Linear Regression Assignment

Shared Bikes Statistics

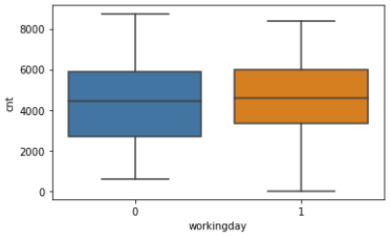
Contents

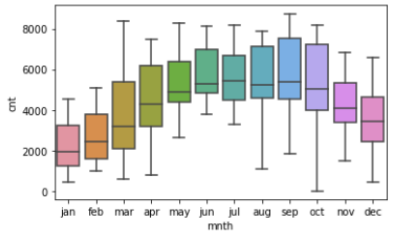
BOOMBIKES

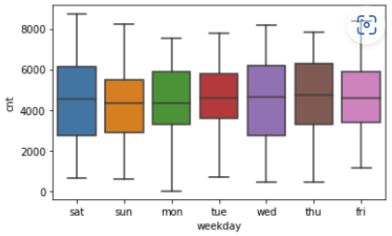
BOOMBIKES

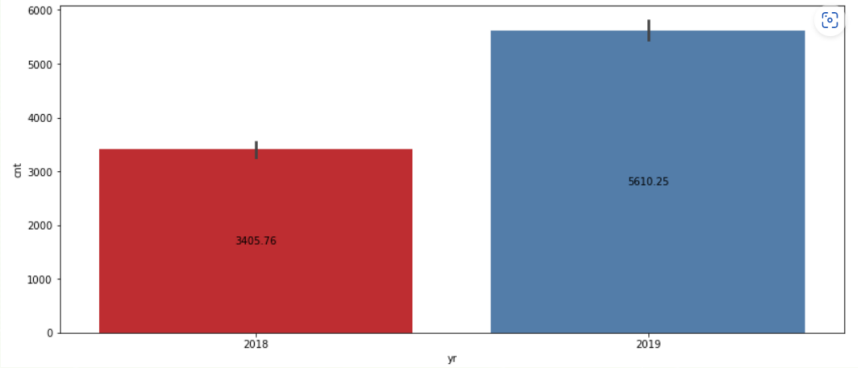
Q1a: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?(3 marks)

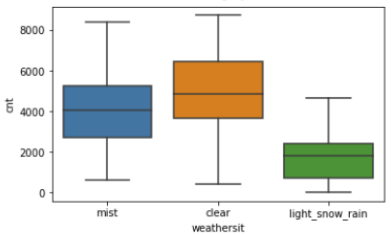
Categorical Variable	Inferences on the dependent variable (cnt)
yr	Very positive demand growth (64%) in 2019 compare to 2018
mnth	Business is with increasing trend from Jan to Sep and later decreasing trend till Dec. Lowest demands in Jan for both years whereas demands peak observed in Jun for 2018 and in Sep for 2019
season	Major demand yearly surge (113%) in Spring compare to other seasons; though maximum demands in fall season
holiday	More demands during working days compare to holidays
workingday	almost consistent demands for workingdays and weekends
weekday	Lowest demands on Tues, overall low demands for start of the week and later demands increase till Sat and again drop from Sun
weathersit	Low demands when light snow or rains, moderate demands when misty and high demands when clear sky. No demands for heavy snow or rains

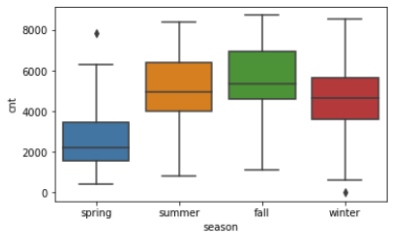


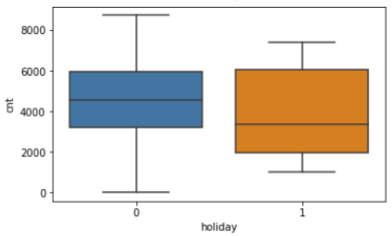


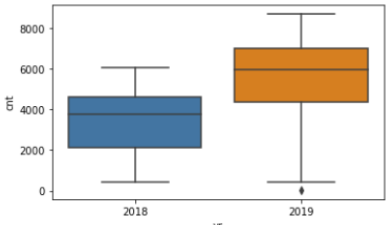












Q1b: Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer:

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.

If ‘n’ number of attributes for a categorical variable then (n-1) dummy columns are optimistic to represent the data. If we don’t drop the first column then by default, n number of columns will be created and complexity will be increased.

Example: weathersit variable ‘with 3 attributes as – (1) clear (2) mist (3) light_snow_rain
Here for 3 attributes, only 2 columns are sufficient – 1. mist 2. light_snow_rain; clear attribute will be represented by having 0 value for the both the columns.

attributes	mist	light_snow_rain
clear	0	0
mist	1	0
light_snow_rain	0	1

In the assignment, below categorical variables where dummy variables are created

1. Weather: 3 attributes are represented with 2 dummy columns; ‘clear’ status is dropped
2. Season : 4 seasons are represented with 3 dummy columns; ‘fall’ season is dropped
3. Weekday: 7 days are represented with 6 dummy columns; ‘fri’ is dropped
4. Month: 12 months are represented with 11 dummy columns; ‘apr’ is dropped

BOOMBIKES

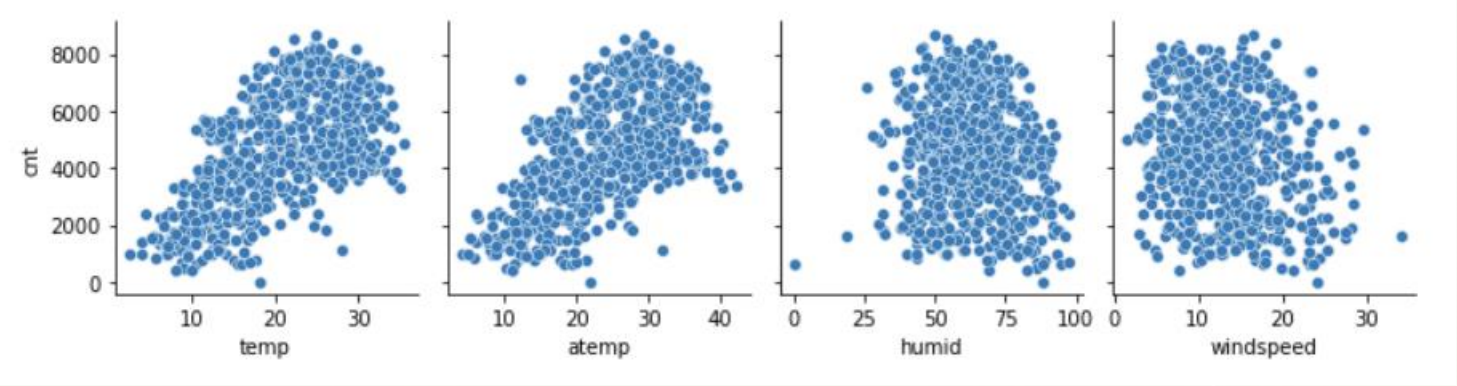
Q1c: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Among the 8 numerical columns, below 4 columns are as not giving any additional details

1. Instant
2. Dteday
3. Casual
4. Registered

Looking at the pair-plot with 4 remaining columns (temp, atemp, humid, windspeed), highest correlation observed in **'temp'** and **'atemp'** variables with the target variable 'cnt'.



Q1d: How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Figure #	Assumptions	Outcome	Expectations Met?
1d.1	Linear Relationship	linear relationship between actual and predicted target values on train data	✓
1d.2	No Multicollinearity	all VIF values < 5 which represents no multicollinearity exists	✓
1d.3	Homoscedasticity	the residuals have constant variance at every level of x	✓
1d.4	Normal distribution of error terms	error terms are normally distributed with mean = 0	✓

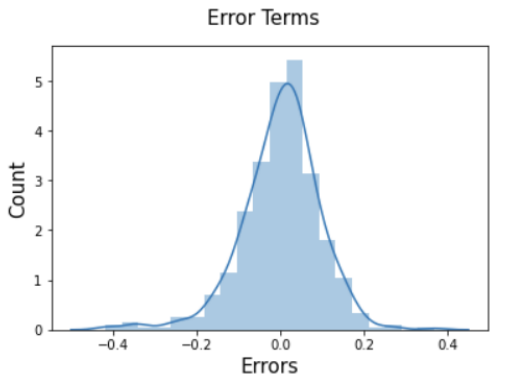


Fig 1d.4

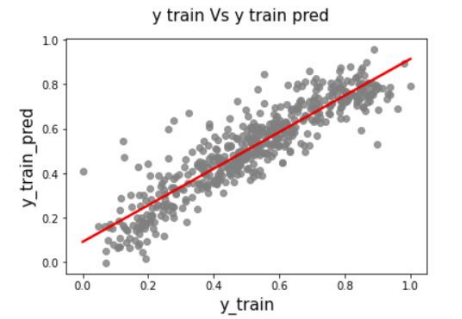


Fig 1d.1

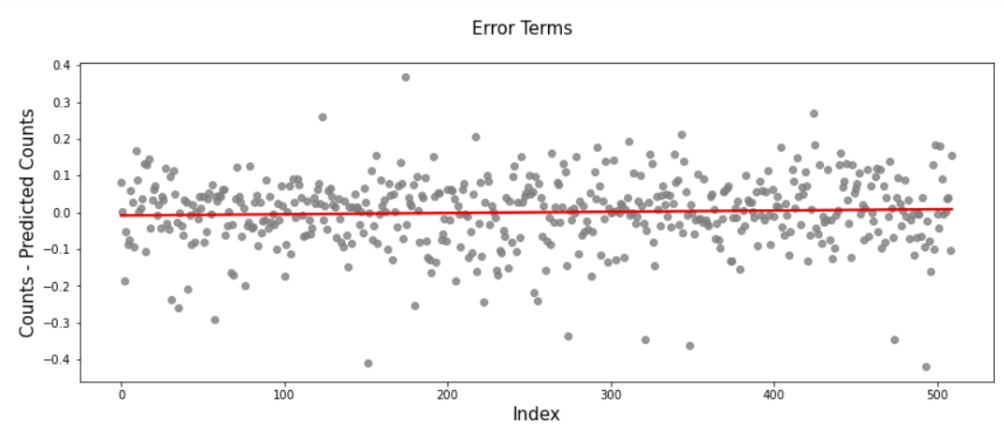


Fig 1d.3

	features	VIF
1	temp	4.04
0	yr	1.94
7	summer	1.79
2	aug	1.56
8	winter	1.47
6	mist	1.45
3	sep	1.29
4	sun	1.16
5	light_snow_rain	1.06

Fig 1d.2

Q1d: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

equation of the best fitted line is as below

cnt = 0.071052 + (0.231555 * yr) + (0.544107 * temp) + (0.057489 * aug) + (0.118258 * sep) - (0.044770 * sun) - (0.296800 * light_snow_rain) - (0.079591 * mist) + (0.097684 * summer) + (0.145187 * winter)

Top 3 features contributing significantly towards the demand of the shared bikes are as below along with the weightage/units –

1. **temp** (+ 0.544107) – positive impact
2. **yr** (+ 0.231555) – positive impact
3. **low_snow_rain (**- 0.296800**)** – negative impact

Q2a: Explain the Linear Regression Algorithm in detail. (4 marks)

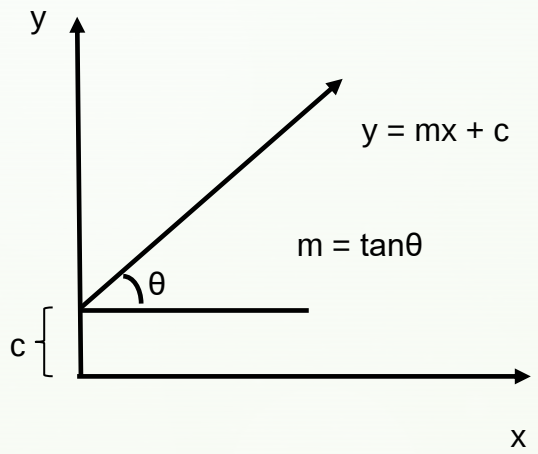
Regression is one of the machine learning models

Linear Regression is supervised learning algorithm. It is a form of predictive modelling technique which tells us the relationship between dependent (target) variable and independent variables (predictors)

Type of Linear Regression: (1) Simple Linear Regression (2) Multiple Linear Regression

(1) **Simple Linear Regression:** This model explains the relationship between a dependant variable and one independent variable using a straight line
(2) **Multiple Linear Regression:** This model explains the relationship between a dependant variable and **two or more** independent variables

Simple Linear Regression

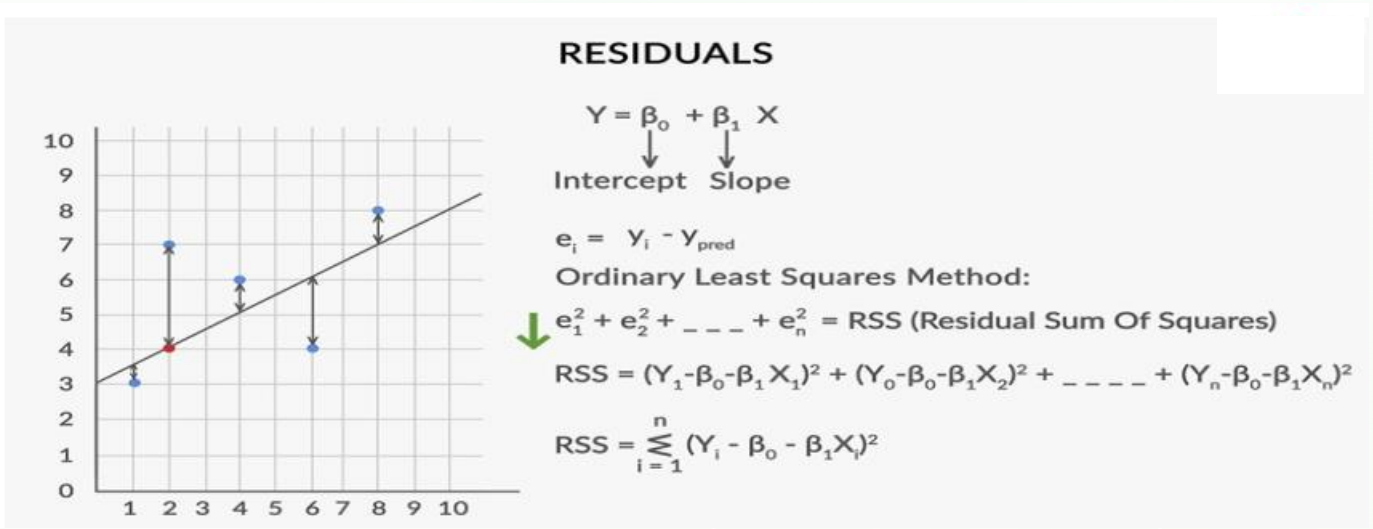


In case of Simple Linear Regression, you fit a straight line between the available data points and later with the straight line model, new data points can be predicted.

Straight Line Equation is y = mx + c where m is slope and c is intercept
It is also represented as **Y = β0 + β1X** where β0 is intercept and β1 is slope
m signifies how strong relationship between y & x whereas c is the value of y when x = 0
Positive Linear Relationship: Dependent variable increases as independent variable increases
Negative Linear Relationship: Dependent variable decreases as independent variable decreases

Best-Fit Line:

The line which best fits the given scatter plot in the best way. Best fit line equation can be identified by minimizing RSS (Residual Sum of Squares) which is equal to the sum of squares of the residuals for each point in the plot. Residuals for any data point is found by subtracting predicate value of dependent variable from actual value of dependent variable:



Strength of Linear Regression:

Strength of Linear Regression model can be assessed using 2 metrics:

1. R2 or coefficient of determination
2. Residual Standard Error (RSE)

R2 or coefficient of determination

R2 = 1 – (RSS/TSS)

RSS: Residual Sum of Squares, **TSS**: Total Sum of Squares (sum of errors of the data from mean)

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

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Multiple Linear Regression

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

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times x_1 + \hat{\beta}_2 \times x_2 + \hat{\beta}_3 \times x_n \ldots . \hat{\beta}_n \times x_n$$

Standard Error:

It measures the variability in the estimates of the coefficients

p-value:

In regression model building, the **null hypothesis** corresponding to each p-value is that the corresponding independent variable **does not impact dependent variable**. The **alternate hypothesis** is that the corresponding independent variable **impacts on dependent variable.** Therefore, a low p-value (<0.05) indicates that you can reject the null hypothesis. In other words, the independent variable is meaningful addition for the model.

R-squared vs Adjusted R-squared

Adjusted R-squared is better metric than R-squared to assess how good the model fits the data. R-squared always increases when variables are added even when they are not related to dependent variable. Whereas Adjusted R-squared penalizes R-squared on unnecessary addition of variables.

Multicollinearity:

Some variables could have some relation amongst themselves. A large value in the correlation matrix would indicate a pair of highly correlated variables. Unfortunately, not all collinearity can be detected using correlation matrix. It is possible for collinearity to exist between 3 or 4 variables but no pair of variables with high correlation. This situation is called multicollinearity.

Variance Inflation Factor(VIF):

It measures degree of collinearity or multicollinearity in the regression model. Higher the VIF, the higher the multicollinearity.

Model Building Algorithm:

1. Build a model containing all variables
2. Check VIF and summary
3. Remove variables with high VIF (>2 generally) and which are significant (p>0.05), one by one
4. If the model has variables which have a high VIF and are significant, check and remove other insignificant variables
5. After removing the insignificant variables, the VIFs should decline
6. If some variables still have a high VIF, remove the variable which is relatively less significant
7. Now, variables must be significant. If the number of variables is still high, remove in order of insignificance until you arrive at a limited number of variables that explain the model well.

Train the Model:

You get a model that is trained on the training data set. This model should be able to accurately predict dependent variable value in the test data.

Model Validation:

It is desired that the R-squared between the predicted value and the actual value in the test set should be high. There are many other metrics for validation like RFE (Recursive Feature Elimination)

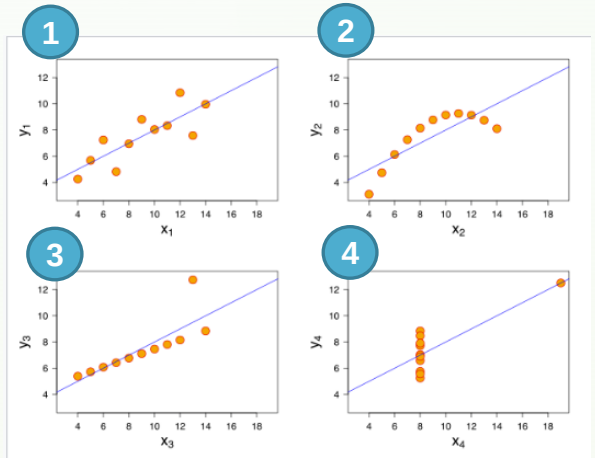
Key Points:

- In statistical modelling, Linear Regression is a process of estimating the relationship among variables. The focus is to establish the relationship between a dependent variable and one or more independent variables (predictors)
- Simple Linear Regression does not allow to change all predictors at a time and measure the impact. You can only change one at a time.
- Regression shows relationship that is co-relation and not causality. Correlation doesn't imply causation.
- Regression is widely for 2 purposes: (1) Forecasting (2) Prediction
- Regression guarantees 'interpolation' but not necessarily 'extrapolation'
- Linear Regression is a form of parametric regression
-

Q2b: Explain the Anscombe's quartet in detail. (3 marks)

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



Anscombe's quartet comprises of 4 datasets 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 are constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing 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"

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

Explanation on scatter plots [refer numbers tagged to each graph]

1. Linear relationship between x & y
2. Non linear relationship between x & y
3. Perfect linear relationship except 1 outlier
4. One high-leverage point is enough to produce a high correlation coefficient

Summary: Anscombe's quartet shows why visualization our data is important as summary statistics can be same while data distribution can be very different

Q2c: What is Pearson's R? (3 marks)

The **Pearson correlation coefficient (r)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

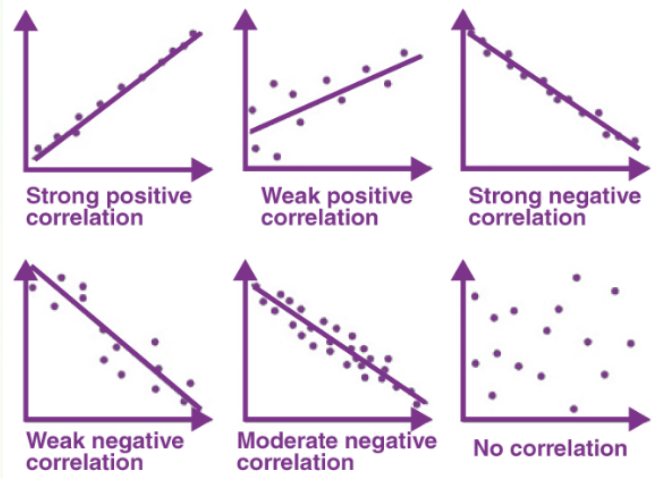
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

It is also known as Bivariate Correlation and the correlation coefficient

r Value	Correlation Type	Interpretation	Example
0 to 1	Positive Correlation	One variable changes then other variable also changes in the same direction	Baby length and weight
0	No Correlation	No relationship between variables	Car price & width of windshield wipers
0 to -1	Negative Correlation	One variable changes then other variable also changes in the opposite direction	Elevation & air pressure

General rules of Thumb:

r Value	Strength	Direction
> 0.5	Strong	Positive
0.3 to 0.5	Moderate	Positive
0 to 0.3	Weak	Positive
0 to -0.3	Strong	Negative
-0.3 to 0.5	Moderate	Negative
< 0.5	Weak	Negative



Summary:

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

BOOMBIKES

Q2d: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?(3 marks)

What? 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.

Why? 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.

Note: Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference Between Normalized & Standardized Scaling:

	Normalized Scaling	Standardized Scaling
1	Minimum and Maximum value of features are used for scaling	Mean and standard deviation is used for scaling
2	X_new = (X - X_min)/(X_max - X_min)	X_new = (X - mean)/Std
3	It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation
4	Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
5	It is really affected by outliers	It is much less affected by outliers
6	Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization
7	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
8	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian
9	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization

Q2e: You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

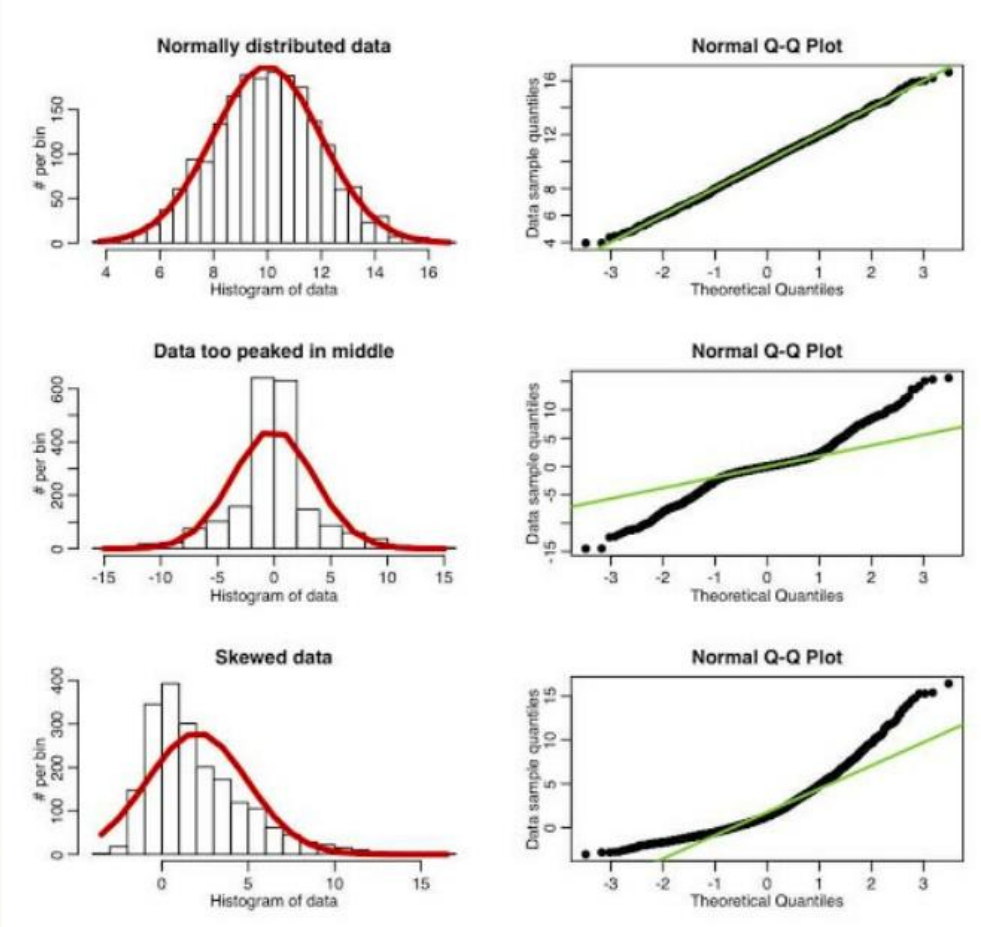
When? If there is perfect correlation between 2 independent variables then VIF is infinite.

Why? In case of perfect correlation, R2 =1 and VIF = 1/(1-R^2) which is infinite. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Note: To solve this infinite VIF problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q2f: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

In statistics, a Q–Q plot (**quantile-quantile plot**) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

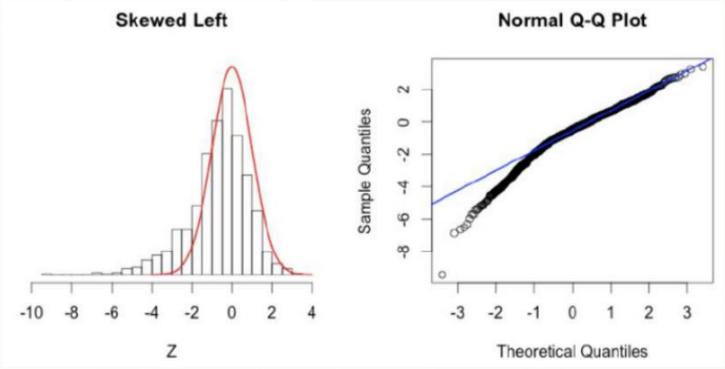


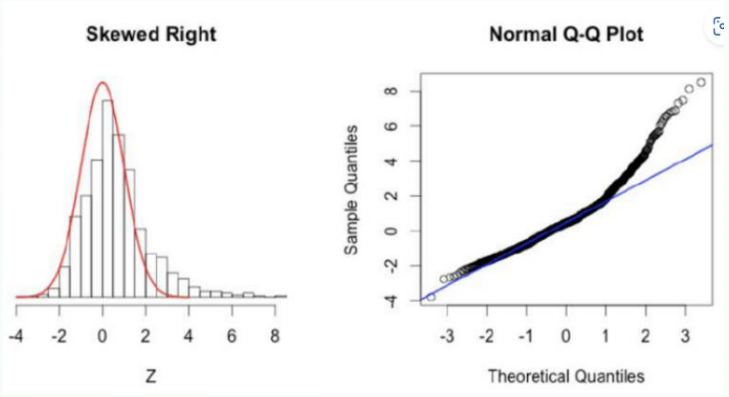
Q-Q plots for Normal Distribution:

We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.

Now we have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are not Normally distributed.

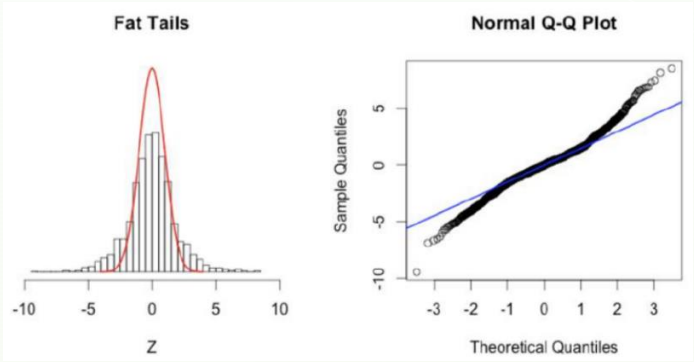
If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.

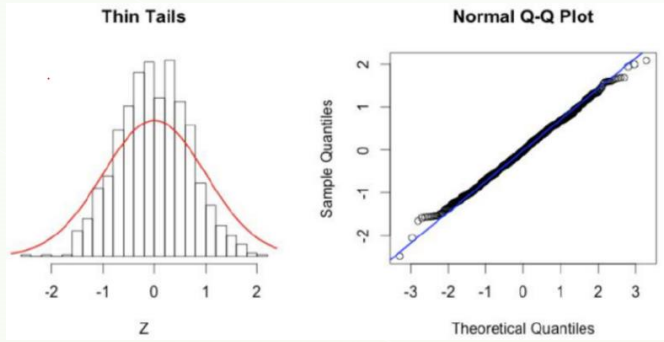




Skewed Q-Q Plots:

Q-Q plots are also used to find the Skewness (a measure of “asymmetry”) of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution we want to know on the y-axis then we see a very peculiar shape of a Normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is left-skewed (or negatively skewed) but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer till to its right and it is right-skewed (or positively skewed).





Trailed Q-Q Plots:

The distribution with a fat tail will have both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line, whereas a thin-tailed distribution will form a Q-Q plot with a very less or negligible deviation at the ends thus making it a perfect fit for the Normal Distribution.

The Goal was to build linear regression model to provide inputs for business strategy