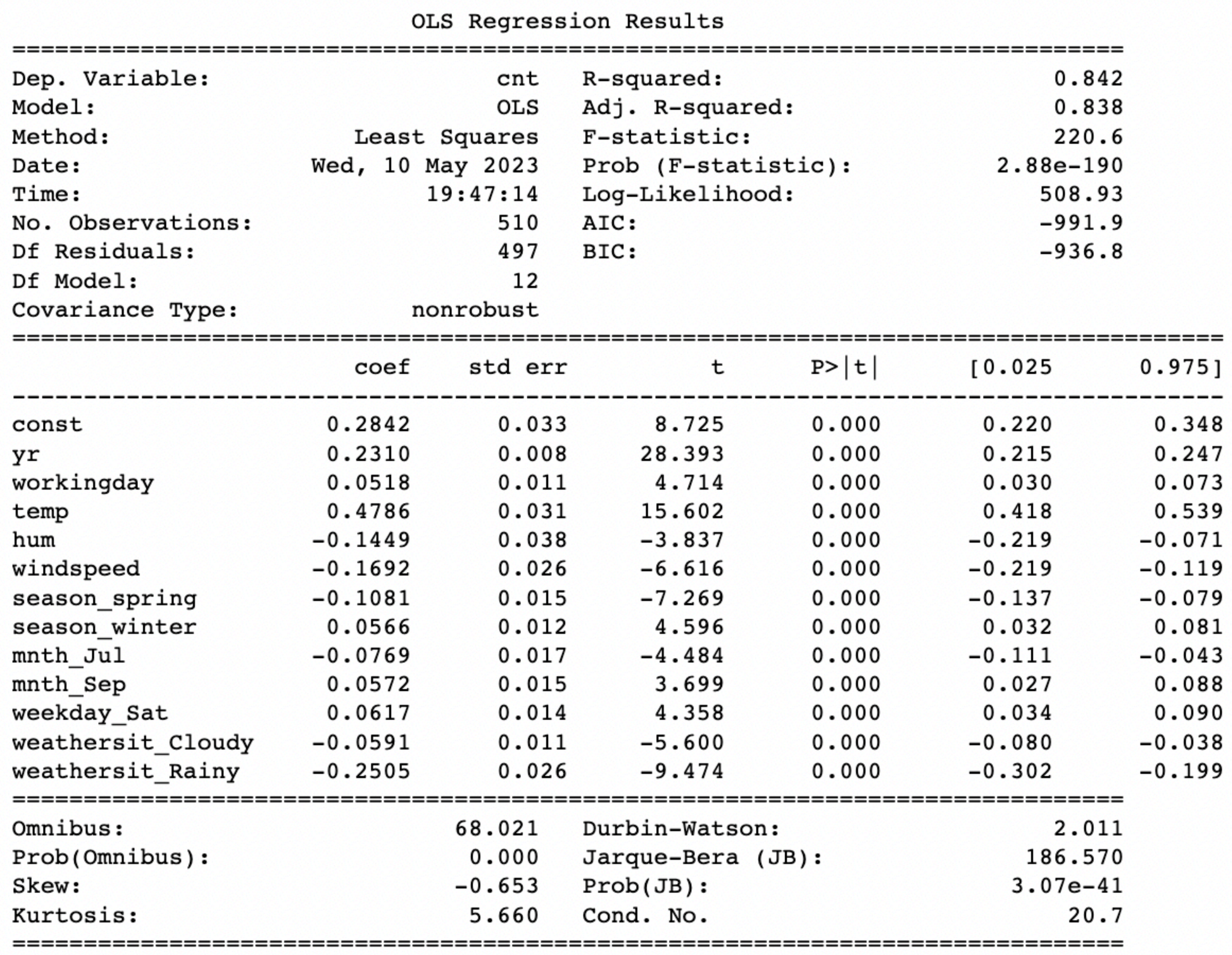
Assignment-based Subjective Questions

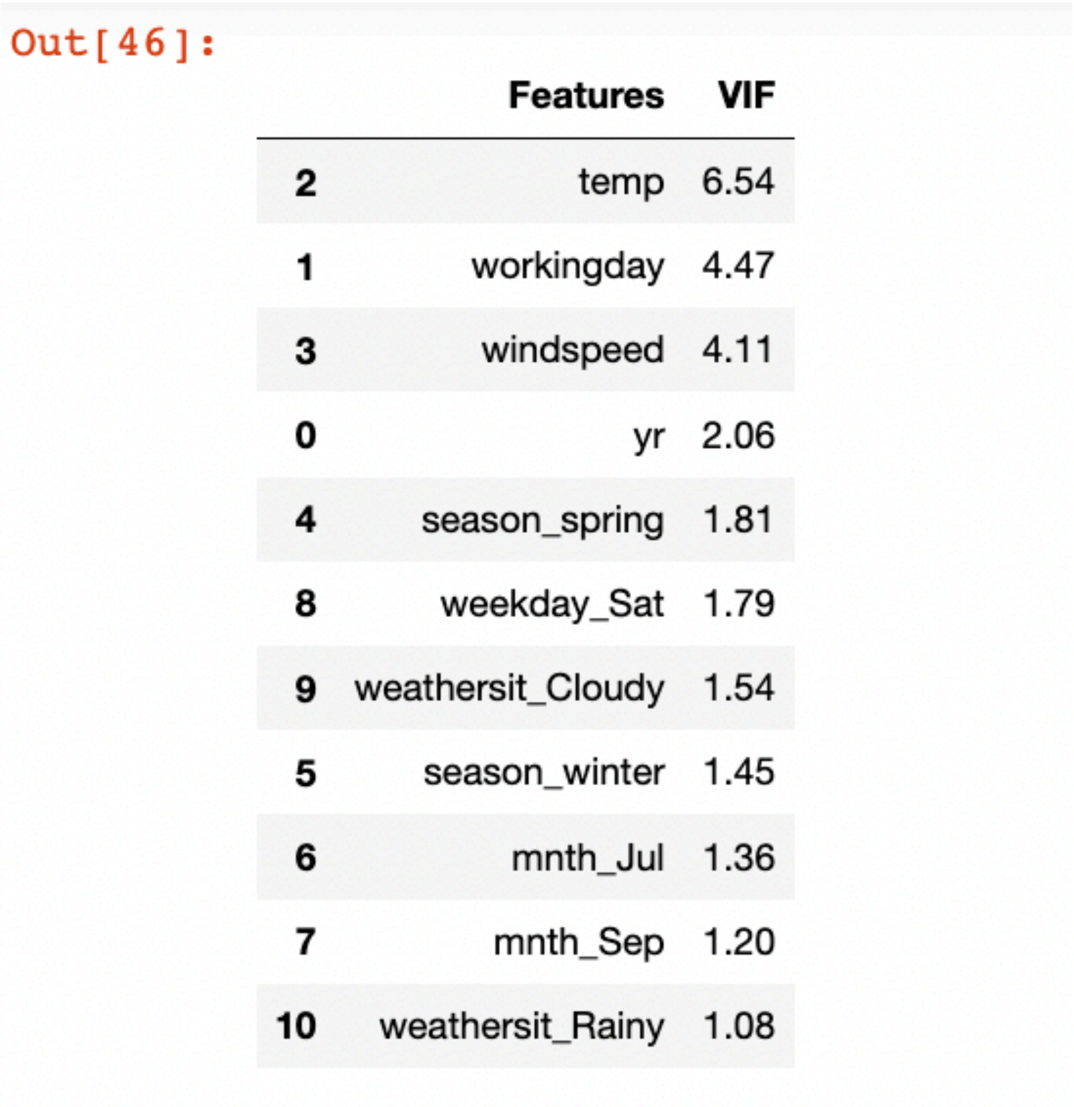
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

In dataset these are categorical variables list : [‘season’,'yr','mnth','holiday','weekday','workingday','weathersit']

Further proceeding with model building steps we followed creating dummy variables for these categorical variables and finally came up with p-value and VIF values which are u
about their effect.
on the dependent variable ‘cnt’(count of total rental bikes including both casual and registered):



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.842
Model:                            OLS   Adj. R-squared:                  0.838
Method:                 Least Squares   F-statistic:                     220.6
Date:                Wed, 10 May 2023   Prob (F-statistic):          2.88e-190
Time:                        19:47:14   Log-Likelihood:                 508.93
No. Observations:                 510   AIC:                            -991.9
Df Residuals:                     497   BIC:                            -936.8
Df Model:                          12
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.2842      0.033      8.725      0.000       0.220       0.348
yr                    0.2310      0.008     28.393      0.000       0.215       0.247
workingday            0.0518      0.011      4.714      0.000       0.030       0.073
temp                  0.4786      0.031     15.602      0.000       0.418       0.539
hum                  -0.1449      0.038     -3.837      0.000      -0.219      -0.071
windspeed            -0.1692      0.026     -6.616      0.000      -0.219      -0.119
season_spring        -0.1081      0.015     -7.269      0.000      -0.137      -0.079
season_winter         0.0566      0.012      4.596      0.000       0.032       0.081
mnth_Jul             -0.0769      0.017     -4.484      0.000      -0.111      -0.043
mnth_Sep              0.0572      0.015      3.699      0.000       0.027       0.088
weekday_Sat           0.0617      0.014      4.358      0.000       0.034       0.090
weathersit_Cloudy    -0.0591      0.011     -5.600      0.000      -0.080      -0.038
weathersit_Rainy     -0.2505      0.026     -9.474      0.000      -0.302      -0.199
==============================================================================
Omnibus:                       68.021   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              186.570
Skew:                          -0.653   Prob(JB):                     3.07e-41
Kurtosis:                       5.660   Cond. No.                         20.7
==============================================================================
```



Out[46]:

	Features	VIF
2	temp	6.54
1	workingday	4.47
3	windspeed	4.11
0	yr	2.06
4	season_spring	1.81
8	weekday_Sat	1.79
9	weathersit_Cloudy	1.54
5	season_winter	1.45
6	mnth_Jul	1.36
7	mnth_Sep	1.20
10	weathersit_Rainy	1.08

Above screenshots shows following:

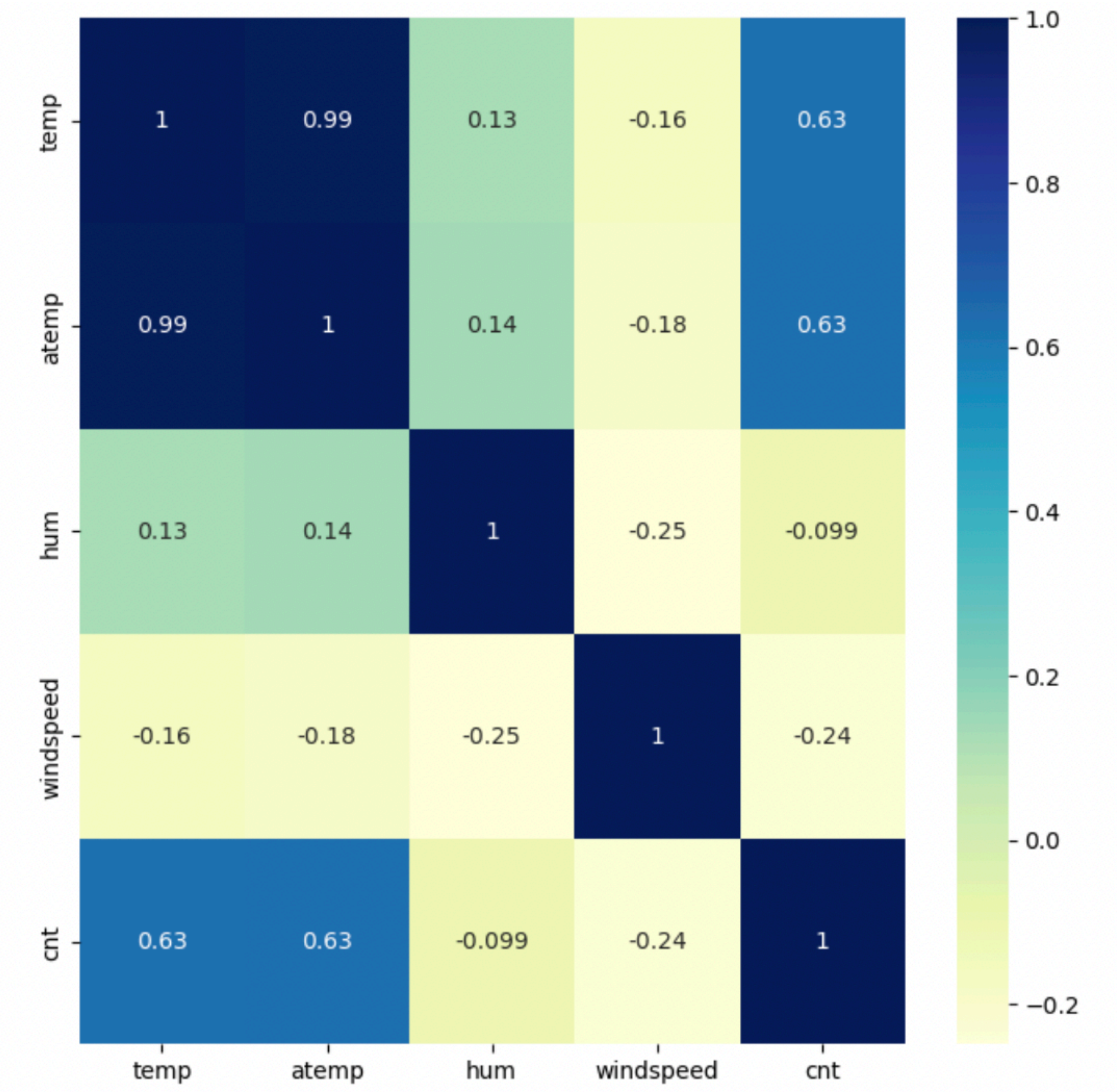
‘**Season**’ : variable has an effect on bike demand where summer and winter will show more demands than spring and fall
‘**temp**’ : variable has an effect on bike demand where on its increase there will be increase in demand too
‘**Yr**’ : variable has an effect on bike demand where each passing year demand will increase
‘**Windspeed**’ : variable has an effect on bike demand where when slow demand will increase
‘**Weathersit**’ : variable has an effect on bike demand where Rainy and Cloudy weather will decrease demand
‘**Weekday**’ : variable has an effect on bike demand where there can be positive impact on demand over weekends

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Variables that assume values such as 0 and 1 are called dummy variables.
Value 1 represents presence and 0 represents absence.
In its definitive term , Dummy variables (also known as binary, indicator, dichotomous, discrete, or categorical variables) are a way of incorporating qualitative information into
data, unlike continuous data, tell us simply whether the individual observation belongs to a particular category.
Dummy variables are another way in which the flexibility of regression can be demonstrated.

Example : Gender (male=1, female=0 or vice versa)

If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollineari
Dummy Variable Trap.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable? (1 mark)
Variables ‘temp’ (temperature in Celsius) and ‘atemp’(feeling temperature in Celsius) have highest correlation.
This can be seen in heat map also as well as pair plots too. Here providing heat map snapshot:
As it can be seen its 0.99 value here which is higher.

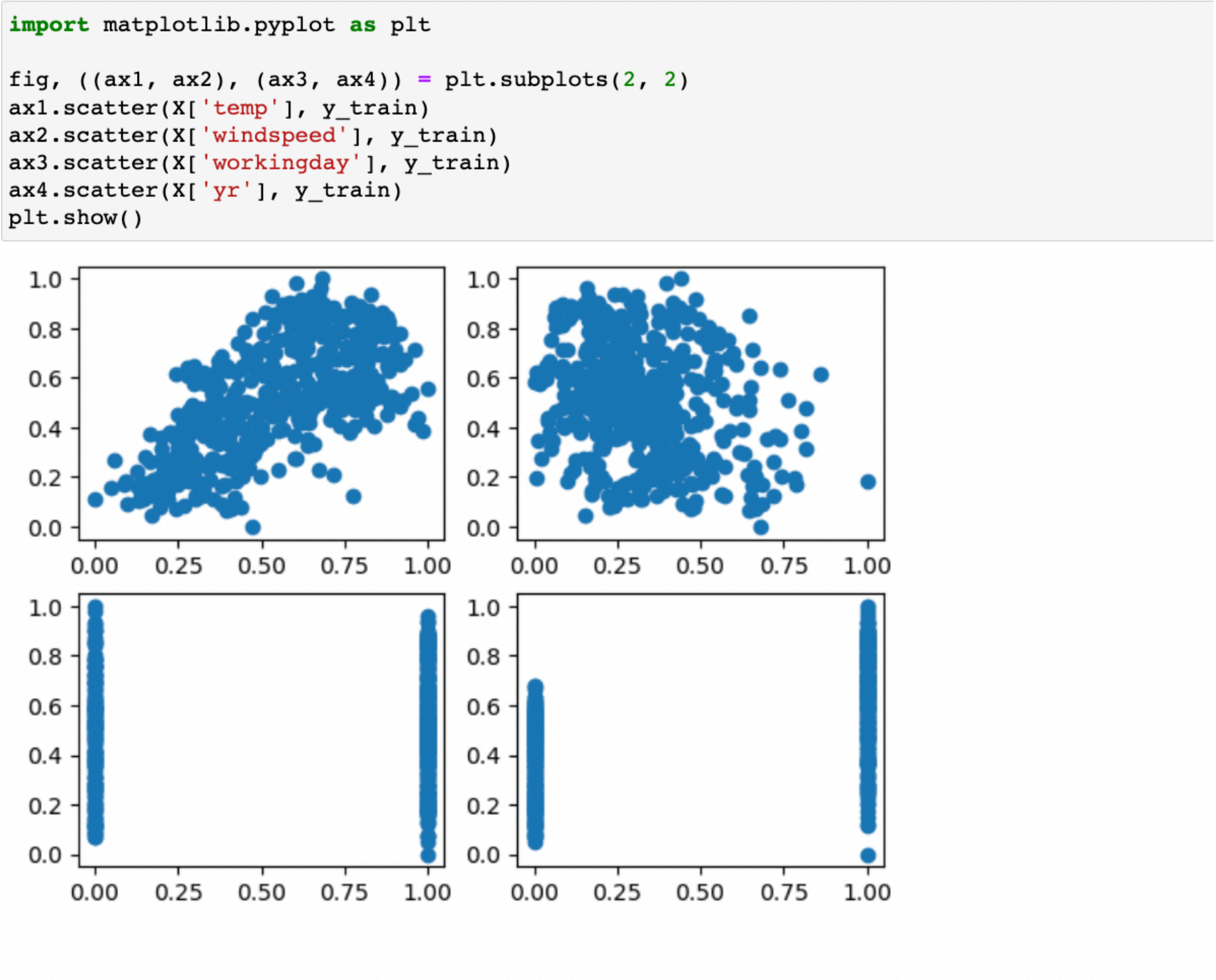


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Linear Regression Assumptions validation is done as follows based on below 5 points :

- Linear relationship
- No or little Multicollinearity
- Homoscedasticity
- Residual autocorrelation
- Normality of the residuals

A. Linear relationship was validated using scatterplot to identify linear relationship between numeric variables and dependent variables which was found linear only.

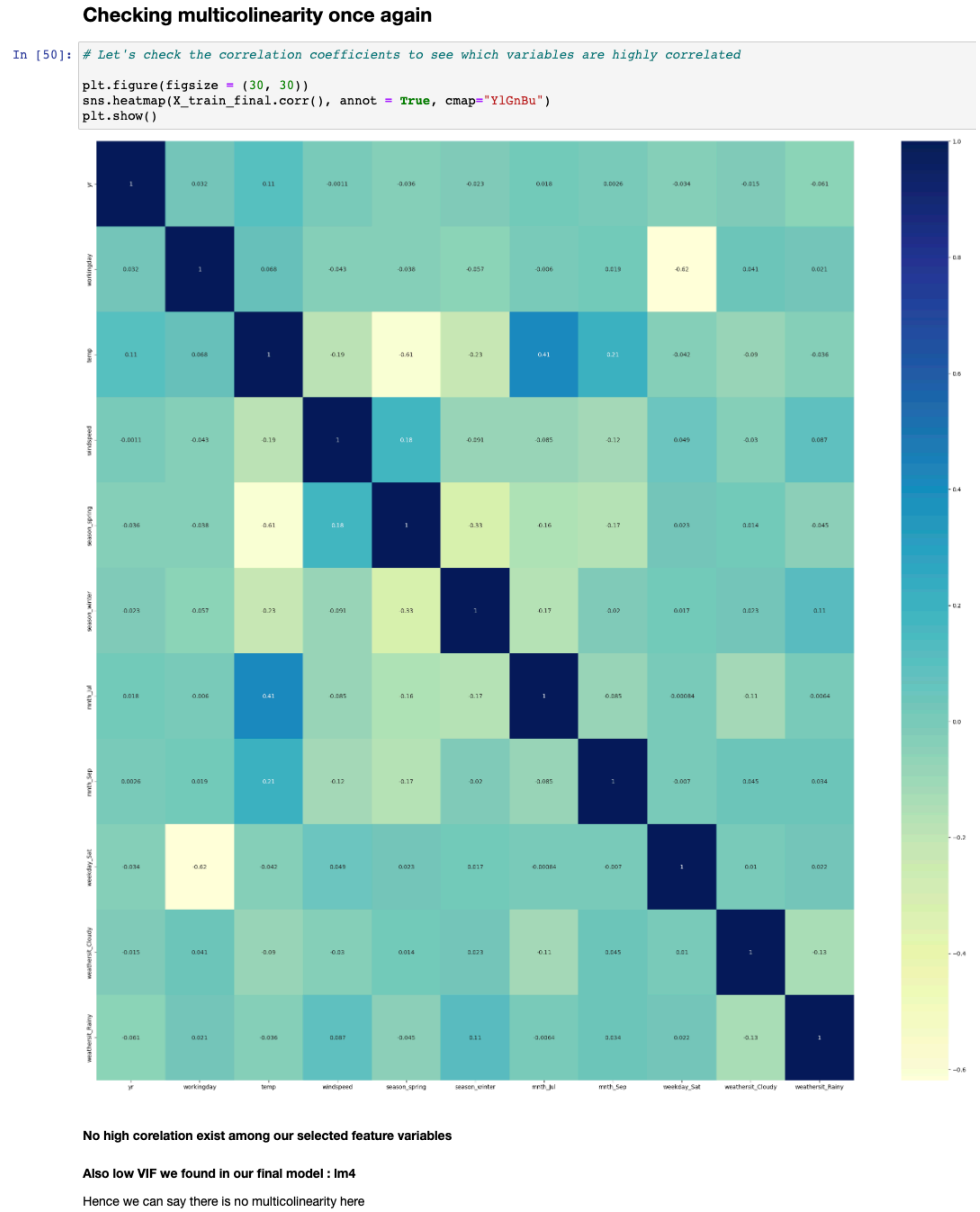


```
import matplotlib.pyplot as plt

fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2)
ax1.scatter(X['temp'], y_train)
ax2.scatter(X['windspeed'], y_train)
ax3.scatter(X['workingday'], y_train)
ax4.scatter(X['yr'], y_train)
plt.show()
```

Linear relationshp exist : There isn't any clear non-linear pattern and a linear model may work well on this.

B. No or little Multicollinearity was validated using heat map and we found that with final model lm4 there was no high correlation exist in selected features. While prior final model it was existing most among ‘temp’ and ‘temp’.



Checking multicolinearity once again

```
In [50]: # Let's check the correlation coefficients to see which variables are highly correlated

         plt.figure(figsize = (30, 30))
         sns.heatmap(X_train_final.corr(), annot = True, cmap="YlGnBu")
         plt.show()
```

	yr	workingday	temp	windspeed	season_spring	season_winter	mnth_Jul	mnth_Sep	weekday_Sat	weathersit_Cloudy	weathersit_Rainy
yr	1	0.032	0.11	-0.0011	-0.036	-0.023	0.018	0.0026	-0.034	-0.015	-0.061
workingday	0.032	1	0.068	-0.043	-0.038	-0.057	-0.006	0.019	-0.62	0.041	0.021
temp	0.11	0.068	1	-0.19	-0.61	-0.23	0.41	0.21	-0.042	-0.09	-0.036
windspeed	-0.0011	-0.043	-0.19	1	0.18	-0.091	-0.085	-0.12	0.049	-0.03	0.087
season_spring	-0.036	-0.038	-0.61	0.18	1	-0.33	-0.16	-0.17	0.023	0.014	-0.045
season_winter	-0.023	-0.057	-0.23	-0.091	-0.33	1	-0.17	-0.02	0.017	0.023	0.11
mnth_Jul	0.018	-0.006	0.41	-0.085	-0.16	-0.17	1	-0.085	0.00084	-0.11	-0.0064
mnth_Sep	0.0026	0.019	0.21	-0.12	-0.17	-0.02	-0.085	1	-0.007	0.045	0.034
weekday_Sat	-0.034	-0.62	-0.042	0.049	0.023	0.017	-0.00084	-0.007	1	0.01	0.022
weathersit_Cloudy	-0.015	0.041	-0.09	-0.03	0.014	0.023	-0.11	0.045	0.01	1	-0.13
weathersit_Rainy	-0.061	0.021	-0.036	0.087	-0.045	0.11	-0.0064	0.034	0.022	-0.13	1

No high corelation exist among our selected feature variables

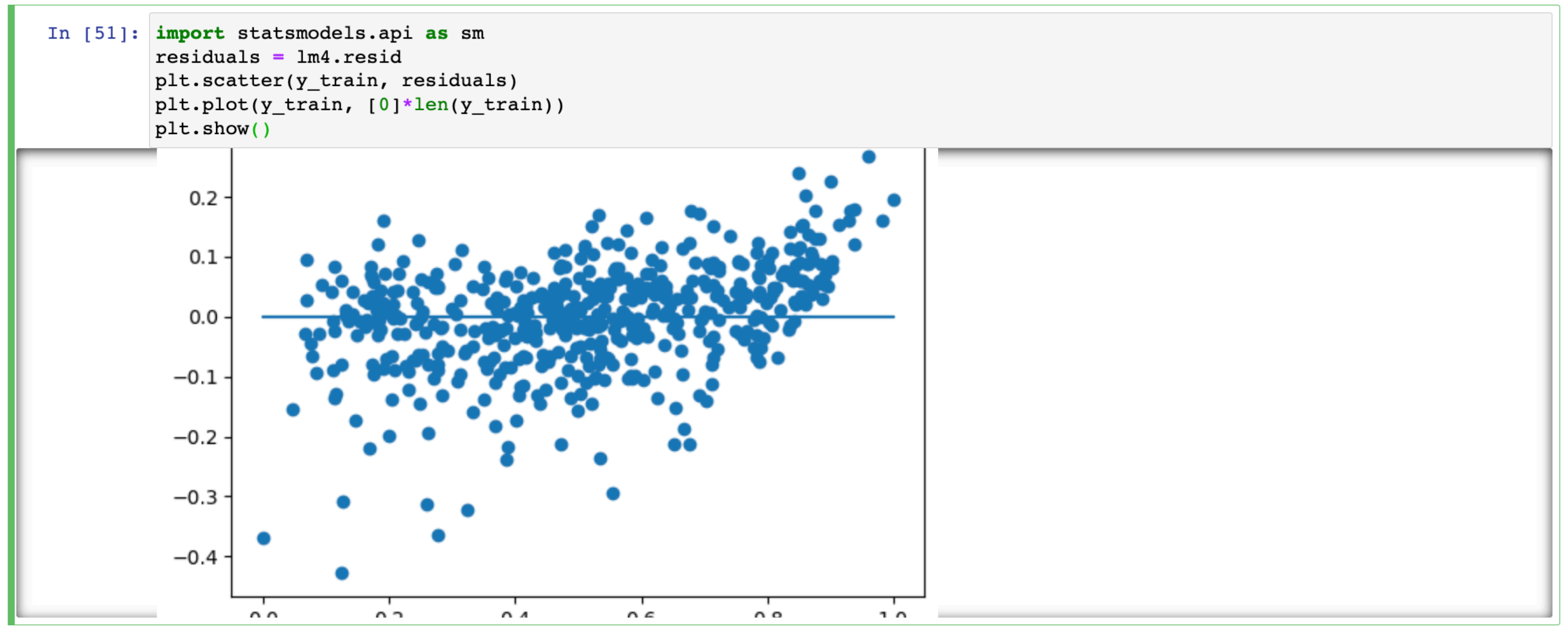
Also low VIF we found in our final model : lm4

Hence we can say there is no multicolinearity here

C. Homoscedasticity

It means that the error is constant along the values of the dependent variable.

Homoscedasticity is present as there is constant deviation of the points from the zero-line. This was done using stats model.api as follows:



```
In [51]: import statsmodels.api as sm
         residuals = lm4.resid
         plt.scatter(y_train, residuals)
         plt.plot(y_train, [0]*len(y_train))
         plt.show()
```

D. Residual autocorrelation

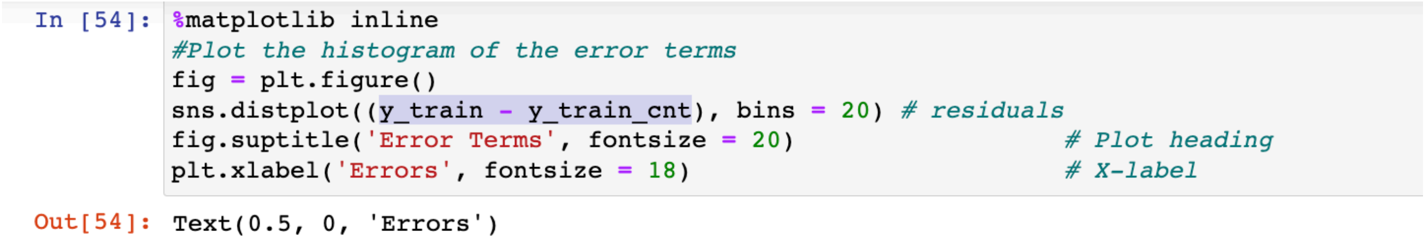
This was done using durbin_watson and its value is fine between 1.5-2.5 and for our model it was found 2.004. Hence its fine.

E. Normality of the residuals

To check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression), let us plot the histogram of the error terms and see what it looks like.

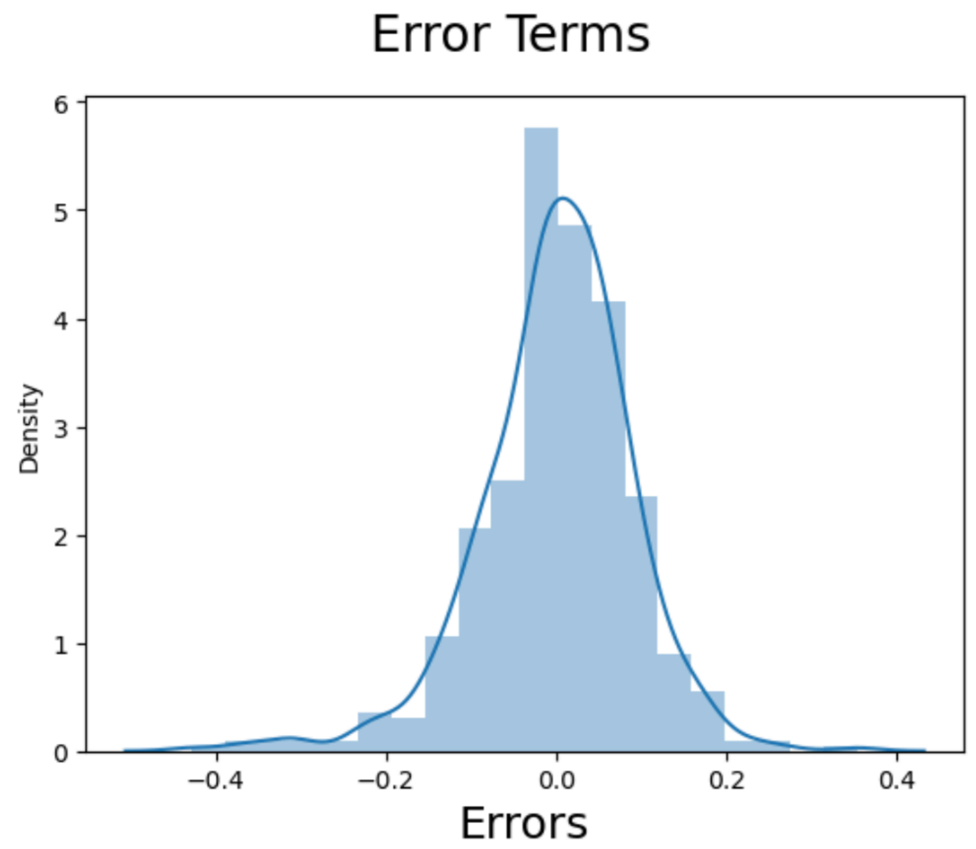
This was done using distplot of seaboard for error_term = y_train - y_train_cnt(or y predicted)

The graph was found normalised and hence it was fine.



```
In [54]: %matplotlib inline
         #Plot the histogram of the error terms
         fig = plt.figure()
         sns.distplot((y_train - y_train_cnt), bins = 20) # residuals
         fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
         plt.xlabel('Errors', fontsize = 18)                         # X-label

Out[54]: Text(0.5, 0, 'Errors')
```



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

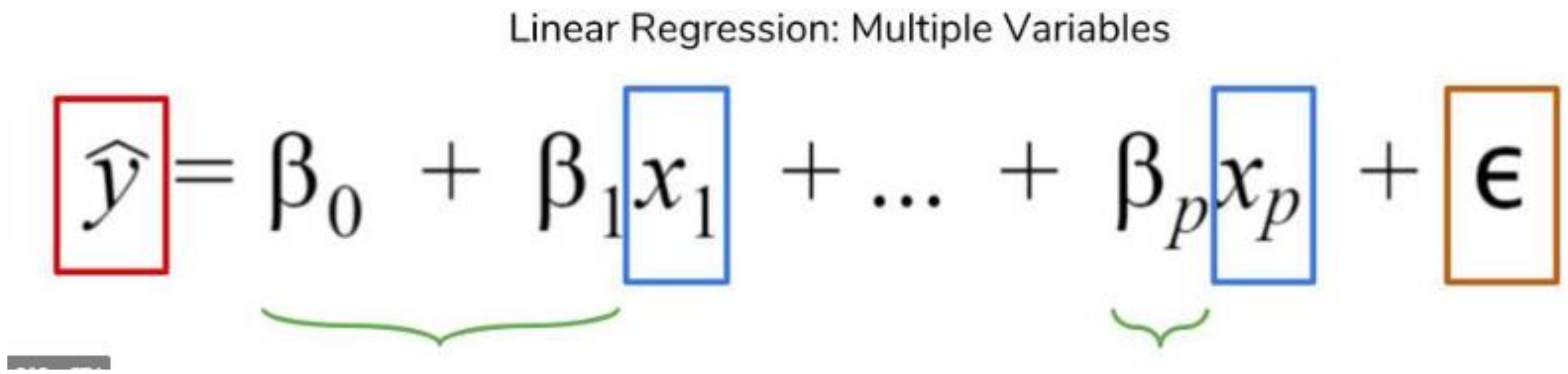
‘temp’ : regression coefficient = 0.4510 , this variable has an effect on bike demand where on its increase there will be increase in demand too
‘Yr’ : regression coefficient = 0.2344 , variable has an effect on bike demand where each passing year demand will increase
‘Weathersit_rainy’ : regression coefficient = -0.2904 , variable has an effect on bike demand where Rainy weather will decrease demand

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a type of supervised machine learning algorithm, it is used for predictive analysis. Linear regression makes predictions for continuous variables such as sales, salary, age, product price, etc.

It shows a linear relationship between dependent variable (y) and independent variables (x1,x2...xn) in the form of following equation where (β0,β1...βn) are coefficients/weights to depict the relationship:



Linear Regression: Multiple Variables

$$\boxed{\hat{y}} = \underbrace{\beta_0 + \beta_1}_{}\boxed{x_1} + \dots + \underbrace{\beta_p}_{}\boxed{x_p} + \boxed{\epsilon}$$

Linear regression can be further divided into two types of the algorithm: o Simple Linear Regression: one independent variable

o Multiple Linear regression multiple independent variables
The Algorithms finds the best fit line by using a cost function which is least for best fit line

which is given by R square:

o It is a measure of goodness of fit.
o It is a relative value.
o Its value varies from 0 to 1, where is 1 is best.

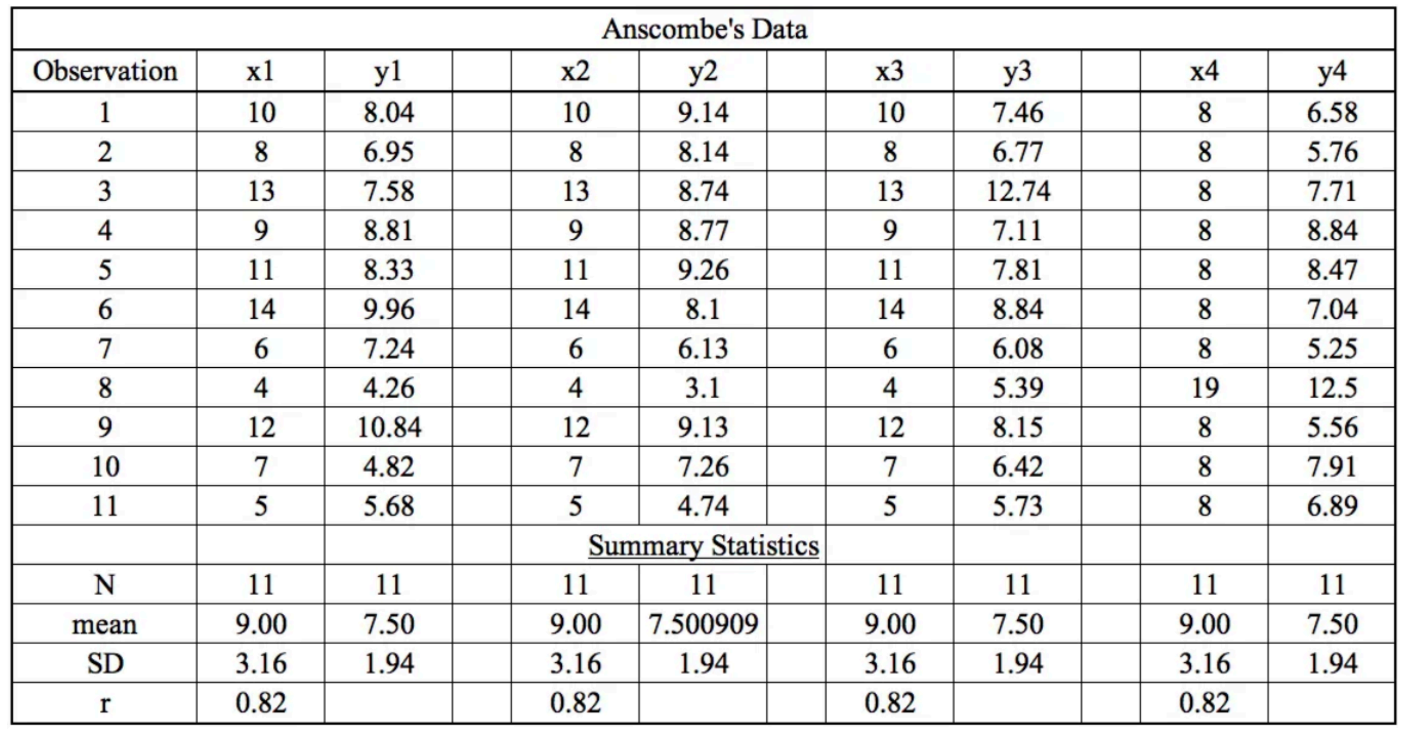
There are some assumptions related to linear regression given as follows:

o Linear relationship between the features and target
o Small or no multicollinearity between the features
o Homoscedasticity Assumption: Error terms should not follow any pattern. o Normal distribution of error term
o No autocorrelations

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analysing data.
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.

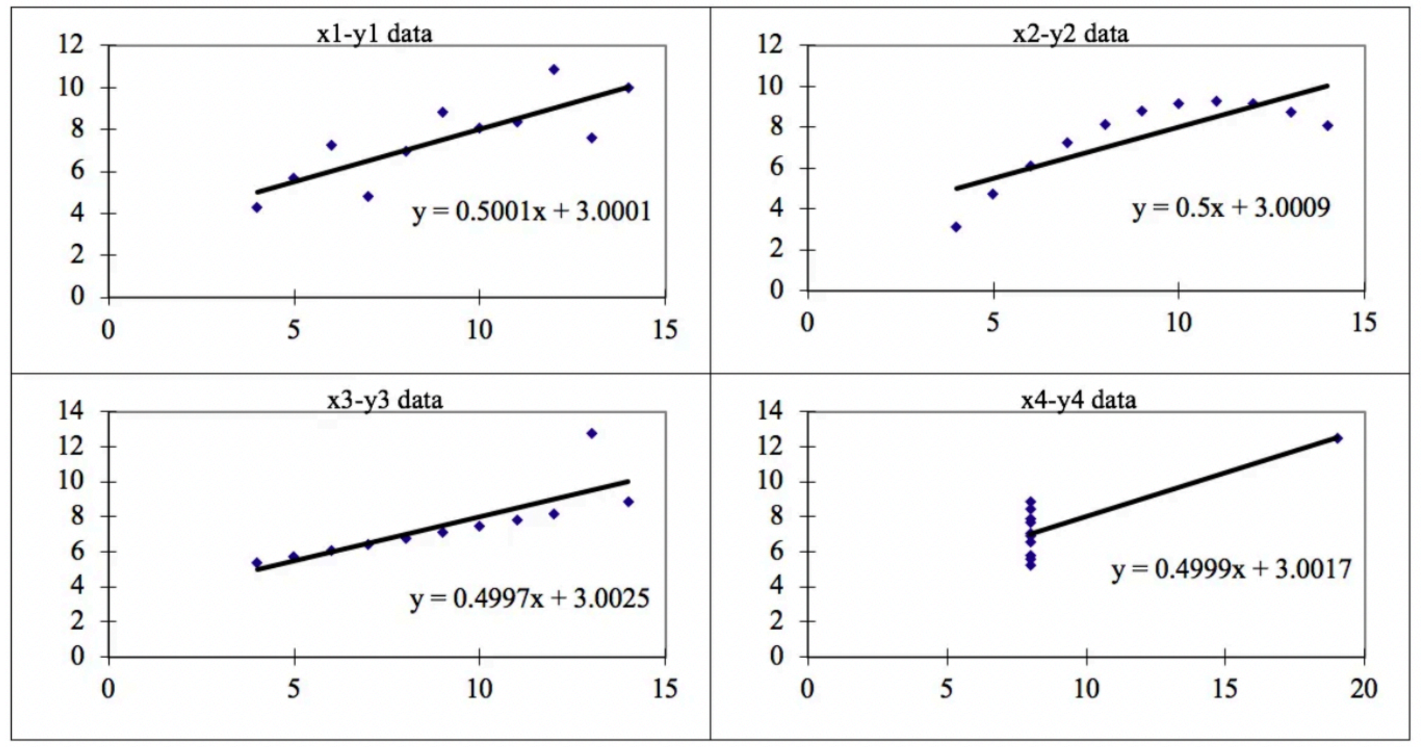
The statistical information for all these four datasets are approximately similar and can be computed as follows:



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:

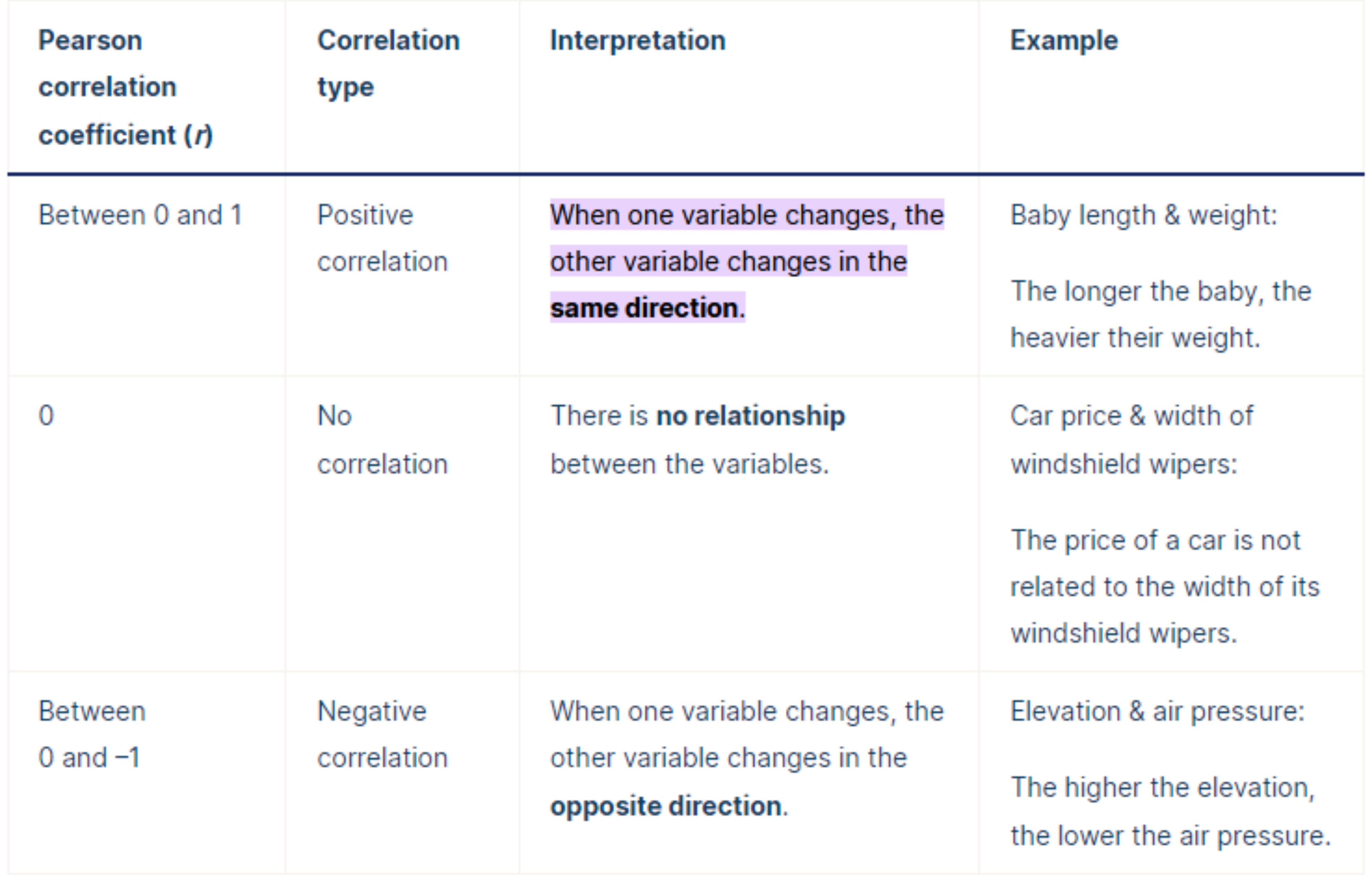


The four datasets can be described as:

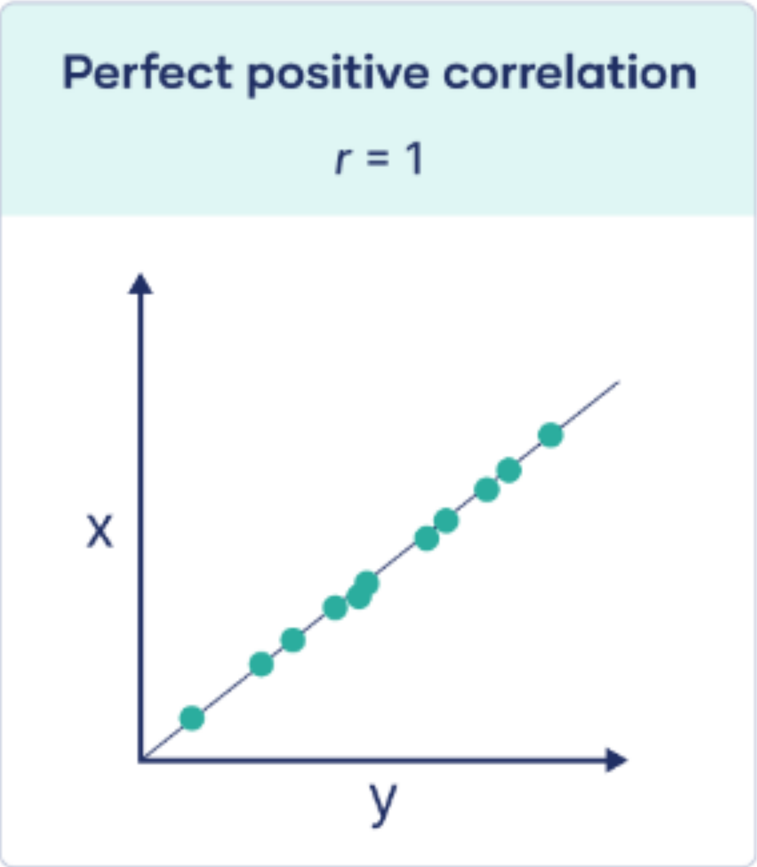
- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model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

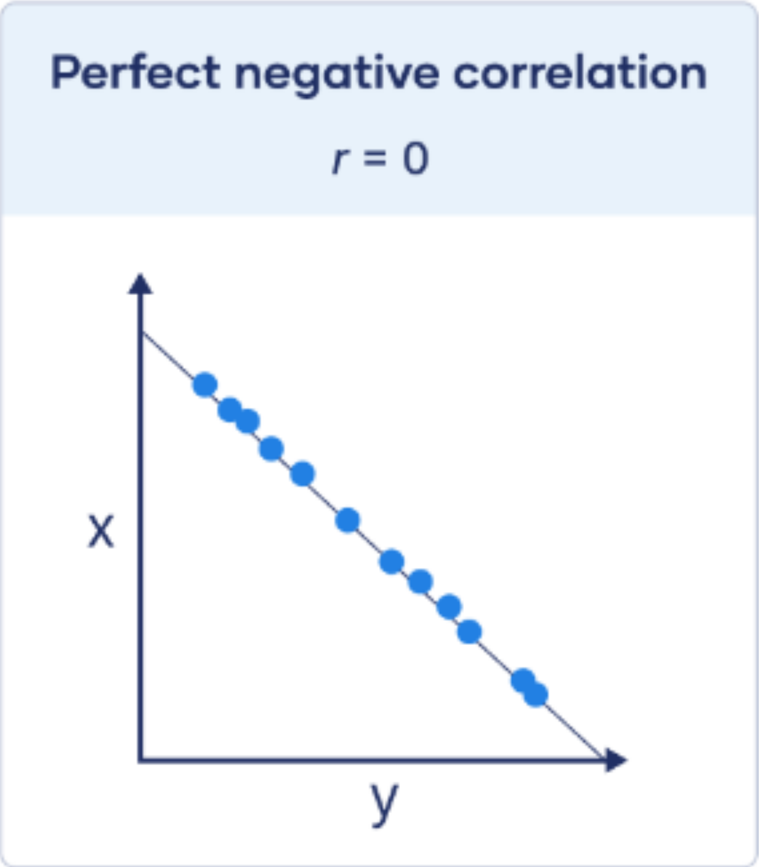
3. What is Pearson's R?

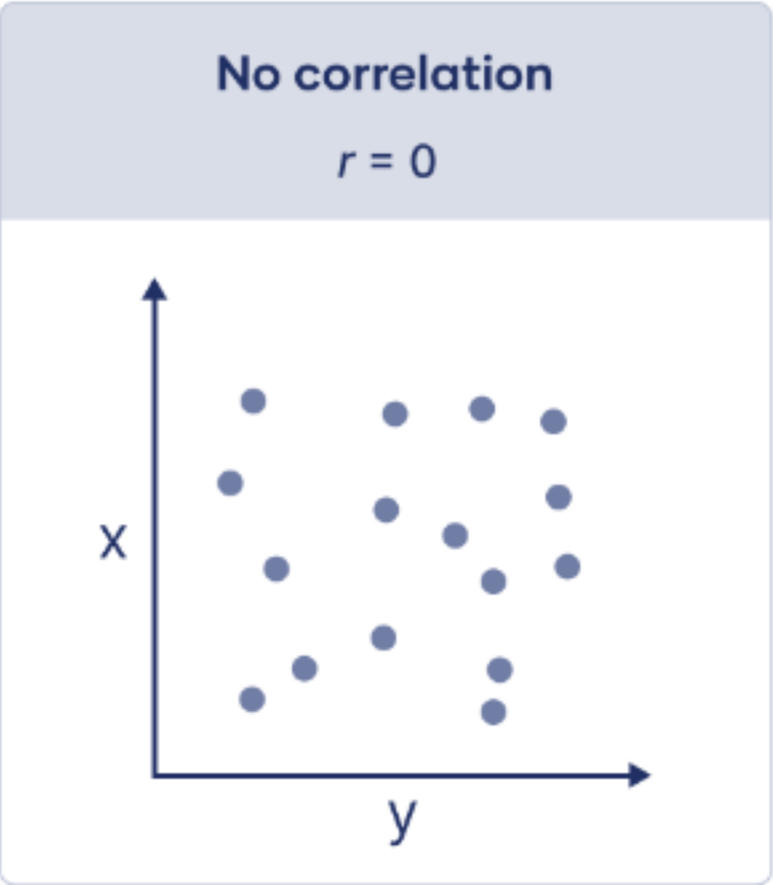
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

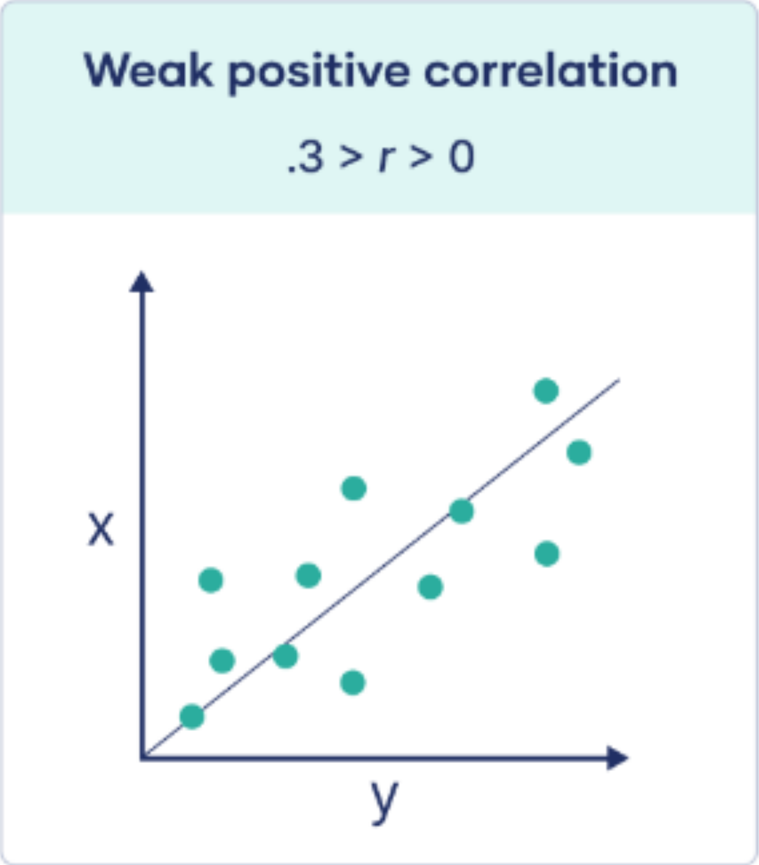


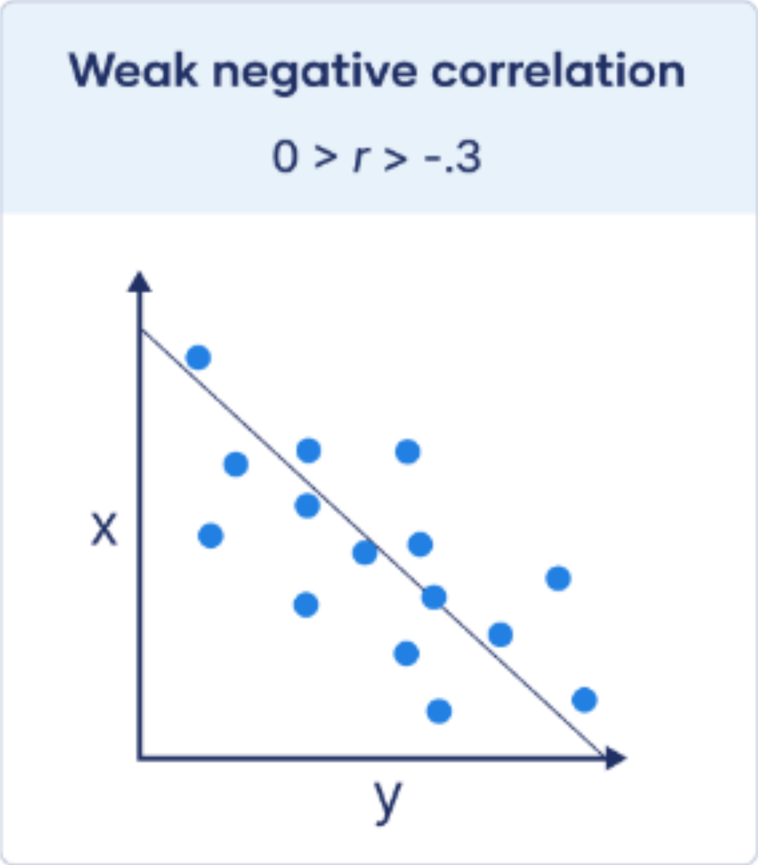
Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

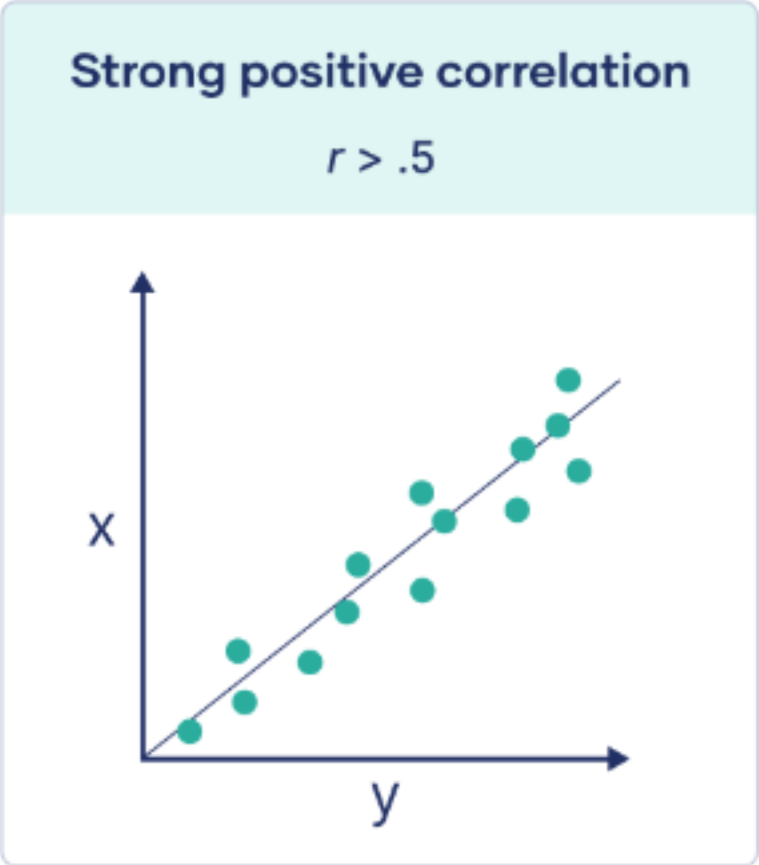


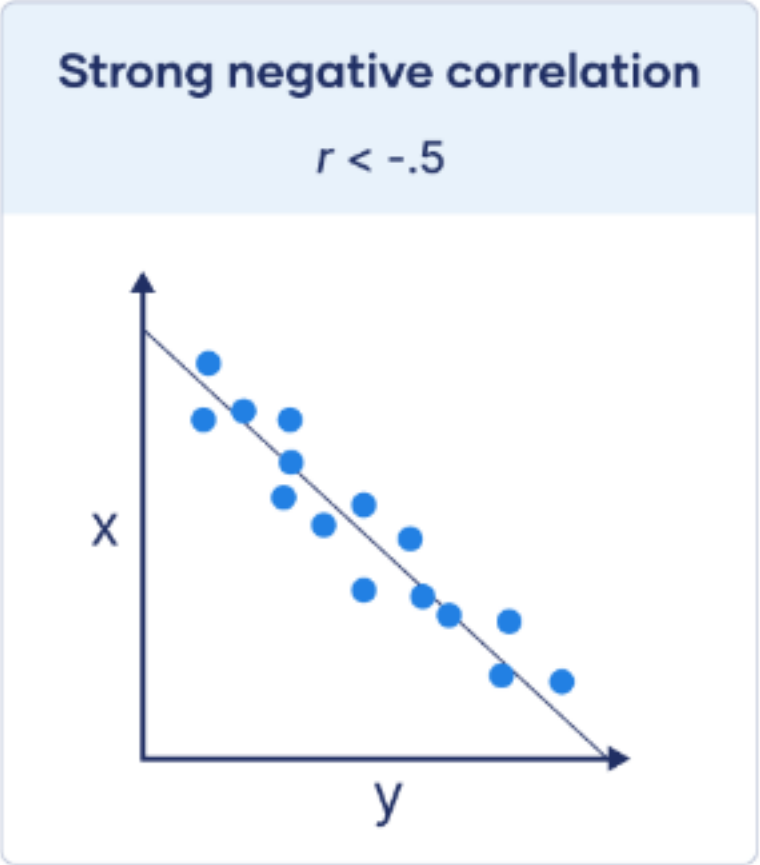












Assumptions with the Pearson's R are as follows:

o Both variables are quantitative
o The variables are normally distributed

o The data have no outliers
o The relationship is linear

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

Scaling also called Feature Scaling is a method to standardize the independent features present in the data in a fixed range. It is part of data pre-processing to handle highly varying features which can cause bias in the model towards large values regardless of units which lead to wrong predictions.

Scaling is used to reduce the columns range to similar range across all columns to prevent bias generation due to very high values.

Two most common Scaling method are as follows:

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

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

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

$$x = \frac{x - mean(x)}{sd(x)}$$

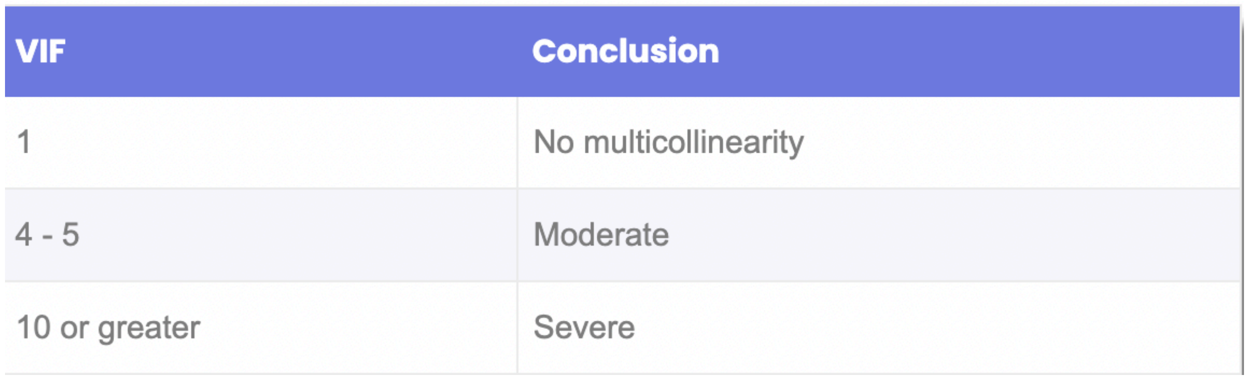
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

If there is perfect correlation, then VIF = infinity.

A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

In case of perfect correlation, R2=1 and we get VIF=1/(1-R2) which is Infinity, this means that variable can be explained by linear of other variables. This is perfect multicollinearity, which is resolved by removing one of the features causing this and iterating till VIF is lowered enough.



VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

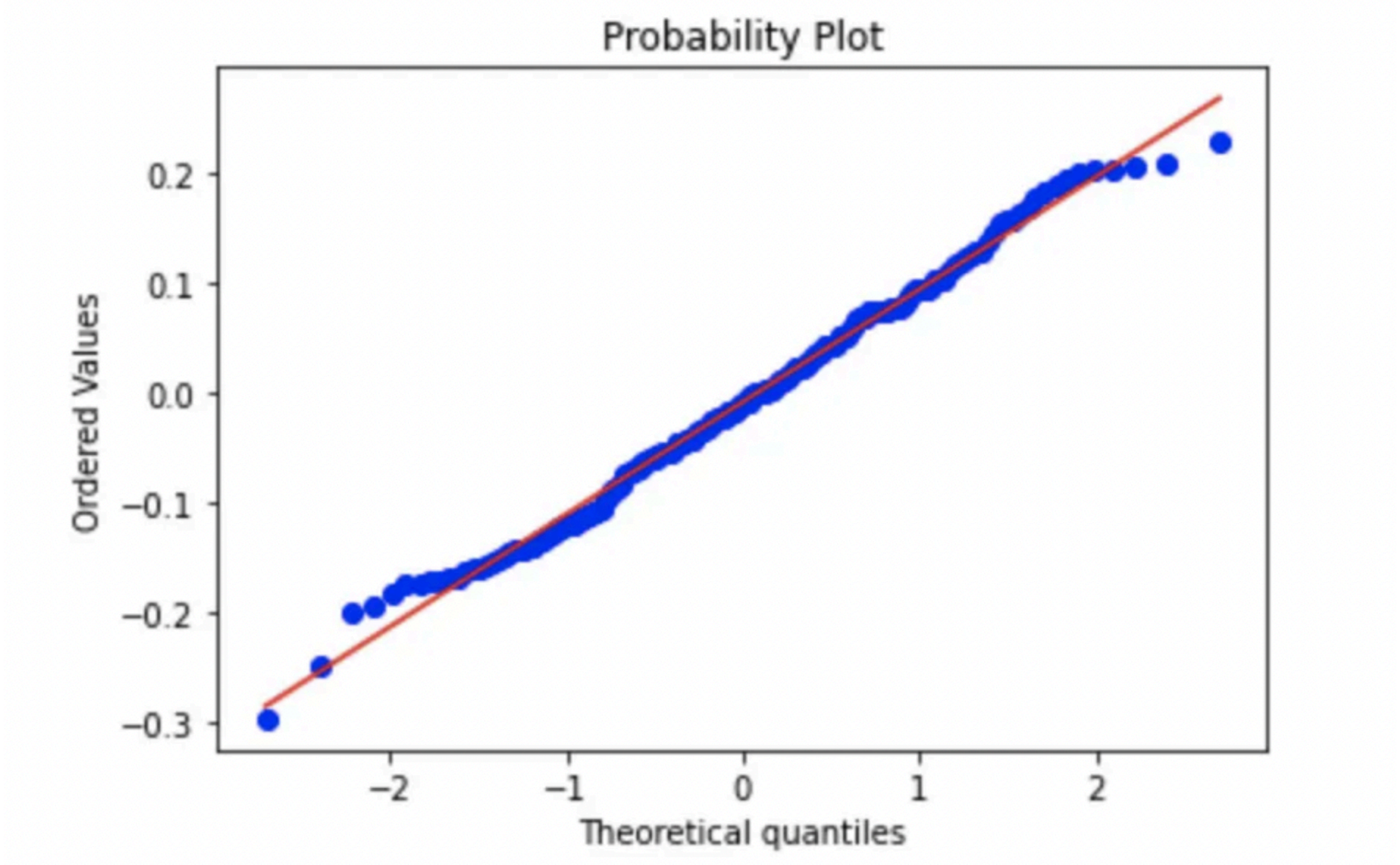
In summary, A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.

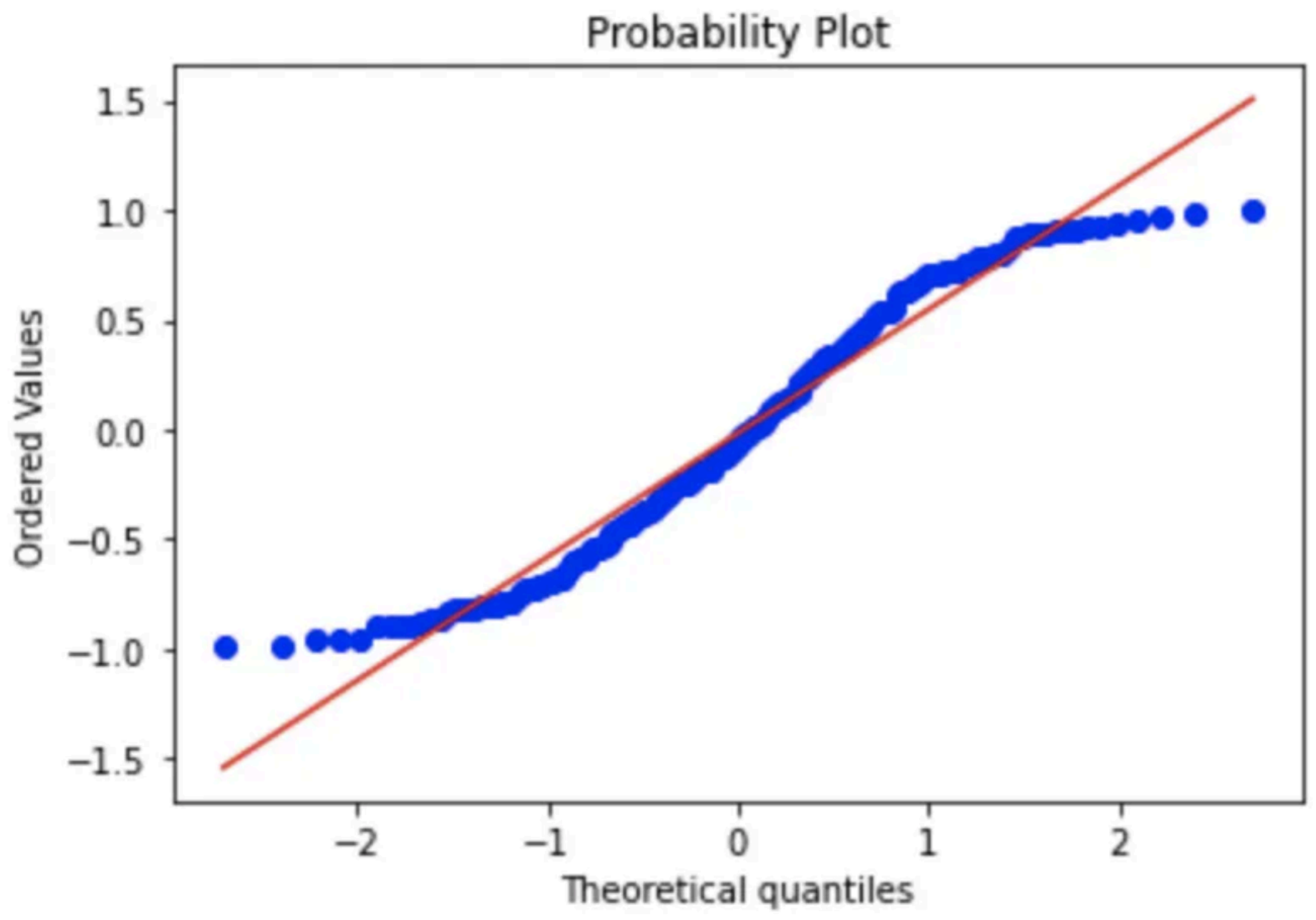
How is it generated?

Below are the steps to generate a Q-Q plot for team members age to test for normality

- Take your variable of interest (team member age in this scenario) and sort it from smallest to largest value. Let's say you have 19 team members in this scenario.
- Take a normal curve and divide it into 20 equal segments (n+1; where n=#data points)
- Compute z score for each of these points
- Plot the z-score obtained against the sorted variables. Usually, the z-scores are in the x-axis (also called theoretical quantiles since we are using this as a base for comparison) and the variable quantiles are in the y-axis (also called ordered values)

- Observe if data points align closely in a straight 45-degree line
- If it does, the age is normally distributed. If it is not, you might want to check it against other possible distributions


Probability Plot
Ordered Values
Theoretical quantiles


Probability Plot
Ordered Values
Theoretical quantiles