How changing weather affects attendance bridges over the East River through cyclists in New York City?

Statistics, data visualization and prediction number of cyclists in relation to the weather conditions

Presentation of the problem

In the following presentation, I will analyze the change in the number of cyclists on the East River bridges in New York City in April 2016.

The bridges are located between Manhattan and Brooklyn. Traffic between these two important and populated areas is mainly carried out over the bridges included in the dataset on which the research will be conducted. These bridges are a sensitive place in this area due to the accumulated traffic and are an excellent source of information

The combination of weather conditions and the number of cyclists on the bridges is a real reflection of the relationship that arises between the weather conditions and the number of cyclists in the city of New York on the East River.

The aim of the research is to determine the relationship between the weather and the number of cyclists. I will use the Python programming language and its libraries for this. Then I will build a model that will be able to predict the number of cyclists under certain weather conditions.

Data presentation

	Unnamed: 0	Date	Day	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
0	0	2016-04-01 00:00:00	2016-04-01 00:00:00	78.1	66.0	0.01	1704.0	3126	4115.0	2552.0	11497
1	1	2016-04-02 00:00:00	2016-04-02 00:00:00	55.0	48.9	0.15	827.0	1646	2565.0	1884.0	6922
2	2	2016-04-03 00:00:00	2016-04-03 00:00:00	39.9	34.0	0.09	526.0	1232	1695.0	1306.0	4759
3	3	2016-04-04 00:00:00	2016-04-04 00:00:00	44.1	33.1	0.47 (S)	521.0	1067	1440.0	1307.0	4335
4	4	2016-04-05 00:00:00	2016-04-05 00:00:00	42.1	26.1	0	1416.0	2617	3081.0	2357.0	9471
...	...	...	...	...	...	...	...	...	...	...	...
205	205	2016-04-26 00:00:00	2016-04-26 00:00:00	60.1	46.9	0.24	1997.0	3520	4559.0	2929.0	13005
206	206	2016-04-27 00:00:00	2016-04-27 00:00:00	62.1	46.9	0	3343.0	5606	6577.0	4388.0	19914
207	207	2016-04-28 00:00:00	2016-04-28 00:00:00	57.9	48.0	0	2486.0	4152	5336.0	3657.0	15631
208	208	2016-04-29 00:00:00	2016-04-29 00:00:00	57.0	46.9	0.05	2375.0	4178	5053.0	3348.0	14954
209	209	2016-04-30 00:00:00	2016-04-30 00:00:00	64.0	48.0	0	3199.0	4952	5675.0	3606.0	17432

210 rows × 11 columns

```
Wejściowy zbiór danych zawiera : 210 rekordów oraz 11 kolumn
```

That's how the received dataset presents like

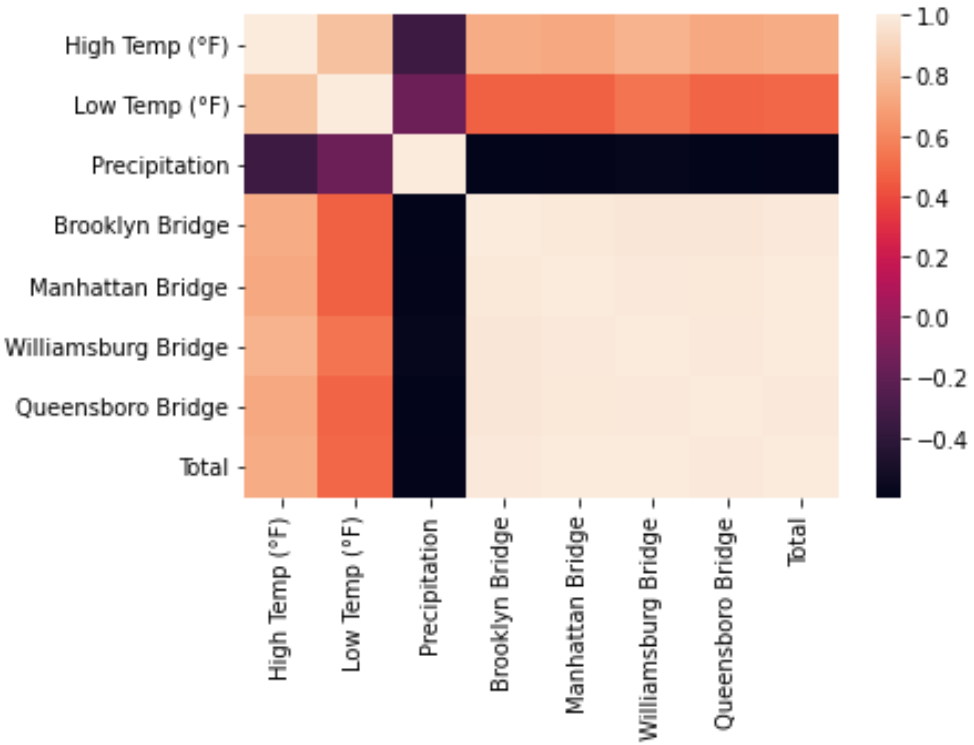
Data preprocessing

Date	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
2016-04-01	78.1	66.0	0.01	1704.0	3126.0	4115.0	2552.0	11497.0
2016-04-02	55.0	48.9	0.15	827.0	1646.0	2565.0	1884.0	6922.0
2016-04-03	39.9	34.0	0.09	526.0	1232.0	1695.0	1306.0	4759.0
2016-04-04	44.1	33.1	0.47	521.0	1067.0	1440.0	1307.0	4335.0
2016-04-05	42.1	26.1	0.00	1416.0	2617.0	3081.0	2357.0	9471.0

Zmodifikowany zbiór danych zawiera : 30 rekordów oraz 8 kolumn

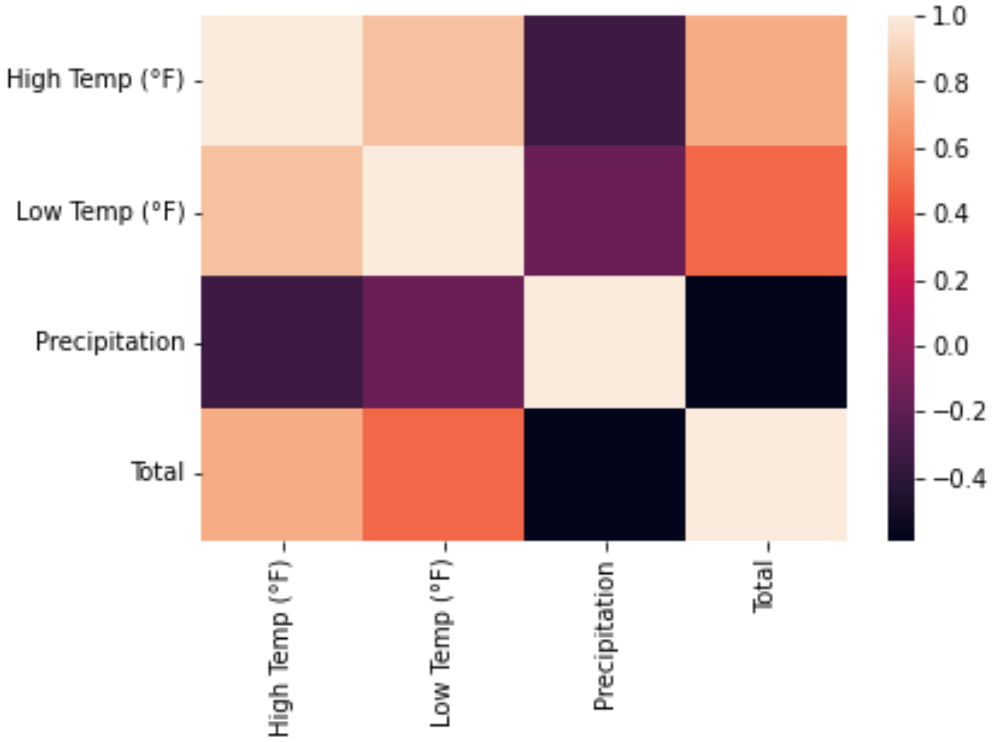
Duplicate records and unnecessary columns have been removed. Above are the first 5 rows from the current dataset

Data preprocessing



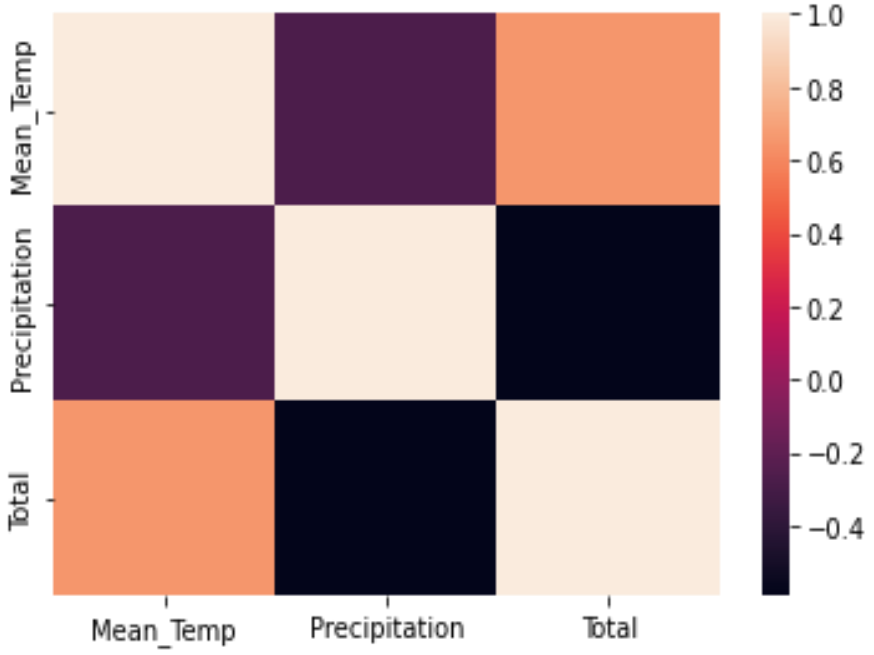
The feature correlation chart shows a strong correlation between the values from the columns describing the number of cyclists on the bridges

Data preprocessing



After removing the strongly correlated features, I will average the temperature for each day

Data preprocessing



	Mean_Temp	Precipitation	Total
Mean_Temp	1.000000	-0.269981	0.657221
Precipitation	-0.269981	1.000000	-0.590035
Total	0.657221	-0.590035	1.000000

Above presents how the correlation matrix of the current dataset looks like in the form of a chart and a table

Data preprocessing

Date	Mean_Temp	Precipitation	Total
2016-04-01	72.05	0.01	11497.0
2016-04-02	51.95	0.15	6922.0
2016-04-03	36.95	0.09	4759.0
2016-04-04	38.60	0.47	4335.0
2016-04-05	34.10	0.00	9471.0
2016-04-06	37.50	0.00	11919.0
2016-04-07	55.05	0.09	9596.0
2016-04-08	45.50	0.01	12744.0
2016-04-09	40.45	0.09	4510.0
2016-04-10	39.90	0.00	9126.0

Zbiór danych zawiera : 30 rekordów oraz 3 kolumn

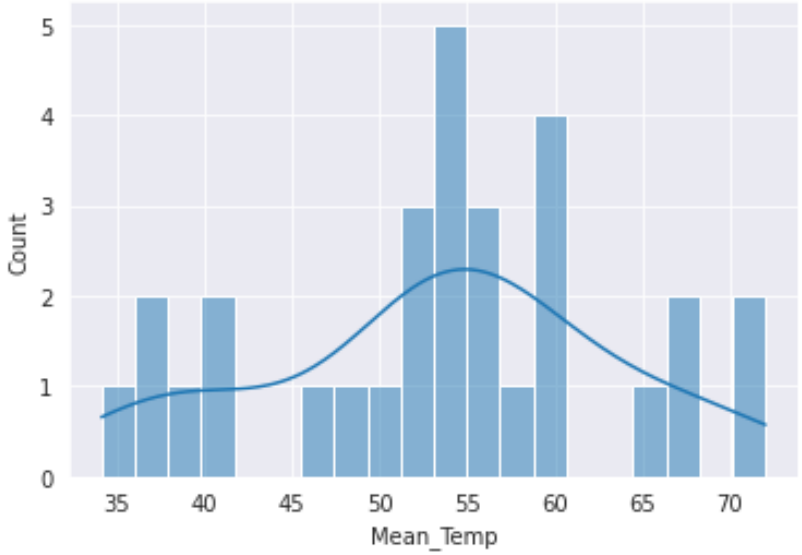
The final dataset will contain information on the date of measurement, average temperature, amount of precipitation and the total number of cyclists moving on the bridges

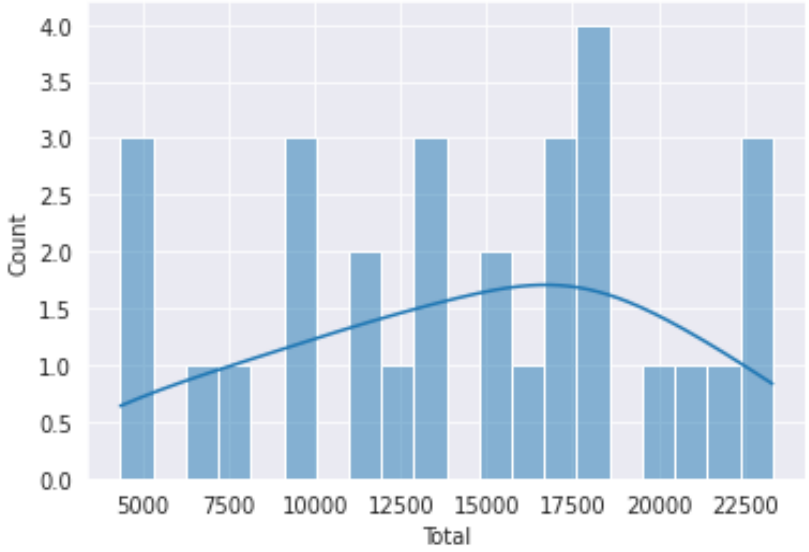
Data visualisation

	Mean_Temp	Precipitation	Total
count	30.000000	30.000000	30.000000
mean	53.496667	0.052667	14534.500000
std	10.034749	0.103489	5650.877227
min	34.100000	0.000000	4335.000000
25%	49.087500	0.000000	10071.250000
50%	54.050000	0.000000	15292.500000
75%	59.712500	0.080000	18281.250000
max	72.050000	0.470000	23318.000000

These are the basic statistics for the dataset shown on the slide earlier

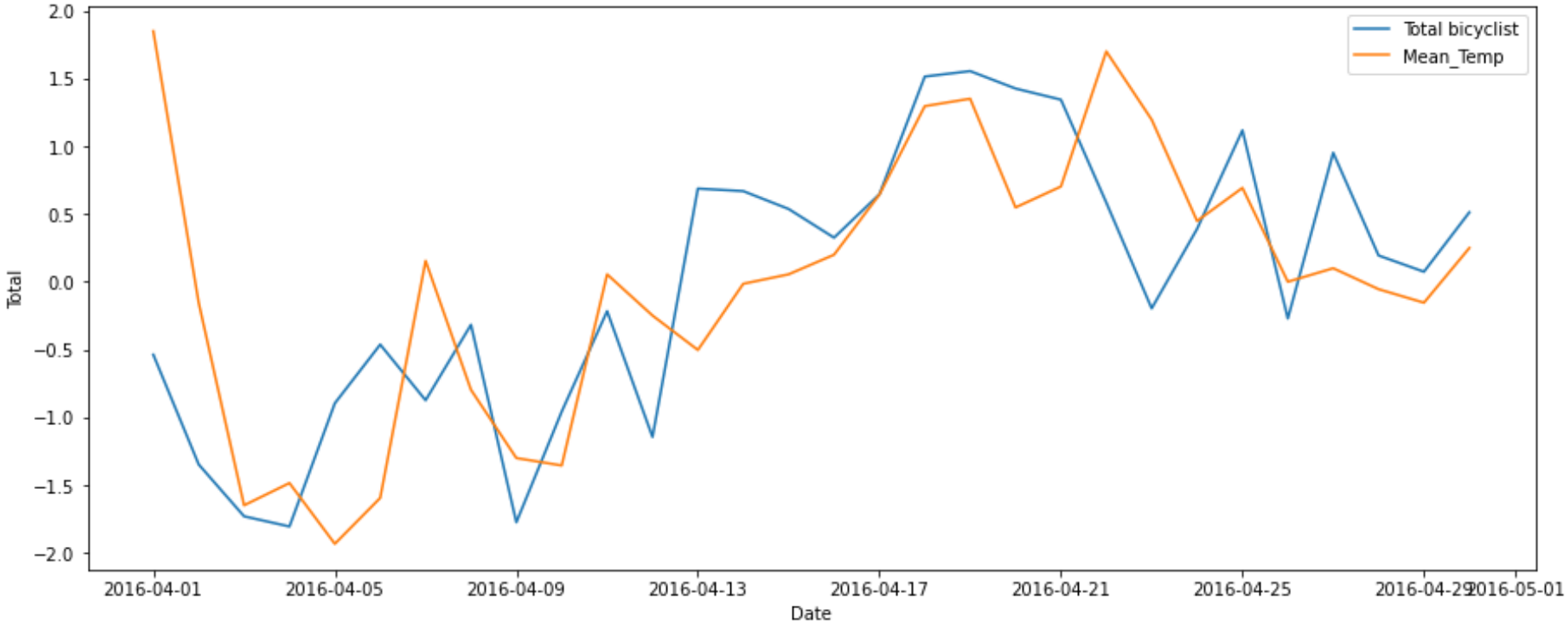
Data visualisation





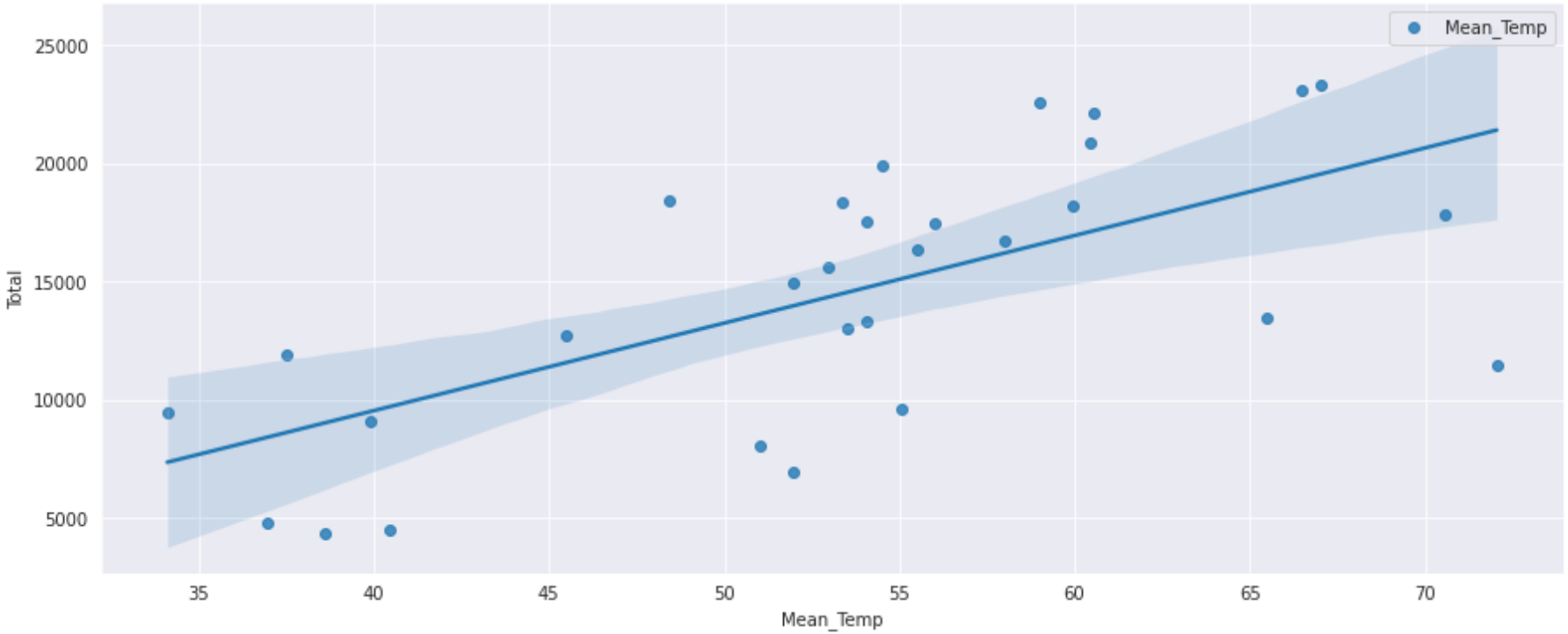
The histograms of the average temperature and the number of cyclists are shown above

Data visualisation



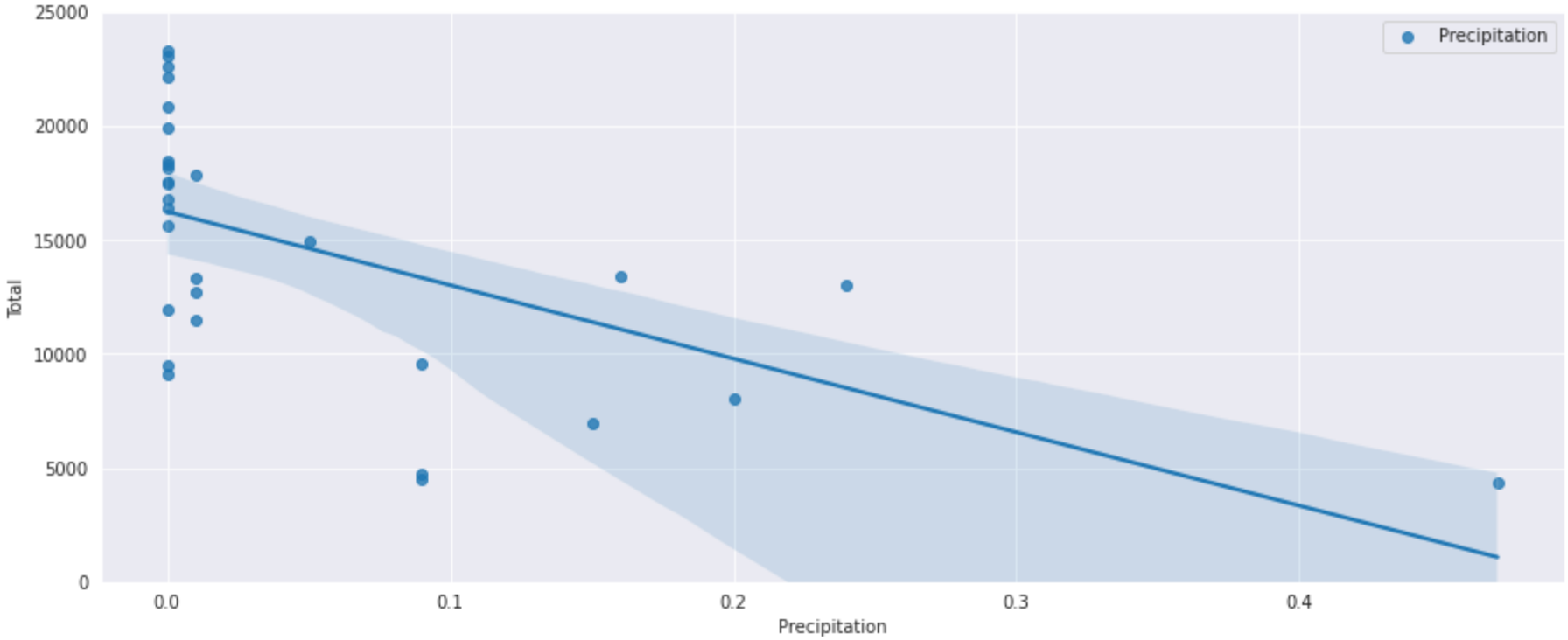
The chart above shows how the average temperature values and the number of cyclists behave in comparison to each other. It can be seen that the number of cyclists changes in a similar way with the temperature change.

Data visualisation



The chart above shows the relationship between the average temperature and the number of cyclists. There is a clear upward trend in the number of cyclists relative to the temperature value

Data visualisation



The chart above shows the relationship between precipitation and the number of cyclists. There is a clear downward trend in the number of cyclists to an increase in precipitation

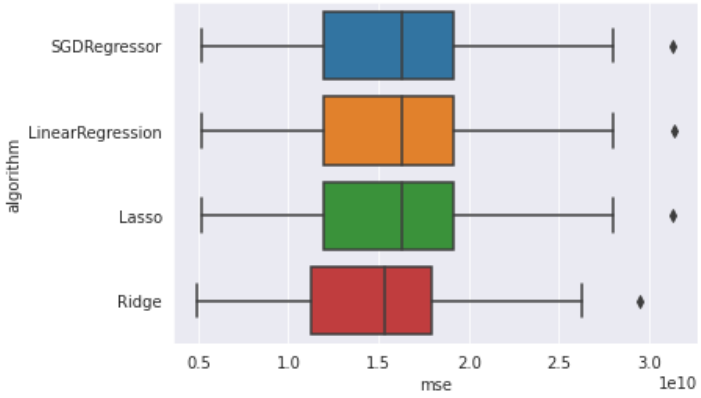
Data visualisation

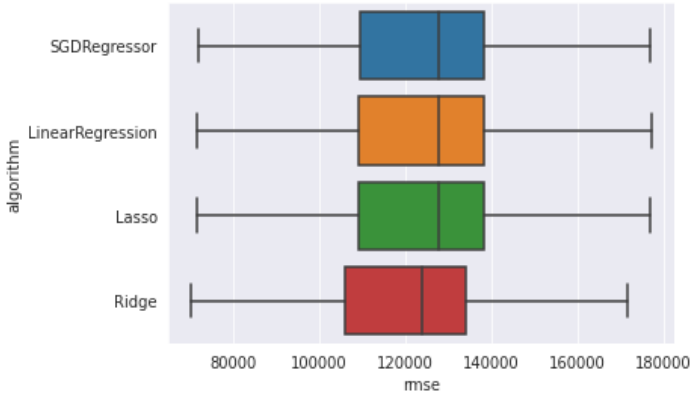
From the data visualization above, it can be seen that the increase in temperature and decrease in precipitation determine the increase in the number of cyclists

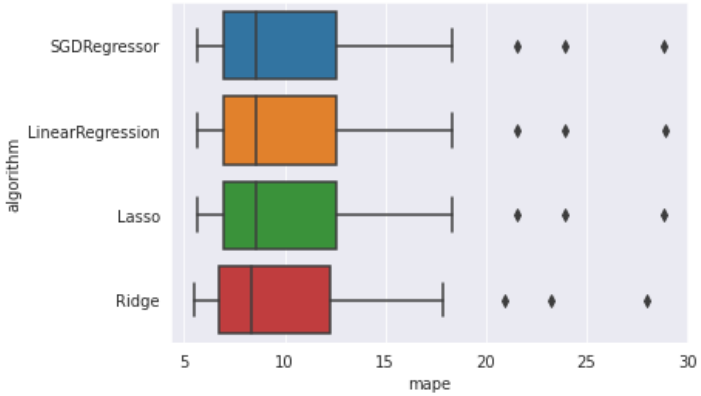
Modelling

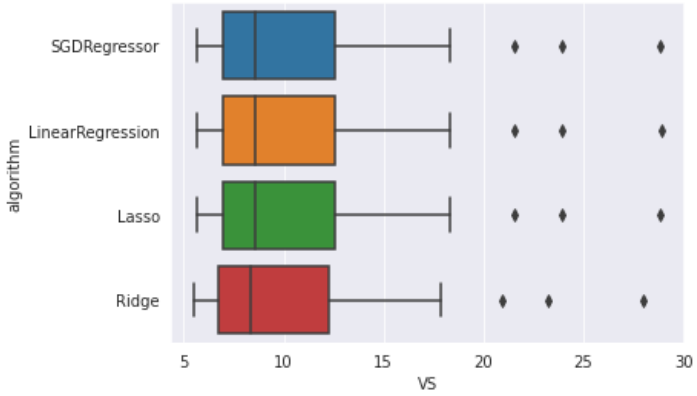
Four types of models will be compared to choose the most optimal type. The decision will be based on MSE, RMSE, MAPE and VS metrics . Due to dataset containing only 30 records, the validation will be preformed by using the Leave One Out method

Modelling



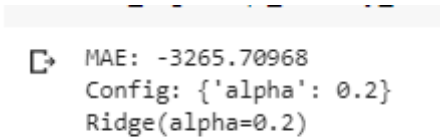






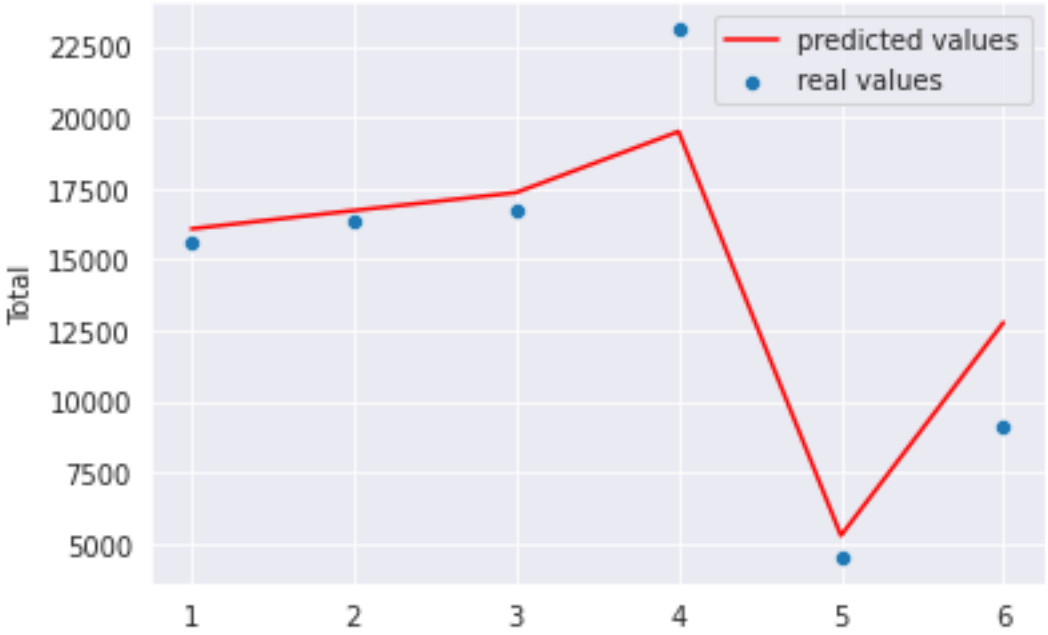
The compared models types are linear regression, ridge regression, lasso method and SGDRegressor. It turns out that the most accurate model is ridge regression model

Modelling



The parameter alpha of ridge regression is set on the value of 0.2

Modelling




```
MSE : 4572566.749097046
RMSE : 2138.356085664183
MAPE : 8.395125958162499
VS : 0.15010923476676685
R2 : 0.8705213643146981
```

To validate the operation of the model, test data were entered into it and the metrics that were used earlier to select the type of model were calculated. Blue points in the chart are the actual values for test data and red line is the estimated values for this data

Modelling

```
MSE : 4572566.749097046
RMSE : 2138.3560856641 83
MAPE : 8.395125958162499
VS : 0.150109234766766 85
R2 : 0.8705213643146981
```

It is worth looking again at the value of R2 coefficient. It proves that the selected model explains 87 % of the actual values

Modelling

	Mean_Temp	Precipitation	Prediction_of_number_of_cyclists
0	52.95	0.01	13917.695103
1	55.50	0.00	18801.129455
2	54.05	0.47	11101.550441

This is what the number of cyclists between Manhattan and Brooklyn would look like for a randomly generated dataset

Summary

- Thanks to the analysis of the dataset, it can be seen that the number of cyclists increases with the increase in temperature and decrease in precipitation
- The created Ridge regression model, which takes as an input average daily temperature value and precipitation value, can estimate number of cyclists on the bridges between Manhattan and Brooklyn
- A possible improvement of the model in the future (in the context of describing the relationship between the number of cyclists and weather conditions) could be adding also the wind force, which would also depend on the number of cyclists
- Possible model extension is to add particular days of week, if there would be need to answer the hypothesis but in the context of days of week

Thank you for your attention

Made by Kacper Gudalewski