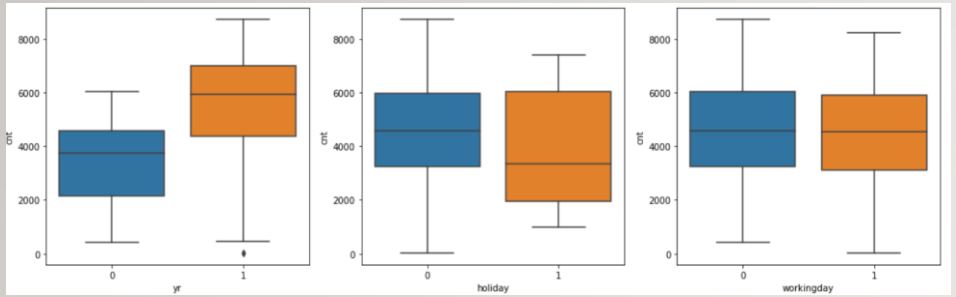
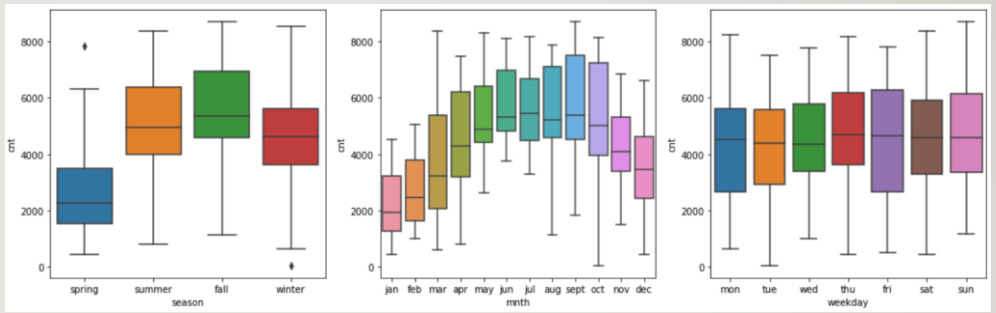
LINEAR REGRESSION ASSIGNMENT

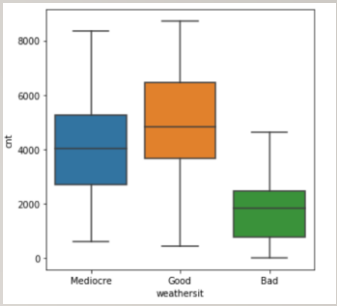
UDAY KUMAR L

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)







Insights/observations based on categorical variables

- The second year 2019 has better rentals compared to first year 2018 probably because of good feedback from customers and branding/establishment.
- Rentals are less on holidays.
- Working day doesn't share clarity on bike rentals.
- Bike rentals are high in 3rd season (fall) relative to other seasons
- Bike rentals go higher from June to September at its peak after that it depreciates
- The Bike rentals are mostly with high IQR on every friday of the week.
- Bike rentals are high for clear(good) weather.
- Bike sharing/rentals are less at the end of the year after september due to ice pallets/snow/fog winter conditions.

Assignment based subjective -Questions

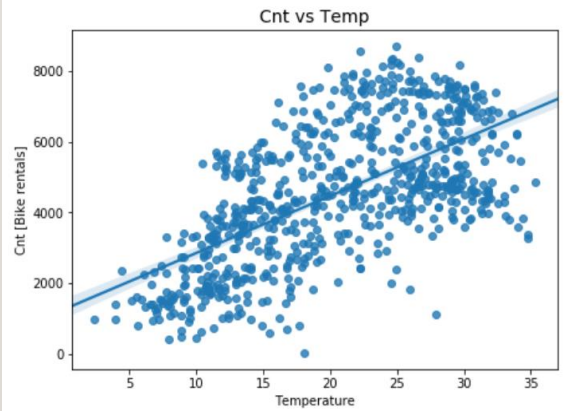
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

The drop_first is true it removes the first column which is created for the first unique value of a column. example: If there are 8 columns using drop first = True will reduce the columns to 7

- when all the other columns are zero that means the first column is 1.

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

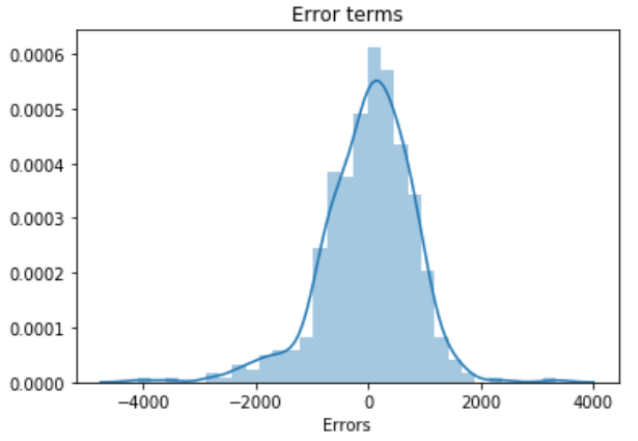
Temperature

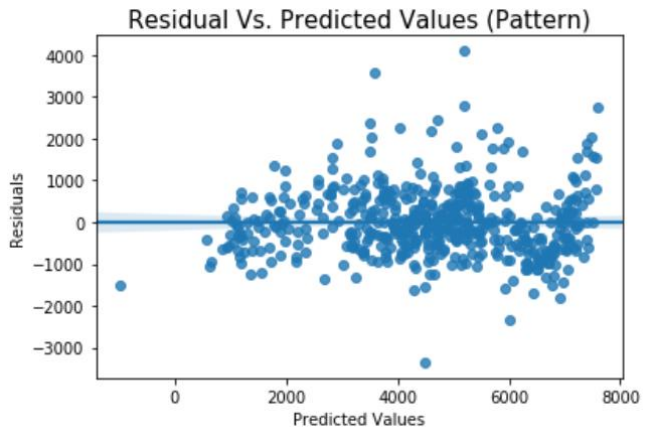


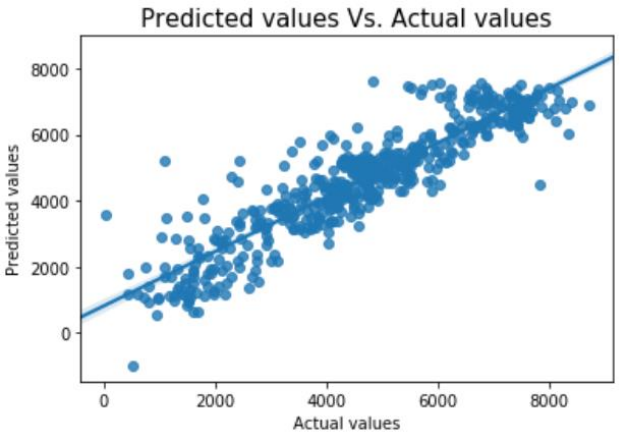
Assignment based subjective -Questions

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

- Error terms should be normally distributed.
- Residuals (Errors) should be independent of each other.
- Error terms have constant variance (Homoscedasticity)
- Obviously linear relationship between X variables and y variable







Assignment based subjective -Questions

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

Good weather(Moderate temperature/low wind speed), Seasons (Summer and fall),months(Aug – October)

- The company should be ready cater more rentals if the temperature is moderate around 25-30 C and windspeed is low

- The company should focus more rentals/offers during summer and fall as this is evident for high rentals during this months

- The company should offer more rental bikes in the months of september to october

- The company should use good weather conditions for bike rentals with subsidised offers and bad weather to regear the bikes

- Weekends (friday/sat/sun) are observed with more rentals , some offers are recommended here to customers.

- During peak months (Aug - Oct), this is the season from summer to fall(autumn), Rental companies can conduct marathons to certain tourist spots.

List of significant variables

- Temperature
- Weekends
- Holidays
- windspeed
- season
- Months
- Year 2019
- weathersit

General subjective -Questions

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

Linear regression is a predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).

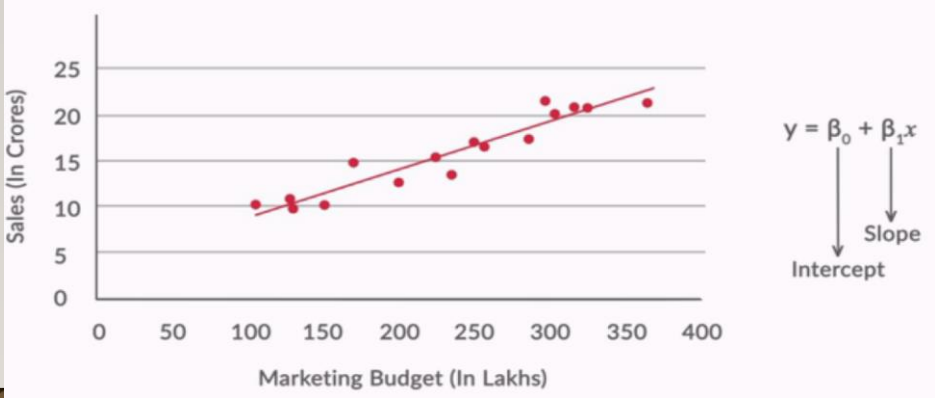
Kinds of regression

- Simple linear regression
- Multiple linear regression

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 straight line. The straight line is plotted on the scatter plot of these two points.

standard equation of the regression line is given by the following expression $Y = \beta_0 + \beta_1.X$



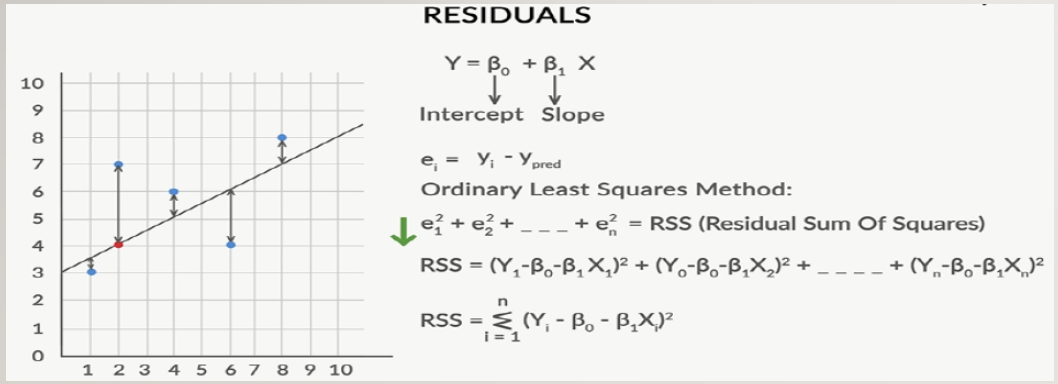
Independent variable – Marketing budget

Dependent variable - Sales

General subjective -Questions

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

Evaluation of regression : best-fit line is found by minimizing 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:



strength of the linear regression model can be assessed using 2 metrics:

1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

R2 Formula

- $R2 = 1 - \frac{RSS}{TSS}$

Where

RSS= Residual sum of square

TSS= Sum of errors of the data from mean

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

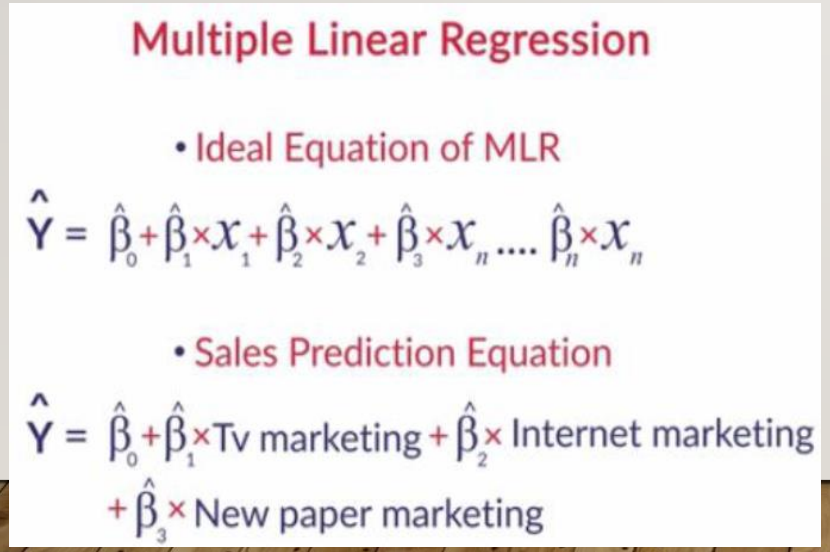
$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

General subjective -Questions

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

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).
The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, Radio marketing, and Newspaper marketing.
Thus, the equation of multiple linear regression would be as follows:



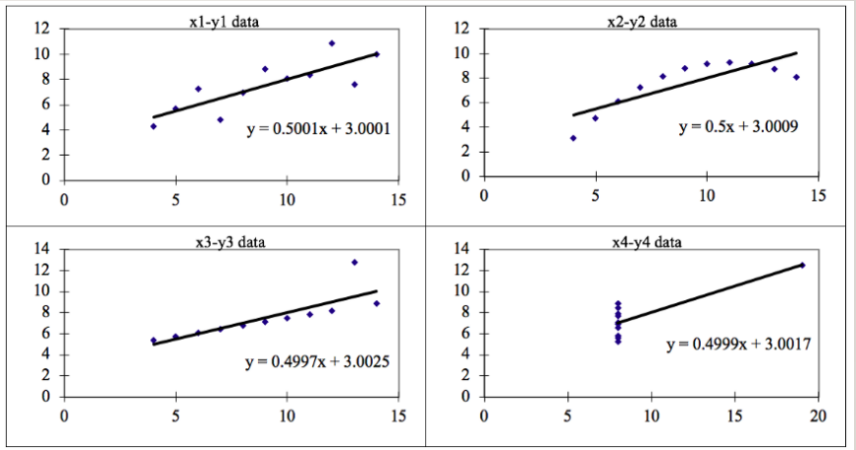
General subjective -Questions

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

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 Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Inference:
All the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.



Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this did not fit linear regression model on the data well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

General subjective -Questions

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

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation.
It shows the linear relationship between two sets of data.

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.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

General subjective -Questions

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

- Scaling is a technique to standardize the independent features/variables present in the data in a fixed range.
- Scaling is performed to make sure all the variables value fall within certain boundary range for example: all values within 0 and 1
- Having features on a same scale can help the gradient descent converge more quickly towards the minima.
- Standardization (or Z-score normalization) scaling is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1
- Normalization (Min-Max scaling), 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

General subjective -Questions

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

A better way to assess multicollinearity is to compute the **variance inflation factor** (VIF).
Since one of the major goals of linear regression is identifying the important explanatory variables, it is important to assess the impact of each and then keep those which have a significant impact on the outcome. This is the major issue with multicollinearity. Multicollinearity makes it difficult to assess the effect of individual predictors.

A variable with a high/infinite VIF means it can be largely explained by other independent variables. Thus, removing the variable with a high VIF would make it easier to assess the impact of other variables, while making little difference to the predicted outcome.

The higher the VIF, the higher the multicollinearity. variables with a high VIF or multicollinearity may be statistically significant $p<0.05$, in which case you will first have to check for other insignificant variables before removing the variables with a higher VIF and lower p-values. The value of VIF thershold will depend on the case requirements.

General subjective -Questions

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

Quantile-Quantile (Q-Q) plot is a graph that helps to 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.
we can confirm using Q-Q plot that both the training and test data sets are from populations with same distributions. Typically train and test dataset should show linear behavior when plotted on Q-Q graph.
Similarly applies for predicted and actual dataset
We check for 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