RELATION BETWEEN DAILY AVERAGE TEMPERATURE AND DAILY NUMBER OF CYCLISTS IN FARIA LIMA'S CYCLE PATG IN THE CITY OF SAO PAULO IN BRAZIL

Mascarenhas Alexandre
University of Tukuba, Tsukuba – Ibaraki, Japan
e-mail: mascarenhasav@gmail.com

Abstract – **The objective of this paper it is to show how is the relation of the average temperature of the day and the daily number of cyclists riding on the Faria Lima's cycle path in Sao Paulo city in Brazil in this day, testing whether there is a linear relation between the two variables and using Pearson's correlation coefficient for this.**

Keywords – **Cyclists, Cycle paths**

I. INTRODUCTION

The big cities, such as city of Sao Paulo in Brazil have been constantly faced with problems related to the number of cars on the streets. This shows the needs in encourage other modes of transport. In the last ten years, the number of cyclist in Sao Paulo city has increased sharply. This increase has several reasons, such as public policies to encourage the use of alternatives ways of locomotion, for example, the bike.

II. ORGANIZATION OF THE PAPER

This paper will be organized like follows:

- How to obtain datasets: Will be described how to obtain the datasets used in the experiment;
- Pre-processing of the data: Usually a dataset has some problems in the data, such as data missing and outliers. Will be described how was made the treatment of this issues;
- Analaysis and presentation of the data: What results were obtained in the analysis of the datas;
- Conclusion

III. EXPERIMENT

A. Obtaining the data

The data used in this experiment are all public and were obtained in the website of responsible institution. They are composed of two datasets. One containing the daily number of cyclists passing in the counter installed on the cycle path of Faria Lima Avenue in the city of Sao Paulo. Other containing, among other data, the daily average temperature of the city. Both datasets are files with extension .CSV that it is a tabular form of organise the data, with the columns separated by comma.

The way to get each of the dataset is as follows:

1) Dataset of the number of cyclists: The dataset of the number of cyclists in the Faria Lima's cycle path it is available on the page `https://data.eco-counter.com/public2/?id=100027495`. In this page it is possible select the desired period, since the start of counting on 18 January of 2016 to the current day. After select the period just click in button "export to csv" to download the .CSV file. In this experiment

2) Dataset of the temperature: The dataset of temperature of the city of Sao Paulo it is available on the page `https://bdmep.inmet.gov.br/`. On the page it is necessary click in the "prosseguir", on the next page it is asked to enter an email address which one the dataset will be sent. After this, some settings must be made to set which variables will be included in the file as well as the desired period. Once all the settings have been set, the file will be sent to the email. In this experiment the period of the data which was from 1 January 2016 to

3) Both Datasets: Other way to obtain the datasets used in this experiment is through the link `https://www.github.com/mascarenhasav/master/tree/main/courses/experimental-design-in-computer-science/report-1`, where the files of number of cyclists and temperature were named respectively as "faria-lima.csv" and "temperature-sp.csv".

B. Pre-processing

Before effectively starting the data analysis, it was made a pre-processing of the datas. This is important, because allows to identify some issues with the dataset, such as, missing data, input error, incompatibility. For the purpose of this article, the period considered in data analysis was from 18 January 2018 to 1 January 2020. This choice was made mainly to avoid abrupt behavior caused by the Covid 19. It was noticed that there is two days missing in dataset of number of cyclists, the day 10 march 2016 and 15 november 2017. The approach to this missing datas was remove this two day in the dataset of temperature, so that the number of days both dataset would be the same. Once the data poneeriod is selected and there is no more issues with then, it becomes possible to do some analysis.

C. Analysis and presentation

There are a number of possible approaches to do the relation of average daily temperature and the daily number of cyclists. Naturally, many factors can influence the numbers, such as the day of the week, if it is a weekend, a holiday, and if it is raining or not. As a first analysis, will be make a relation without any of this considerations, which would allow to see if the temperature is one of the major factor. The Figure 1 shows

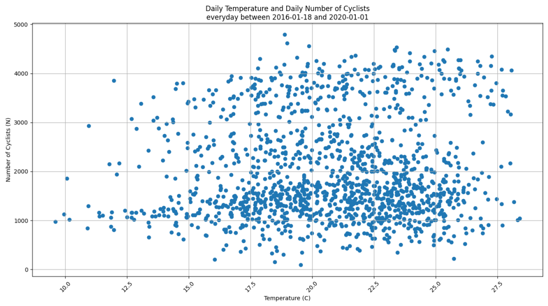

Fig. 1. Daily number of cyclists vs Daily average temperature everyday between 18 January 2016 and 01 January 2020.

the daily number of cyclists vs daily average temperature. As it is possible notice, visually, there is no clear linear relation between the two variables.

The idea is evaluate the Pearson's correlation coefficient of daily number of cyclists and daily average temperature. This coefficient measure a statistical relation between two continuous variables. It can be a range of values which goes from -1 to +1, in this scale, a zero value indicates no association between the variables, as well a positive value indicates a positive relation, which means that if one variable grows, the other grows too. On the other hand, if the coefficient turns out to be a negative value, means that if one variable decreases the other also decreases. Then, calculate the confidence interval of this coefficient with 95% of confidence, which represents a level of certainty about our estimate. The table 1 shows an interpretation about the values which correlation coefficient can assume.

TABLE I
Interpretation of Correlation coefficient

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Firstly, was considered the data representing all days between 18 january 2016 and 01 january 2022, so the number of samples (n) is 1443. Following are the steps to calculate the Pearson's correlation coefficient and the Confidence interval of the coefficient:

- Correlation coefficient:

  It will be used the following formula:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{1}$$

  where:

  r_{xy} - correlation coefficient;

  n - number of samples;

  x - number of cyclists;

  y - temperature.

  And the value found was $r_{xy} = 0.064$.

- Confidence interval:

  To calculate the confidence interval, It starts with the Fisher transformation:

$$z_r = \frac{1}{2}\log\frac{1+r}{1-r} \tag{2}$$

  Then calculates the lower and upper confidence limits with:

$$r_L = \frac{\exp(2z_L) - 1}{\exp(2z_L) + 1} \tag{3}$$

  and

$$r_U = \frac{\exp(2z_U) - 1}{\exp(2z_U) + 1} \tag{4}$$

  where:

$$z_L = z_r - z_{1-\alpha/2}\sqrt{\frac{1}{n-3}} \tag{5}$$

  and

$$z_U = z_r + z_{1-\alpha/2}\sqrt{\frac{1}{n-3}} \tag{6}$$

  with $\alpha = 0.05$.

  And the values found was $r_L = 0.006$ and $r_U = 0.121$.

As the value of the correlation coefficient was close to zero, it reinforces the visual idea that there is no strong relation between this two variables and taking a look at Table I it is possible to see that the value of correlation coefficient implies in a week correlation of type positive.

In order to mitigate some influences, as a natural variation according to the day of the week, the same graph and calculations were made but considering only specific day of the week, in this case, the Wednesdays. And following were the results:

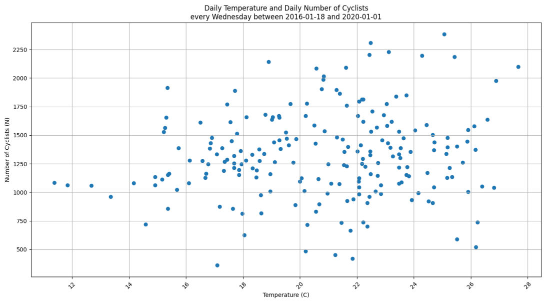

Fig. 2. Daily number of cyclists vs Daily average temperature every Wednesday between 18 January 2016 and 01 January 2020.

- Correlation coefficient:

The value found was $r_{xy} = 0.130$.

- Confidence interval:

  The values found was $r_L = -0.023$ and $r_U = 0.277$.

The value of the correlation coefficient is higher than the previous one, but still close to zero which implies a low correlation between the two variables, this value according to Table I it is considered a week relation of the type positive. In Table II it is possible to see the correlation coefficient as well the confidence interval and number of samples to each day of the week.

TABLE II
Correlation coefficient per the day of the week

Day of the week	Correlation coefficient	Confidence interval	N Samples
Mondays	0.060	[-0.093:0.210]	207
Tuesdays	0.047	[-0.106:0.197]	206
Wednesdays	0.130	[-0.023:0.277]	206
Thursdays	0.192	[0.041:0.334]	206
Fridays	0.014	[-0.139:0.165]	206
Saturdays	0.036	[-0.117:0.187]	206
Sundays	0.091	[-0.062:0.240]	206

D. Other representations

In possession of the data, it is possible to do some other graphs, which allow to see others way that the variables are related to each other. In the following will be showed line graphs and bar graphs of everyday and all Wednesdays between 18 January 2016 and 01 January 2020. First, in Figure 3a line graph of all data with two y-axes, along the time.

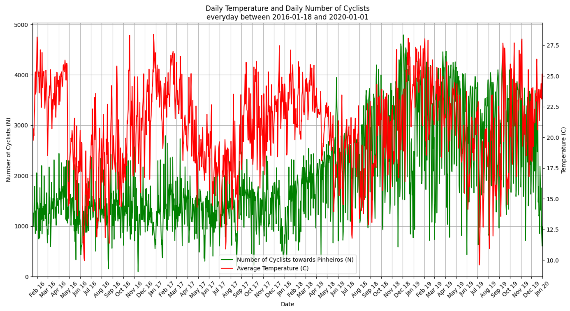

Fig. 3. Daily number of cyclists and Daily average temperature vs Time everyday between 18 January 2018 and 01 January 2020.

Now a bar graph of this same conditions:

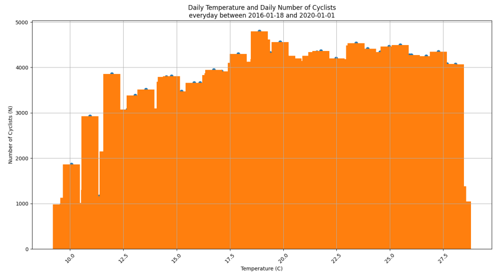

Fig. 4. Daily number of cyclists vs Daily average temperature everyday between 18 January 2016 and 01 January 2020.

Here, the same two graph but considering only the Wednesdays.

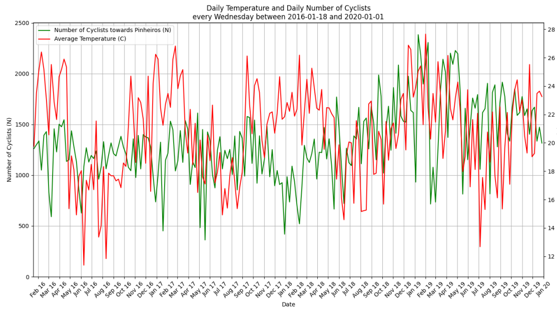

Fig. 5. Daily number of cyclists and Daily average temperature vs Time every Wednesday between 18 January 2016 and 01 January 2020.

And the bar graph:

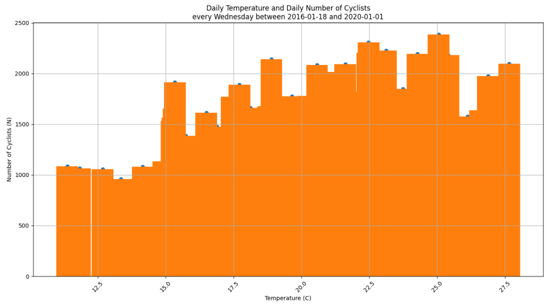

Fig. 6. Daily number of cyclists vs Daily average temperature every Wednesday between 18 January 2016 and 01 January 2020.

IV. OTHER INSTRUCTIONS

All these graphs and calculations were made by a python script. This script is available in the same link to access the datasets `https://github.com/mascarenhasav/master/tree/main/courses/experimental-design-in-computer-science/report-1` named code.py. This script allows to choice the time period which the data will be considered, if it will be everyday in this period or only one day of the week and a cut-off temperature which will be taken into account only the data with temperature lower than that.

V. CONCLUSIONS

The values obtained in the analysis, it is suggests that there is not a linear relation between the daily average temperature and daily number of cyclists in Faria Limas's Cycle path. Given that the parameter used, the Pearson's correlation coefficient, had a value close to zero, which implies that there is a weak, practically null, linear relation between the considered variables.

However, considerations must be taken into account, because it can have a considerable influence in the results. Some of the main ones that would be interesting to be considered in future studies are:

- **Average Temperature:** It was considered the average temperature of the day, and it can vary greatly throughout the day, not reflecting very well the temperature in the moment when the cyclist is leaving home, a moment that could possibly influence the cyclist's decision.

- **Rain:** Rain is a factor that is likely to have a great influence in the cyclist's decision. So, even the temperature is high, if it is raining, there is a big chance that the cyclist will not leaving home by bike.