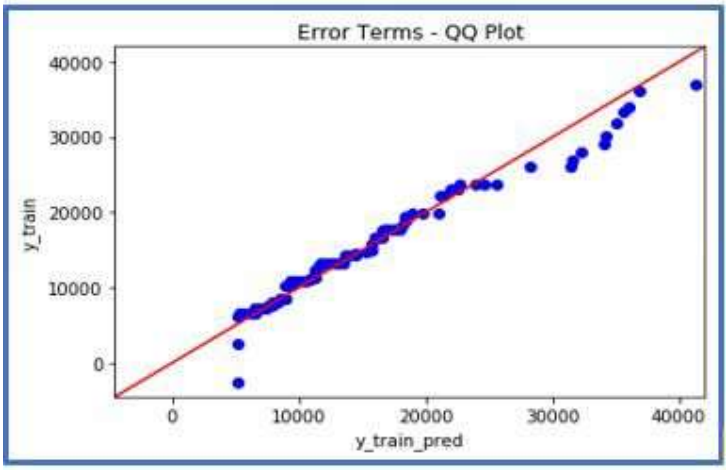
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

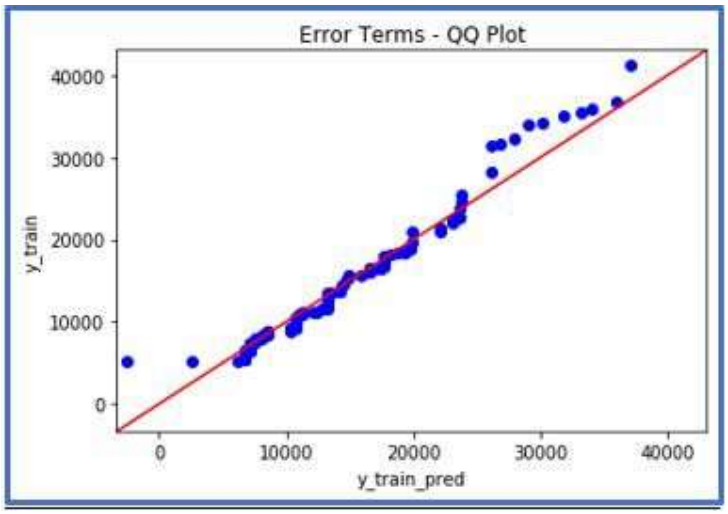
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.