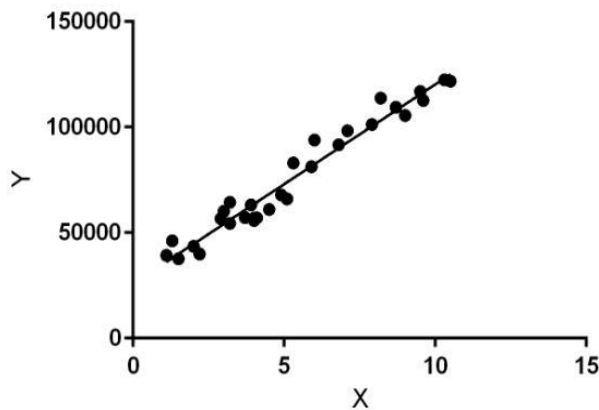


General Subjective Questions

@1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regress and. The independent variables can be called exogenous variables, predictor variables, or regressors



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model. Hypothesis function for Linear Regression:

Hypothesis function for Linear Regression: $y = B_1 + B_2 * x$

While training the model we are given:

x: input training data (univariate – one input variable(parameter))

y: labels to data (Supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values. θ_1 : intercept θ_2 : coefficient of x. Once we find the best θ_1 and θ_2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x

How to update θ_1 and θ_2 values to get the best fit line?

Cost Function (J): By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$\text{minimize } \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2 \quad J = \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y). Gradient Descent: To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

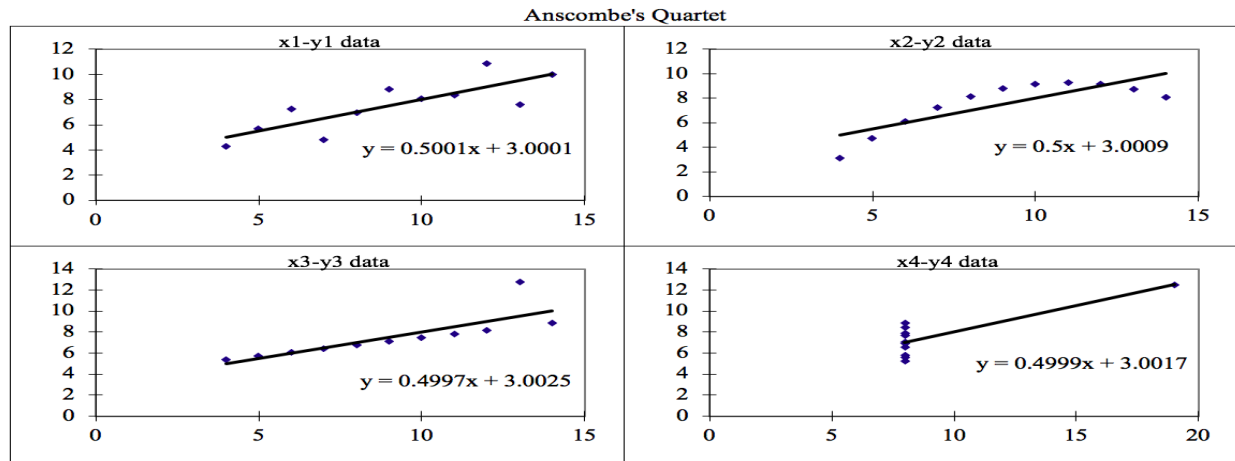
@2. Explain the Anscombe's quartet in detail.

Ans: 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.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only

be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



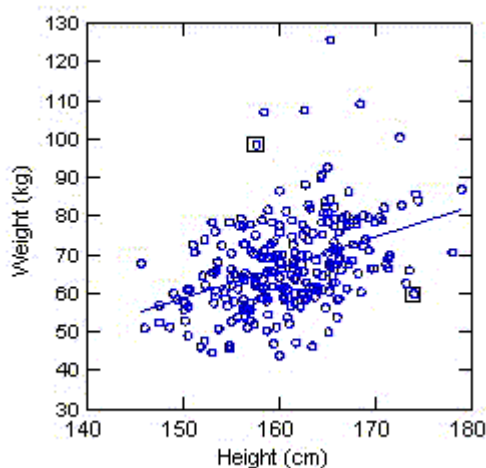
The four datasets can be described as:

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

@3. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.



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

Formula is

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

@4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

What Scaling?:

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Why Scaling?:

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.

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.

$$v_i = \max_{a_i} \min_{a_{-i}} v_i(a_i, a_{-i})$$

i = index of the player of interest

$-i$ = denotes all other players except player i

a_i = action taken by player i

a_{-i} = actions taken by all other players

v_i = value function of player i

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).

`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.

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

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

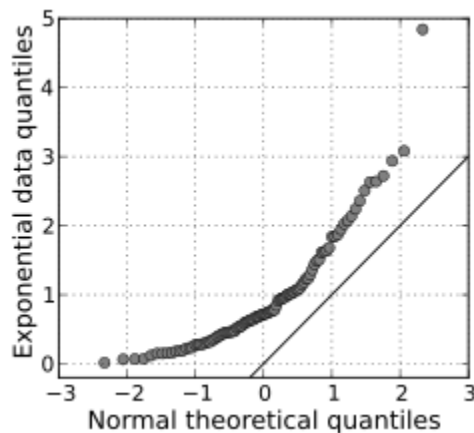
Ans: If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which 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.

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

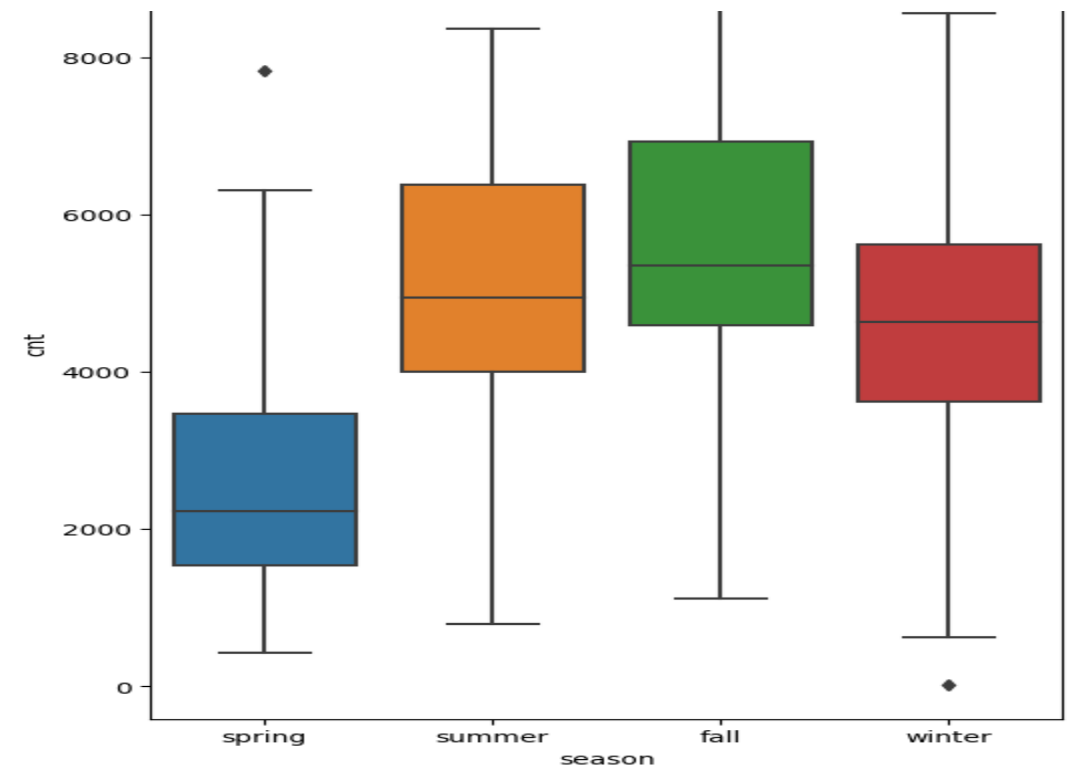
A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

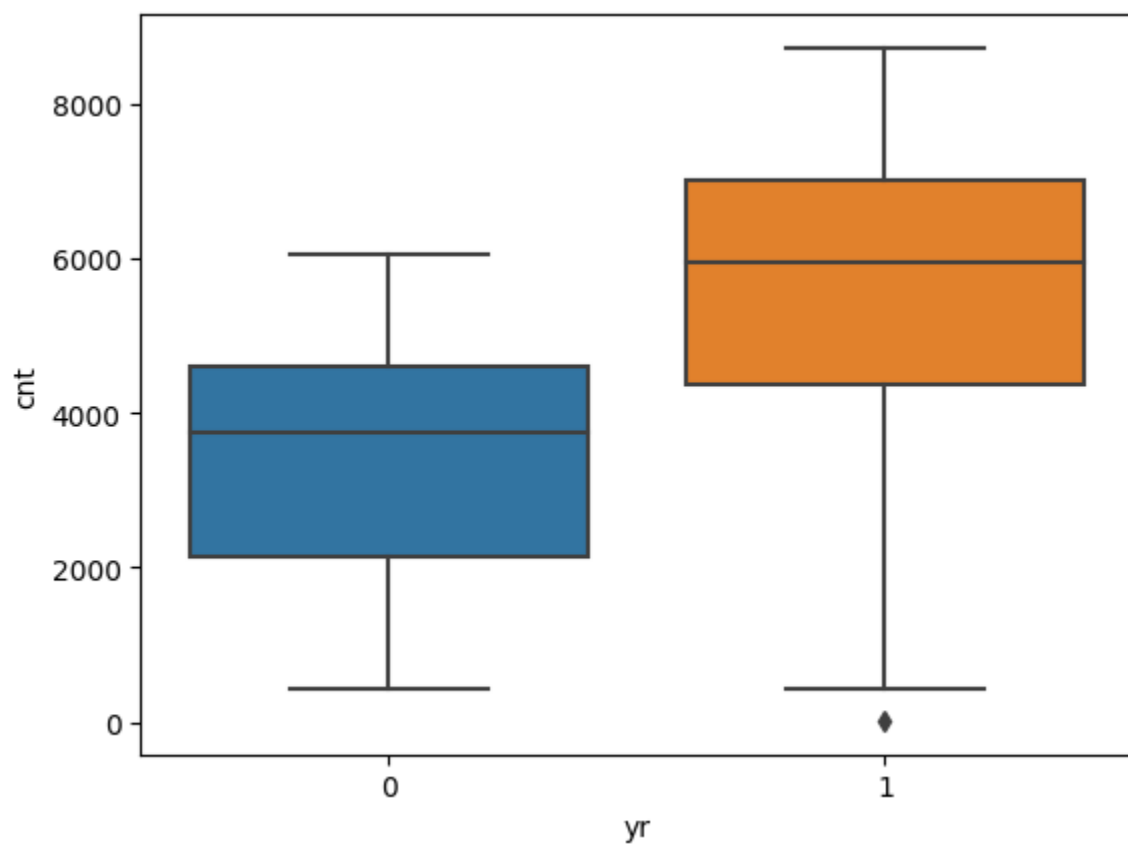
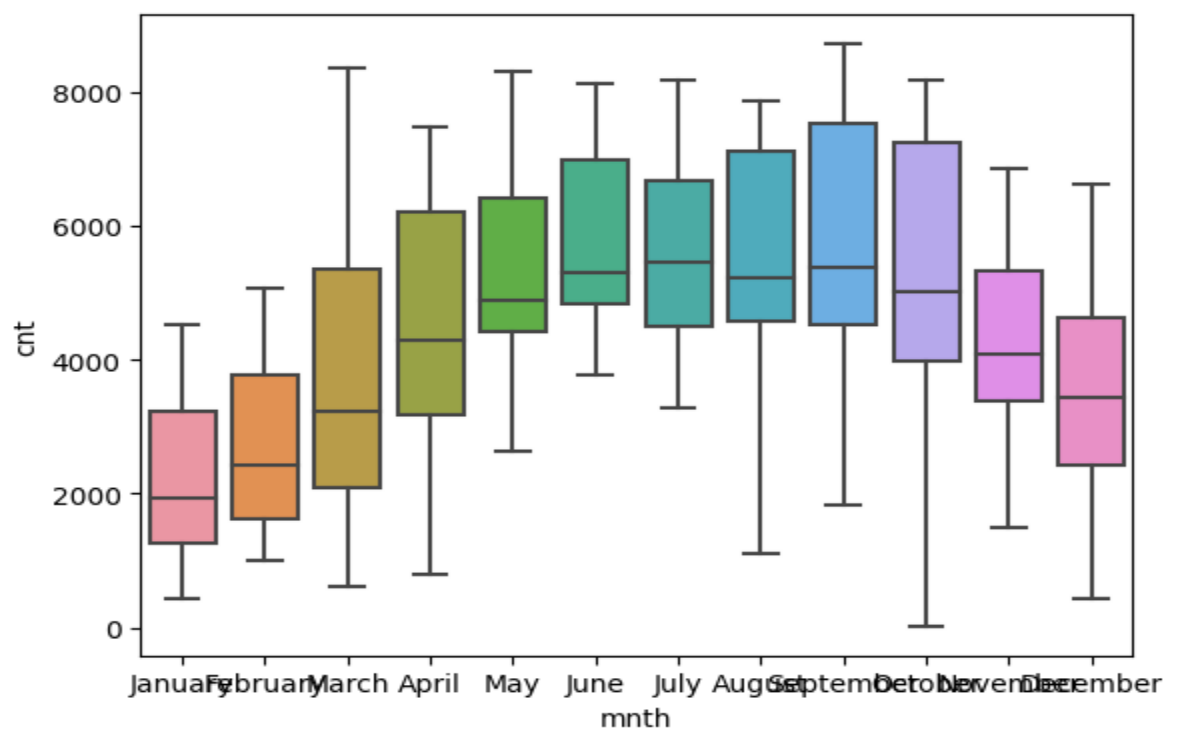
Assignment-based Subjective Questions

@1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: I found few categorical variable shows linear relation between target variable such as temp shows that.

- if temperature increases demands are also increases
- Months also playing good role here between April to October shows linear demands in bike
- Year shows a good growth in demands of bikes





@2 Why is it important to use drop first=True during dummy variable creation?

Ans: 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.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

It can be necessary for some situations, while not applicable for others. The goal is to reduce the number of columns by dropping the column that is not necessary. However, it is not always true. For some situations, we need to keep the first column.

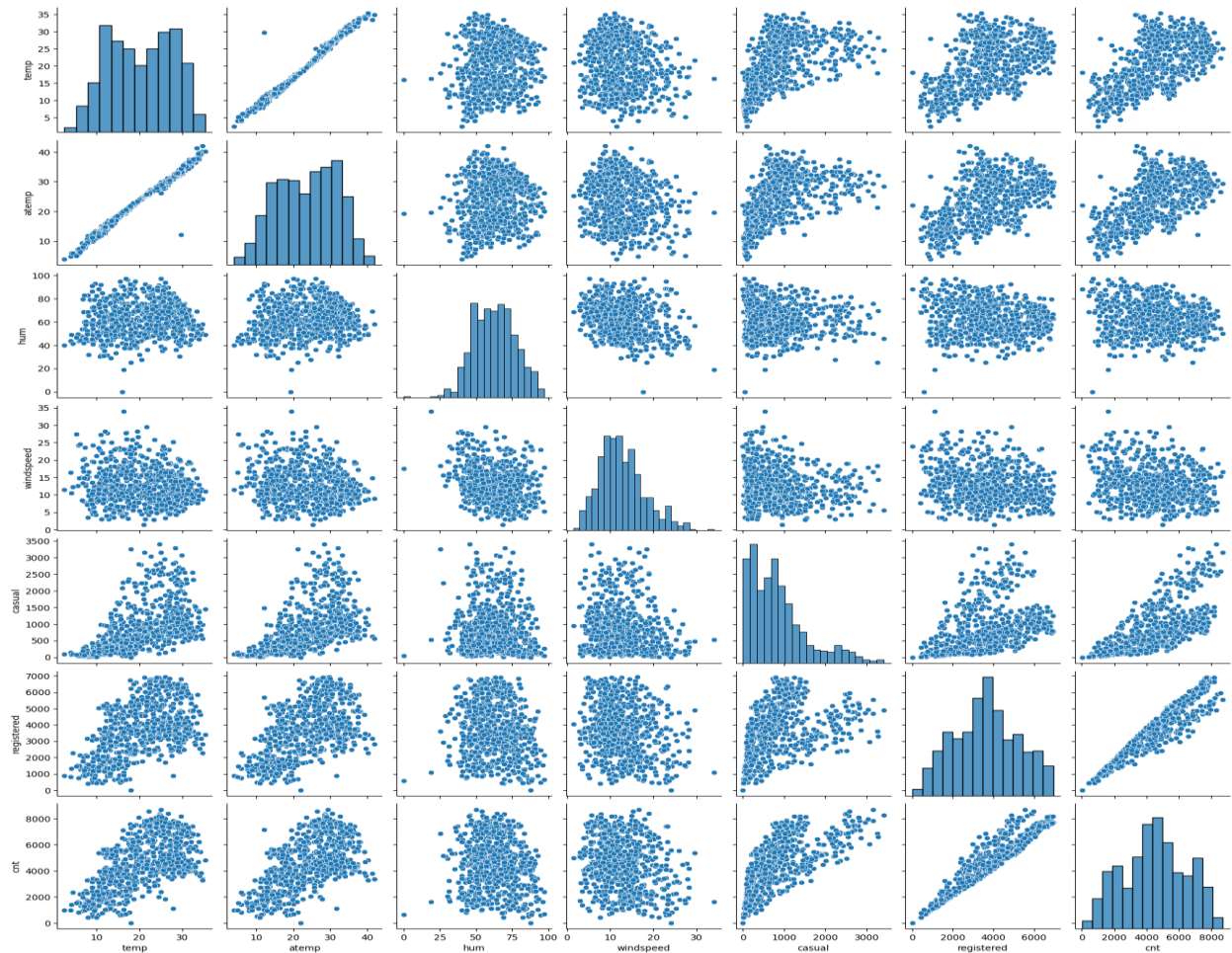
Example

Suppose, we have 5 unique values in a column called "Fav_genre"- "Rock", "Hip hop", "Pop", "Metal", "Country" This column contains value While dummy variable creation, we usually generate 5 columns. In this case, drop_first=True is not applicable. A person may have more than one favorite genres. So dropping any of the columns would not be right. Hence, drop_first=False is the default parameter.

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

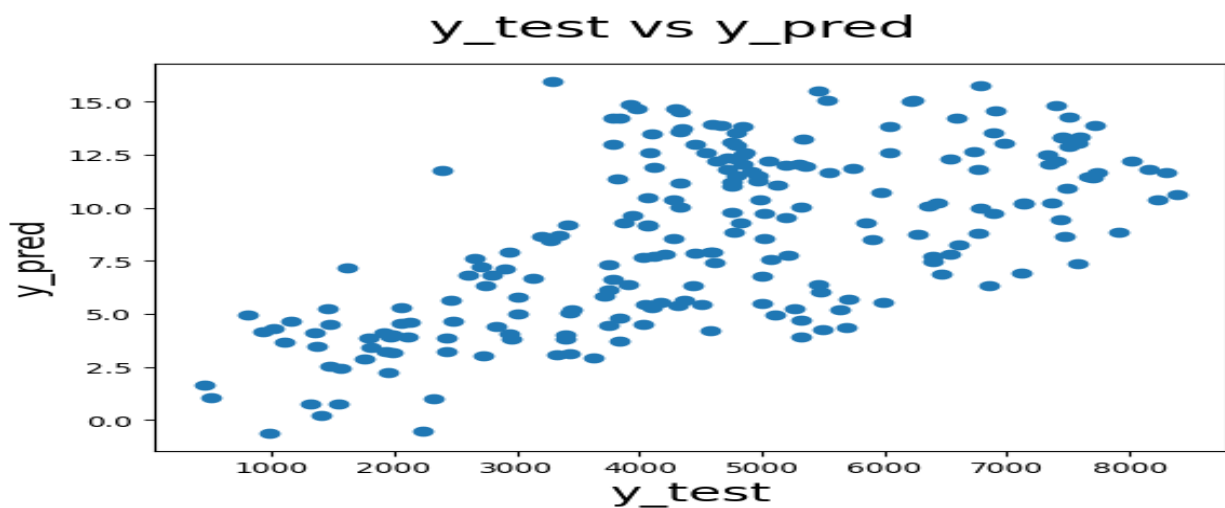
Looking at pair plot I can Identify that few numerical variables has highest coefficient with cnt among others are

1. atemp
2. temp
3. registered

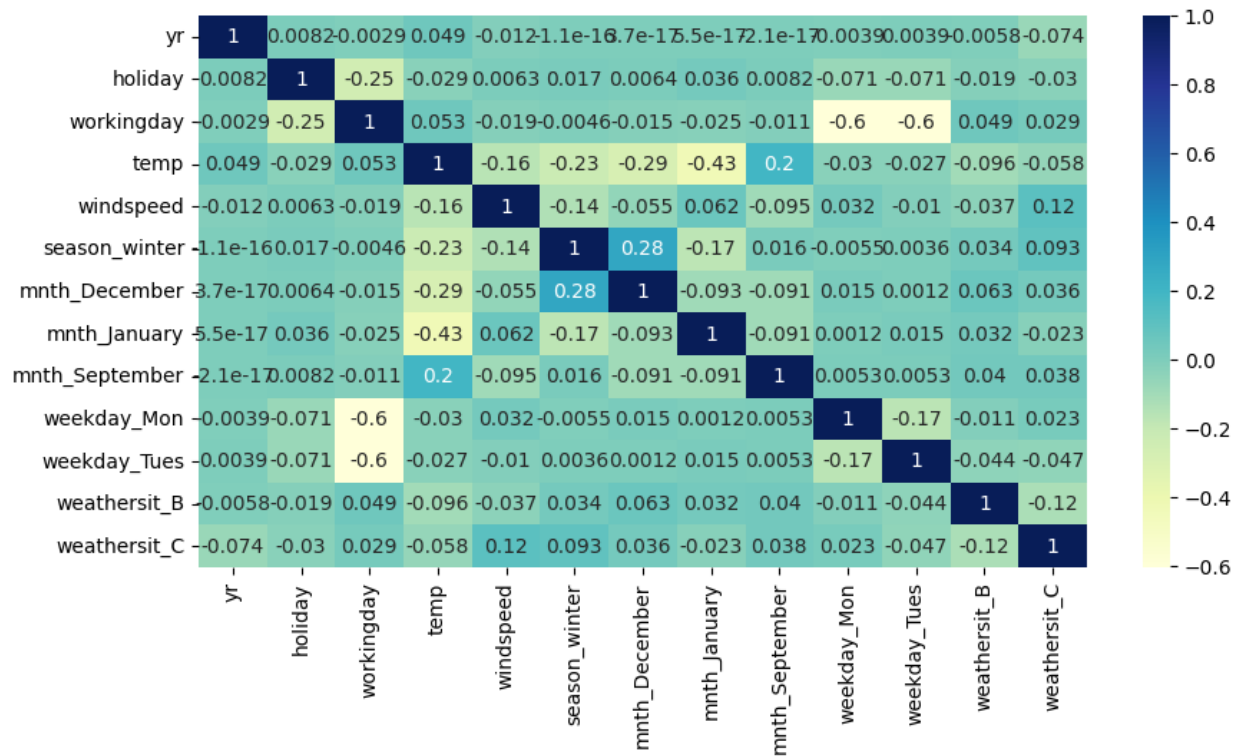


@4. How did you validate the assumptions of Linear Regression after building the model on the training set?

by creating a scatter plot x vs y . If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variable



Checking the correlations with heatmap between the final predictor variables



@4. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. Temperature
2. Working Day
3. Season