# paper04_imputaciones_sinaica

October 9, 2021

# 1 Proyecto Final

# 2 Datos de SINAICA.

```
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
/home/jaa6766/.conda/envs/cuda/lib/python3.7/importlib/_bootstrap.py:219:
RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility.
Expected 192 from C header, got 216 from PyObject
```

Listing data files from: /home/jaa6766/Documents/jorge3a/itam/deeplearning/dlfinal/data/sinaica2/…

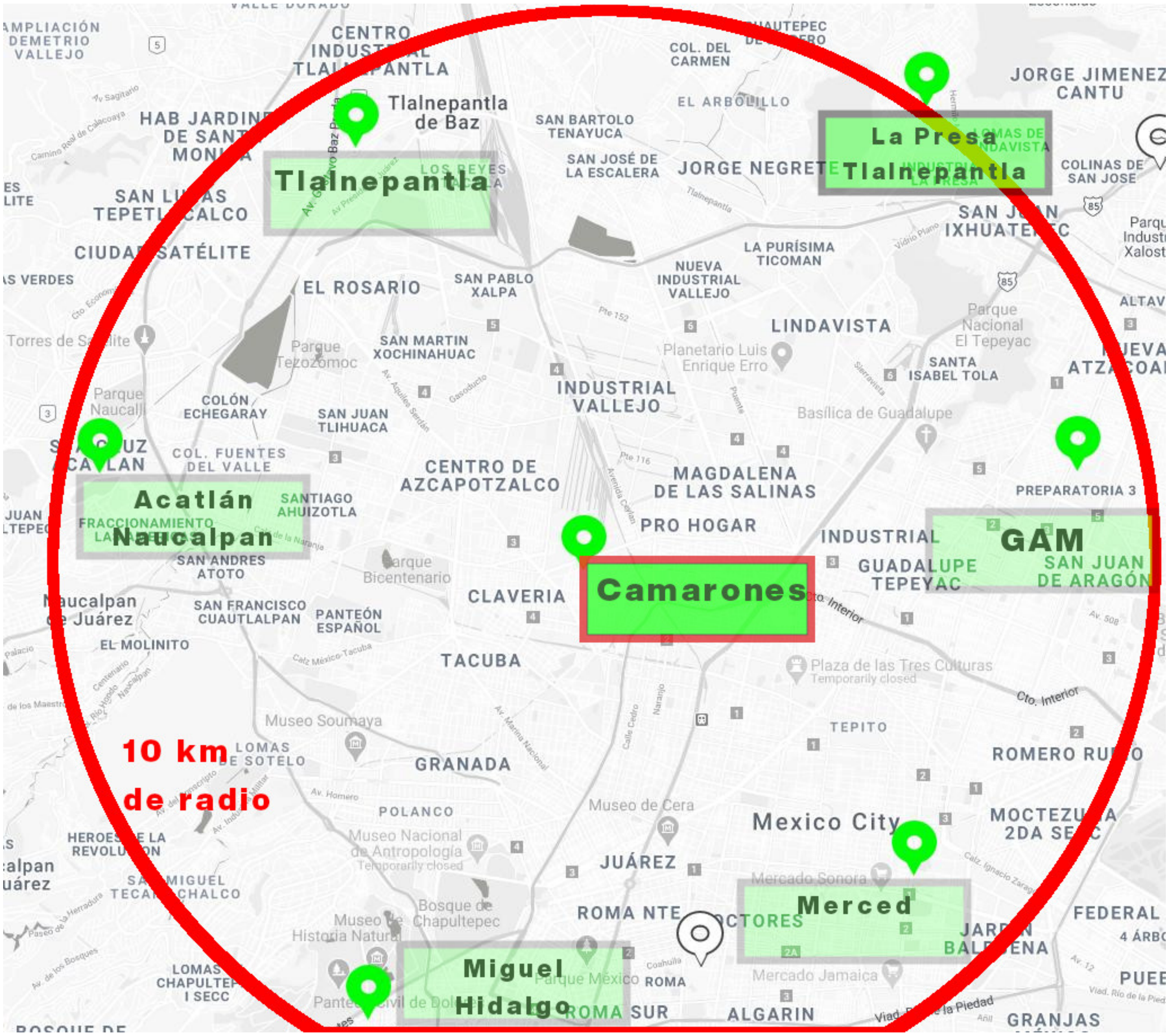
Done!

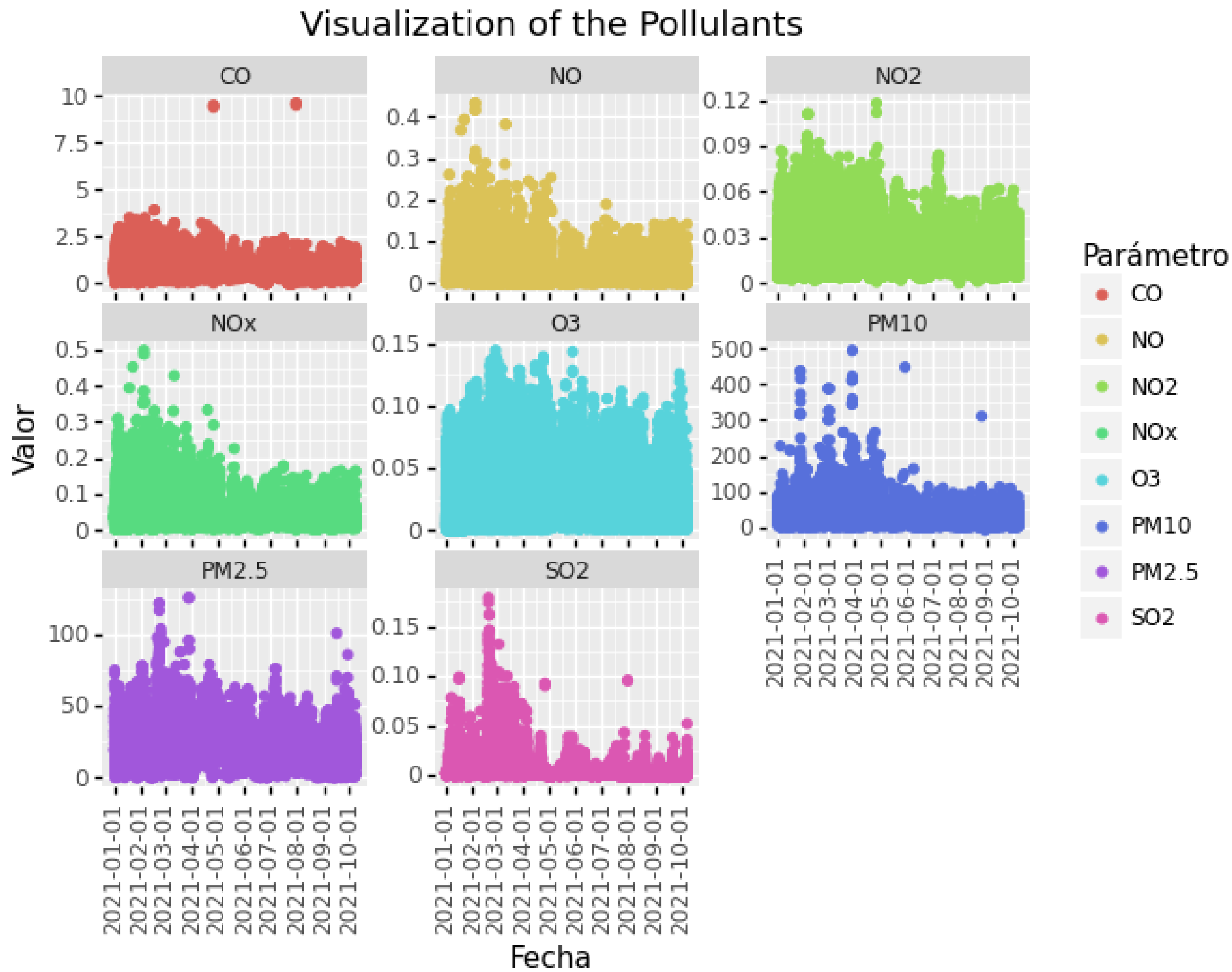
| | Parámetro | Fecha | Valor | Unidad | Estacion |
|---|---|---|---|---|---|
| 1 | CO | 2021-01-01 | 0.600 | ppm | Camarones |
| 1 | NO | 2021-01-01 | 0.006 | ppm | Camarones |
| 1 | NO2 | 2021-01-01 | 0.029 | ppm | Camarones |
| 1 | NOx | 2021-01-01 | 0.034 | ppm | Camarones |
| 1 | O3 | 2021-01-01 | 0.011 | ppm | Camarones |
| .. | … | … | … | … | … |
| 34 | SO2 | 2021-10-08 | 0.002 | ppm | Merced |
| 35 | SO2 | 2021-10-08 | 0.001 | ppm | Merced |
| 36 | SO2 | 2021-10-08 | 0.001 | ppm | Merced |
| 37 | SO2 | 2021-10-08 | 0.000 | ppm | Merced |
| 38 | SO2 | 2021-10-08 | 0.001 | ppm | Merced |

[196289 rows x 5 columns]

### 2.0.1 Nearby Air Quality Monitoring Stations

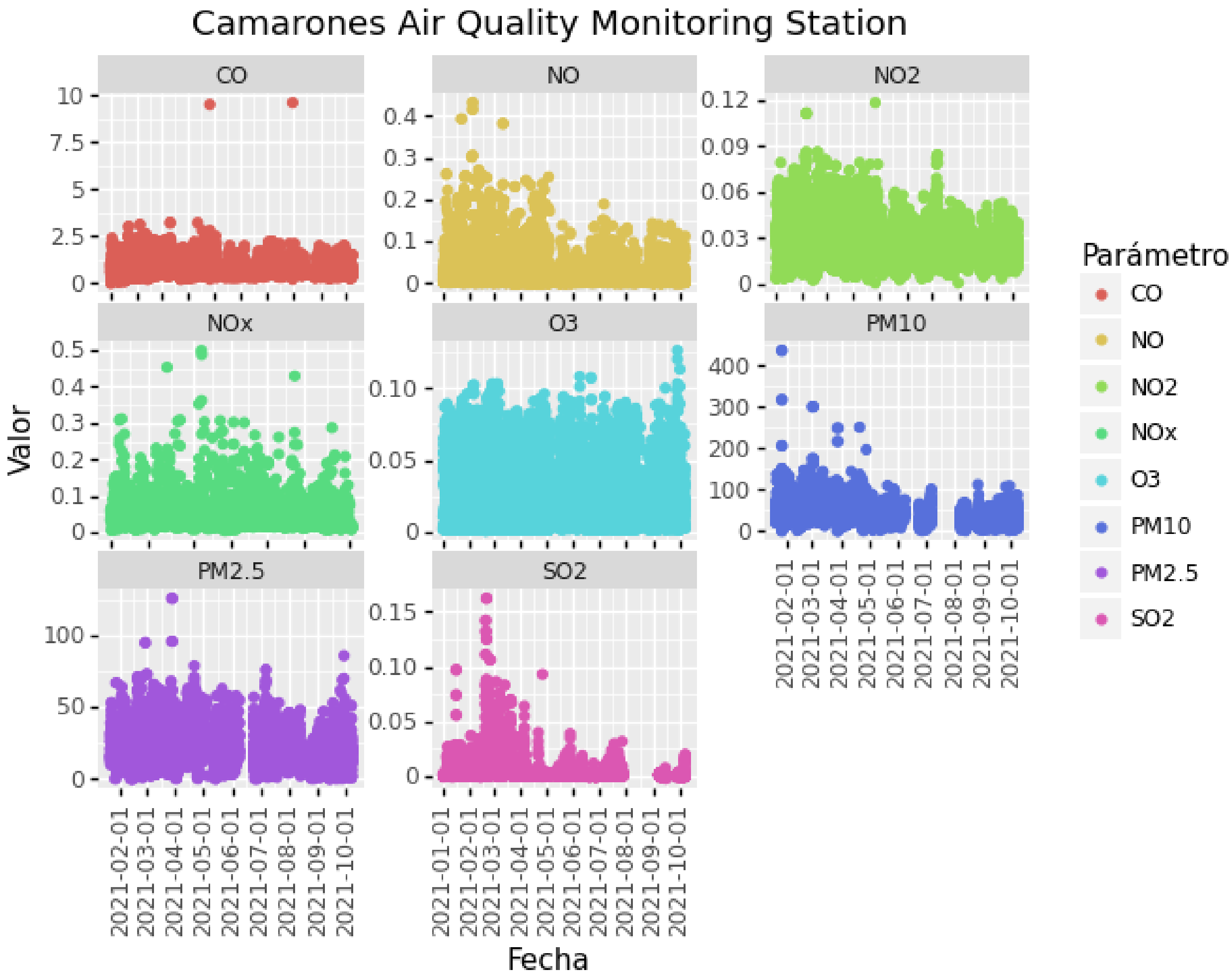
Here you may find the most proximate stations to "Camarones" which is the closest one to our sensor.


Tlalnepantla
La Presa Tlalnepantla
Acatlán Naucalpan
Camarones
GAM
10 km de radio
Merced
Miguel Hidalgo




```
<ggplot: (8760311468333)>
```

### 2.0.2 Camarones Air Quality Monitoring Station




```
<ggplot: (8760311483357)>
```

## 2.1 Imputation: Missing Data from the Air Quality Monitoring Stations.

Some of the missisng observatinos are caused by maintenance on the monitoring systems. So we could try to fill out the missing data with nearby government sesnsors. Then we propose to evaluate how the imputations work.

```
                     Fecha  Camarones_CO  Camarones_NO  Camarones_NO2  \
0    2021-01-01 00:00:00      0.600000      0.006000       0.029000
1    2021-01-01 01:00:00      1.000000      0.021000       0.038000
2    2021-01-01 02:00:00      0.800000      0.013000       0.035000
3    2021-01-01 03:00:00      1.000000      0.031000       0.034000
4    2021-01-01 04:00:00      0.600000      0.005000       0.029000
…                      …            …             …             …
2347 2021-10-04 00:00:00      0.441667      0.008292       0.015833
2348 2021-10-05 00:00:00      0.490000      0.010000       0.017000
2349 2021-10-06 00:00:00      0.542857      0.007571       0.022571
2350 2021-10-07 00:00:00      0.582609      0.011565       0.023130
2351 2021-10-08 00:00:00      0.738889      0.023778       0.026778

      Camarones_NOx  Camarones_O3  Camarones_PM10  Camarones_PM2.5  \
0             0.034      0.011000             NaN              NaN
1             0.059      0.002000             NaN              NaN
```

| | | | | |
|---|---|---|---|---|
| 2 | 0.049 | 0.003000 | NaN | NaN |
| 3 | 0.065 | 0.002000 | NaN | NaN |
| 4 | 0.034 | 0.005000 | NaN | NaN |
| … | … | … | … | … |
| 2347 | NaN | 0.017167 | 22.173913 | 10.952381 |
| 2348 | NaN | 0.013947 | 22.142857 | 8.736842 |
| 2349 | NaN | 0.014333 | 25.150000 | 10.150000 |
| 2350 | NaN | 0.021304 | 33.500000 | 15.428571 |
| 2351 | NaN | 0.019667 | 41.266667 | 18.800000 |

| | Camarones_SO2 | FES Acatlán_CO | … | Miguel Hidalgo_O3 \ |
|---|---|---|---|---|
| 0 | 0.002000 | 0.400000 | … | 0.009 |
| 1 | 0.002000 | 0.600000 | … | 0.006 |
| 2 | 0.001000 | 0.900000 | … | 0.003 |
| 3 | 0.001000 | 0.800000 | … | 0.004 |
| 4 | 0.001000 | 1.000000 | … | 0.006 |
| … | … | … | … | … |
| 2347 | 0.000125 | 0.315000 | … | NaN |
| 2348 | 0.000000 | 0.466667 | … | NaN |
| 2349 | 0.000000 | 0.347619 | … | NaN |
| 2350 | 0.001783 | 0.447826 | … | NaN |
| 2351 | 0.010500 | 0.566667 | … | NaN |

| | Miguel Hidalgo_SO2 | Tlalnepantla_CO | Tlalnepantla_NO | Tlalnepantla_NO2 \ |
|---|---|---|---|---|
| 0 | 0.003 | 0.6 | NaN | 0.030 |
| 1 | 0.003 | 0.6 | NaN | 0.026 |
| 2 | 0.002 | 0.7 | NaN | 0.032 |
| 3 | 0.002 | 0.7 | NaN | 0.033 |
| 4 | 0.002 | 0.7 | NaN | 0.032 |
| … | … | … | … | … |
| 2347 | NaN | NaN | NaN | NaN |
| 2348 | NaN | NaN | NaN | NaN |
| 2349 | NaN | NaN | NaN | NaN |
| 2350 | NaN | NaN | NaN | NaN |
| 2351 | NaN | NaN | NaN | NaN |

| | Tlalnepantla_NOx | Tlalnepantla_O3 | Tlalnepantla_PM10 \ |
|---|---|---|---|
| 0 | 0.034 | 0.012 | 37.0 |
| 1 | 0.029 | 0.013 | 42.0 |
| 2 | 0.036 | 0.006 | 58.0 |
| 3 | 0.039 | 0.004 | 59.0 |
| 4 | 0.038 | 0.004 | 64.0 |
| … | … | … | … |
| 2347 | NaN | NaN | NaN |
| 2348 | NaN | NaN | NaN |
| 2349 | NaN | NaN | NaN |
| 2350 | NaN | NaN | NaN |
| 2351 | NaN | NaN | NaN |

| | Tlalnepantla_PM2.5 | Tlalnepantla_SO2 |
|---|---|---|
| 0 | 19.0 | 0.002 |
| 1 | 29.0 | 0.003 |
| 2 | 43.0 | 0.002 |
| 3 | 41.0 | 0.002 |
| 4 | 46.0 | 0.002 |
| … | … | … |
| 2347 | NaN | NaN |
| 2348 | NaN | NaN |
| 2349 | NaN | NaN |

```
2350                 NaN               NaN
2351                 NaN               NaN

[2352 rows x 45 columns]

                   Fecha  Camarones_CO  Camarones_NO  Camarones_NO2  \
0    2021-01-01 00:00:00      0.600000      0.006000       0.029000
1    2021-01-01 01:00:00      1.000000      0.021000       0.038000
2    2021-01-01 02:00:00      0.800000      0.013000       0.035000
3    2021-01-01 03:00:00      1.000000      0.031000       0.034000
4    2021-01-01 04:00:00      0.600000      0.005000       0.029000
…                    …             …             …             …
2347 2021-10-04 00:00:00      0.441667      0.008292       0.015833
2348 2021-10-05 00:00:00      0.490000      0.010000       0.017000
2349 2021-10-06 00:00:00      0.542857      0.007571       0.022571
2350 2021-10-07 00:00:00      0.582609      0.011565       0.023130
2351 2021-10-08 00:00:00      0.738889      0.023778       0.026778

      Camarones_NOx  Camarones_O3  Camarones_PM10  Camarones_PM2.5  \
0             0.034      0.011000             NaN              NaN
1             0.059      0.002000             NaN              NaN
2             0.049      0.003000             NaN              NaN
3             0.065      0.002000             NaN              NaN
4             0.034      0.005000             NaN              NaN
…               …            …              …               …
2347            NaN      0.017167       22.173913        10.952381
2348            NaN      0.013947       22.142857         8.736842
2349            NaN      0.014333       25.150000        10.150000
2350            NaN      0.021304       33.500000        15.428571
2351            NaN      0.019667       41.266667        18.800000

      Camarones_SO2  FES Acatlán_CO  …  Miguel Hidalgo_O3  \
0          0.002000        0.400000  …              0.009
1          0.002000        0.600000  …              0.006
2          0.001000        0.900000  …              0.003
3          0.001000        0.800000  …              0.004
4          0.001000        1.000000  …              0.006
…               …              …  …                …
2347       0.000125        0.315000  …                NaN
2348       0.000000        0.466667  …                NaN
2349       0.000000        0.347619  …                NaN
2350       0.001783        0.447826  …                NaN
2351       0.010500        0.566667  …                NaN

      Miguel Hidalgo_SO2  Tlalnepantla_CO  Tlalnepantla_NO  Tlalnepantla_NO2  \
0                  0.003              0.6              NaN             0.030
1                  0.003              0.6              NaN             0.026
2                  0.002              0.7              NaN             0.032
3                  0.002              0.7              NaN             0.033
4                  0.002              0.7              NaN             0.032
…                    …              …               …               …
2347                 NaN              NaN              NaN               NaN
2348                 NaN              NaN              NaN               NaN
2349                 NaN              NaN              NaN               NaN
2350                 NaN              NaN              NaN               NaN
2351                 NaN              NaN              NaN               NaN

      Tlalnepantla_NOx  Tlalnepantla_O3  Tlalnepantla_PM10  \
0                0.034            0.012               37.0
```

```
1               0.029           0.013                42.0
2               0.036           0.006                58.0
3               0.039           0.004                59.0
4               0.038           0.004                64.0
…               …               …               …
2347              NaN             NaN                 NaN
2348              NaN             NaN                 NaN
2349              NaN             NaN                 NaN
2350              NaN             NaN                 NaN
2351              NaN             NaN                 NaN

     Tlalnepantla_PM2.5  Tlalnepantla_SO2
0                  19.0             0.002
1                  29.0             0.003
2                  43.0             0.002
3                  41.0             0.002
4                  46.0             0.002
…                 …               …
2347                NaN               NaN
2348                NaN               NaN
2349                NaN               NaN
2350                NaN               NaN
2351                NaN               NaN

[1694 rows x 45 columns]
```

### 2.1.1 Missing Data in Camarones

We can tell that "Camarones", the closest one, has missing data on all variables.

```
     Camarones_CO  Camarones_NO  Camarones_NO2  Camarones_NOx  Camarones_O3  \
0        0.600000      0.006000       0.029000          0.034      0.011000
1        1.000000      0.021000       0.038000          0.059      0.002000
2        0.800000      0.013000       0.035000          0.049      0.003000
3        1.000000      0.031000       0.034000          0.065      0.002000
4        0.600000      0.005000       0.029000          0.034      0.005000
…             …             …             …             …             …
2347     0.441667      0.008292       0.015833            NaN      0.017167
2348     0.490000      0.010000       0.017000            NaN      0.013947
2349     0.542857      0.007571       0.022571            NaN      0.014333
2350     0.582609      0.011565       0.023130            NaN      0.021304
2351     0.738889      0.023778       0.026778            NaN      0.019667

     Camarones_PM10  Camarones_PM2.5  Camarones_SO2
0               NaN              NaN       0.002000
1               NaN              NaN       0.002000
2               NaN              NaN       0.001000
3               NaN              NaN       0.001000
4               NaN              NaN       0.001000
…             …               …               …
2347      22.173913        10.952381       0.000125
2348      22.142857         8.736842       0.000000
2349      25.150000        10.150000       0.000000
2350      33.500000        15.428571       0.001783
2351      41.266667        18.800000       0.010500

[791 rows x 8 columns]
```

Then we can look forward to avoid loosing a big portion of data: 33.63% by using imputation. Our goal is to evaluate the different imputation methods in order to have data to back our decision.

### 2.1.2 Complete Observations in Camarones.

| | Camarones_CO | Camarones_NO | Camarones_NO2 | Camarones_NOx | Camarones_O3 |
|---|---|---|---|---|---|
| 478 | 0.933333 | 0.035333 | 0.03275 | 0.028 | 0.0195 |
| 480 | 0.500000 | 0.021000 | 0.02400 | 0.046 | 0.0020 |
| 481 | 0.600000 | 0.017000 | 0.02300 | 0.039 | 0.0020 |
| 482 | 0.500000 | 0.023000 | 0.02200 | 0.046 | 0.0020 |
| 483 | 0.600000 | 0.030000 | 0.02100 | 0.051 | 0.0020 |
| … | … | … | … | … | … |
| 2167 | 0.400000 | 0.003000 | 0.01100 | 0.013 | 0.0160 |
| 2168 | 0.400000 | 0.002000 | 0.01100 | 0.012 | 0.0180 |
| 2169 | 0.400000 | 0.002000 | 0.01300 | 0.015 | 0.0160 |
| 2170 | 0.400000 | 0.002000 | 0.01900 | 0.021 | 0.0120 |
| 2171 | 0.400000 | 0.001000 | 0.01400 | 0.015 | 0.0210 |

| | Camarones_PM10 | Camarones_PM2.5 | Camarones_SO2 |
|---|---|---|---|
| 478 | 43.047619 | 21.333333 | 0.004708 |
| 480 | 30.000000 | 14.000000 | 0.008000 |
| 481 | 28.000000 | 16.000000 | 0.006000 |
| 482 | 38.000000 | 26.000000 | 0.005000 |
| 483 | 35.000000 | 21.000000 | 0.003000 |
| … | … | … | … |
| 2167 | 69.000000 | 7.000000 | 0.001000 |
| 2168 | 71.000000 | 9.000000 | 0.001000 |
| 2169 | 37.000000 | 9.000000 | 0.001000 |
| 2170 | 19.000000 | 0.000000 | 0.001000 |
| 2171 | 61.000000 | 21.000000 | 0.001000 |

[1561 rows x 8 columns]

### 2.1.3 Train and Test Split

- Complete observations: 1561 (100%).
    - Complete observations on Training Set: 1092 (~70%).
    - Complete observations on Test Set: 469 (~30%).
- Incomplete Observations: 791.

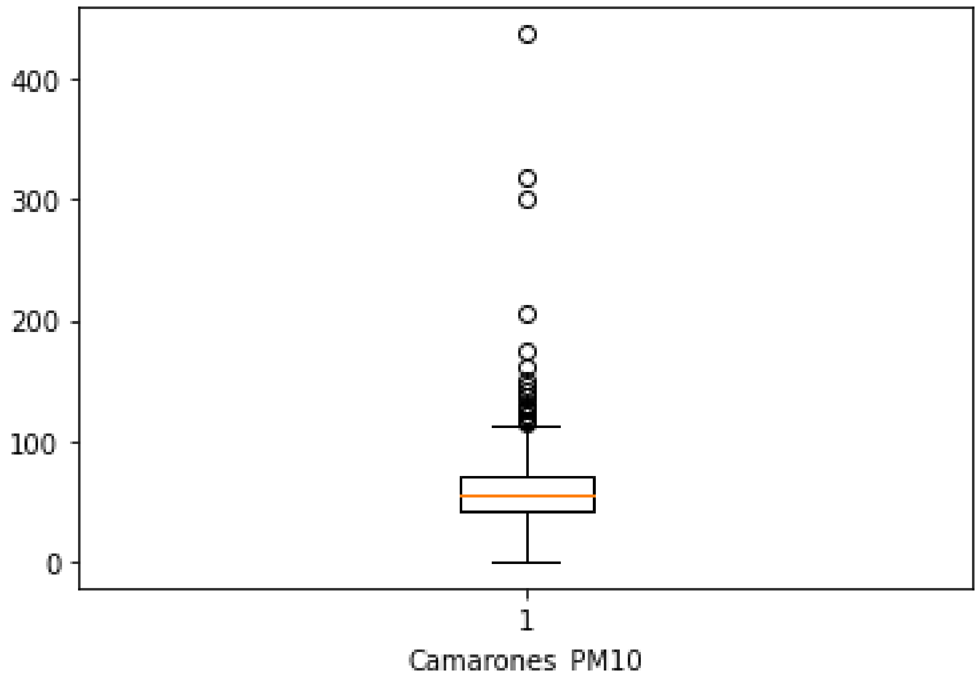
### 2.1.4 Data Distribution

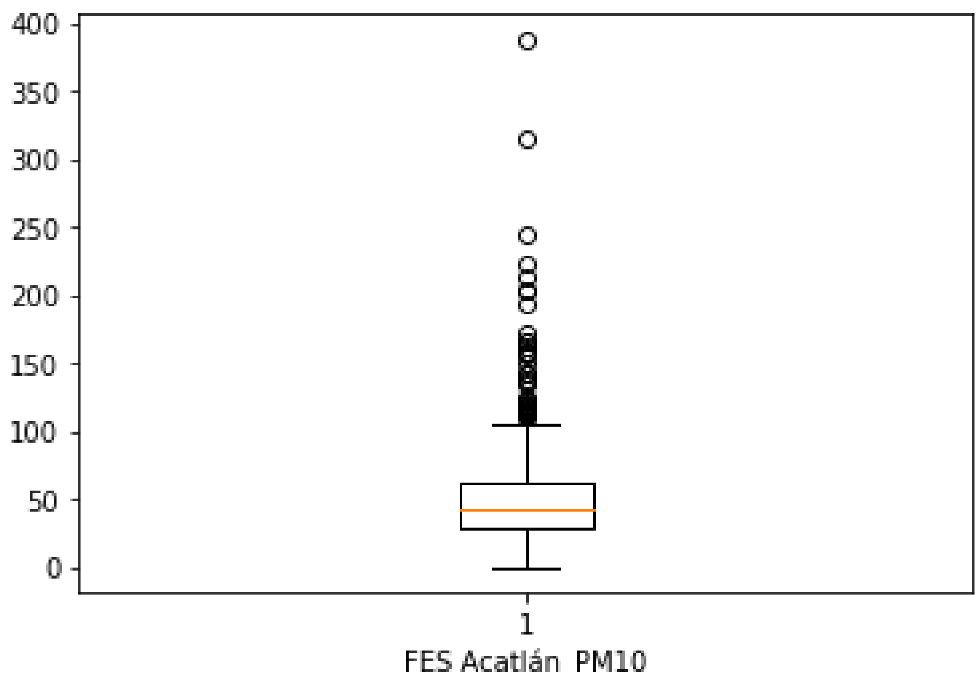
**PM10** These are the Air Quality Monitoring Stations that measure PM10 pollutant.

| | Camarones_PM10 | FES Acatlán_PM10 | Gustavo A. Madero_PM10 | Merced_PM10 |
|---|---|---|---|---|
| 478 | 43.047619 | 34.0 | 32.0 | 39.727273 |
| 480 | 30.000000 | 15.0 | 33.0 | 34.000000 |
| 481 | 28.000000 | 11.0 | 32.0 | 29.000000 |
| 482 | 38.000000 | 15.0 | 28.0 | 33.000000 |
| 483 | 35.000000 | 15.0 | 25.0 | 32.000000 |
| … | … | … | … | … |
| 2166 | 71.000000 | 178.0 | 69.0 | 49.000000 |
| 2167 | 69.000000 | 162.0 | 33.0 | 36.000000 |
| 2168 | 71.000000 | 49.0 | 32.0 | 24.000000 |
| 2170 | 19.000000 | 14.0 | 19.0 | 21.000000 |
| 2171 | 61.000000 | 56.0 | 44.0 | 44.000000 |

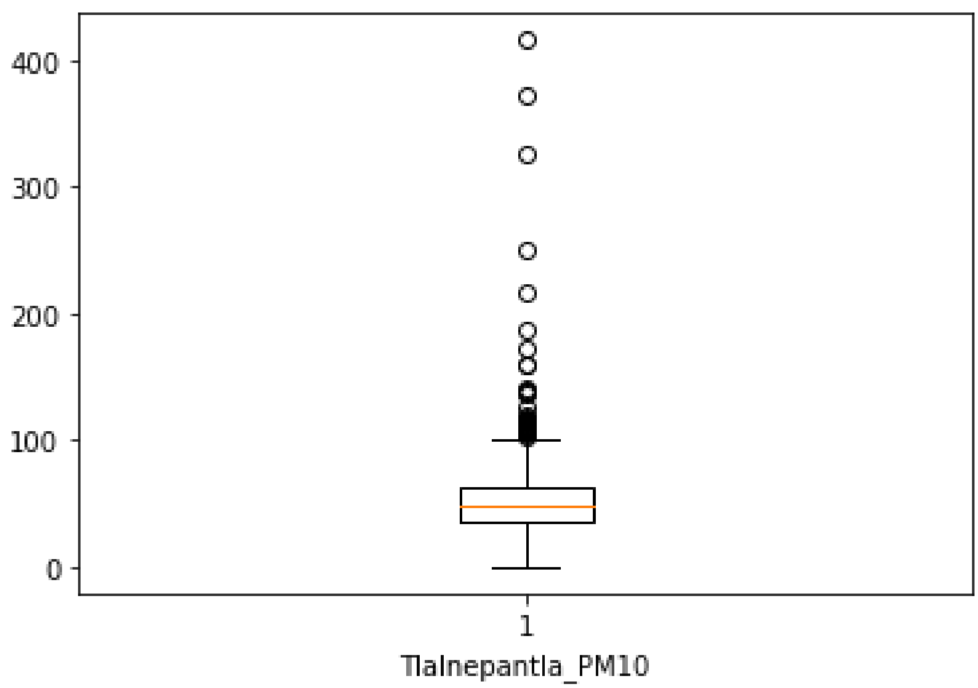
| | Tlalnepantla_PM10 |
|---|---|
| 478 | 23.0 |
| 480 | 32.0 |
| 481 | 24.0 |
| 482 | 41.0 |

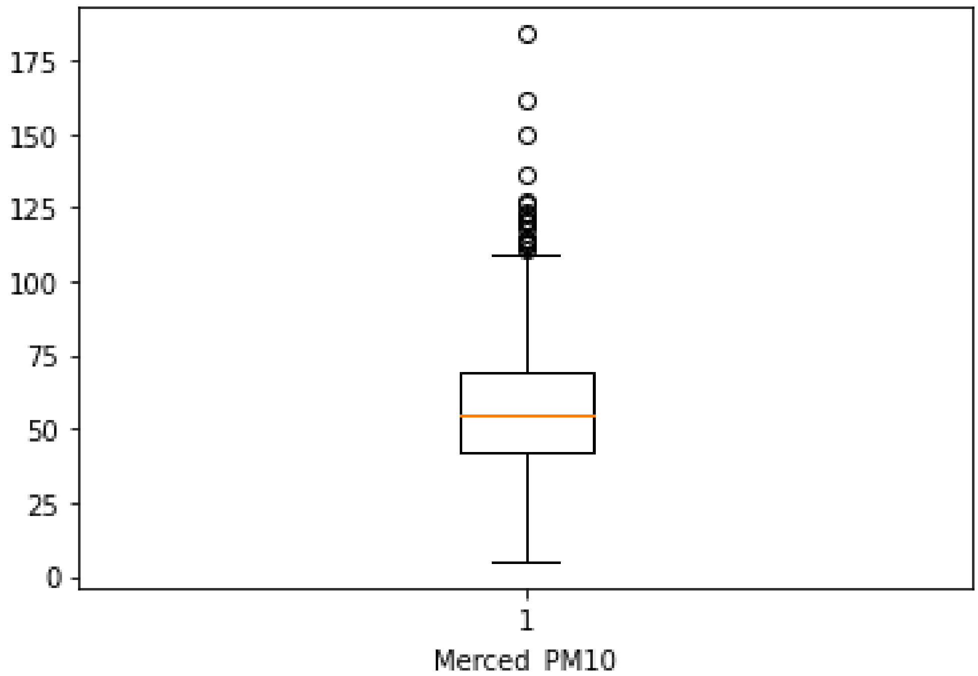
```
483                 20.0
…                   …
2166                90.0
2167                52.0
2168                22.0
2170                11.0
2171                55.0

[1305 rows x 5 columns]

                                     Estacion   count       mean        std  \
Camarones_PM10                 Camarones_PM10  1092.0  58.163765  27.817189
FES Acatlán_PM10             FES Acatlán_PM10  1011.0  48.669366  32.624587
Gustavo A. Madero_PM10  Gustavo A. Madero_PM10  1029.0  52.906706  26.733354
Merced_PM10                       Merced_PM10  1084.0  56.798977  21.410434
Tlalnepantla_PM10           Tlalnepantla_PM10  1034.0  52.254352  29.446065

                        min   25%   50%   75%    max   NAs
Camarones_PM10          0.0  42.0  55.0  71.0  437.0   0.0
FES Acatlán_PM10        0.0  29.0  43.0  61.0  388.0  81.0
Gustavo A. Madero_PM10  0.0  34.0  50.0  67.0  352.0  63.0
Merced_PM10             5.0  42.0  55.0  69.0  184.0   8.0
Tlalnepantla_PM10       0.0  36.0  48.0  63.0  416.0  58.0

                        Estacion   count       mean        std       min  \
Camarones_CO        Camarones_CO  2241.0   0.767037   0.412628  0.000000
Camarones_NO        Camarones_NO  2227.0   0.024599   0.043639  0.000000
Camarones_NO2      Camarones_NO2  2227.0   0.031139   0.015144  0.003000
Camarones_NOx      Camarones_NOx  2042.0   0.056961   0.054418  0.004000
Camarones_O3        Camarones_O3  2233.0   0.026094   0.022991  0.001000
Camarones_PM10    Camarones_PM10  1749.0  56.638261  26.729443  0.000000
Camarones_PM2.5  Camarones_PM2.5  1765.0  24.872841  12.620210  0.000000
Camarones_SO2      Camarones_SO2  2190.0   0.006284   0.012464 -0.000048

                       25%     50%     75%      max    NAs
Camarones_CO      0.500000   0.700   0.900    3.200  111.0
Camarones_NO      0.003000   0.007   0.026    0.432  125.0
Camarones_NO2     0.020000   0.029   0.040    0.111  125.0
Camarones_NOx     0.022000   0.039   0.071    0.499  310.0
Camarones_O3      0.005000   0.021   0.039    0.103  119.0
Camarones_PM10   40.000000  54.000  70.000  437.000  603.0
Camarones_PM2.5  16.000000  24.000  32.000  126.000  587.0
Camarones_SO2     0.001217   0.003   0.005    0.162  162.0
```

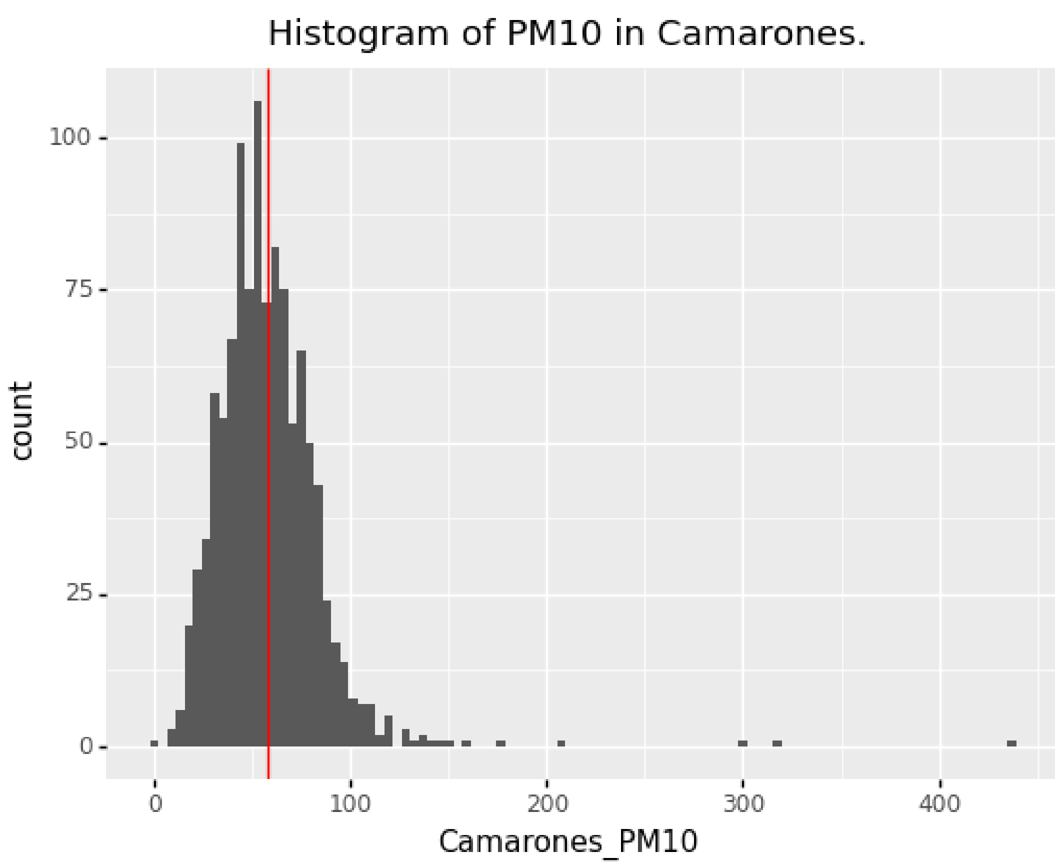

400
300
200
100
0
1
Camarones_PM10


400
350
300
250
200
150
100
50
0
1
FES Acatlán_PM10


350
300
250
200
150
100
50
0
1
Gustavo A. Madero_PM10


400
300
200
100
0
1
Tlalnepantla_PM10

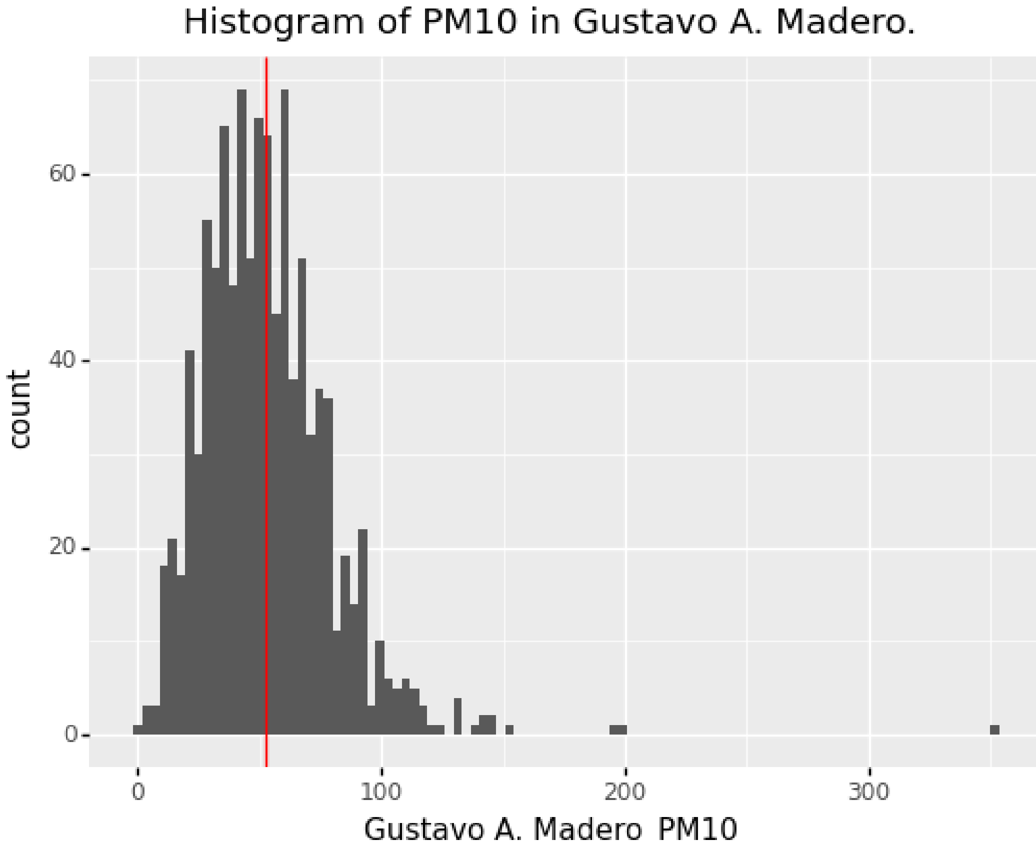



```
<ggplot: (8760308721261)>
```

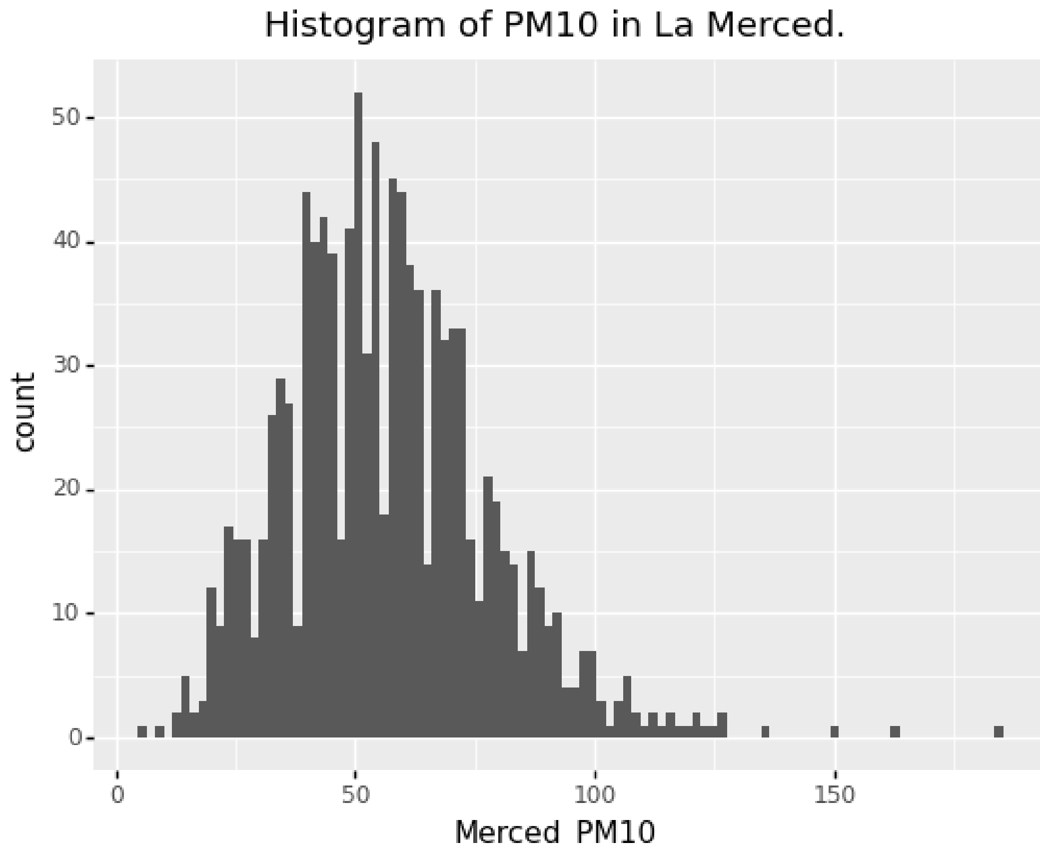
```
/home/jaa6766/.conda/envs/cuda/lib/python3.7/site-
packages/plotnine/layer.py:372: PlotnineWarning: stat_bin : Removed 63 rows
```

containing non-finite values.



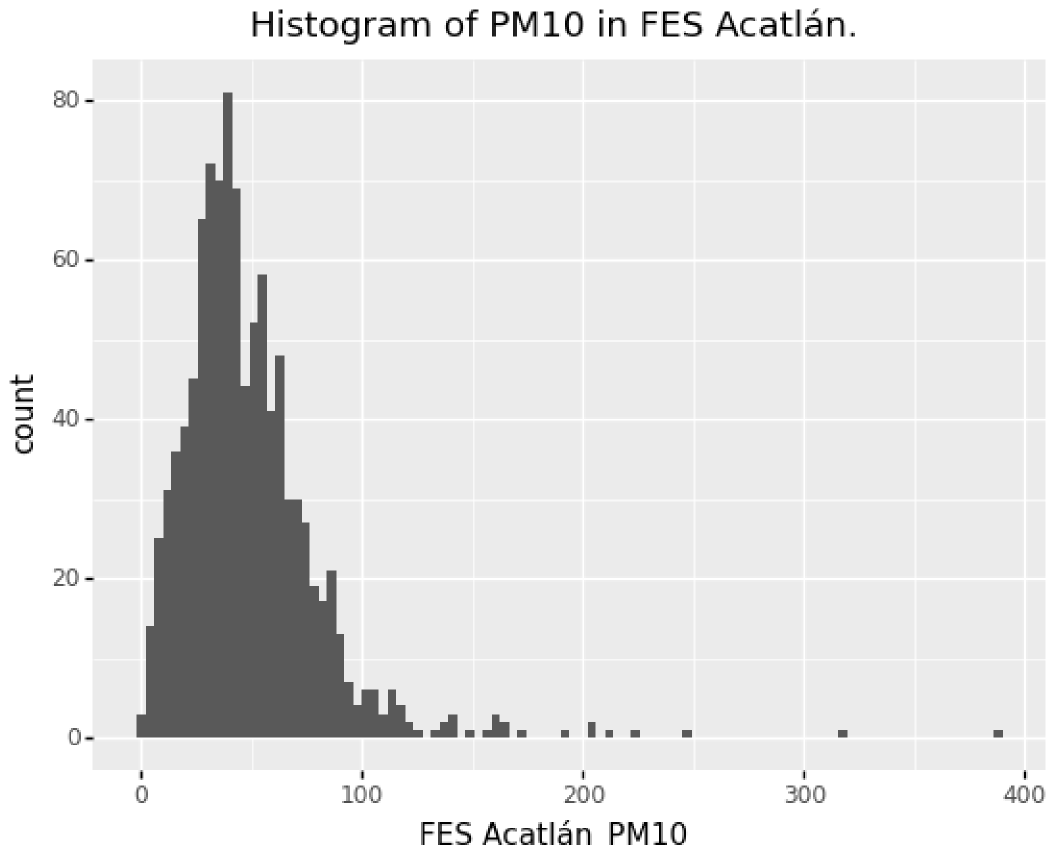

<ggplot: (8760301794865)>

/home/jaa6766/.conda/envs/cuda/lib/python3.7/site-packages/plotnine/layer.py:372: PlotnineWarning: stat_bin : Removed 8 rows containing non-finite values.



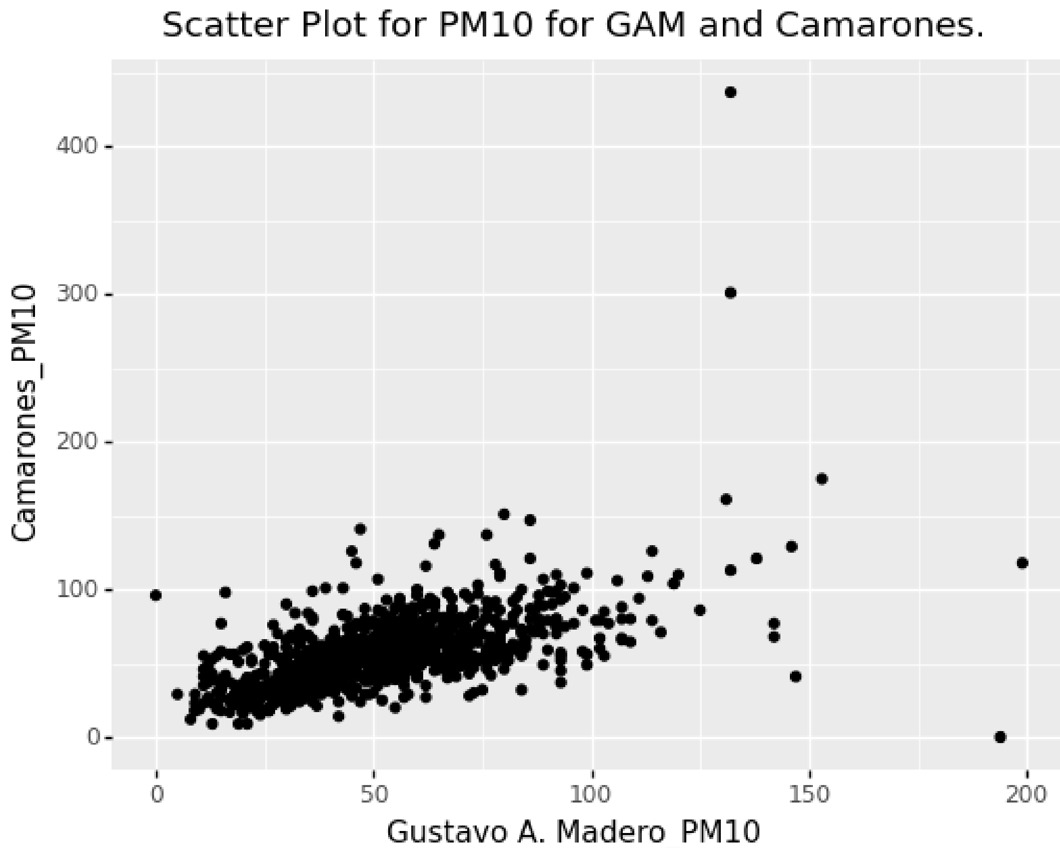

```
<ggplot: (8760308678081)>
```

```
/home/jaa6766/.conda/envs/cuda/lib/python3.7/site-
packages/plotnine/layer.py:372: PlotnineWarning: stat_bin : Removed 81 rows
containing non-finite values.
```

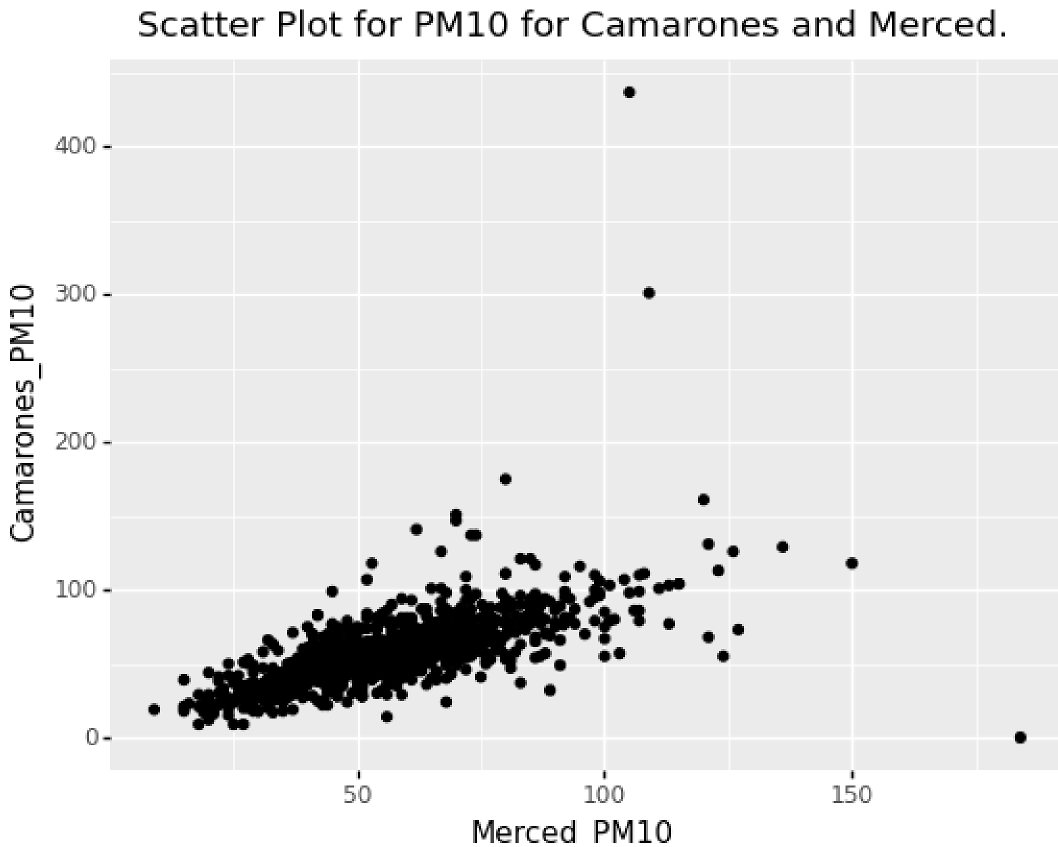



```
<ggplot: (8760301750401)>
```

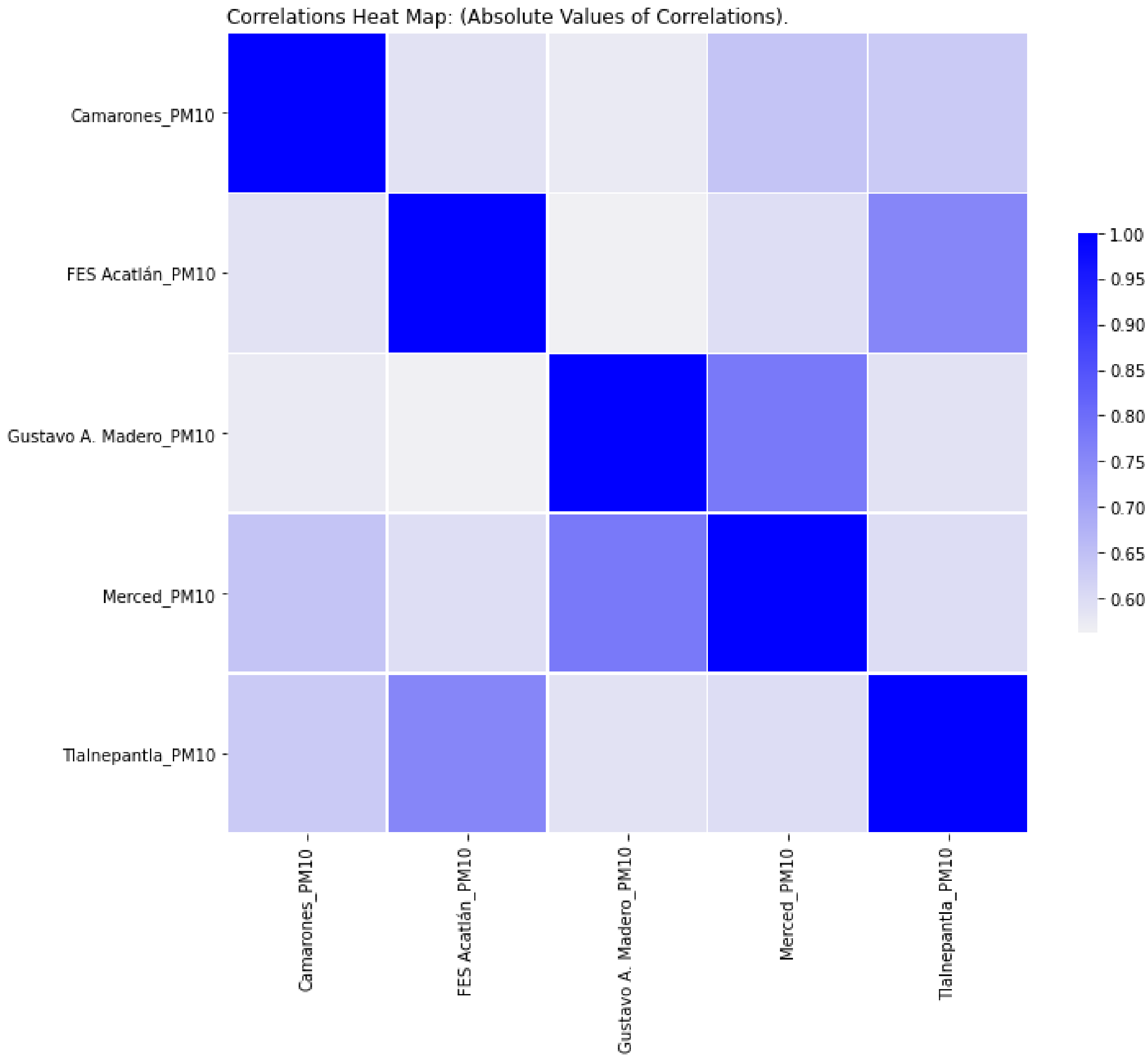
**Comparing Stations**




```
<ggplot: (8760311503345)>
```



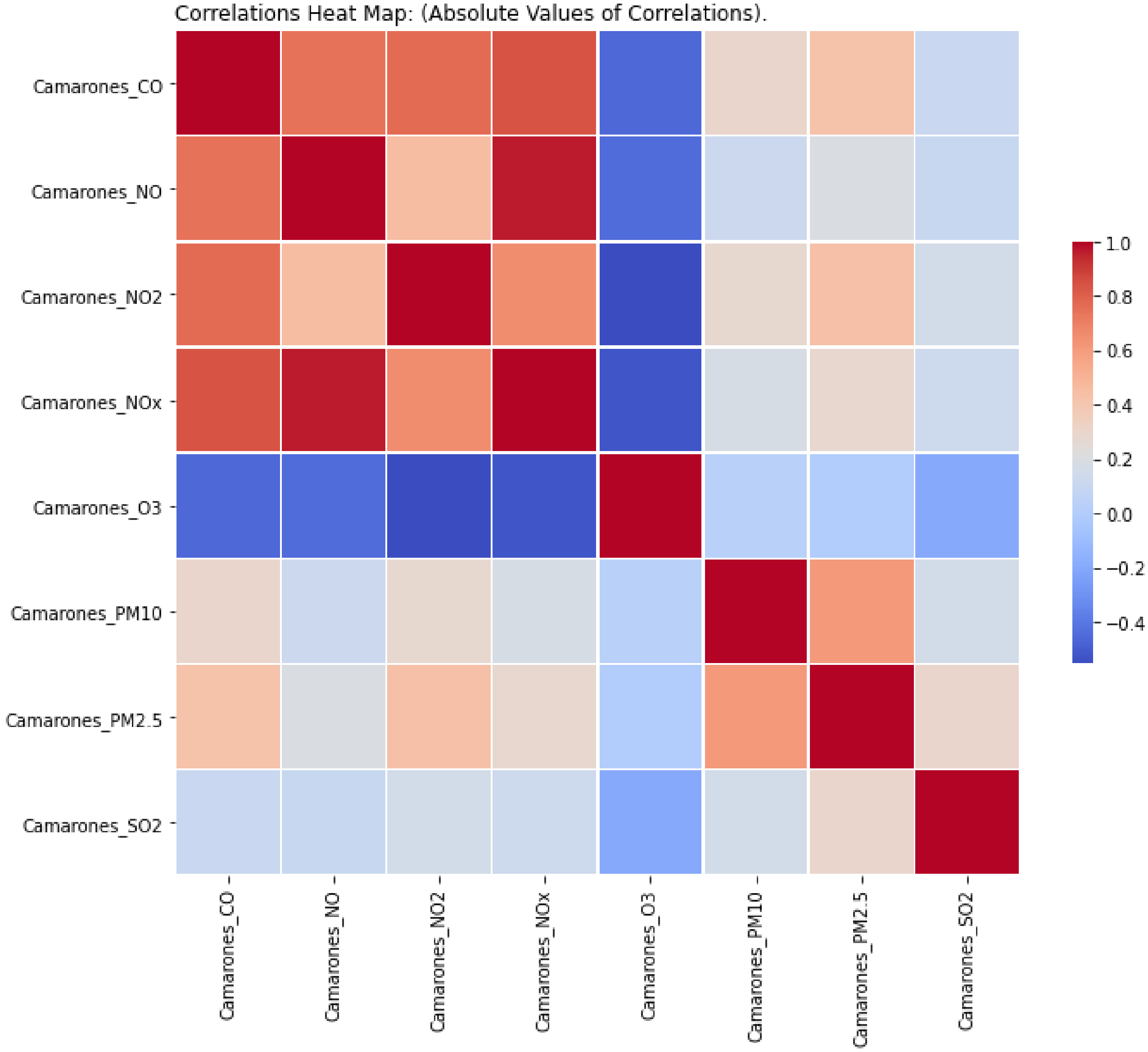

```
<ggplot: (8760312234929)>

                      Camarones_PM10  FES Acatlán_PM10  \
Camarones_PM10              1.000000          0.588032
FES Acatlán_PM10            0.588032          1.000000
Gustavo A. Madero_PM10      0.573777          0.561526
Merced_PM10                 0.642624          0.596102
Tlalnepantla_PM10           0.633015          0.755112

                      Gustavo A. Madero_PM10  Merced_PM10  Tlalnepantla_PM10
Camarones_PM10                      0.573777     0.642624           0.633015
FES Acatlán_PM10                    0.561526     0.596102           0.755112
Gustavo A. Madero_PM10              1.000000     0.779445           0.588039
Merced_PM10                         0.779445     1.000000           0.599155
Tlalnepantla_PM10                   0.588039     0.599155           1.000000
```



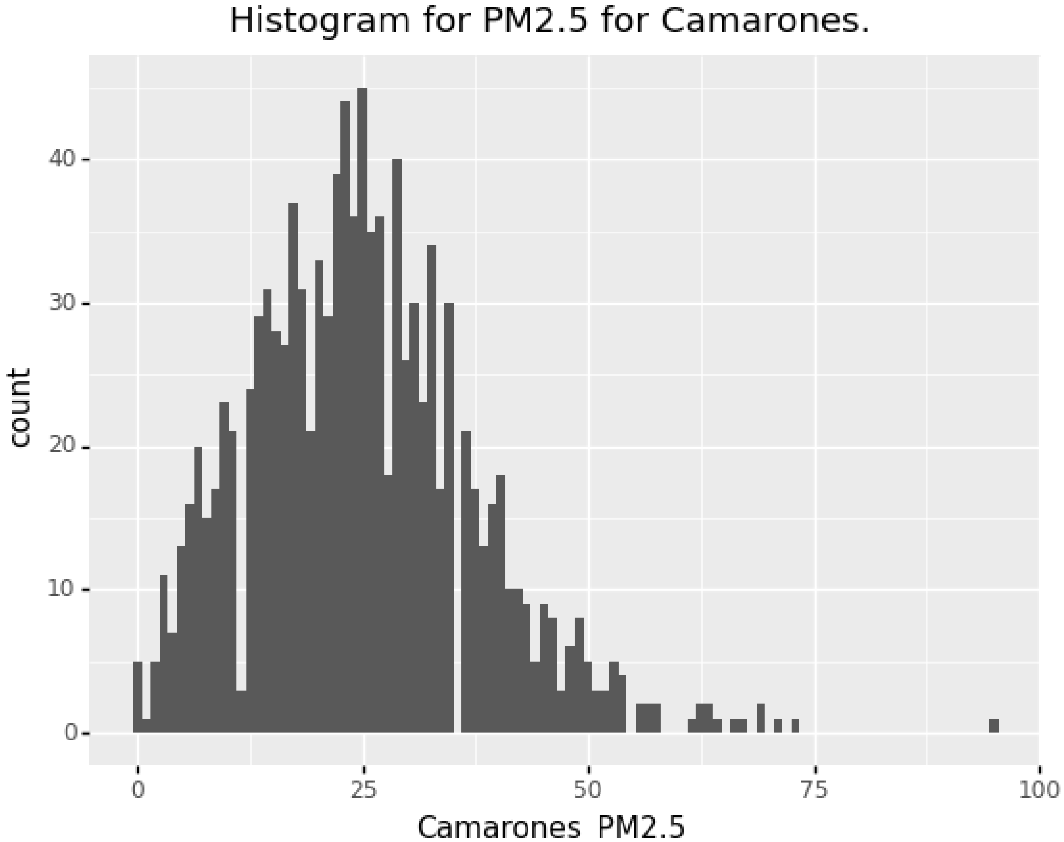

| | Camarones_CO | Camarones_NO | Camarones_NO2 | Camarones_NOx | \ |
|---|---|---|---|---|---|
| Camarones_CO | 1.000000 | 0.745593 | 0.771172 | 0.839178 | |
| Camarones_NO | 0.745593 | 1.000000 | 0.456594 | 0.965661 | |
| Camarones_NO2 | 0.771172 | 0.456594 | 1.000000 | 0.656534 | |
| Camarones_NOx | 0.839178 | 0.965661 | 0.656534 | 1.000000 | |
| Camarones_O3 | -0.455741 | -0.440542 | -0.554310 | -0.522438 | |
| Camarones_PM10 | 0.297277 | 0.124081 | 0.279483 | 0.181161 | |
| Camarones_PM2.5 | 0.425596 | 0.200851 | 0.435905 | 0.288952 | |
| Camarones_SO2 | 0.109512 | 0.105557 | 0.160720 | 0.133070 | |

| | Camarones_O3 | Camarones_PM10 | Camarones_PM2.5 | Camarones_SO2 |
|---|---|---|---|---|
| Camarones_CO | -0.455741 | 0.297277 | 0.425596 | 0.109512 |
| Camarones_NO | -0.440542 | 0.124081 | 0.200851 | 0.105557 |
| Camarones_NO2 | -0.554310 | 0.279483 | 0.435905 | 0.160720 |
| Camarones_NOx | -0.522438 | 0.181161 | 0.288952 | 0.133070 |
| Camarones_O3 | 1.000000 | 0.030864 | 0.003132 | -0.192407 |
| Camarones_PM10 | 0.030864 | 1.000000 | 0.610515 | 0.157361 |
| Camarones_PM2.5 | 0.003132 | 0.610515 | 1.000000 | 0.298078 |

```
Camarones_SO2        -0.192407        0.157361        0.298078        1.000000
```




**PM2.5**




```
<ggplot: (8760312112321)>
```

## 2.2 Regresión Lineal

Removemos observaciones incompletas para realizar la regresión.

```
Index(['Merced_PM10', 'Tlalnepantla_PM10'], dtype='object')
```

```
array([0.53026139, 0.36873538])
```

```
8.861413306258605
```

```
0.5089006183870153
```

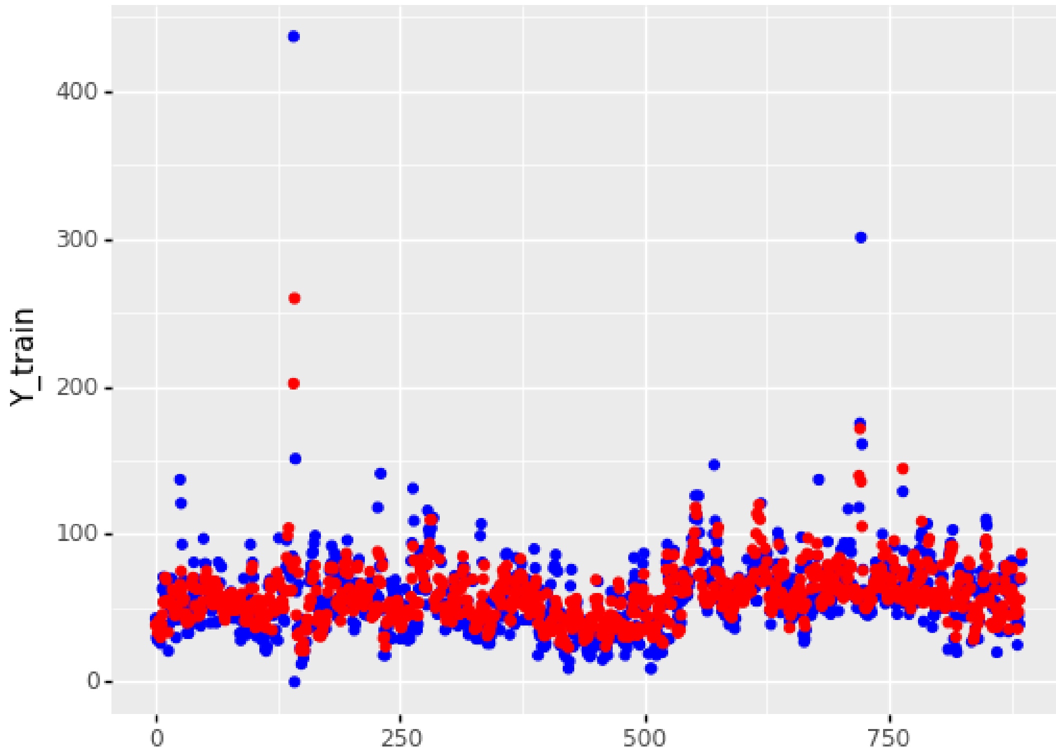
```
-0.4162199659052408
```

```
371.5820988359095
```

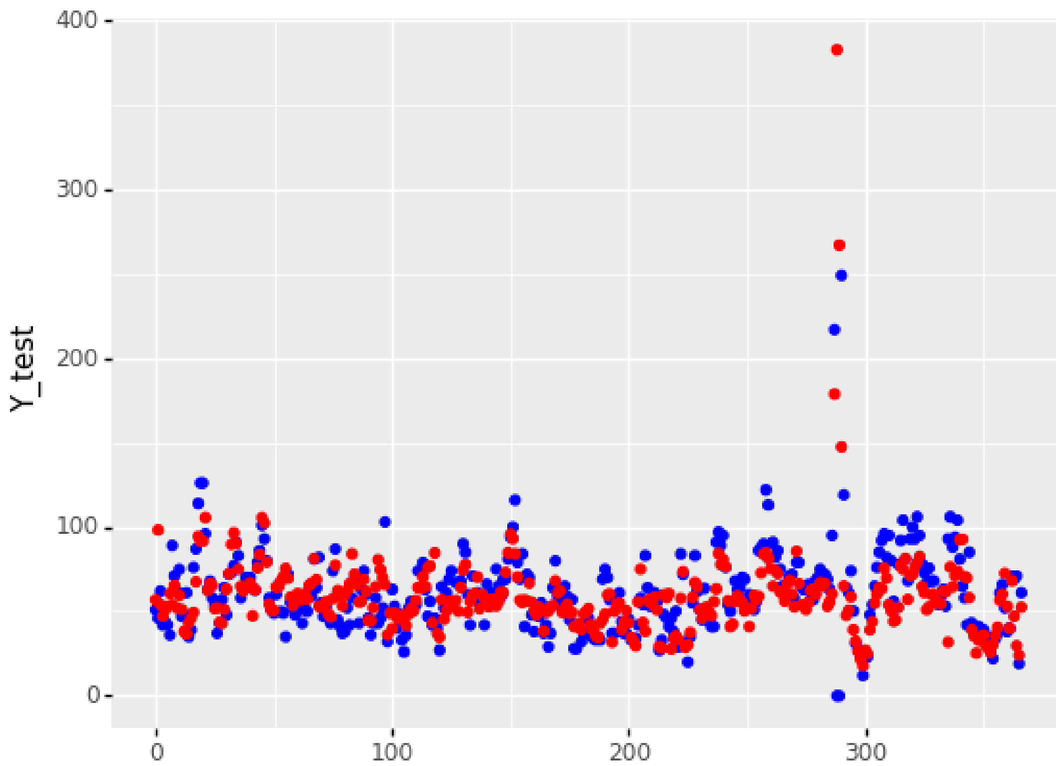
```
845.7794576326868
```

```
10.931473014049846
```

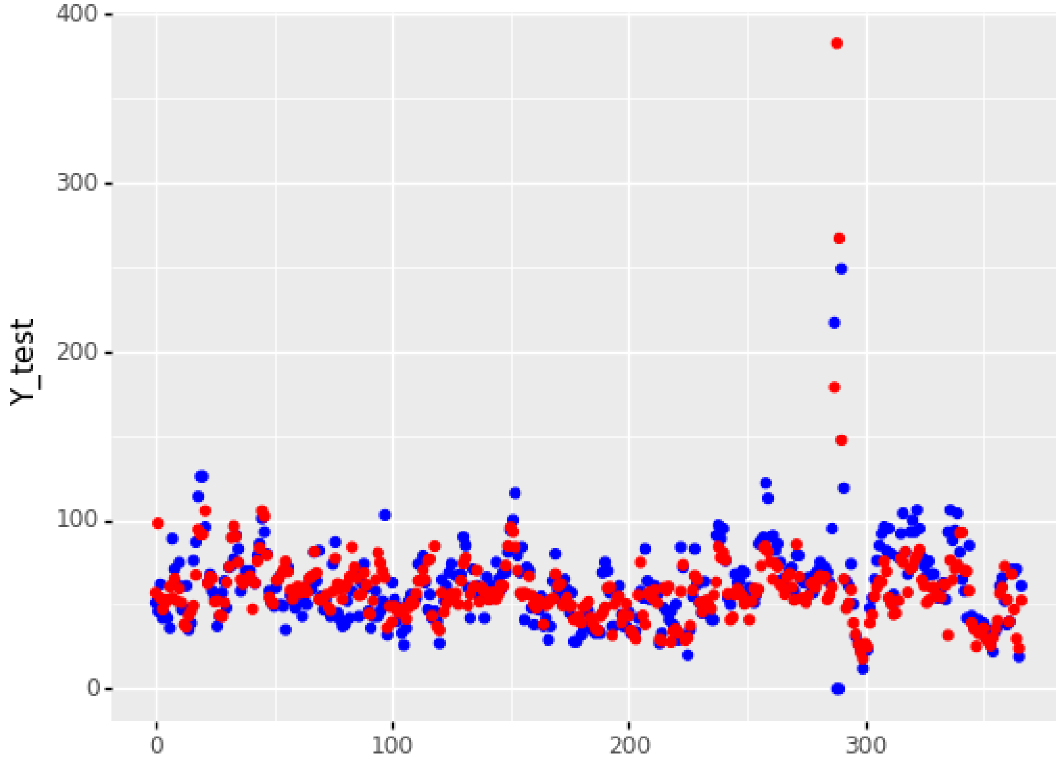
```
13.047752185462492
```




<ggplot: (8760311750473)>




<ggplot: (8760311667537)>




```
<ggplot: (8760311752613)>
```

## 2.3 Lasso

```
array([0.10494418, 0.0689564 , 0.43803925, 0.28309869])
```

```
9.793830115973797
```

```
0.5169119069776114
```

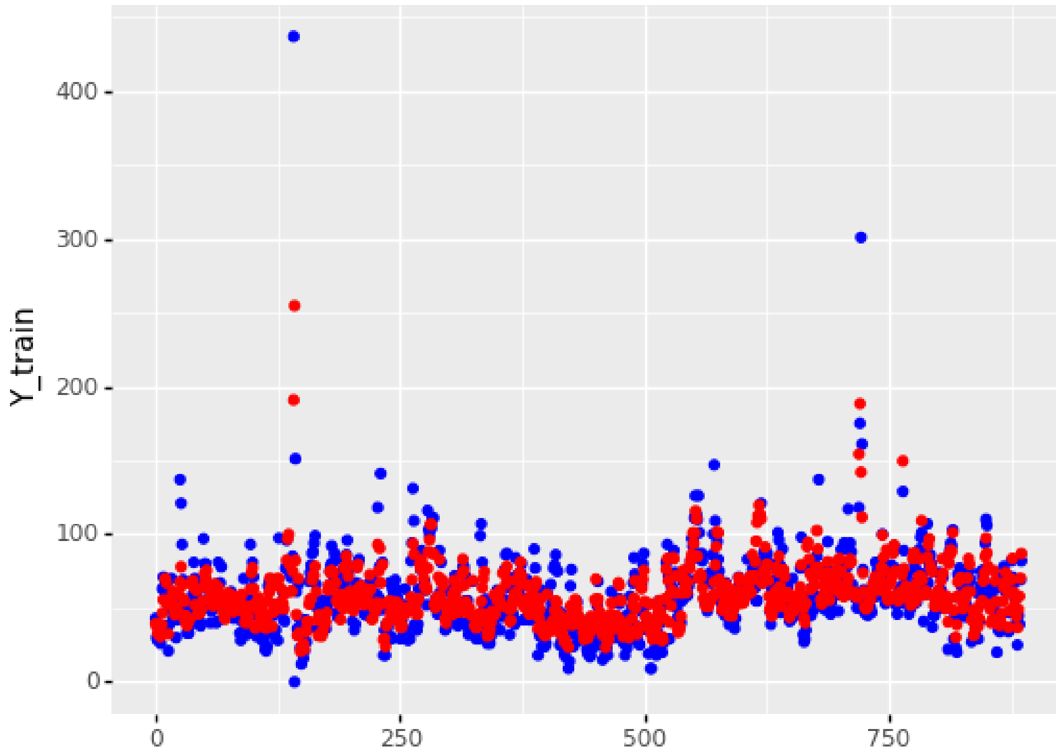
```
-0.11224862731990415
```

```
365.52049187746337
```

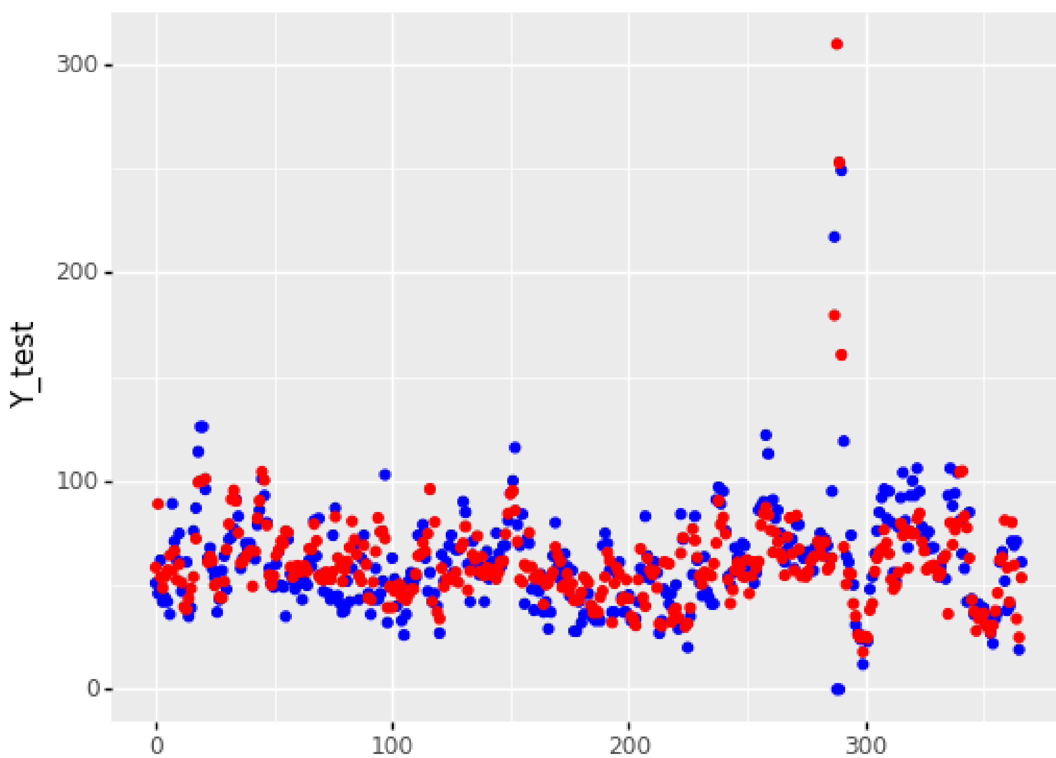
```
664.2450067183082
```

```
10.791035566246128
```

```
12.489847967051261
```




<ggplot: (8760311573617)>




<ggplot: (8760311673881)>

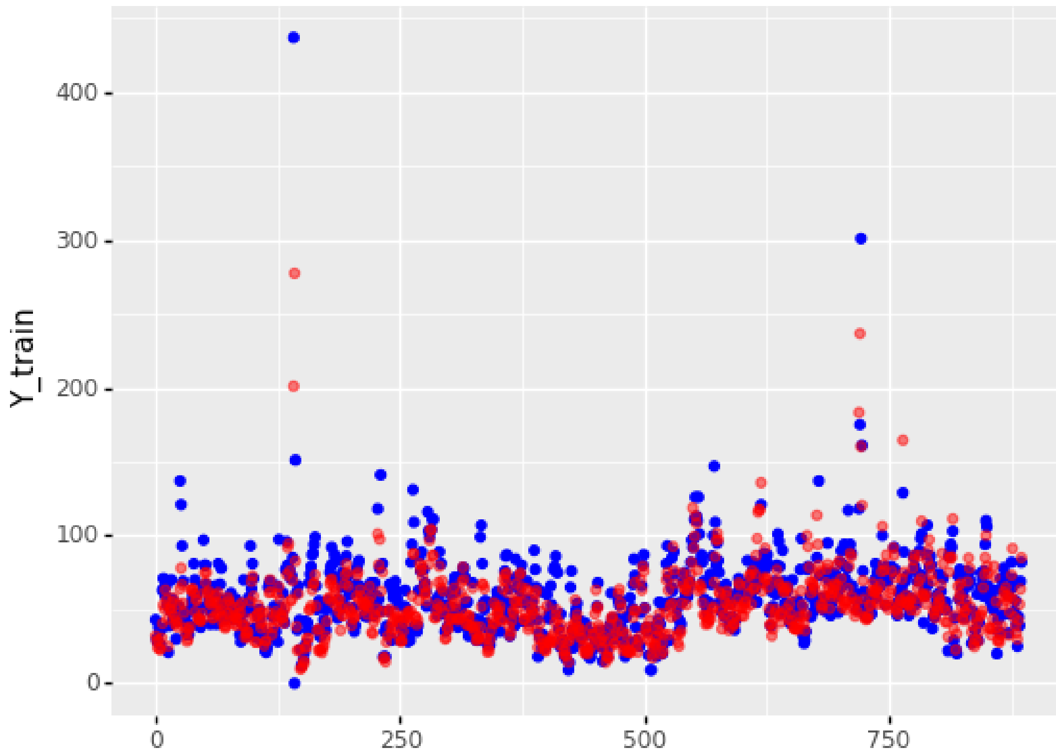
## 2.4 Mean

```
422.1441318590655
```

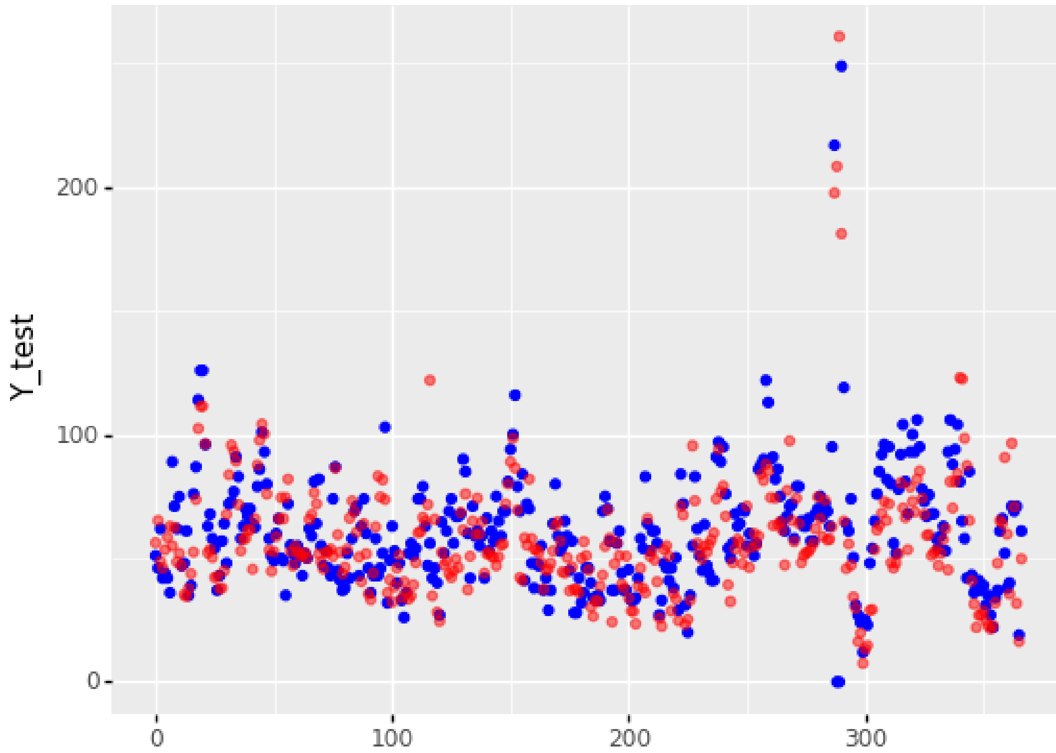
```
579.0378415931376
```

```
12.22969669074124
```

```
13.812958507896164
```




```
<ggplot: (8760311646597)>
```




<ggplot: (8760311419145)>

## 2.5 Generalized Linear Models: GLM

((886,), (886,))

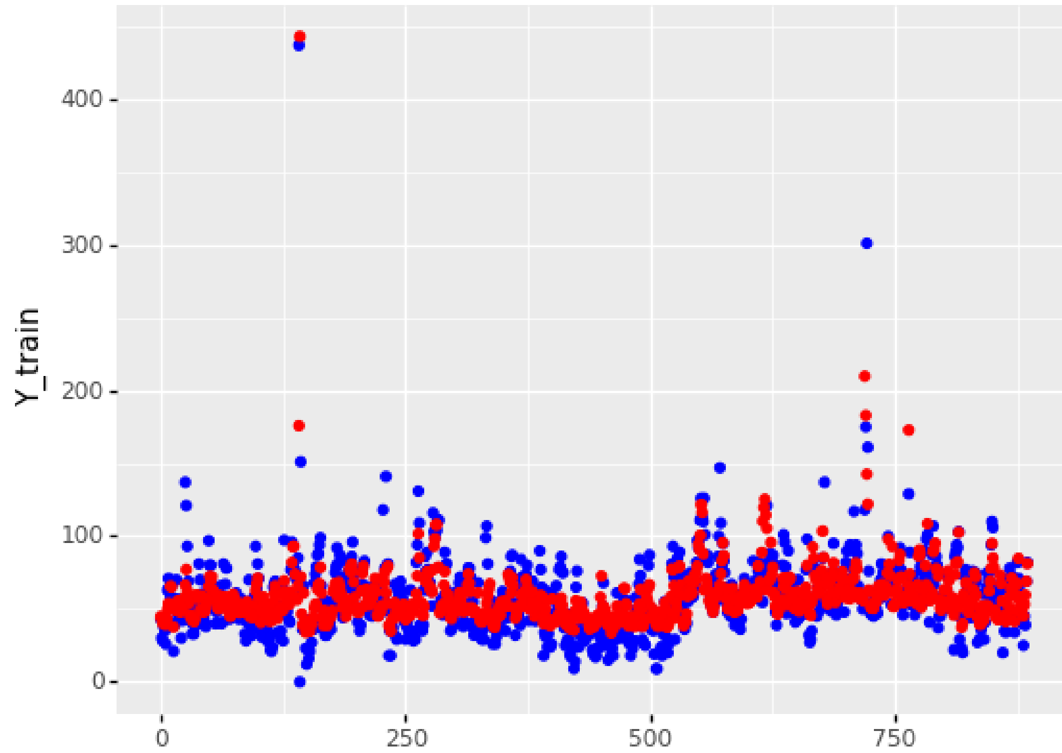
((367,), (367,))

562.8146939425658

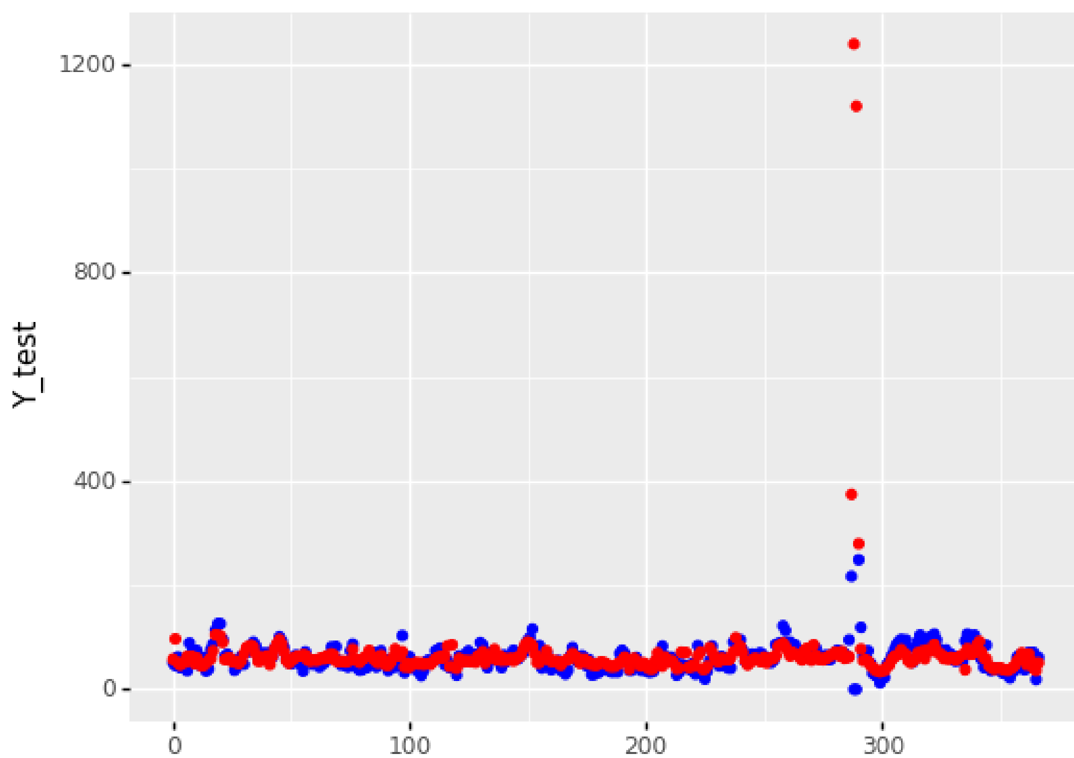
7875.849124574941

12.15806111211628

17.957714790509712




<ggplot: (8760311419265)>




<ggplot: (8760311437605)>

## 2.6 K-Nearest Neighbors

```
((886,), (886,))
```

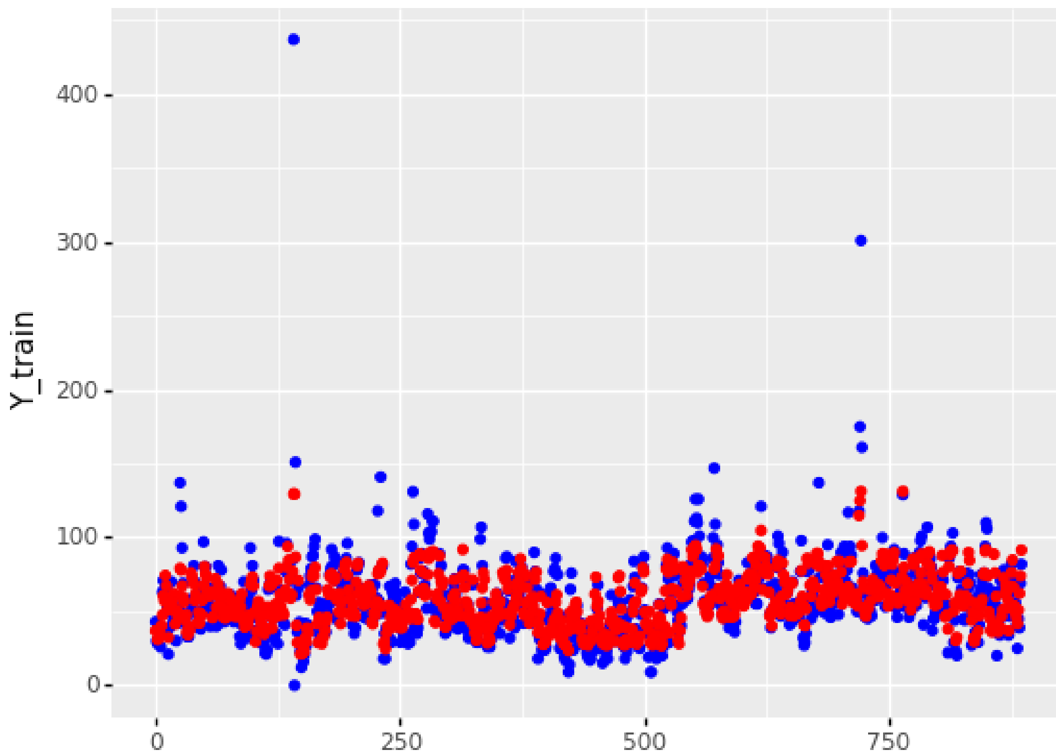
```
((367,), (367,))
```

```
347.29259904240126
```

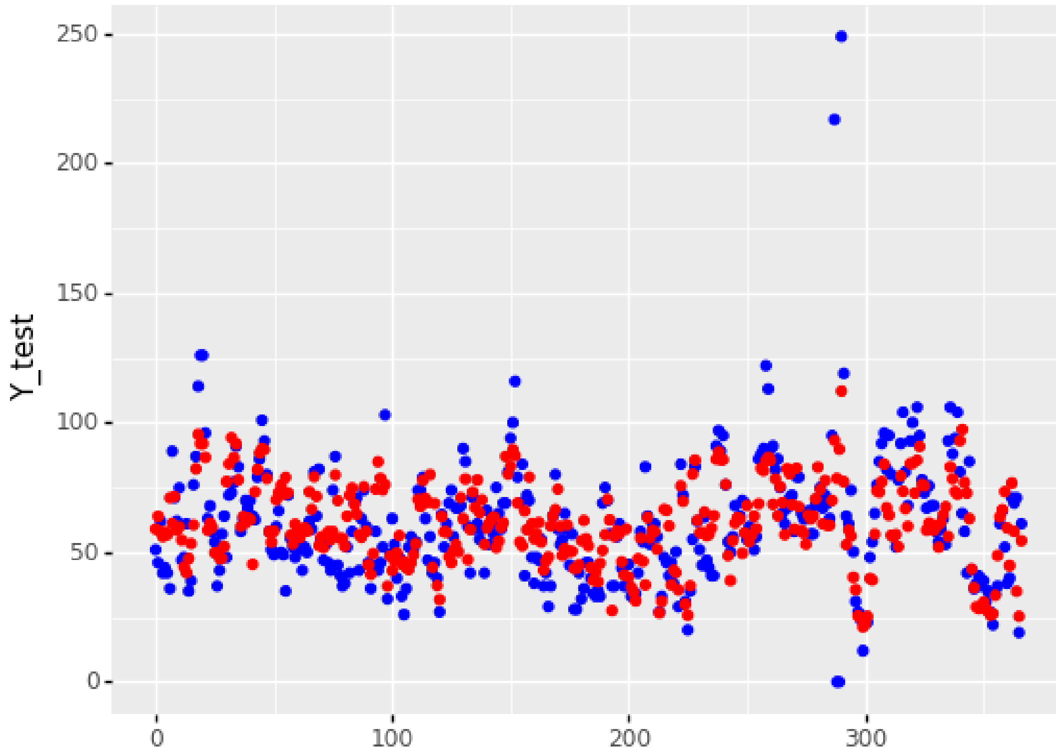
```
325.37804270928683
```

```
10.748652093480683
```

```
11.676590585214306
```




```
<ggplot: (8760311568649)>
```




```
<ggplot: (8760311448033)>
```

### 2.7 Evaluation

```
                   Model  MSE (Train Set)  MAE (Train Set)  MSE (Test Set)  \
4                1.7 KNN       347.292599        10.748652      325.378043
2              1.5 Media       422.144132        12.229697      579.037842
1              1.3 Lasso       365.520492        10.791036      664.245007
0  1.2 Regresión Lineal       371.582099        10.931473      845.779458
3                1.6 GLM       562.814694        12.158061     7875.849125

   MAE (Test Set)
4       11.676591
2       13.812959
1       12.489848
0       13.047752
3       17.957715
```

### 2.8 Early Conclusions

Given that we all further treatment to use the data should be in a sequential fashion, ie as timeseries: we found that linear interpolation is adequate.

Then in the next section we are detailing it.

### 2.9 Interpolation

We found in the EDA and in previous sections that Merced has similar data as Camarones, and it has fewer incomplete observations (missing data).

```
Merced_CO        70.0
Merced_NO       971.0
```

```
Merced_NO2       88.0
Merced_NOx       88.0
Merced_O3        82.0
Merced_PM10      31.0
Merced_PM2.5     36.0
Merced_SO2       64.0
Name: count, dtype: float64
```

Camarones has more incomplete observations:

```
Camarones_CO       111.0
Camarones_NO       125.0
Camarones_NO2      125.0
Camarones_NOx      310.0
Camarones_O3       119.0
Camarones_PM10     603.0
Camarones_PM2.5    587.0
Camarones_SO2      162.0
Name: count, dtype: float64
```

Using those results we create a new dataframe with those imputations.

The data in the dataframe have the following columns (vars) in the following manner:

- CO: Merced
- NO: Camarones
- NO2: Merced
- NOx: Merced
- O3: Merced
- PM10: Merced
- PM2.5: Merced

We used a lag to use the missing hours and found these gaps in the time line:

```
                   Fecha        CO        NO       NO2       NOx        O3  \
0    2021-01-01 02:00:00  1.100000  0.013000  0.032000  0.039000  0.004000
1    2021-01-01 03:00:00  1.200000  0.031000  0.033000  0.043000  0.001000
2    2021-01-01 04:00:00  1.200000  0.005000  0.031000  0.039000  0.002000
3    2021-01-01 05:00:00  1.200000  0.016000  0.028000  0.036000  0.002000
4    2021-01-01 06:00:00  1.400000  0.024000  0.029000  0.060000  0.001000
…                    …         …        …         …         …         …
2129 2021-10-04 00:00:00  0.545833  0.008292  0.019083  0.026875  0.015833
2130 2021-10-05 00:00:00  0.563158  0.010000  0.019722  0.030500  0.012278
2131 2021-10-06 00:00:00  0.672222  0.007571  0.026111  0.035611  0.011000
2132 2021-10-07 00:00:00  0.713636  0.011565  0.028636  0.040318  0.017909
2133 2021-10-08 00:00:00  0.758824  0.023778  0.029412  0.050588  0.017941

            PM10      PM2.5       SO2            datetime  year  month  day  \
0      37.000000  24.000000  0.003000 2021-01-01 02:00:00  2021      1    1
1      49.000000  39.000000  0.003000 2021-01-01 03:00:00  2021      1    1
2      80.000000  65.000000  0.003000 2021-01-01 04:00:00  2021      1    1
3      89.000000  75.000000  0.003000 2021-01-01 05:00:00  2021      1    1
4      75.000000  64.000000  0.003000 2021-01-01 06:00:00  2021      1    1
…            …          …        …                    …     …    …    …
2129   11.826087   7.913043  0.000750 2021-10-04 00:00:00  2021     10    4
2130   11.090909   6.772727  0.000556 2021-10-05 00:00:00  2021     10    5
2131   18.722222  11.833333  0.000111 2021-10-06 00:00:00  2021     10    6
2132   26.772727  17.000000  0.001045 2021-10-07 00:00:00  2021     10    7
2133   29.000000  17.705882  0.008176 2021-10-08 00:00:00  2021     10    8
```

```
      hour          datetime-1  delta  imputated
0        2 2021-01-01 00:00:00    2.0      False
1        3 2021-01-01 02:00:00    1.0      False
2        4 2021-01-01 03:00:00    1.0      False
3        5 2021-01-01 04:00:00    1.0      False
4        6 2021-01-01 05:00:00    1.0      False
…    …                  …    …        …
2129     0 2021-10-03 00:00:00    0.0      False
2130     0 2021-10-04 00:00:00    0.0      False
2131     0 2021-10-05 00:00:00    0.0      False
2132     0 2021-10-06 00:00:00    0.0      False
2133     0 2021-10-07 00:00:00    0.0      False

[2134 rows x 17 columns]
```

These are the missing gaps:

```
                 Date  Missing observations
0 2021-01-09 13:00:00                  13.0
1 2021-03-11 23:00:00                  12.0
2 2021-03-17 09:00:00                   9.0
3 2021-03-17 00:00:00                   8.0
4 2021-01-21 23:00:00                   8.0
5 2021-02-26 08:00:00                   8.0
6 2021-02-26 00:00:00                   7.0
7 2021-02-22 15:00:00                   6.0
8 2021-02-26 14:00:00                   6.0
9 2021-01-01 18:00:00                   5.0
```

Realizamos una interpolación quedando los datos así:

```
Skipping 0
CPU times: user 737 ms, sys: 3.47 ms, total: 740 ms
Wall time: 739 ms

               CO        NO       NO2       NOx        O3       PM10  \
987      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
988      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
989      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
990      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
991      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
…             …        …        …        …        …        …
1435385  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435386  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435387  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435388  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435389  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333

             PM2.5       SO2  month  day  hour                      datetime  \
987      22.000000  0.004000      2   12     6 2021-02-12 06:05:35.846304417
988      22.000000  0.004000      2   12     6 2021-02-12 06:05:38.837326527
989      22.000000  0.004000      2   12     6 2021-02-12 06:05:47.812360048
990      22.000000  0.004000      2   12     6 2021-02-12 06:05:50.803695202
991      22.000000  0.004000      2   12     6 2021-02-12 06:05:53.795462847
…              …         …     …    …    …                             …
1435385  51.041667  0.004292      4   24     0 2021-04-24 00:59:45.958476782
1435386  51.041667  0.004292      4   24     0 2021-04-24 00:59:48.954332352
1435387  51.041667  0.004292      4   24     0 2021-04-24 00:59:51.949791908
1435388  51.041667  0.004292      4   24     0 2021-04-24 00:59:54.945475817
1435389  51.041667  0.004292      4   24     0 2021-04-24 00:59:57.941107273
```

```
        minute  temperature  pressure  humidity  gasResistance    IAQ
987       35.0        21.51    777.41     44.04       152149.0   34.7
988       34.0        21.51    777.41     43.98       152841.0   33.6
989       32.0        21.54    777.41     43.73       153259.0   31.5
990       32.0        21.53    777.41     43.70       152841.0   31.5
991       30.0        21.52    777.41     43.70       153399.0   30.2
…          …           …        …         …              …      …
1435385  265.0        29.84    778.10     25.91       328422.0  264.8
1435386  265.0        29.87    778.12     25.86       327446.0  264.9
1435387  265.0        29.85    778.14     25.85       326799.0  265.0
1435388  265.0        29.83    778.14     25.81       328096.0  264.8
1435389  264.0        29.83    778.16     25.80       329076.0  264.5

[1434403 rows x 18 columns]
```

Hemos imputado con éxito todos los datos:

```
Empty DataFrame
Columns: [Fecha, Horas Faltantes]
Index: []
```

Reconocemos que tal vez no es el mejor método, pero requerimos más tiempo para explorar más sobre imputaciones en series de tiempo.

```
               CO        NO       NO2       NOx        O3       PM10  \
987      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
988      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
989      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
990      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
991      2.200000  0.205000  0.031000  0.207000  0.002000  45.000000
…             …         …         …         …         …          …
1435385  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435386  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435387  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435388  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333
1435389  1.245833  0.027667  0.034125  0.052833  0.050542  83.458333

             PM2.5       SO2  month  day  hour                      datetime  \
987      22.000000  0.004000      2   12     6 2021-02-12 06:05:35.846304417
988      22.000000  0.004000      2   12     6 2021-02-12 06:05:38.837326527
989      22.000000  0.004000      2   12     6 2021-02-12 06:05:47.812360048
990      22.000000  0.004000      2   12     6 2021-02-12 06:05:50.803695202
991      22.000000  0.004000      2   12     6 2021-02-12 06:05:53.795462847
…              …         …    …    …    …                             …
1435385  51.041667  0.004292      4   24     0 2021-04-24 00:59:45.958476782
1435386  51.041667  0.004292      4   24     0 2021-04-24 00:59:48.954332352
1435387  51.041667  0.004292      4   24     0 2021-04-24 00:59:51.949791908
1435388  51.041667  0.004292      4   24     0 2021-04-24 00:59:54.945475817
1435389  51.041667  0.004292      4   24     0 2021-04-24 00:59:57.941107273

        minute  temperature  pressure  humidity  gasResistance    IAQ
987       35.0        21.51    777.41     44.04       152149.0   34.7
988       34.0        21.51    777.41     43.98       152841.0   33.6
989       32.0        21.54    777.41     43.73       153259.0   31.5
990       32.0        21.53    777.41     43.70       152841.0   31.5
991       30.0        21.52    777.41     43.70       153399.0   30.2
…          …           …        …         …              …      …
1435385  265.0        29.84    778.10     25.91       328422.0  264.8
1435386  265.0        29.87    778.12     25.86       327446.0  264.9
```

```
1435387   265.0         29.85    778.14     25.85       326799.0  265.0
1435388   265.0         29.83    778.14     25.81       328096.0  264.8
1435389   264.0         29.83    778.16     25.80       329076.0  264.5

[1434403 rows x 18 columns]
```

## 2.10 References


- https://scikit-learn.org/stable/modules/linear_model.html#generalized-linear-regression
- https://pythonhealthcare.org/2018/05/03/81-distribution-fitting-to-data/
- https://medium.com/@amirarsalan.rajabi/distribution-fitting-with-python-scipy-bb70a42c0aed
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KernelDensity.html?highlight=kernel%20density#sklearn.neighbors.KernelDensity