Assessment 2: Exploratory Analysis

Programming for Data Analytics
Karla Aniela Cepeda Zapata
D00242569

General description

Website: freemeteo.com
Code with scraping process: Code/weather_scraping_proc.py
Final dataset: Dataset/weather_2019.csv
Conditions: weather information from Mexico City and Dublin in 2019.

Description of Dataset

In this section, the dataset is described in terms of variables and observations (fig 1.) The Description of variables is displayed in Table I. The dataset is compound by 18612 observations and 20 variables from which: 6344 are observations from Mexico City, and 11946 are from Dublin.

TABLE I. Description of dataset weather_2019.csv

Variable name	Type	Units	Data type	Comments
country	Categorical N.		String	Created by myself
city	Categorical N.		String	Created by myself
station	Categorical N.		String	
date	Categorical O.		Date	
time	Categorical N.		String	format ##:##, where # is any integer from 1-9
day	Numerical D.		Integer	Created by myself
month	Numerical D.		Integer	Created by myself
year	Numerical D.		Integer	Created by myself
hour	Numerical D.		Integer	Created by myself to manipulate time
min	Numerical D.		Integer	Created by myself to manipulate time since time
temperature	Numerical C.	C°	Float	
relative_temperature	Numerical C.	C°	Float	
wind	Numerical C.	Km/hr	Float	
wind_dir	Categorical N.	°	String	The direction is measure in degrees, but in freemeteo when the wind is absent, they place world “Calm”.
wind_gust				N/A
rel_humidity	Numerical C.	%	Integer	
dew_point	Numerical C.	C°	Float	
pressure	Numerical C.	mb	Float	
icon				Image
description	Categorical N.		String	

```
In [4]: w.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18612 entries, 0 to 18611
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   country               18612 non-null  object
 1   city                  18612 non-null  object
 2   station               18612 non-null  object
 3   date                  18612 non-null  object
 4   time                  18398 non-null  object
 5   day                   18398 non-null  float64
 6   month                 18398 non-null  float64
 7   year                  18398 non-null  float64
 8   hour                  18398 non-null  float64
 9   min                   18398 non-null  float64
 10  temperature           18398 non-null  float64
 11  relative_temperature  18398 non-null  float64
 12  wind                  18398 non-null  float64
 13  wind_dir              18398 non-null  object
 14  wind_gust             0 non-null      float64
 15  rel_humidity          18398 non-null  float64
 16  dew_point             18398 non-null  float64
 17  pressure              18398 non-null  float64
 18  icon                  0 non-null      float64
 19  description           18023 non-null  object
dtypes: float64(13), object(7)
memory usage: 2.8+ MB
```

Fig 1. Method info called in dataframe “w” which stands for weather

There are some issues in this dataset:

1. Column "*wind_gust*" has missing values. **Solution:** I removed the column. All cells have no values.
2. Column "*icon*" has missing values. **Solution**: I removed the column. All cells have no values.
3. In `info()` data does not have the correct type as mentioned in the table above. **Solution**: I had to use `map()` to set the proper type in each variable. Before this, I had to solve problem number 4.
4. Missing values in columns: "*time", "day", "month", "year", "hour", "min", "temperature", "relative_temperature", "wind", "wind_dir", "rel_humidity", "dew_point",* and *"pressure"*. **Solution:** above.

In point number four (see above), this problem was identified during the web scrapping process. Most of these are well known to be missing dates in the website freemeteo.com (randomly I selected 10, loaded date on the website, and found out that table was empty, fig. 2). To try to understand how this could affect further analysis of e.g. temperature, plots were created to display the number of missing dates by month and city (fig. 3.)

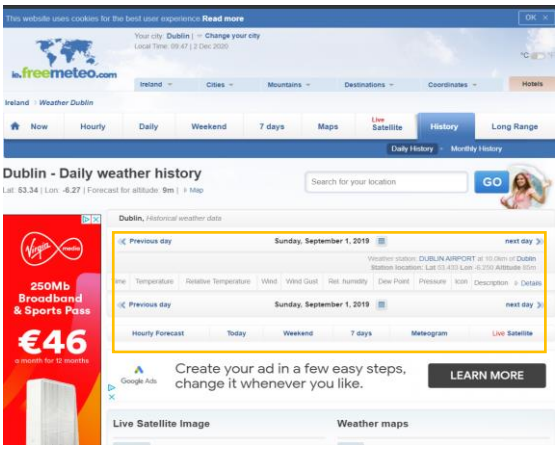

Fig 2. Missing data from website freemeteo.com

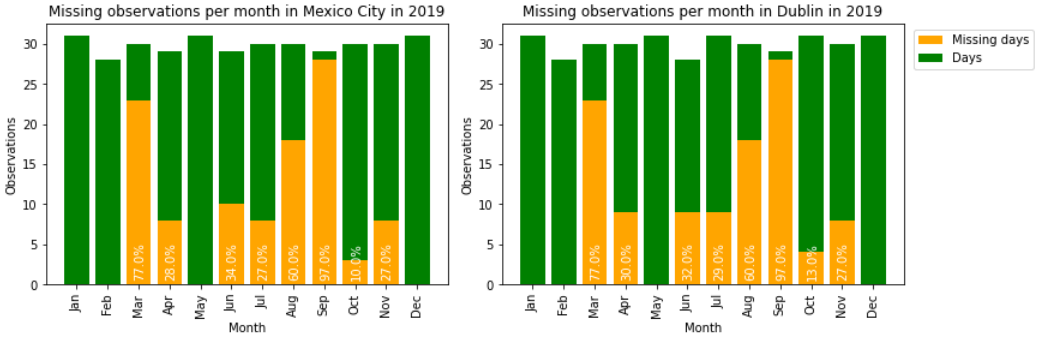

Fig 3. Missing daily observations of weather per month from Mexico City and Dublin in 2019

Both charts (Mexico City and Dublin) have roughly the same amount of days missing per month. What is more, the top 3 are the same: March, August, and September (and the same proportion too). **Solution:** to carry out the Continuous Assessment 2 of Programming for Data Analytics, I decided to drop these dates from the dataset. In fig. 4, the samples seem to be normally distributed after removing missing dates, it seems it is fine. However, for a deeper analysis in the future, I would prefer to look for another source to collect the missing dates to have a complete dataset.

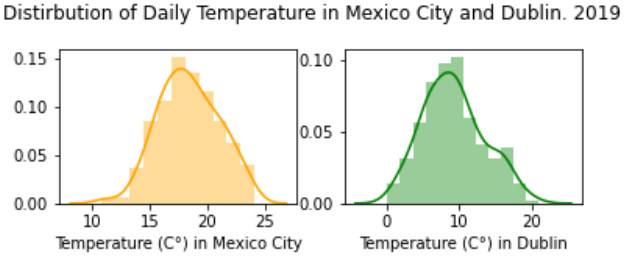

Fig 4. Histogram plot from Mexico City and Dublin weather in 2019.

Summary statistics

By using `describe()`, I got the following summary regarding the weather of Mexico City in 2019 (Table II) and the weather of Dublin in 2019 (Table III). Also, the information was plotted into a boxplot to examine summary information and compare between cities (fig. 5.)

TABLE II. Summary Statistics from weather of Mexico City in 2019

	temperature	relative_temperature	wind	rel_humidity	dew_point	pressure
count	6344	6344	6344	6344	6344	6344
mean	18.42765	18.11996	11.39392	49.43648	6.006778	1025.308
std	4.679998	4.517831	7.106303	21.0626	5.154047	2.661919
min	4	1	0	5	-18	1017
25%	15	15	7	33	3	1023
50%	18	18	9	49	6	1025
75%	22	21	15	66	10	1027
max	30	28	52	100	16	1034

TABLE III. Summary Statistics from weather of Dublin in 2019

	temperature	relative_temperature	wind	rel_humidity	dew_point	pressure
count	11946	11946	11946	11946	11946	11946
mean	9.222083	7.369915	16.87142	82.13511	6.113929	1011.069
std	4.783481	6.17992	8.674412	11.74038	4.147934	13.29274
min	-5	-10	0	33	-6	971
25%	6	3	11	76	3	1003
50%	9	6	15	83	6	1011
75%	12	12	22	93	9	1020
max	24	24	61	107	18	1044

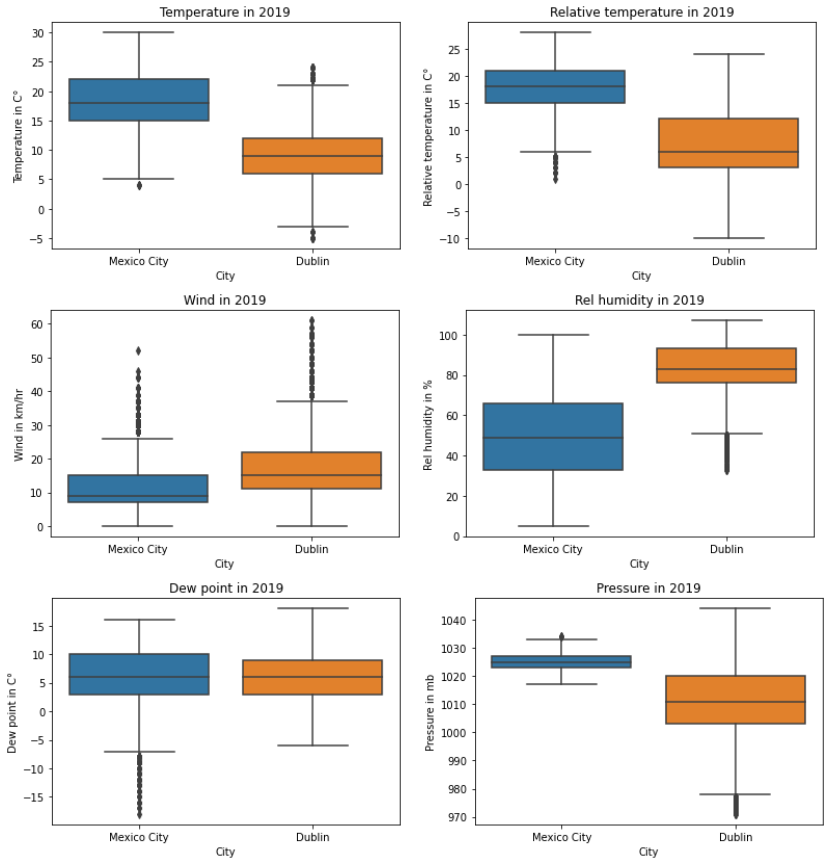

Fig 5. Boxplot graphs from summary statistics in tables II and III.

What stands out from the summary is:

- Mexico City's maximum temperature recorded in 2019 was 30C° whereas in Dublin was 24°C.
- The minimum temperature in Mexico City recorded in 2019 was 4C° and in Dublin was -5C°, which is 10°C lower than Mexico's.
- The median temperature in Mexico City is significantly higher than Dublin's. This means Mexico is weather is most of the times hotter than Dublin's.
- The relative temperature comparing with the temperature in Mexico City has roughly the same median.
- The relative temperature mean is 9.22 C° and the temperature means is 7.36C° in Dublin. It means, on average it feels colder due to other factors like wind or rain.
- Dublin recorded a max. wind speed of 61km/hr whereas 52km/hr in Mexico City.
- The median wind speed in Mexico City is slightly lower than Dublin's
- The relative humidity median in Dublin is higher than Mexico's. This means that Dublin's air tends to be more humid than Mexico's.
- Surprisingly, dew point in Mexico City and Dublin have the same median and the same spread.
- Mexico City's Pressure value is more constant than Dublin's

Temperature

The following plot shows the temperature of Mexico City and Dublin in 2019 (fig. 6). These plots show how the mean temperature per month had changed thorough 2019. *Note: September has a narrow range change between max and min temperature due to the lack of observations in this month (see fig. 3, page 2).*

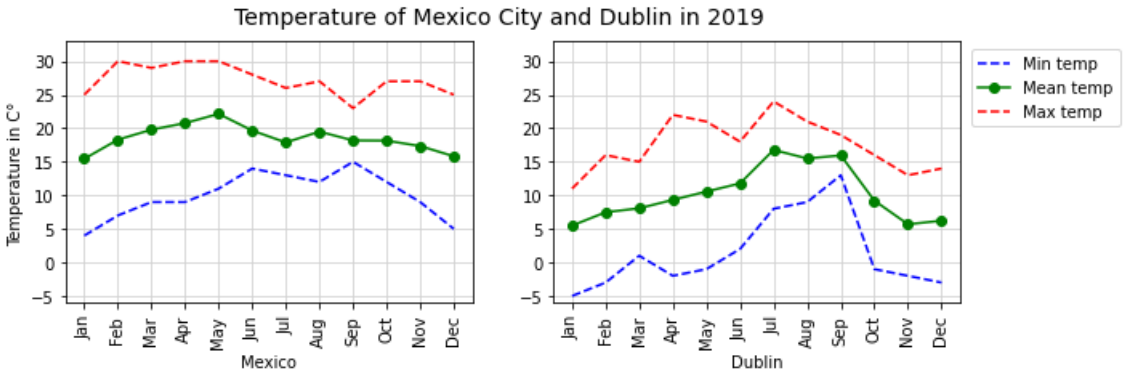

Fig 6. Temperature of Mexico City and Dublin in 2019.

By looking at the chart of Mexico City, December and January are the months which have the coldest days, but it seems that throughout the year the max temperature is roughly 24C°. On the other hand, Dublin tended to be less cold after April, when it started to show an increase until September. After this month, the temperature decreases until January. The next thing to compare is the temperatures but in the same plot (fig, 7.)

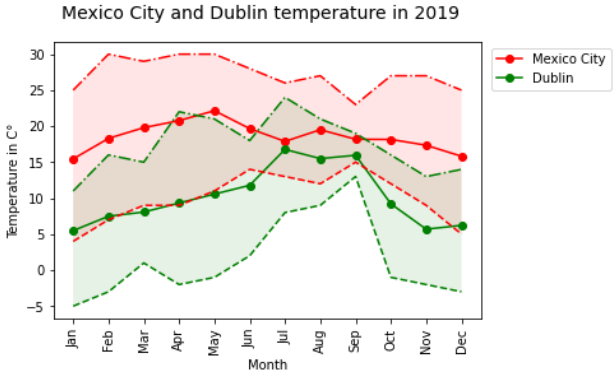

Fig 7. Overlap temperatures of Mexico City and Dublin in 2019.

In this graph (fig. 7), it is clear that the min. temperature from Mexico City overlapped the Max temperature of Dublin throughout 2019. July, August, and September show a close mean temperature between these cities; however, these could be misleading since the absence of observations in these months, especially in September (see fig. 3, page 2.)

Following, I will focus on scatter plots within the temperature and another continuous variable. In fig. 8, daily temperature and relative temperature in both cities had a positive strong linear relationship. In other words, when the temperature (C°) increased, the relative temperature (C°) also increased. Note: relative temperature is the temperature perceived by people.

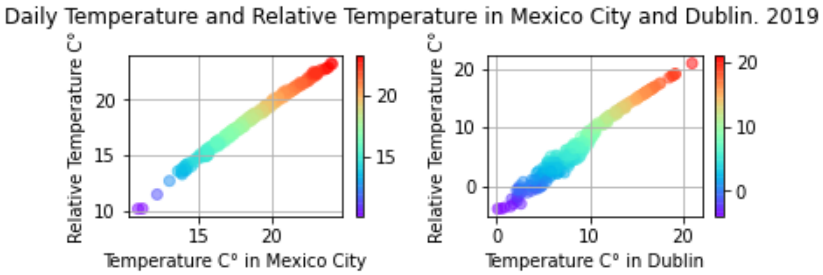

Fig 8. Scatter plots. Daily temperature VS Relative Temperature.

By comparing daily temperature and daily pressure (fig. 9), in Mexico City, it seems that there was a negative linear relationship whereas in Dublin there is a weak positive relationship. Interestingly, these relationships have different directions, Mexico City's recorded a negative slope and Dublin's a weak positive one.

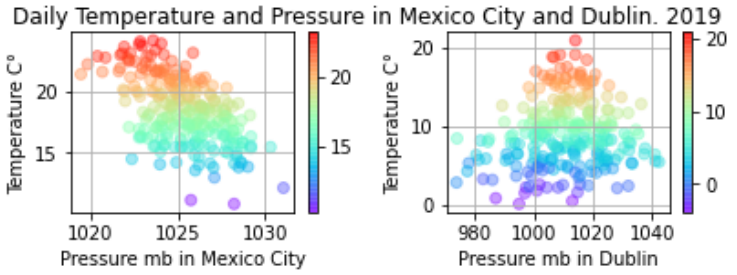

Fig 9. Scatter plots. Daily temperature VS Pressure.

In fig. 10, the plots show there was a positive linear relationship between wind speed and temperature in Mexico City. However, in Dublin, this plot shows there was a negative weak relationship. It means that in Dublin when wind speed increased, the temperature decreased, whereas in Mexico City trended to be hotter when wind speed increased.

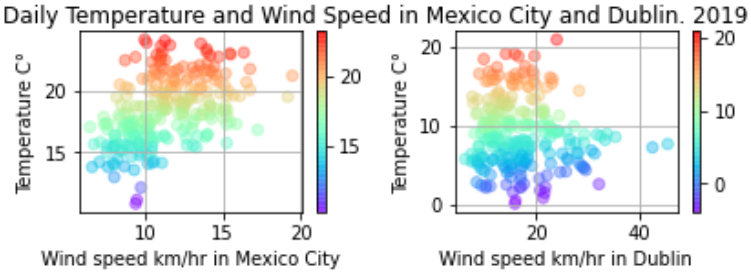

Fig 10. Scatter plots. Daily temperature VS Pressure.

For the last plot fig. 11, for both graphs there is a negative lineal relationship between Temperature and Rel. Humidity (i.e., percentage of water in the air). According to this, when the percentage of water in the air is higher, the temperature decreased in both cities.

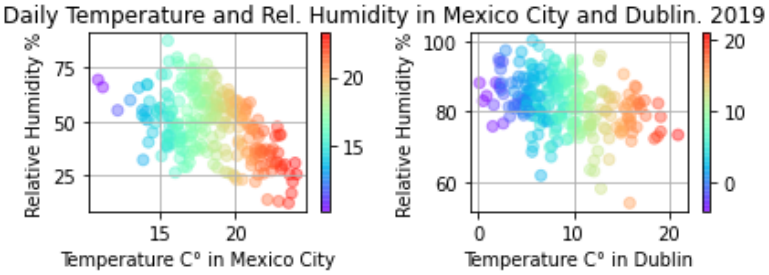

Fig 11. Scatter plots. Daily temperature VS Pressure.

Conclusion

Web scraping has been an interesting task, especially because it was the first time I know about it. I got so excited about it that I designed an automatic web scraping process to collect data from different dates throughout 2019 in two cities. This exercise was really helpful to understand the web scraping process and the data available in freemeteo (which is the information I want to use for the crossed module project). Unfortunately, there are missing data in the website (not web scraping process fault).

By looking at the weather of Mexico City and Dublin, it was interesting, and an obvious thing, that Mexico City is hotter than Dublin. There are some months of the year which are cooler and perfect to visit Mexico City. For future works, I would like to create a Mexico's weather predictive model to define when it is a good weather time to visit the country according to weather of your city.