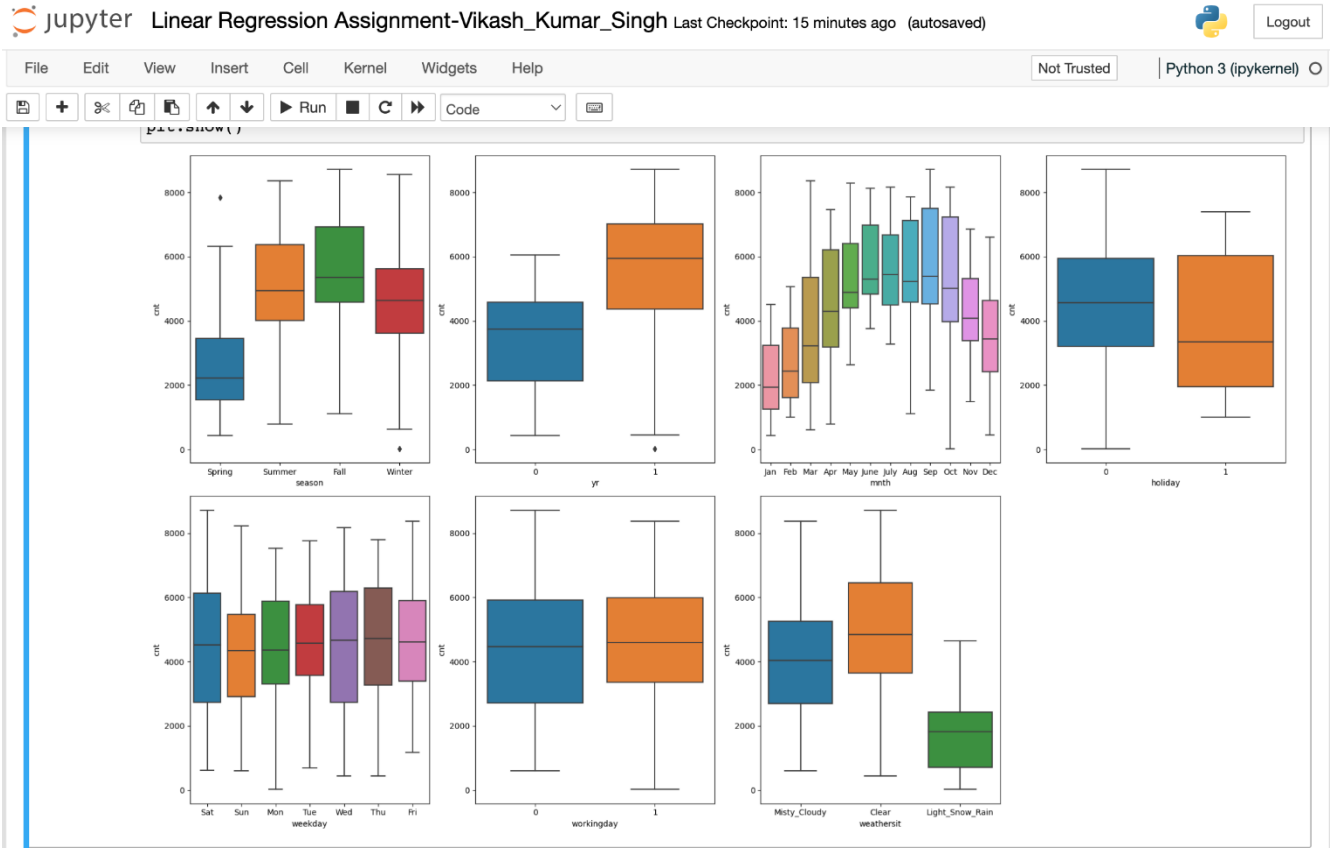
Assignment-based Subjective Questions

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



Ans: season, yr, mnth, holiday, weekday, workingday and weathersit are the categorical variable in the boombike dataset. Checking the above visualization through the boxplots we could infer the:

a) Season - The boxplot shows that "spring" season had the least effect on the target variable (cnt) whereas "fall" had maximum effect on the target variable. "summer" and "winter" had intermediate effect on the dependent variable. Business can focus to come up with schemes in "spring" to increase rental counts (cnt).

b) yr - The number of bike rentals (target variable count) in 2019 was more than 2018. This suggests progress growth.

c) weathersit - the boxplot shows less rentals during bad weather like heavy rain or snow. This indicates that this weather is extremely unfavourable and hence affects the target variable. Positive affect was seen on the dependent variable when the weather was clear or partly cloudy.

d) mnth - September has the highest number of rentals followed by October. December sees the least rentals. This observation is in sync with the observation made in weathersit where "fall" has the highest rental count (sep, oct). Due to heavy snow and adverse weather in December and late winter/spring the rental sees a dip.

e) Holiday - rentals (dependent variable count) decreased during holiday.
f) Weekday - The count of rentals (dependent variable count) is almost even throughout the week
g) Workingday – The affect is constant almost throughout the week.

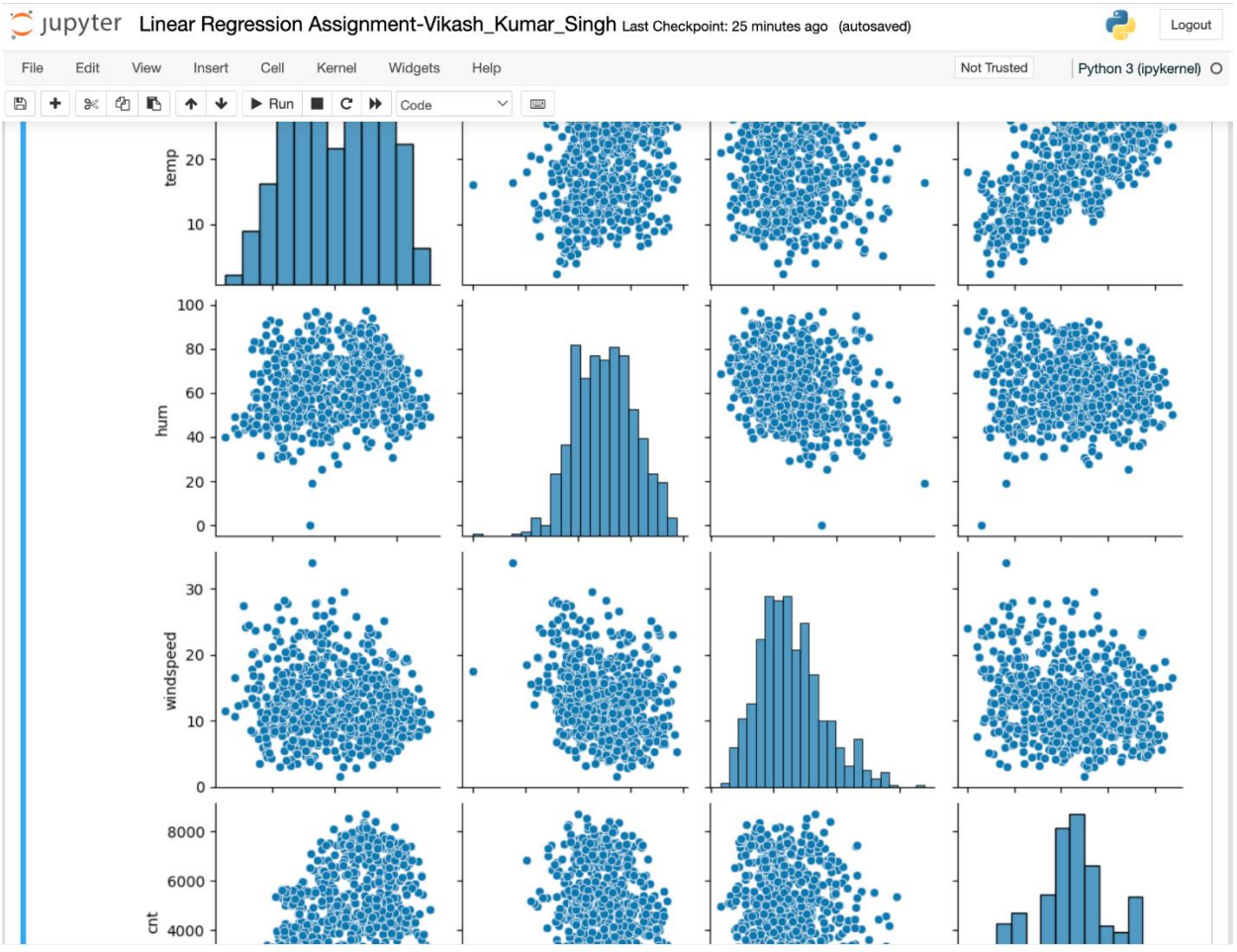
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

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. In case of Multiple Linear Regression keeping k dummies for k levels of a categorical variable is good idea, but there is a redundancy of one level. Since one of the combinations will be uniquely representing the redundant column, hence it is better to drop one of the columns and just have k-1 dummies(columns) to represent k levels.

This Overall approach reduces multi-collinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression and also helps to improve the performance of the model.

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

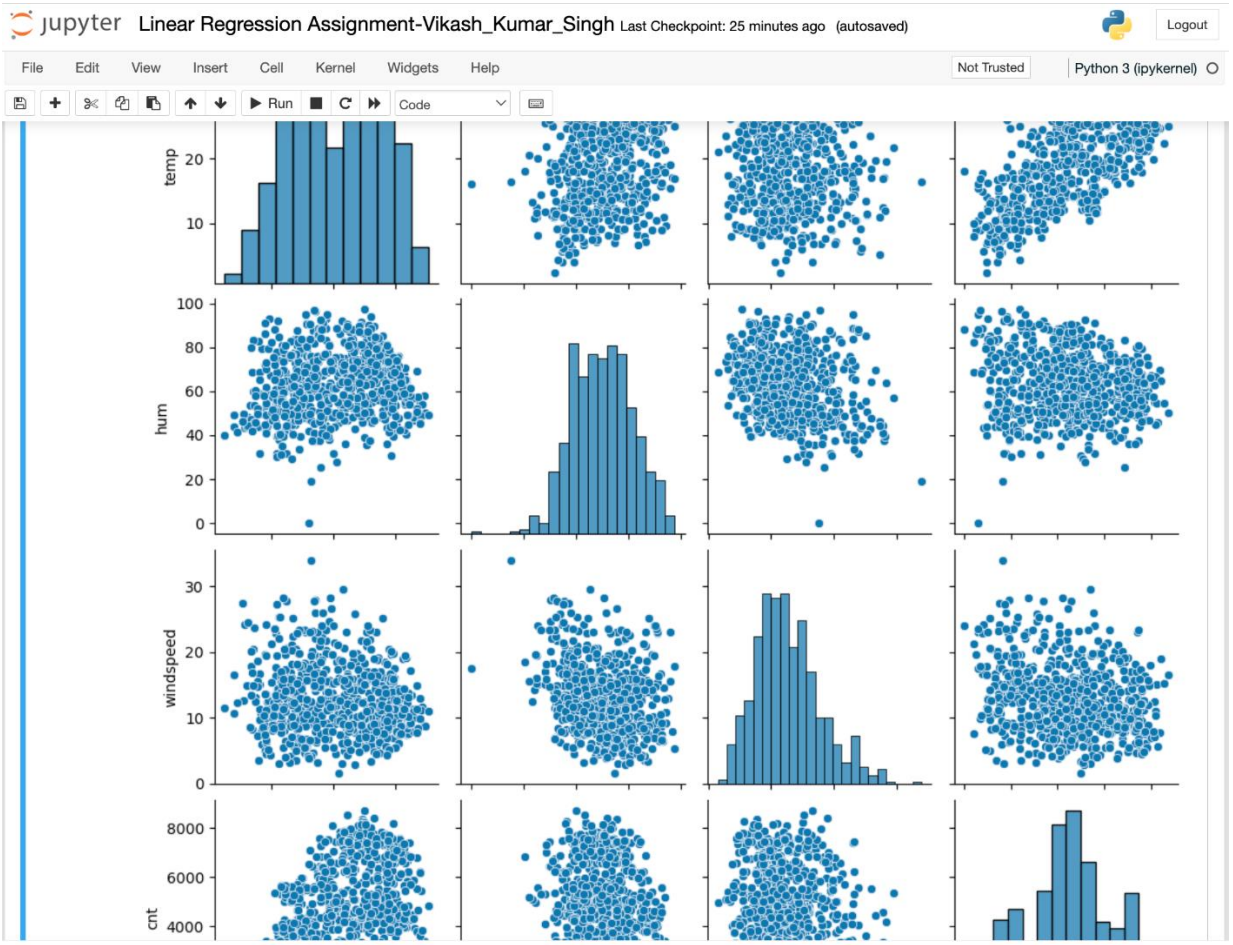
Ans: From the below pairplot we observe that, "temp" is highly correlated with the target variable(cnt)



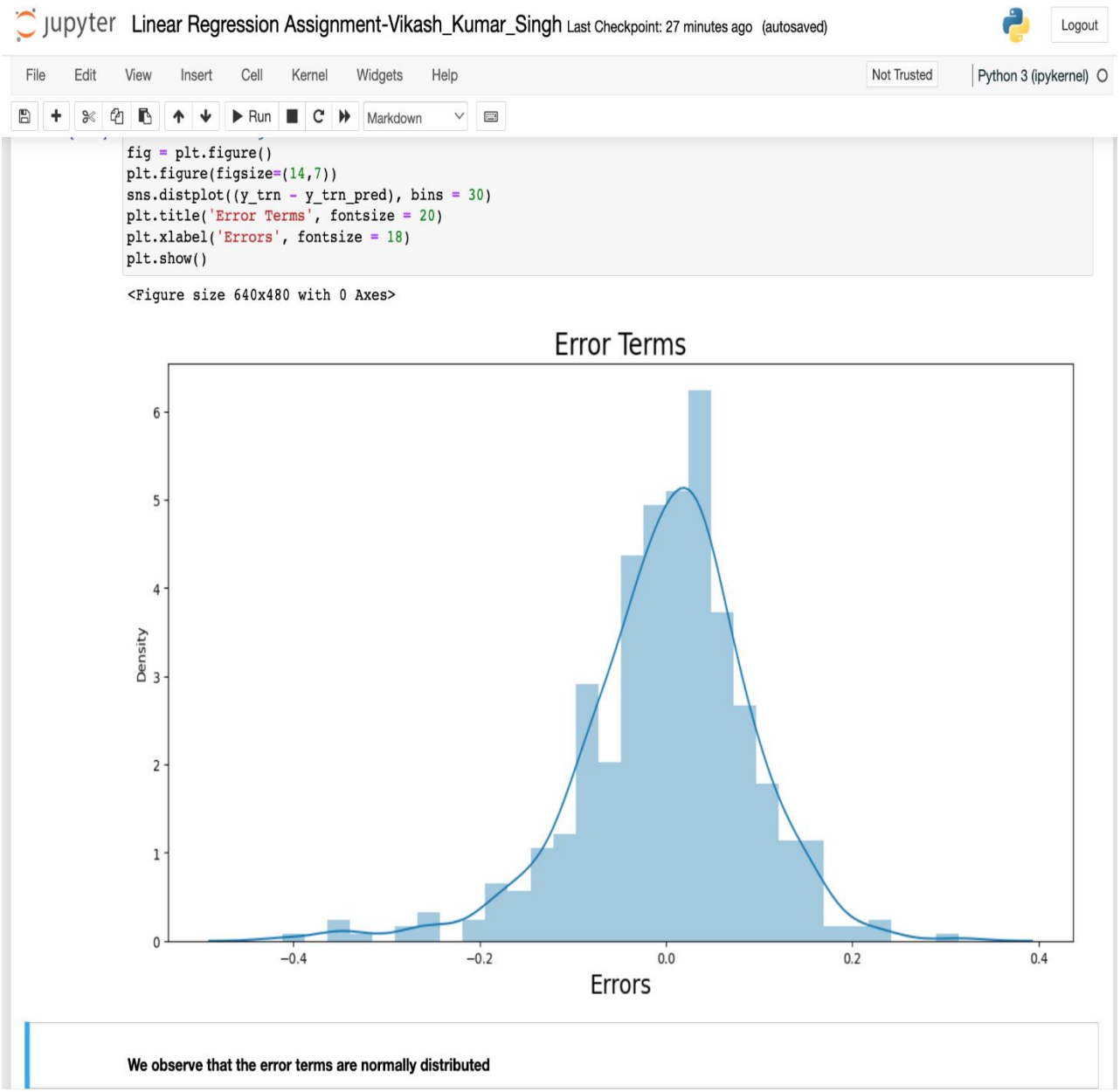
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: The following actions were performed to validate the Assumptions of Linear Regression:

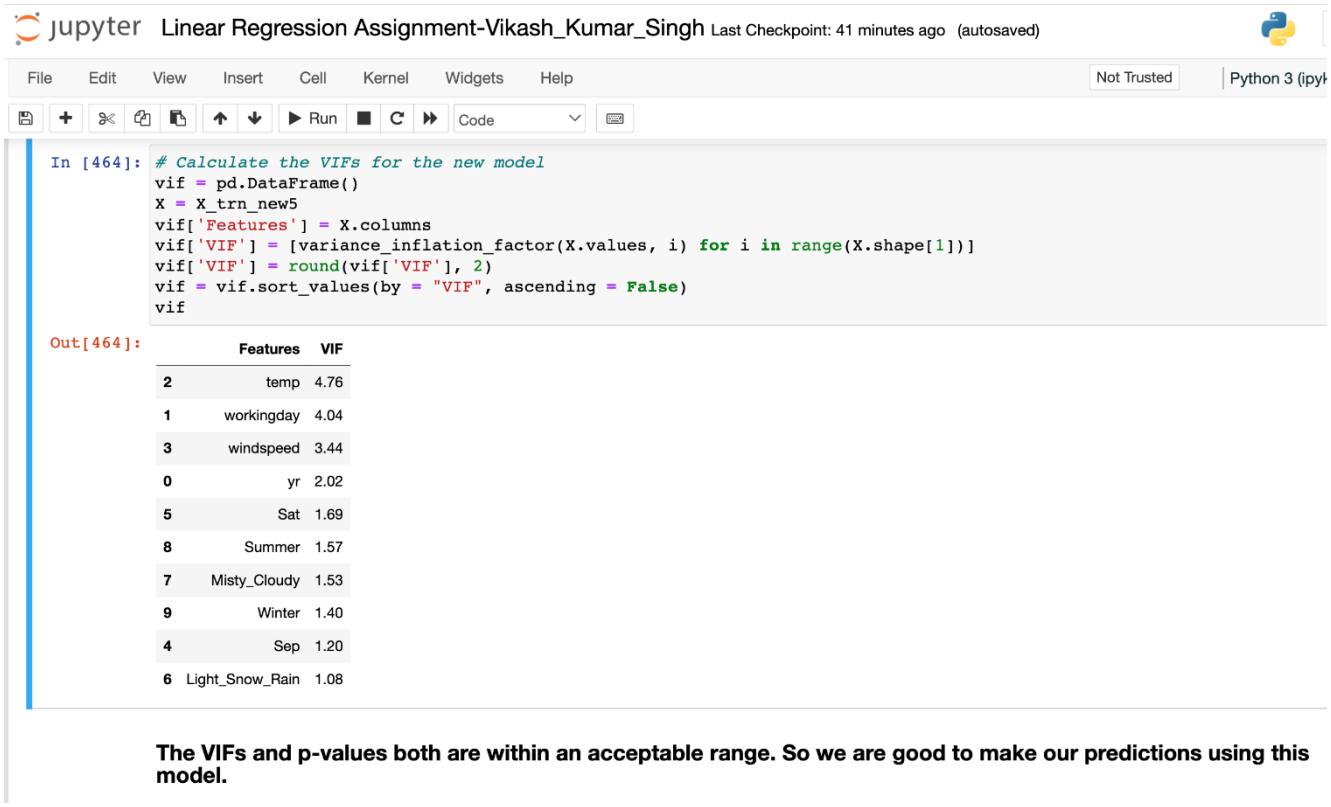
1) **Linear relationship**: First, Linear Regression needs the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatter plots. We visualised the numeric variables using a pairplot where no or little linearity is present (see the attached scatterplot)



2) Secondly, the linear regression analysis requires **all variables to be multivariate normal** (residuals distribution should follow normal distribution and centred around 0 (mean = 0)). This assumption can best be checked with a histogram or a Q-Q-Plot. The diagram below shows that the residuals are distributed about mean = 0.



3) Thirdly, linear regression assumes that there is **little or no multicollinearity in the data**. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model



In [464]:
```
# Calculate the VIFs for the new model
vif = pd.DataFrame()
X = X_trn_new5
vif['Features'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

Out[464]:

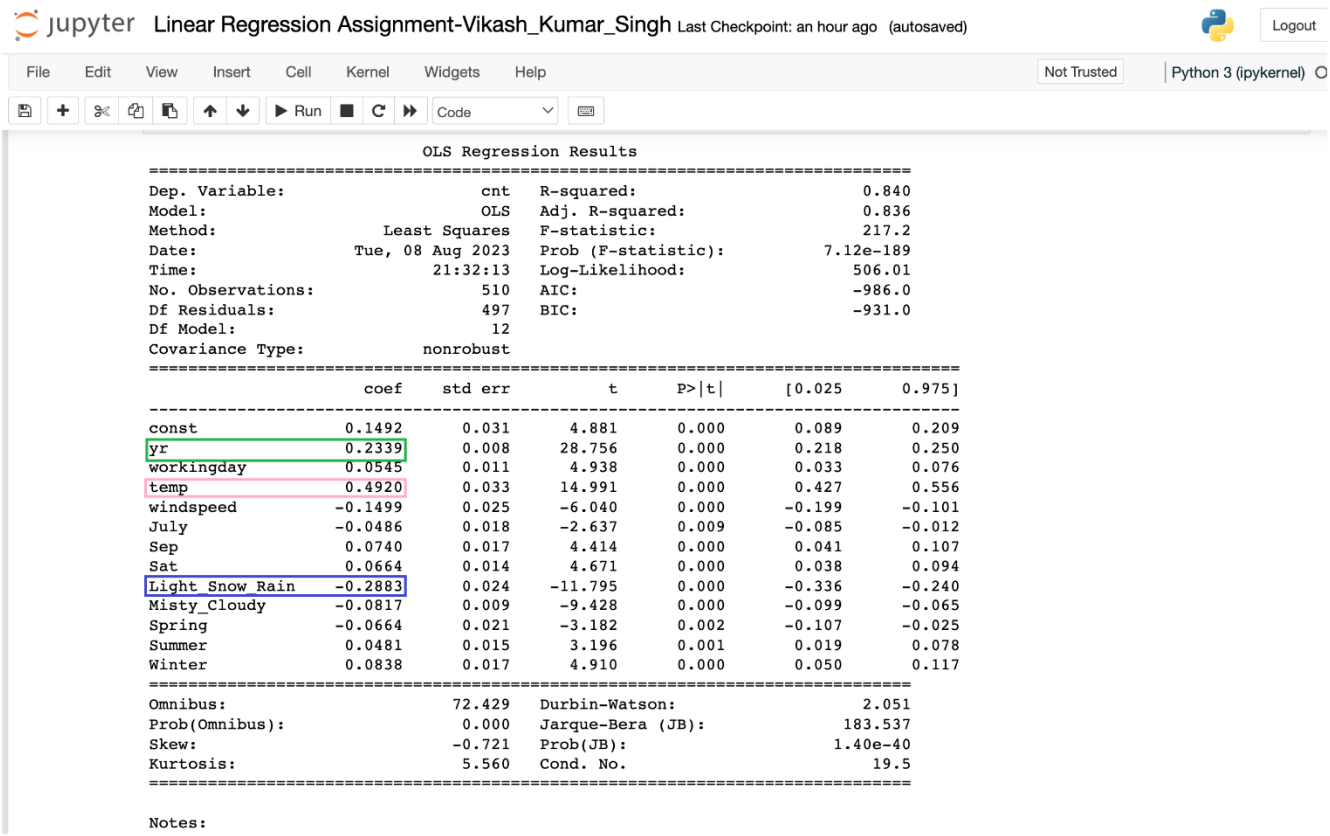
	Features	VIF
2	temp	4.76
1	workingday	4.04
3	windspeed	3.44
0	yr	2.02
5	Sat	1.69
8	Summer	1.57
7	Misty_Cloudy	1.53
9	Winter	1.40
4	Sep	1.20
6	Light_Snow_Rain	1.08

The VIFs and p-values both are within an acceptable range. So we are good to make our predictions using this model.

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

Ans: The top 3 features are:

1. **temp** (coefficient = 0.4920)
2. **yr** (coefficient = 0.2339)
3. **Light_Snow_Rain** (weathersit) →coefficient: -0.2883



OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	217.2
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	7.12e-189
Time:	21:32:13	Log-Likelihood:	506.01
No. Observations:	510	AIC:	-986.0
Df Residuals:	497	BIC:	-931.0
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1492	0.031	4.881	0.000	0.089	0.209
yr	0.2339	0.008	28.756	0.000	0.218	0.250
workingday	0.0545	0.011	4.938	0.000	0.033	0.076
temp	0.4920	0.033	14.991	0.000	0.427	0.556
windspeed	-0.1499	0.025	-6.040	0.000	-0.199	-0.101
July	-0.0486	0.018	-2.637	0.009	-0.085	-0.012
Sep	0.0740	0.017	4.414	0.000	0.041	0.107
Sat	0.0664	0.014	4.671	0.000	0.038	0.094
Light_Snow_Rain	-0.2883	0.024	-11.795	0.000	-0.336	-0.240
Misty_Cloudy	-0.0817	0.009	-9.428	0.000	-0.099	-0.065
Spring	-0.0664	0.021	-3.182	0.002	-0.107	-0.025
Summer	0.0481	0.015	3.196	0.001	0.019	0.078
Winter	0.0838	0.017	4.910	0.000	0.050	0.117

Omnibus:	72.429	Durbin-Watson:	2.051
Prob(Omnibus):	0.000	Jarque-Bera (JB):	183.537
Skew:	-0.721	Prob(JB):	1.40e-40
Kurtosis:	5.560	Cond. No.	19.5

Notes:

General Subjective Questions

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

Ans:

Definition: In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset. Linear Regression is a machine learning technique where the model predicts the output as a continuous numerical value.

Uses: Linear Regression analysis is often used in finance, investing, predicting house price, stock market or salary of an employee and others. It finds out the relationship between a single dependent variable (target variable) dependent on several independent ones.

Types of Linear Regression:

Linear regression models can be classified into two types depending upon the number of independent variables:

1. Simple Linear Regression

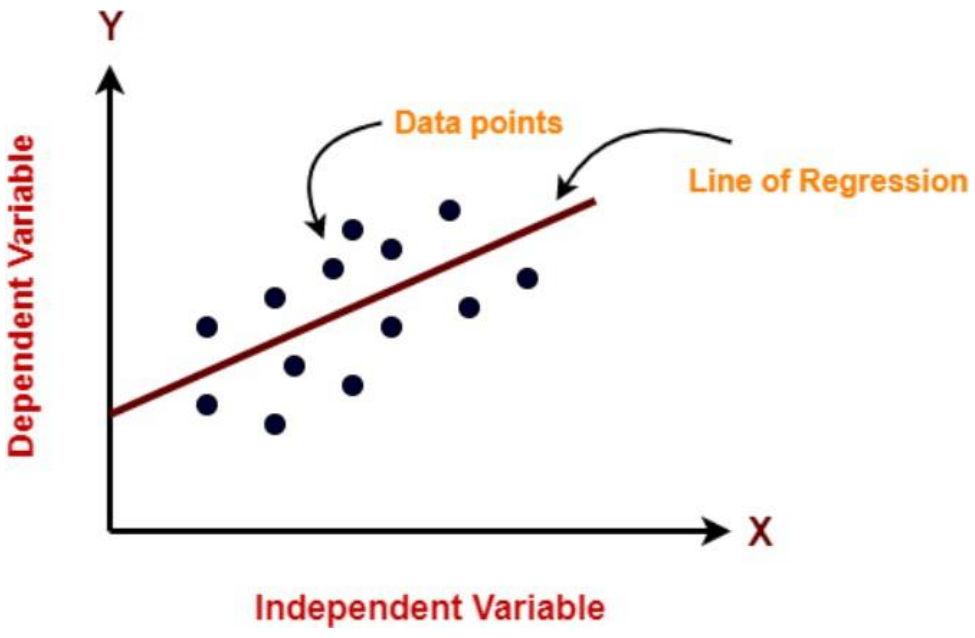
2. Multiple Linear Regression

Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

Simple Linear Regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

One variable, denoted **X** is regarded as the **predicto**r, **explanatory**, or **independent variable**.

The other variable, denoted **y**, is regarded as the **response**, **outcome**, or **dependent variable**.



The standard equation of the Linear Regression line is given by the following expression:

$Y = \beta_0 + \beta_1 X$

Where β_0 is the intercept and β_1

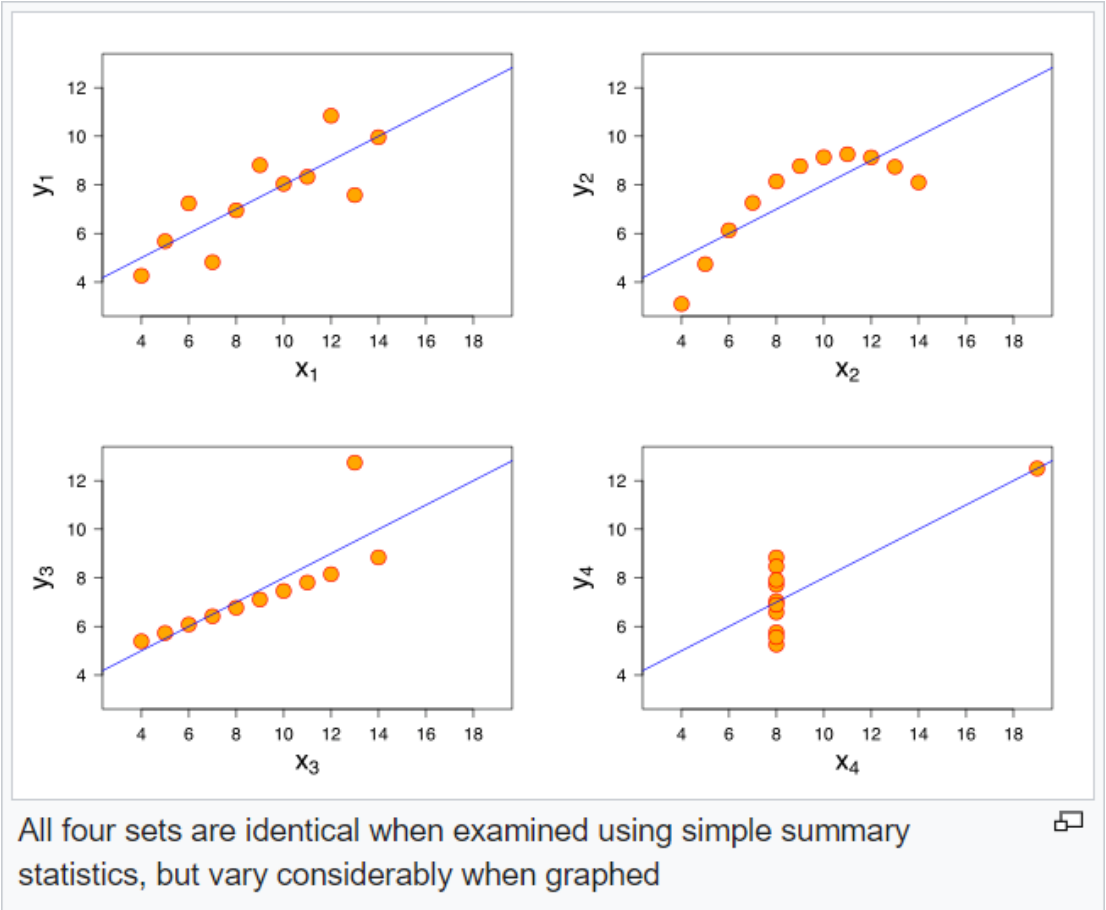
Multiple Linear Regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent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

The formulation for multiple linear regression is also similar to simple linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$y=\beta 0+\beta 1X1+\beta 2X2+\beta 3X3+\ldots+\beta 3X3+\epsilon$

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

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear verydifferent when graphed. Each dataset consists of eleven (x, y) points. They were 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.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

For all four datasets:

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- 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 do not indicate any relationship between the variables.

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

Ans: Pearson correlation coefficient or **Pearson's correlation coefficient** or **Pearson's r** is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative.

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1.

Pearson correlation coefficient formula:

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using the feature scaling method, then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes.

Techniques to perform Feature Scaling

Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature valueso that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

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

Ans: A **variance inflation factor**(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are calculated by taking a predictor and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. "i" is the predictor you're looking at (e.g., x1 or x2):

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient.

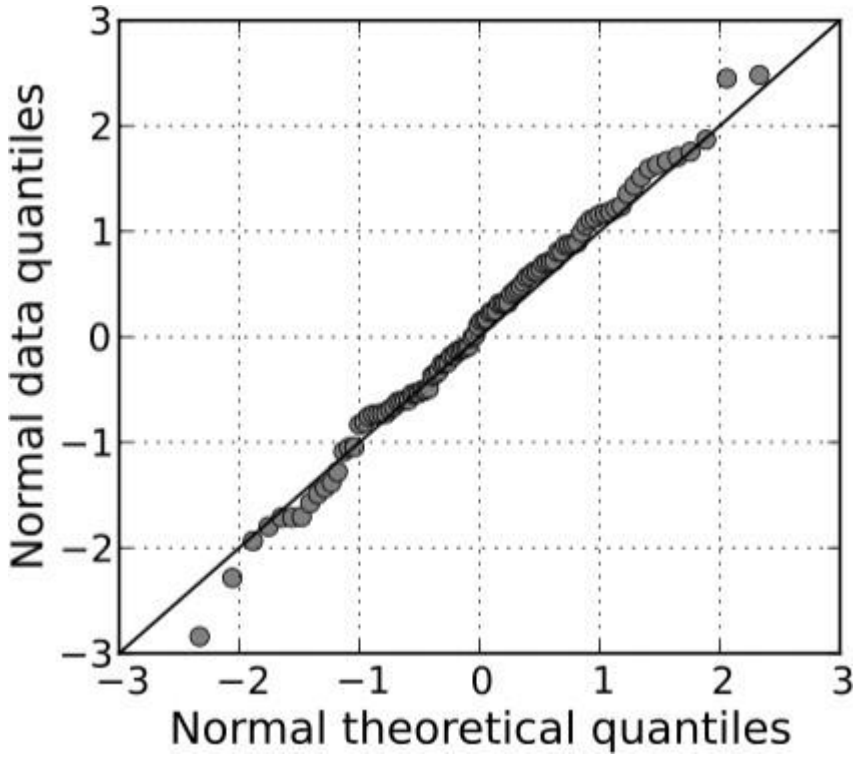
A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

VIF helps to check how well the independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity".

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

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 normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are normally distributed.