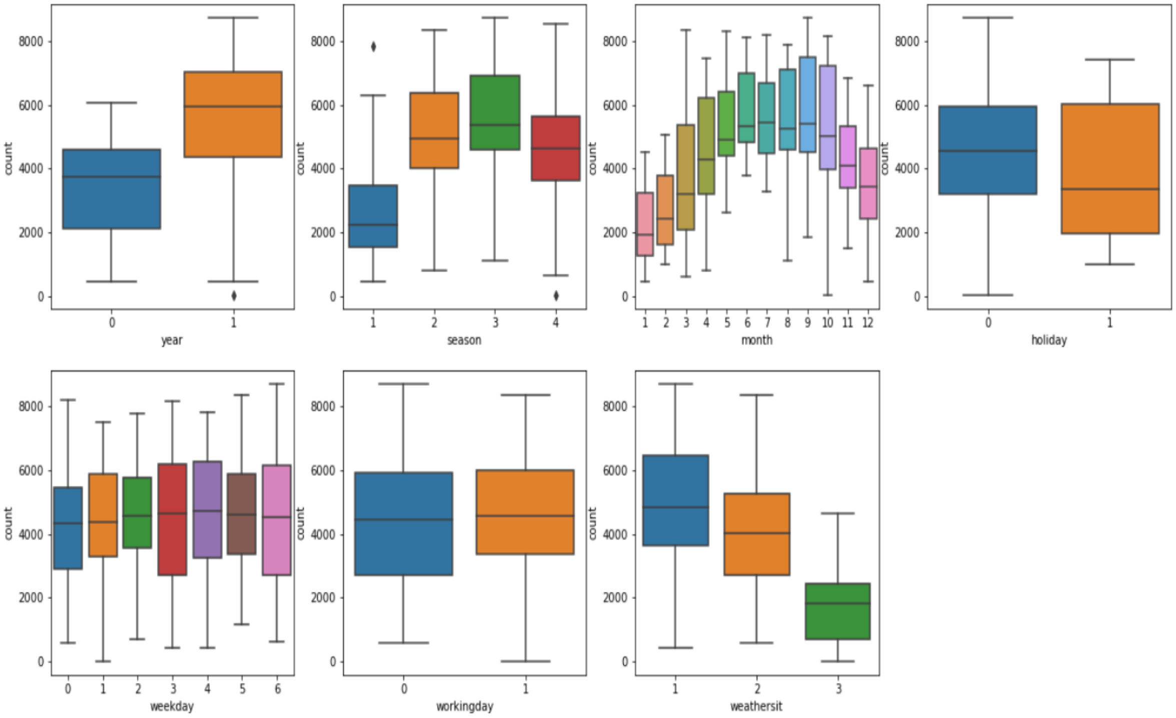
Assignment-based Subjective Questions

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. ‘year’, 'month', 'season', 'weekday', ‘holiday’, ‘workingday’, ‘weathersit' are the categorical variables in the dataset.



`Year and count`: We see that second has overall increase in the count

`Season and count`: Spring has low count whereas summer and fall has high count of users which is reasonable and also Winter has slightly less count.

`Month and count`: It follows similar distribution of season and count because the previous one is binned boxplot of the latter.

`Weekday and count`: Weekday median is almost similar however the range differs. The median is highest on day 3.

`Workingday and count` and `Holiday and count` are inverse to each other.

`Weathersit and Count`: count is highest in Clear, few clouds, Partly cloudy, Partly cloudy days and followed by Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist days and at last Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds days

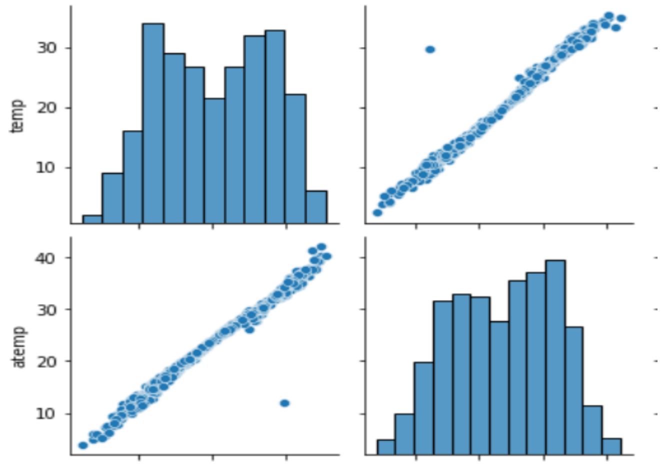
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans. Any variable with ‘n’ levels can be represented with ‘n-1’ dummy variables. If we know that there are 7 known levels and if we label 6 with dummy variables then the remaining one has to be the 7th one and we don’t need it to be encoded into dummy variable.

Coming to the importance, it helps in reducing that extra column created while creating dummy variables and therefore results in reduction of the correlations among dummy variables.

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

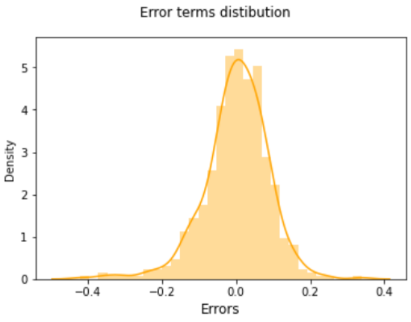
Ans. atemp and temp has the highest correlation among the numerical variables as we see in pairplot.



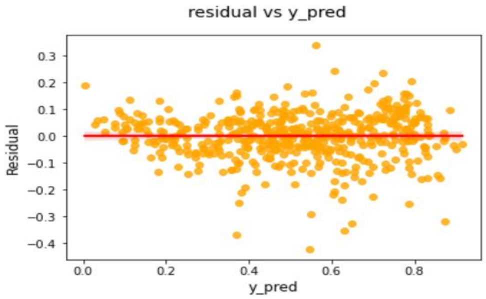
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. I have validated the model in residual analysis section.

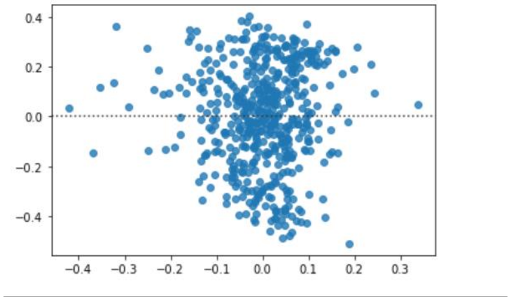
- Checking if error terms follow normal distribution



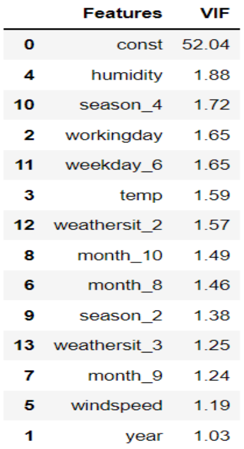
- Checking for multicollinearity.



- Check to confirm there is no pattern in residuals



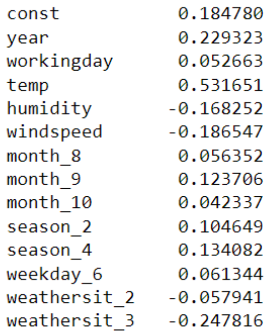
- Checking to confirm if there is multicollinearity.



	Features	VIF
0	const	52.04
4	humidity	1.88
10	season_4	1.72
2	workingday	1.65
11	weekday_6	1.65
3	temp	1.59
12	weathersit_2	1.57
8	month_10	1.49
6	month_8	1.46
9	season_2	1.38
13	weathersit_3	1.25
7	month_9	1.24
5	windspeed	1.19
1	year	1.03

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.
```
temp             0.531651
weathersit_3    -0.247816
year             0.229323
```
These are top 3 features contributing greatly to the demand of shared bikes.



```
const           0.184780
year            0.229323
workingday      0.052663
temp            0.531651
humidity       -0.168252
windspeed      -0.186547
month_8         0.056352
month_9         0.123706
month_10        0.042337
season_2        0.104649
season_4        0.134082
weekday_6       0.061344
weathersit_2   -0.057941
weathersit_3   -0.247816
```

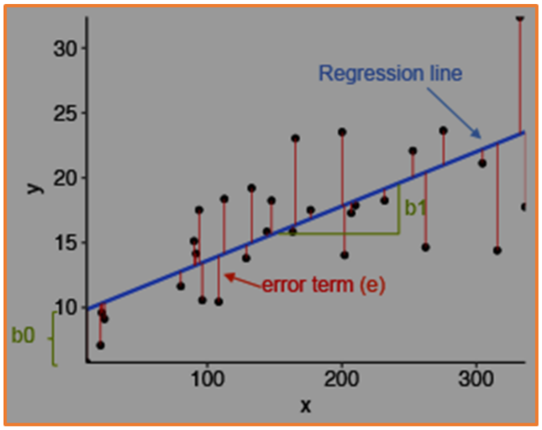
General Subjective Questions

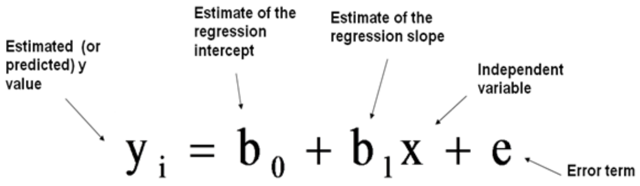
1. Explain the linear regression algorithm in detail. (4 marks)

Ans. A Linear regression represents a linear relationship between one or more independent variable x with single target variable x. It is ML Algorithm based on supervised learning.

Hypothesis of univariate regression function: $h\theta(x)=\theta_0+\theta_1x$

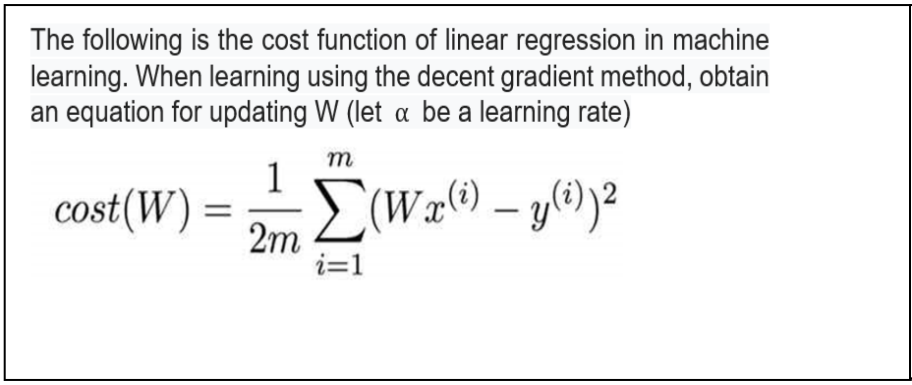
Hypothesis of Multivariate regression function: $h\theta(x)=\theta_0+\theta_1x_{1+}\ \theta_2x_{2+}\ \theta_3x_3+\dots\dots\dots\ +\theta_nx_n$




Estimated (or predicted) y value
Estimate of the regression intercept
Estimate of the regression slope
Independent variable
$y_i = b_0 + b_1x + e$
Error term

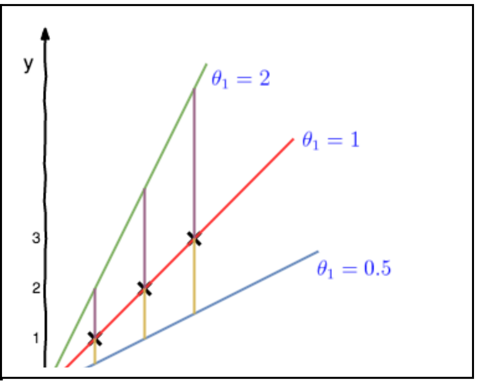
We fit the best line while updating parameter of the X, we do it by training the model with the train data with both y values and X values. After building model based on various controls and metrics, we come to conclude the best possible **Cost Function**.

Cost Function:



The following is the cost function of linear regression in machine learning. When learning using the decent gradient method, obtain an equation for updating W (let α be a learning rate)

$$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$



Our objective here will be to minimize the cost function.

Assumptions of linear regression:

- There is linear relation between X & y.
- Error terms are normally distributed.
- Error terms are independent to each other.
- Error terms have constant variance, Homoscedasticity.

In multiple linear regression we also consider the following aspects:

- Overfitting, model should not memorize the train data set.
- Multicollinearity check with Variance inflation factor.
- Feature selection from the pool of features/variables

After we train the data with either **scikit library** or **statsmodel** or **RFE,** we check the assumptions and proceed to evaluate the model**.**

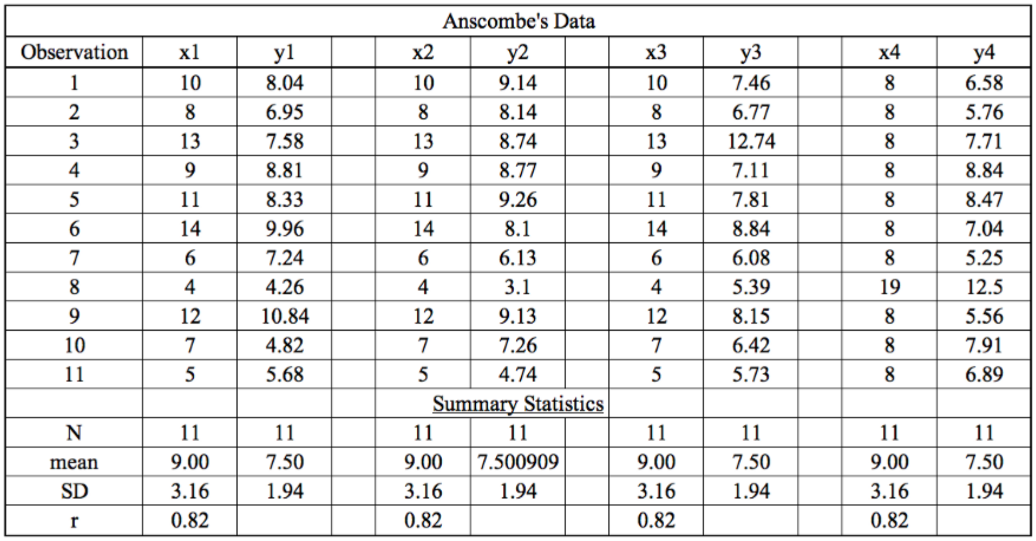
Metrics for evaluating best model:

- Adjusted R^2
- AIC-Akaike Info criterion
- BIC

The basic idea of these metrics is to explain most of the variance of the data with as few variables as possible, so we penalize the data for keeping large no. of variables.

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

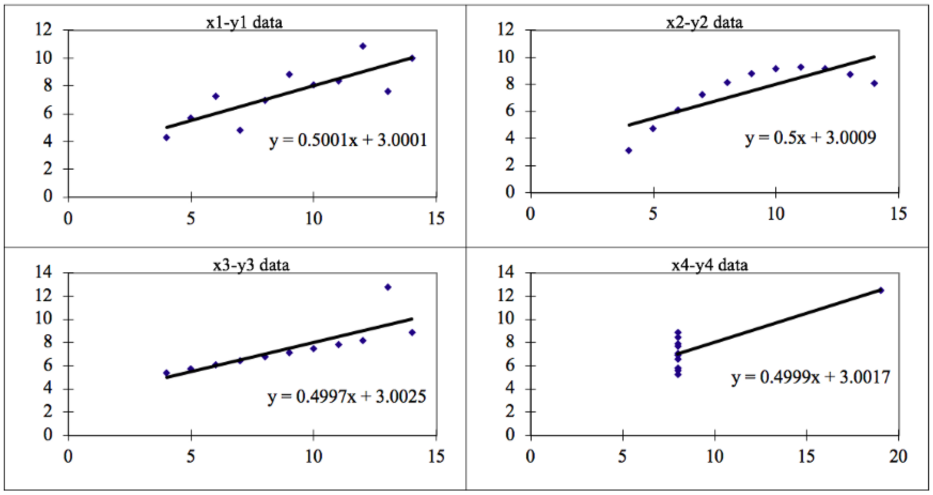
Ans. This is illustrated by statistician **Francis Anscombe** in 1973, The idea of the Anscombe's quartet is to tell the importance of plotting the graphs before analysing, modelling. Anscombe's Quartet can be defined as a group of four data sets that almost identical in simple descriptive statistics wise but there are some anomalies in the dataset that fools the regression model if built. Example:



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	

Here though the data sets are completely different, statistics show same numbers for all the four datasets, which can be pretty misleading. Lets plot the datasets and see the difference.



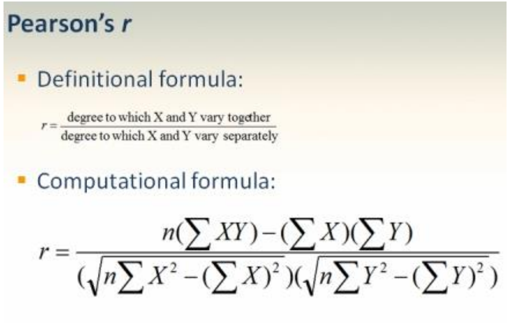
Now we understand that the datasets are different and also plotting them is very crucial before analysis and modelling.

3. What is Pearson’s R? (3 marks)

Ans. The Pearson correlation method is the most common method used for finding the correlation between the numerical variables, it ranges between − 1 and 1.

- 1 indicates complete positive correlation
- 0 indicating zero correlation
- -1 indication complete negative correlation.

A positive correlation means that if 1 variable goes up, then 2nd will also go up, whereas if the correlation is negative, then if 1 increase, 2 decreases. Zero correlation on the other hand means the change in one variable will not have any impact on the second variable.



Pearson’s *r*

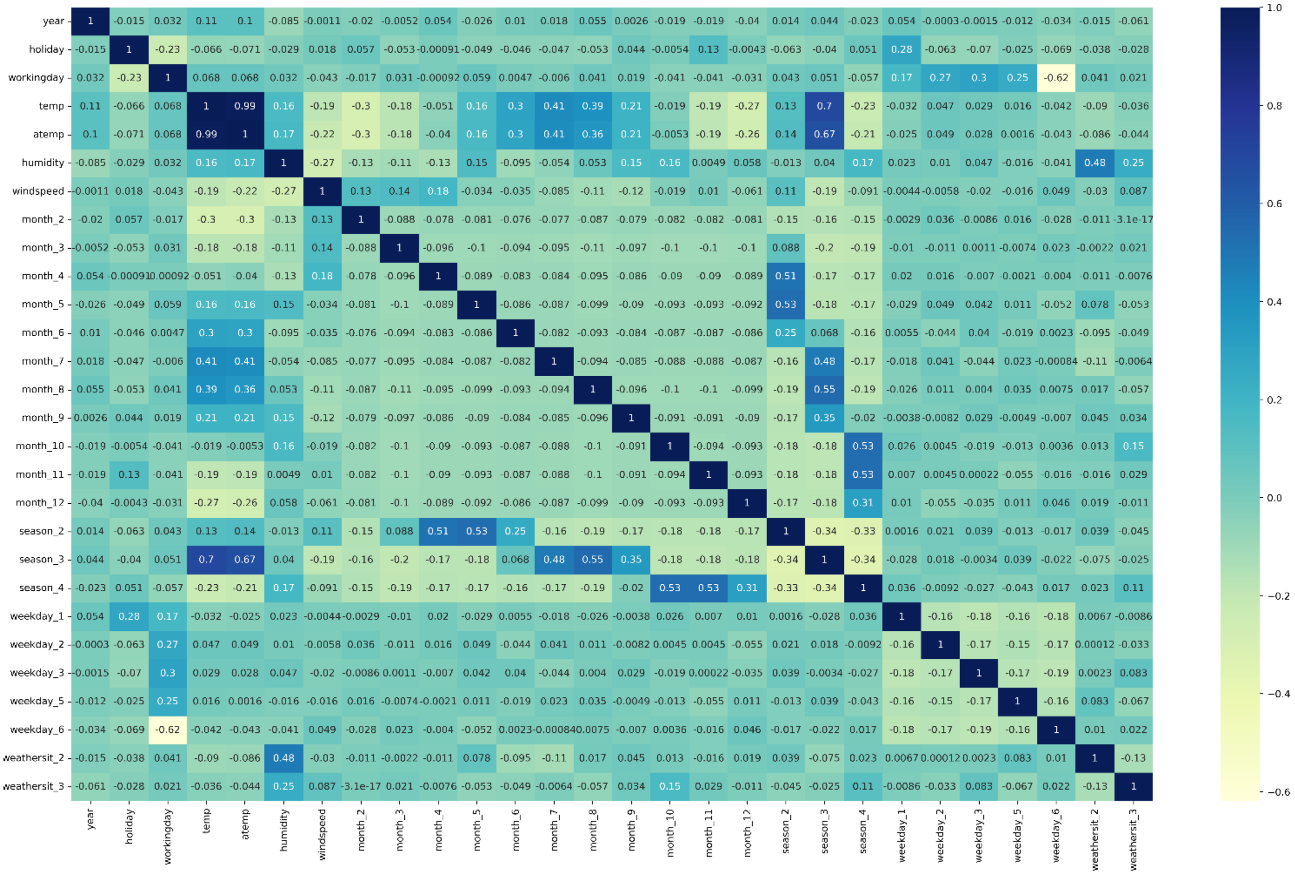
- Definitional formula:

$$r = \frac{\text{degree to which X and Y vary together}}{\text{degree to which X and Y vary separately}}$$

- Computational formula:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{(\sqrt{n\sum X^2 - (\sum X)^2})(\sqrt{n\sum Y^2 - (\sum Y)^2})}$$

Let’s see the correlation matrix or heatmap we plotted in our assignment:



	year	holiday	workingday	temp	atemp	humidity	windspeed	month_2	month_3	month_4	month_5	month_6	month_7	month_8	month_9	month_10	month_11	month_12	season_2	season_3	season_4	weekday_1	weekday_2	weekday_3	weekday_5	weekday_6	weathersit_2	weathersit_3
year	1	-0.015	0.032	0.11	0.1	-0.085	-0.0011	-0.02	-0.0052	0.054	-0.026	0.01	0.018	0.055	0.0026	-0.019	-0.019	-0.04	0.014	0.044	-0.023	0.054	-0.0003	-0.0015	-0.012	-0.034	-0.015	-0.061
holiday	-0.015	1	-0.23	-0.066	-0.071	-0.029	0.018	0.057	-0.053	-0.00091	-0.049	-0.046	-0.047	-0.053	0.044	-0.0054	0.13	-0.0043	-0.063	-0.04	0.051	0.28	-0.063	-0.07	-0.025	-0.069	-0.038	-0.028
workingday	0.032	-0.23	1	0.068	0.068	0.032	-0.043	-0.017	0.031	-0.00092	0.059	0.0047	-0.006	0.041	0.019	-0.041	-0.041	-0.031	0.043	0.051	-0.057	0.17	0.27	0.3	0.25	-0.62	0.041	0.021
temp	0.11	-0.066	0.068	1	0.99	0.16	-0.19	-0.3	-0.18	-0.051	0.16	0.3	0.41	0.39	0.21	-0.019	-0.19	-0.27	0.13	0.7	-0.23	-0.032	0.047	0.029	0.016	-0.042	-0.09	-0.036
atemp	0.1	-0.071	0.068	0.99	1	0.17	-0.22	-0.3	-0.18	-0.04	0.16	0.3	0.41	0.36	0.21	-0.0053	-0.19	-0.26	0.14	0.67	-0.21	-0.025	0.049	0.028	0.0016	-0.043	-0.086	-0.044
humidity	-0.085	-0.029	0.032	0.16	0.17	1	-0.27	-0.13	-0.11	-0.13	0.15	-0.095	-0.054	0.053	0.15	0.16	0.0049	0.058	-0.013	0.04	0.17	0.023	0.01	0.047	-0.016	-0.041	0.48	0.25
windspeed	-0.0011	0.018	-0.043	-0.19	-0.22	-0.27	1	0.13	0.14	0.18	-0.034	-0.035	-0.085	-0.11	-0.12	-0.019	0.01	-0.061	0.11	-0.19	-0.091	-0.0044	-0.0058	-0.02	-0.016	0.049	-0.03	0.087
month_2	-0.02	0.057	-0.017	-0.3	-0.3	-0.13	0.13	1	-0.088	-0.078	-0.081	-0.076	-0.077	-0.087	-0.079	-0.082	-0.082	-0.081	-0.15	-0.16	-0.15	-0.0029	0.036	-0.0086	0.016	-0.028	-0.011	-3.1e-17
month_3	-0.0052	-0.053	0.031	-0.18	-0.18	-0.11	0.14	-0.088	1	-0.096	-0.1	-0.094	-0.095	-0.11	-0.097	-0.1	-0.1	-0.1	0.088	-0.2	-0.19	-0.01	-0.011	0.0011	-0.0074	0.023	-0.0022	0.021
month_4	0.054	-0.00091	0.00092	-0.051	-0.04	-0.13	0.18	-0.078	-0.096	1	-0.089	-0.083	-0.084	-0.095	-0.086	-0.09	-0.09	-0.089	0.51	-0.17	-0.17	0.02	0.016	-0.007	-0.0021	-0.004	-0.011	-0.0076
month_5	-0.026	-0.049	0.059	0.16	0.16	0.15	-0.034	-0.081	-0.1	-0.089	1	-0.086	-0.087	-0.099	-0.09	-0.093	-0.093	-0.092	0.53	-0.18	-0.17	-0.029	0.049	0.042	0.011	-0.052	0.078	-0.053
month_6	0.01	-0.046	0.0047	0.3	0.3	-0.095	-0.035	-0.076	-0.094	-0.083	-0.086	1	-0.082	-0.093	-0.084	-0.087	-0.087	-0.086	0.25	0.068	-0.16	0.0055	-0.044	0.04	-0.019	0.0023	-0.095	-0.049
month_7	0.018	-0.047	-0.006	0.41	0.41	-0.054	-0.085	-0.077	-0.095	-0.084	-0.087	-0.082	1	-0.094	-0.085	-0.088	-0.088	-0.087	-0.16	0.48	-0.17	-0.018	0.041	-0.044	0.023	-0.00084	-0.11	-0.0064
month_8	0.055	-0.053	0.041	0.39	0.36	0.053	-0.11	-0.087	-0.11	-0.095	-0.099	-0.093	-0.094	1	-0.096	-0.1	-0.1	-0.099	-0.19	0.55	-0.19	-0.026	0.011	0.004	0.035	0.0075	0.017	-0.057
month_9	0.0026	0.044	0.019	0.21	0.21	0.15	-0.12	-0.079	-0.097	-0.086	-0.09	-0.084	-0.085	-0.096	1	-0.091	-0.091	-0.09	-0.17	0.35	-0.02	-0.0038	-0.0082	0.029	-0.0049	-0.007	0.045	0.034
month_10	-0.019	-0.0054	-0.041	-0.019	-0.0053	0.16	-0.019	-0.082	-0.1	-0.09	-0.093	-0.087	-0.088	-0.1	-0.091	1	-0.094	-0.093	-0.18	-0.18	0.53	0.026	0.0045	-0.019	-0.013	0.0036	0.013	0.15
month_11	-0.019	0.13	-0.041	-0.19	-0.19	0.0049	0.01	-0.082	-0.1	-0.09	-0.093	-0.087	-0.088	-0.1	-0.091	-0.094	1	-0.093	-0.18	-0.18	0.53	0.007	0.0045	0.00022	-0.055	-0.016	-0.016	0.029
month_12	-0.04	-0.0043	-0.031	-0.27	-0.26	0.058	-0.061	-0.081	-0.1	-0.089	-0.092	-0.086	-0.087	-0.099	-0.09	-0.093	-0.093	1	-0.17	-0.18	0.31	0.01	-0.055	-0.035	0.011	0.046	0.019	-0.011
season_2	0.014	-0.063	0.043	0.13	0.14	-0.013	0.11	-0.15	0.088	0.51	0.53	0.25	-0.16	-0.19	-0.17	-0.18	-0.18	-0.17	1	-0.34	-0.33	0.0016	0.021	0.039	-0.013	-0.017	0.039	-0.045
season_3	0.044	-0.04	0.051	0.7	0.67	0.04	-0.19	-0.16	-0.2	-0.17	-0.18	0.068	0.48	0.55	0.35	-0.18	-0.18	-0.18	-0.34	1	-0.34	-0.028	0.018	-0.0034	0.039	-0.022	-0.075	-0.025
season_4	-0.023	0.051	-0.057	-0.23	-0.21	0.17	-0.091	-0.15	-0.19	-0.17	-0.17	-0.16	-0.17	-0.19	-0.02	0.53	0.53	0.31	-0.33	-0.34	1	0.036	-0.0092	-0.027	-0.043	0.017	0.023	0.11
weekday_1	0.054	0.28	0.17	-0.032	-0.025	0.023	-0.0044	-0.0029	-0.01	0.02	-0.029	0.0055	-0.018	-0.026	-0.0038	0.026	0.007	0.01	0.0016	-0.028	0.036	1	-0.16	-0.18	-0.16	-0.18	0.0067	-0.0086
weekday_2	-0.0003	-0.063	0.27	0.047	0.049	0.01	-0.0058	0.036	-0.011	0.016	0.049	-0.044	0.041	0.011	-0.0082	0.0045	0.0045	-0.055	0.021	0.018	-0.0092	-0.16	1	-0.17	-0.15	-0.17	0.00012	-0.033
weekday_3	-0.0015	-0.07	0.3	0.029	0.028	0.047	-0.02	-0.0086	0.0011	-0.007	0.042	0.04	-0.044	0.004	0.029	-0.019	0.00022	-0.035	0.039	-0.0034	-0.027	-0.18	-0.17	1	-0.17	-0.19	0.0023	0.083
weekday_5	-0.012	-0.025	0.25	0.016	0.0016	-0.016	-0.016	0.016	-0.0074	-0.0021	0.011	-0.019	0.023	0.035	-0.0049	-0.013	-0.055	0.011	-0.013	0.039	-0.043	-0.16	-0.15	-0.17	1	-0.16	0.083	-0.067
weekday_6	-0.034	-0.069	-0.62	-0.042	-0.043	-0.041	0.049	-0.028	0.023	-0.004	-0.052	0.0023	-0.00084	0.0075	-0.007	0.0036	-0.016	0.046	-0.017	-0.022	0.017	-0.18	-0.17	-0.19	-0.16	1	0.01	0.022
weathersit_2	-0.015	-0.038	0.041	-0.09	-0.086	0.48	-0.03	-0.011	-0.0022	-0.011	0.078	-0.095	-0.11	0.017	0.045	0.013	-0.016	0.019	0.039	-0.075	0.023	0.0067	0.00012	0.0023	0.083	0.01	1	-0.13
weathersit_3	-0.061	-0.028	0.021	-0.036	-0.044	0.25	0.087	-3.1e-17	0.021	-0.0076	-0.053	-0.049	-0.0064	-0.057	0.034	0.15	0.029	-0.011	-0.045	-0.025	0.11	-0.0086	-0.033	0.083	-0.067	0.022	-0.13	1

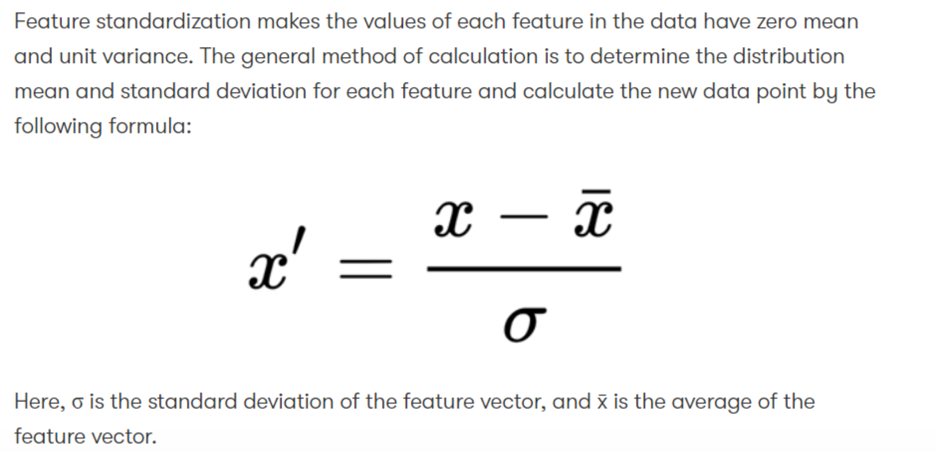
We see the Pearson correlations map of the all the variables which we created in the assignment.

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

Ans. Feature scaling or scaling is the process of converting the values of the variables into one same scale rather than keeping them in different scales which is not good for interpretation. In other words, feature scaling is a method used to normalize the range of independent variables or features of data.

There are two types of scaling methods used:

1. Standardisation:

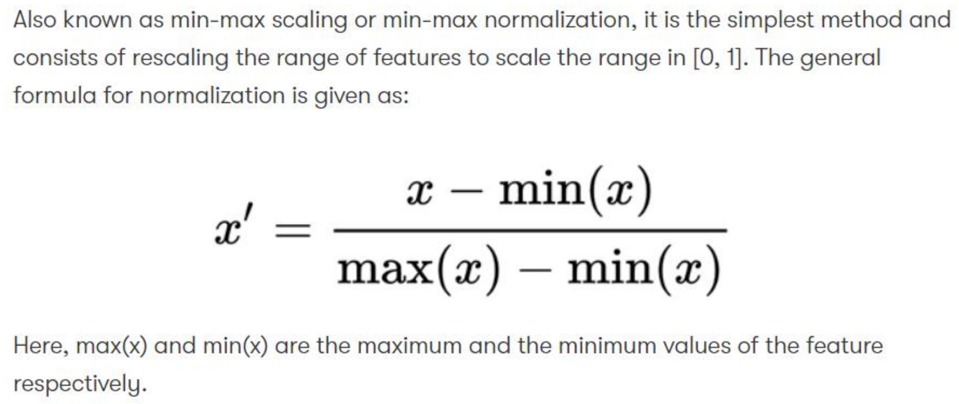


Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

2. Min-Max scaling/ Normalization:



Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

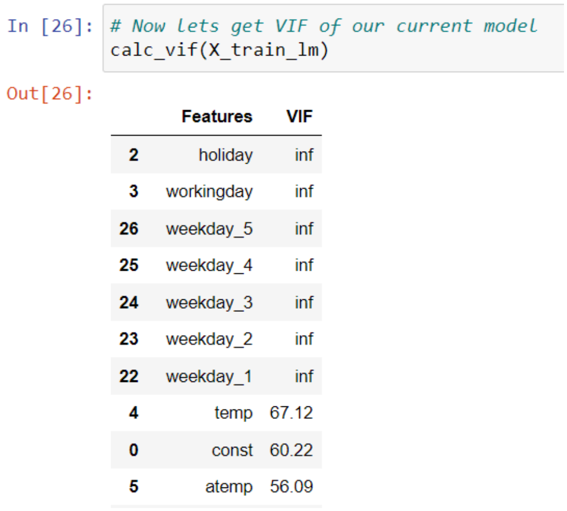
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

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

(3 marks)

Ans.



```
In [26]: # Now lets get VIF of our current model
         calc_vif(X_train_lm)
```

Out[26]:

	Features	VIF
2	holiday	inf
3	workingday	inf
26	weekday_5	inf
25	weekday_4	inf
24	weekday_3	inf
23	weekday_2	inf
22	weekday_1	inf
4	temp	67.12
0	const	60.22
5	atemp	56.09

Infinite VIF value means that the correlation between the two variables i.e., between the target variable and the independent variable is perfect. Correlation=1.

We Know that VIF = $1/(1-R^2)$, if R-Squared is 1, then VIF automatically becomes infinite.
To eradicate this situation, we need to remove the variables that are causing multicollinearity.

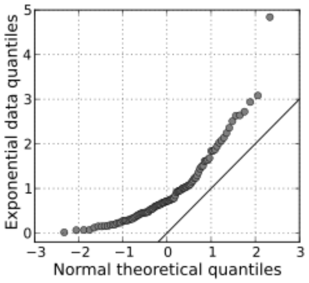
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. Q-Q plots or Quantile-Quantile plots are the quantiles of a sample distribution against theoretical distribution plots. A quantile is a fraction where certain values fall below that quantile. It helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

QQ plots are useful to check:

- Whether two populations are of the same distribution.
- Whether residuals follow a normal distribution. Having a normal error term is an assumption in regression and can be verified with this.
- Skewness of distribution

Ex: Median is a quantile where 50% of the data fall below a point and 50% lie above it. A 45-degree angle is plotted on the Q-Q plot the points will fall on that reference line, if they come from same data set.



References:

- https://machinelearningmedium.com/2017/08/11/cost-function-of-linear-regression/
- https://www.atoti.io/articles/when-to-perform-a-feature-scaling/
- https://www.analyticsvidhya.com/blog/2021/09/q-q-plot-ensure-your-ml-model-is-based-on-the-right-distributions/