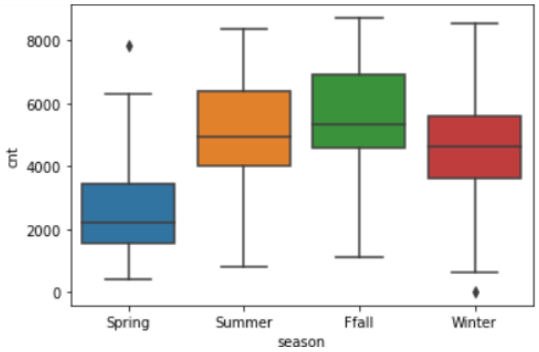
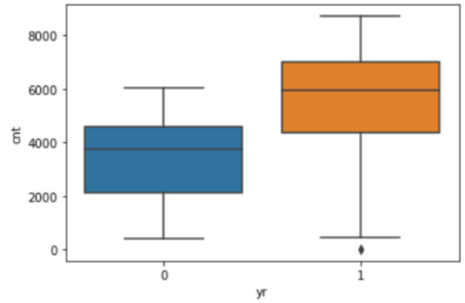
SUBJECTIVE ASSIGNMENT

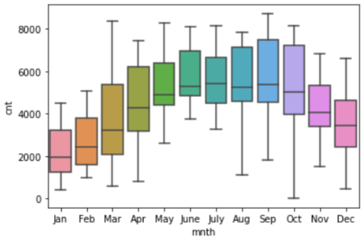
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

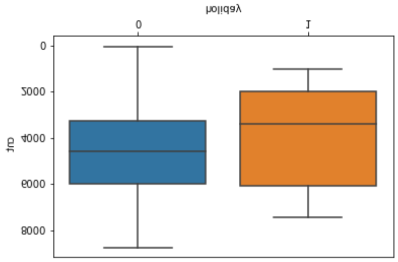
Below is the observation of categorical variable

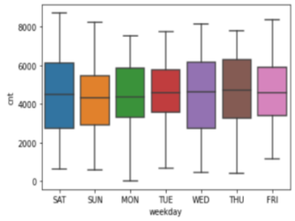
- Count is high in summer and fall season.
- Count is more in 2019 as compared to 2018.
- Count is more in month June to October.
- More people prefer to rent bike on working days and during holiday people prefer to stay home.
- Bike rent out most when the sky is clear or partly clouds
- There is a holiday, demand has decreased.
- During September, bike sharing is more. During the year end and beginning, it may be due to extreme weather condition









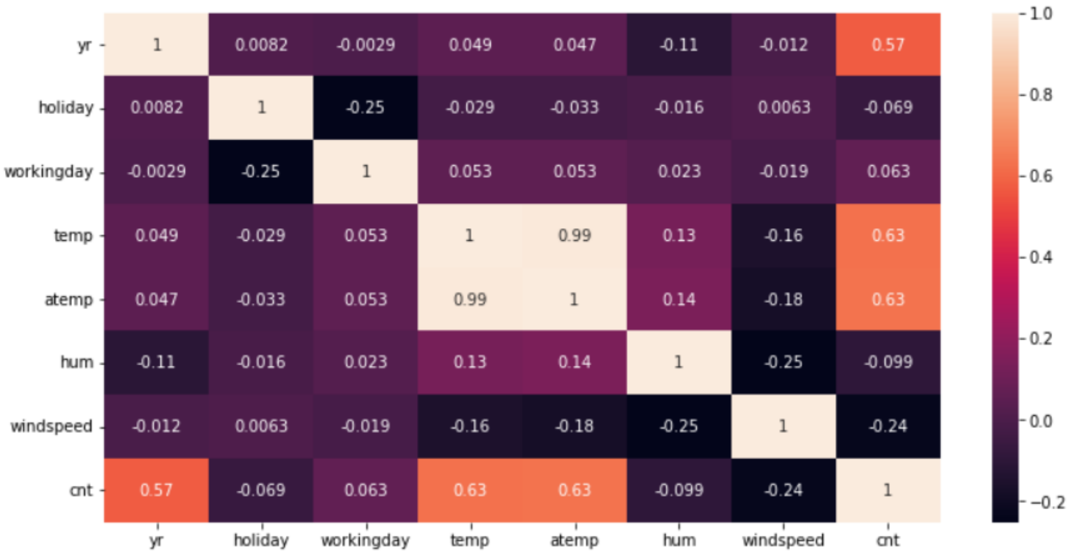


2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

drop_first=True parameter is commonly used in machine learning libraries, particularly with one-hot encoding. When encoding categorical variables, setting drop_first=True removes the first category level to avoid multicollinearity issues. This means that if you have a categorical feature with *n* levels, only *n-1* binary columns will be created during one-hot encoding.

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

temp and **atemp** is highly correlated



	yr	holiday	workingday	temp	atemp	hum	windspeed	cnt
yr	1	0.0082	-0.0029	0.049	0.047	-0.11	-0.012	0.57
holiday	0.0082	1	-0.25	-0.029	-0.033	-0.016	0.0063	-0.069
workingday	-0.0029	-0.25	1	0.053	0.053	0.023	-0.019	0.063
temp	0.049	-0.029	0.053	1	0.99	0.13	-0.16	0.63
atemp	0.047	-0.033	0.053	0.99	1	0.14	-0.18	0.63
hum	-0.11	-0.016	0.023	0.13	0.14	1	-0.25	-0.099
windspeed	-0.012	0.0063	-0.019	-0.16	-0.18	-0.25	1	-0.24
cnt	0.57	-0.069	0.063	0.63	0.63	-0.099	-0.24	1

3. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Below is the assumption for the validation

   a. Error terms are normally distributed with mean 0.
   b. No autocorrelation
   c. Error Terms do not follow any pattern.
   d. Multicollinearity check using VIF(s).
   e. Linearity
   f. Ensured the overfitting by looking the R2 value and Adjusted R2.

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

   The top 3 features contributing significantly towards explaining the demand of the shared bikes
   ➔ season ➔ weekday ➔ temperature

Subjective Questions

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

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.

Why ?

The interpretability of linear regression is a notable strength. The model’s equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.

Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

There are two main types of linear regression:

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$\mathrm{y=\beta0+\beta1X}y=\beta 0+\beta 1X$
where:

- Y is the dependent variable
- X is the independent variable
- $\beta0$ is the intercept
- $\beta1$ is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$\mathrm{y=\beta0+\beta1X+\beta2X+}\ldots\ldots\ldots\mathrm{\beta nX}y=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\ldots\beta nX$
where:

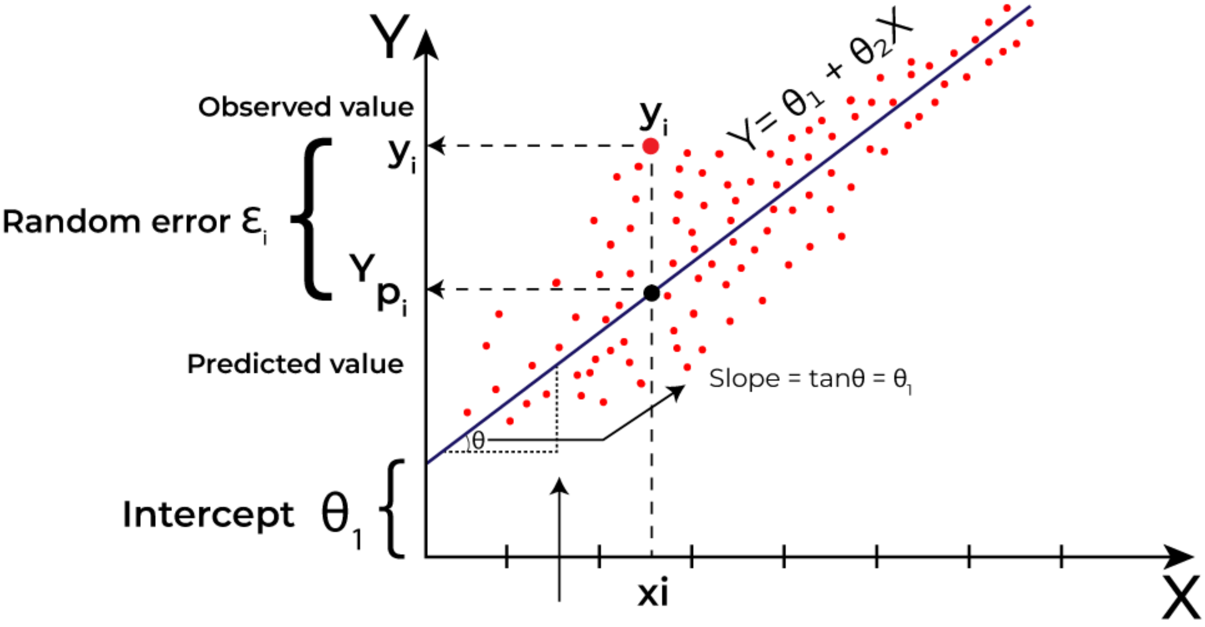
- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- $\beta0$ is the intercept
- $\beta1$, $\beta2$, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



Y
Observed value
y_i
y_i
$Y = \theta_1 + \theta_2 X$
Random error ε_i
Y_{p_i}
Predicted value
θ
Slope = tanθ = θ_1
Intercept θ_1
xi
X

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

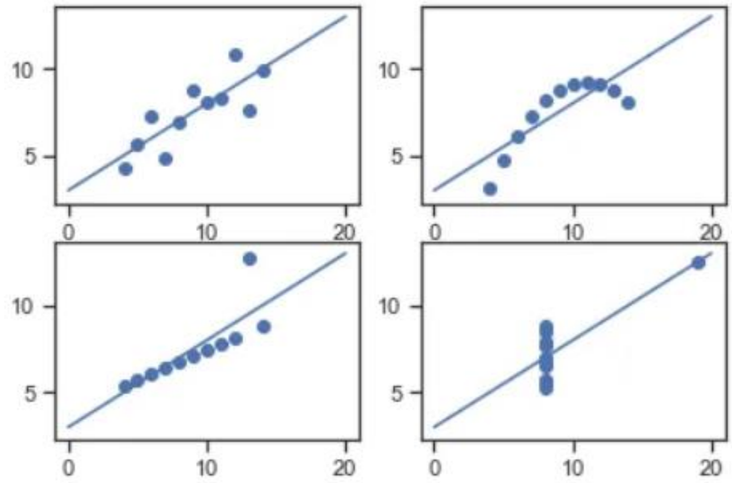
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.



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

The most popular correlation coefficient is Pearson's Correlation Coefficient. It is very commonly used in linear regression

Pearson Correlation Coefficient (r), often denoted as *r*, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

r=1, Perfect positive linear relationship

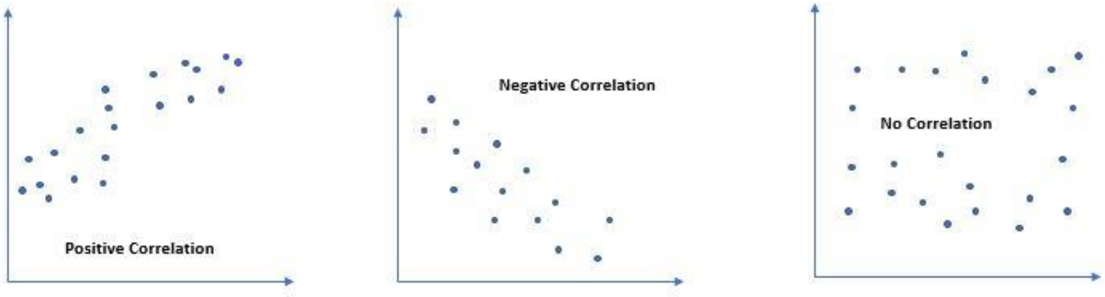
r=-1, Perfect negative linear relationship

r=0, No linear relationship

The formula is

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Value of ‘r’ ranges from ‘-1’ to ‘+1’. Value ‘0’ specifies that there is no relation between the two variables. A value greater than ‘0’ indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable. Value less than ‘0’ indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.



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

Scaling 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.

*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*

The reason of doing scaling is

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.

In Scaling, we’re changing the **range** of the distribution of the data… While in normalizing, we’re changing the **shape** of the distribution of the data.

S.NO.	Normalization	Standardization
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 Standard Scaler 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.

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

Before explain to the answer we need to understand the VIF concept

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. To determine VIF, we fit a regression model between the independent variables.

If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation, then VIF = infinity.** A large value of VIF indicates that there is a correlation between the variables

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

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

his 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

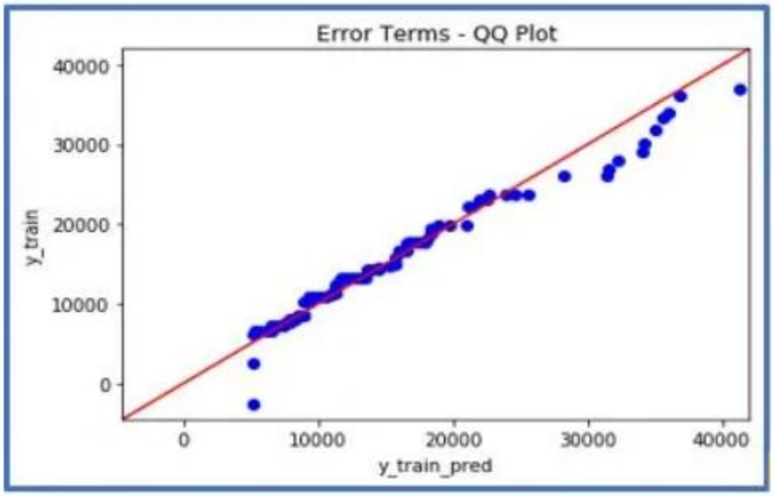
Interpretation:

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

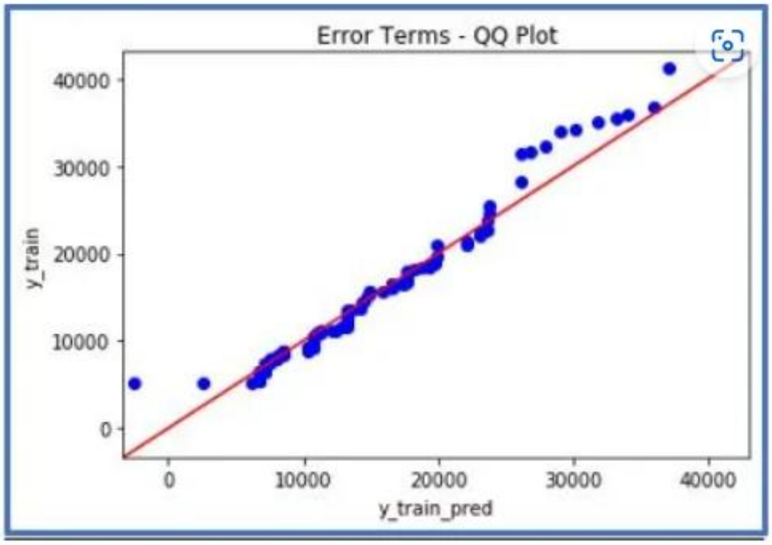
Below are the possible interpretations for two data sets.

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.


Error Terms - QQ Plot
y_train
y_train_pred