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?

Ans:

From the categorical variables, we can infer the following

Season : The demand is usually higher in **fall** and least in **spring**.

Yr : The demand grew significantly in **2019** than that of 2018.

Month : The demand is highest in the month of **sep** and least in the month of **jan**.

Holiday : Peak demand reduces during holidays

Weathersit : The demand is higher during **clear** situation and the least during **light_rain**.

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

Ans:

N levels of values in a categorical variable can be explained by (N-1) dummy variables.

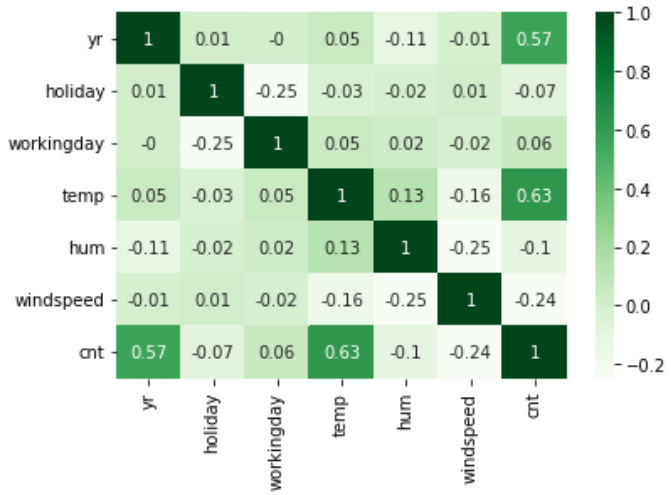
For example – In our bike sharing data set, in season variable, if it's not winter or spring or summer, then it is fall. Hence dropping a column reduces the redundancy in the model.

However, it is not mandatory to always drop the first column. As per business need, any one of the N dummy columns can be dropped, based on the interpretability of the model.

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

Ans:

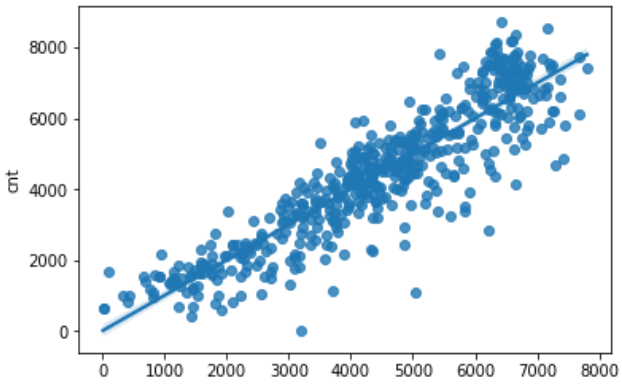
The variable **temp** has the highest correlation (0.63) among the numerical variables.



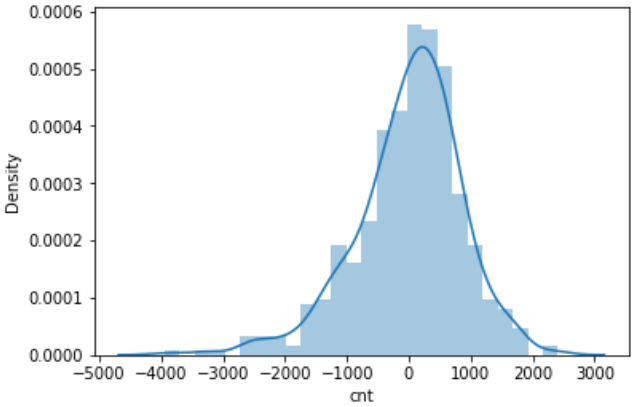
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: The Assumptions of Linear Regression

Linearity : Linear Regression assumes a linear relationship between X and y. In MLR, this can be validated by plotting between y_actual and y_pred.

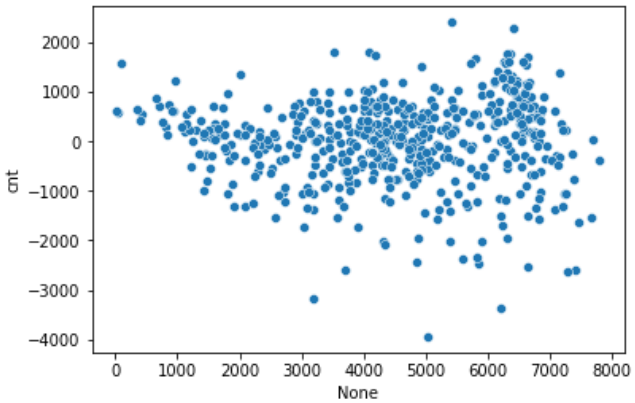


Normality : The assumption that the error terms are normally distributed can be validated by plotting the error terms.



Independence : The score of 2.02 in Durbin_watson test shows that the residuals are independent there is no presence of auto collinearity in them.

Homoscedasticity : The residuals are assumed to have a constant variance without any patterns.



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 that are contributing to the demand of shared bikes are **temp, yr, weathersit_light_rain.** This can be inferred from the Linear Regression Equation.

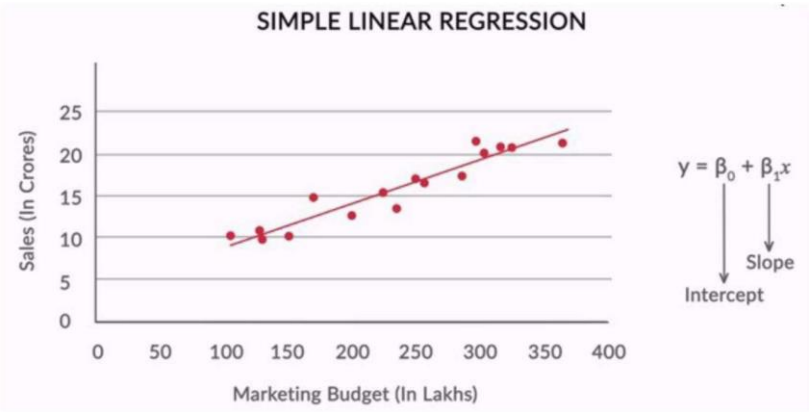
Y = 782.79 + 2029.01(yr) - 753.31(holiday) + 4938.9(temp) - 1264.48(windspeed) + 705.35(season_summer) + 1095.77(season_winter) + 777.92(mnth_sep) - 2203.34(weathersit_light_rain)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is an approach for predicting the relationship between a dependent variable and one or more independent variables or predictors. When there is only one predictor, it is called simple linear regression. In case there are multiple predictors, it is called multiple linear regression. In simple terms – linear regression draws a line or curve through all the predictor variable data points in such a way so that the vertical distance between the data points and the line is minimum.



Simple linear regression formula:

$$y = a_0 + a_1 x$$

Multiple linear regression formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

OLS or Ordinary least squares is one of the methods used in linear regression where parameters of a linear function are chosen by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

Cost function: Cost function helps to predict best possible values of the coefficients of the predictor, hence finding the best fitted line. In Linear regression, Mean squared Error (MSE) is used to calculate cost function. MSE calculates the average of squared error between predicted values and actual values.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Objective of linear regression model is to minimize this cost function. This is done using **Gradient descent**, where coefficients are randomly selected and iteratively updated to reach the minimum cost function.

Assumptions of Linear regression:

- **Linear relationship**: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
- **Normality**: The residuals of the model are normally distributed.
- **Independence**: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
- **Homoscedasticity**: The residuals have constant variance at every level of x.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consists of four different data sets that all have very similar summary statistics, but looks completely different from each other graphically. 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.

The data looks like:

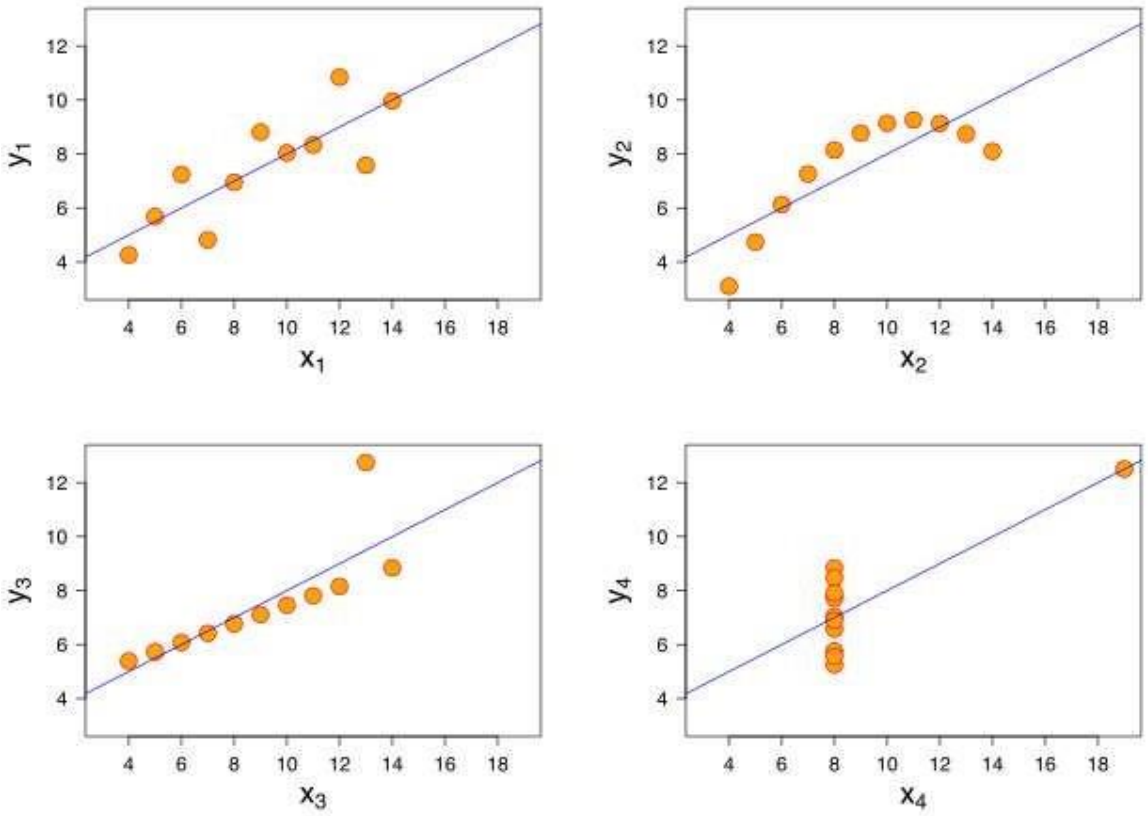
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	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

Summary statistics for all four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regresson line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

Graphical representation of the four data sets:



The first scatter plot (top left) appears to be a simple linear relationship

- The second graph (top right) is not linear.
- In the third graph (bottom left), the relationship is linear, but should have a different regression line. It is biased by the one outlier.
- 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 are not related.

3. What is Pearson's R?

Ans: Person correlation coefficient or Pearson's R or more commonly known as just correlation coefficient is a measure of linear correlation between two sets of data. The measure can only reflect a linear correlation between two variables, and ignores many different types of relations. It has a value between -1 and +1. Correlations equal to -1 or +1 signifies all data points lying exactly on the line.

Formula:

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

r = correlation coefficient

x_i = values of the x-variable in a sample

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

y_i = values of the y-variable in a sample

$\bar{y}$ = 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?

Ans:

Scaling – In machine learning, scaling refers to the transformation of predictor variables in a range, either [0,1] or [-3,3]

Why Scaling is important - Scaling is important as most of the times data sets consists of features with different magnitudes and units of values. It improves the readability of coefficients.

For example: consider a data set consisting of two columns, age and height. Age ranges from 0-100, whereas height ranges from 1-6 foot. In absence of scaling, machine learning model can become biased towards certain features because of their magnitude of values, not taking into account the unit in which those values are represented. Hence, we need to scale and bring the values to the same range before building models.

Difference between normalization and standardization - In normalization, scaling is done using min and max values of the feature. Feature values are mapped into the [0, 1] range.

Formula for normalization :

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

In standardization, we don't enforce the data into a definite range. standardization centralizes the data with mean 0 and a standard deviation of 1.
Standardization formula:

$$z = \frac{x - \mu}{\sigma}$$

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

Ans: Very large values of VIF signifies presence of correlation between variables. So, an infinite VIF would mean perfect correlation between variables. It indicates that the corresponding variable can be exactly expressed as a linear combination of the other variables.
In this case we get R2 =1, hence VIF = 1/(1-R2) is infinity.

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

Ans: A Q-Q (quantile-quantile) plot is a graphical tool to help assess if two sets of data have similar theoretical probability distribution.

It is used to check following scenarios:

If two data sets —

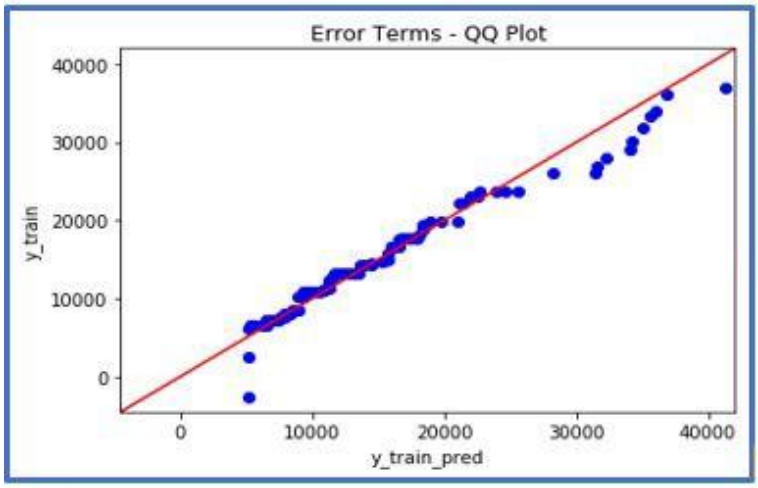
- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

Interpretation

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

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



d) Y-values > X-values: If x-quantiles are lower than the y-quantiles.