Task 5: Generate a More Detailed Profile of the Data

***How sensibly you used the available statistics to profile the data;***

To automatize the profiling of data, created some stored procedures for some of the operations to create a table from Mondays to Sundays to view data more clearly and can find out patterns.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MEDIAN LQ | | Monday | Tuesday | Wednesday | Thursday | Fridays | Saturdays | Sunday |
| 0:00 | 1:00 | -39.0524 | -98.9613 | -44.9687 | -90.2275 | -142.052 | -29.2316 | -136.945 |
| 1:00 | 2:00 | -47.6683 | -205.589 | -185.96 | -174.819 | -57.2891 | -116.534 | -115.227 |
| 2:00 | 3:00 | -117.396 | -363.8 | -54.5605 | -151.782 | -159.51 | -154.005 | -96.7684 |
| 3:00 | 4:00 | -164.552 | -130.464 | -125.304 | -295.018 | -138.031 | -82.1074 | -70.0114 |
| 4:00 | 5:00 | -192.585 | -112.068 | -62.0419 | -329.188 | -135.669 | -136.496 | -117.433 |
| 5:00 | 6:00 | -145.675 | -64.4526 | -202.091 | -1616.91 | -161.644 | -111.374 | -173.477 |
| 6:00 | 7:00 | -166.036 | -129.357 | -16.9682 | -1841.92 | -356.065 | -115.925 | -183.875 |
| 7:00 | 8:00 | -64.6078 | -236.044 | -195.878 | -1335.48 | -364.821 | -105.472 | -229.444 |
| 8:00 | 9:00 | -90.076 | -52.7592 | -38.2264 | -294.941 | -133.316 | -99.9295 | -126.502 |
| 9:00 | 10:00 | -2.87735 | -63.6893 | 107.154 | -265.304 | -120.336 | -44.7852 | -131.581 |
| 10:00 | 11:00 | -40.4536 | -68.945 | 19.5547 | -297.998 | -112.495 | -80.682 | -83.027 |
| 11:00 | 12:00 | -44.8805 | -15.3078 | -144.947 | -242.231 | -53.5986 | -65.4793 | -64.8207 |
| 12:00 | 13:00 | -18.3193 | -47.0226 | 17.7833 | -162.427 | -104.579 | -66.5374 | -65.9119 |
| 13:00 | 14:00 | -72.7433 | -64.6715 | -95.0616 | -212.841 | -47.2569 | -88.5234 | -38.7591 |
| 14:00 | 15:00 | -58.2643 | -67.376 | -26.4293 | -139.024 | -21.8142 | -103.621 | -8.26082 |
| 15:00 | 16:00 | -23.5899 | -63.5833 | 36.107 | -144.048 | -16.2065 | -71.9629 | -49.1644 |
| 16:00 | 17:00 | -17.5752 | -16.4028 | -14.0102 | -133.18 | -21.8254 | -113.452 | -40.6147 |
| 17:00 | 18:00 | -92.8302 | -26.8649 | 41.8803 | -183.878 | -110.694 | -57.1362 | -23.7567 |
| 18:00 | 19:00 | -43.8691 | -60.0368 | 15.9371 | -326.745 | -93.9861 | -45.0893 | -72.9583 |
| 19:00 | 20:00 | -41.2661 | -5.00106 | -16.837 | -275.193 | -94.1113 | -70.6317 | -128.397 |
| 20:00 | 21:00 | -62.3706 | -139.363 | -242.405 | -153.403 | -25.3847 | -38.3872 | -120.575 |
| 21:00 | 22:00 | -104.019 | -22.5567 | -52.4717 | -402.437 | -57.9179 | -54.7217 | -44.3168 |
| 22:00 | 23:00 | -123.174 | -122.763 | -91.1713 | -260.258 | -83.9441 | -55.8242 | -31.0644 |
| 23:00 | 0:00:00 | -170.571 | -128.145 | -86.9533 | -464.129 | -71.9126 | -38.0602 | -42.4656 |

I think the median table give us a pretty good idea about the traffic in Manchester if we take into consideration that occupancy is given to us in LPU and 1 LPU = 17 vehicles, the color code explanation is as follows:

1. Red = Moderate Traffic
2. Yellow = Traffic Flows with the green light
3. Green = No traffic

As you can see with the color codes trying to stablish a connection between the median occupancy and the median LQ,

For example:

In the bracket of 20:00 to 21:00, the Median occupancy is always from 3 to 10 but the most common value is 7. However, when we see the median LQ all are negative values, but some are closer to 0 than others, having a broader spectrum and not creating a connection at all between them.

Is important to point out that for “Sunday”, the stored procedure is taking in consideration all the 5 Sundays relevant on our data, as well as every other day of the week.

***How well you were able to answer the questions described in Table 1.***

Answered most of them with our profiled data, here are some examples of questions

Which sensors are associated with the highest volumes of traffic?

Those are the sensors count when the LQ and BQ are huge and exit is not blocked. Also, the occupancy of the lanes is high.

What time of the day are the busiest for the city location covered by sensor N60311E?

I would say that the busiest hours are 15:00, 16:00, 17:00 for the sensor N60311E.

What range of traffic volume variation can you expect between 15:00 and 16:00 at the location covered by sensor N60311E?

Like in last query I would expect to increase 20% more at 16:00 since it the second busiest hour of the day.

Which days of the week are associated with the highest traffic volumes in general?

Friday, Thursday and Monday seem like the days with highest traffic volume.

Is January a busier month in terms of traffic volume than March?

January is a busier month than march.

Can you compare Occupancy patterns considering a month week day time of day of two different sensors?

Well, this query gave us 25000 records results because it joined all the times – days that these 2 sensors are not in the same area.

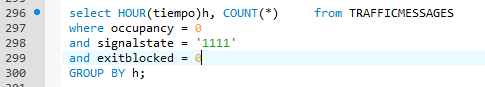
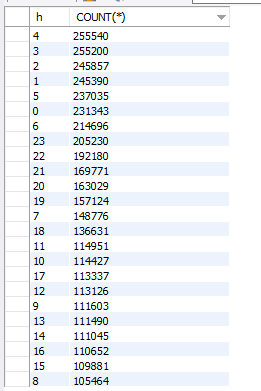
Can you find any faulty Sensor?

Some sensors sent half of the messages the other ones did, but It could be that they were newly installed. Also point out N60442A since it January marked blocked exit every day, every 10 minutes for intervals of 8 secs to 2 minutes.

Is average a Meaningful measure of traffic? Why or why not?

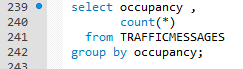
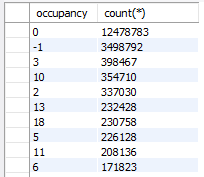
Average is not a meaningful measure of traffic, because it could be full in the morning and at night but very calm during the rest of the day, Average will indicate that the street is semi-busy all-day long.

What times of the day, considering any of the two months, is associated with the lowest traffic?



The early morning is when there is no traffic.

Can you generate the mode of given attribute (Such as occupancy)?

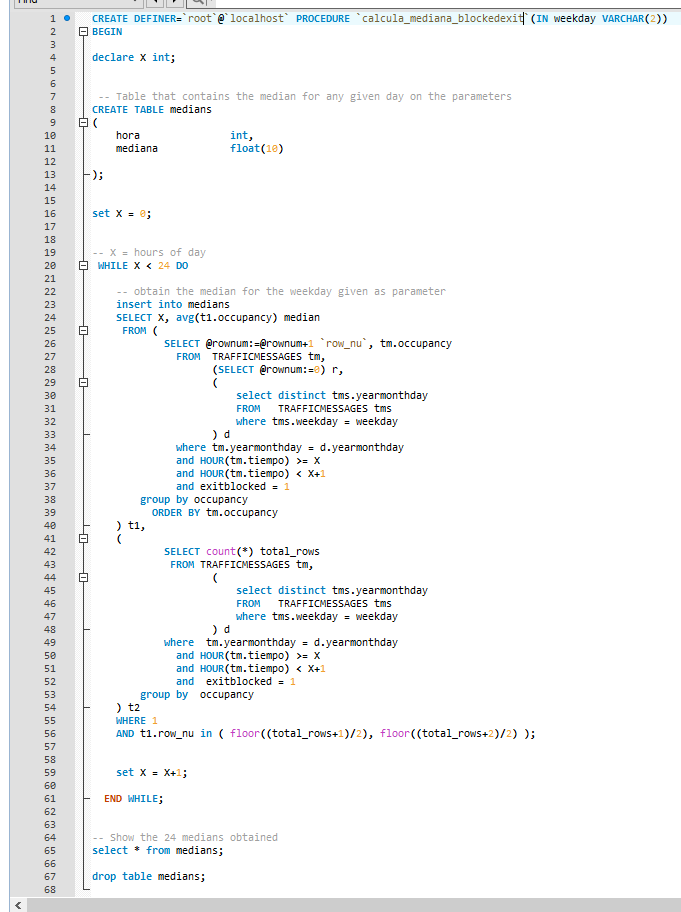
As we can see the mode is 0, the road is full of vehicles up to its capacity and traffic flows normally.

***How well you were able to create your own questions, explain their relevance in the context***

***of this application, and answer them.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Median Occupancy | | Monday | Tuesday | Wednesday | Thursday | Fridays | Saturdays | Sunday |
| 0:00 | 1:00 | - | - | 0 | - | - | - | 0 |
| 1:00 | 2:00 | - | - | - | - | - | -1 | - |
| 2:00 | 3:00 | - | - | - | - | - | - | - |
| 3:00 | 4:00 | - | - | - | - | - | - | - |
| 4:00 | 5:00 | - | - | - | - | - | - | - |
| 5:00 | 6:00 | - | - | - | - | - | - | - |
| 6:00 | 7:00 | - | - | 5.75 | - | - | - | - |
| 7:00 | 8:00 | 8.33333 | 4.44444 | 0.833333 | 0 | 1.25 | - | - |
| 8:00 | 9:00 | 14.8971 | 10.5303 | 6.47059 | 8.29851 | 4.14286 | - | - |
| 9:00 | 10:00 | 7.63636 | 1.60526 | 10.5938 | 0.272727 | 0 | 0 | - |
| 10:00 | 11:00 | 22.1176 | -1 | -1 | 0 | 9.21429 | 8.6 | 7.2 |
| 11:00 | 12:00 | 1.27273 | -1 | -1 | -1 | 17.5625 | 3.18182 | 22.1071 |
| 12:00 | 13:00 | 36.7174 | -1 | 4.33333 | -1 | 27.9615 | 11.75 | 20.1905 |
| 13:00 | 14:00 | 13.2727 | -1 | 0 | 1.2 | 40.2778 | 31.8519 | 4.375 |
| 14:00 | 15:00 | 0 | -0.6087 | -1 | 18.0606 | 17.1224 | 15.2353 | 31.76 |
| 15:00 | 16:00 | 0.705882 | -1 | -1 | 4.08824 | 24.1373 | 4.40541 | 0 |
| 16:00 | 17:00 | 8.12195 | 2.15385 | -0.91667 | 5.38889 | 44.7619 | 20.2174 | 7 |
| 17:00 | 18:00 | 6.70732 | -1 | 3.19355 | -0.76471 | 12.8644 | 19.6889 | - |
| 18:00 | 19:00 | 1.68182 | -1 | -1 | -0.65385 | 19.4444 | 10.6316 | 22 |
| 19:00 | 20:00 | 1.33333 | - | -1 | -1 | 20 | 24 | - |
| 20:00 | 21:00 | -1 | - | -1 | - | -1 | 0 | 3.2 |
| 21:00 | 22:00 | 3 | -1 | - | 0 | - | - | 0 |
| 22:00 | 23:00 | - | -1 | - | - | - | 0 | 11.3333 |
| 23:00 | 0:00:00 | 7.2 | 7.2 | - | - | - | - | - |

Created my own questions with some situations that could be interesting like when the exit blocked = 1, that means that the road is closed and depends of the time interval to make a good judgement of possible outcomes. So here is the median occupancy where the road is blocked.



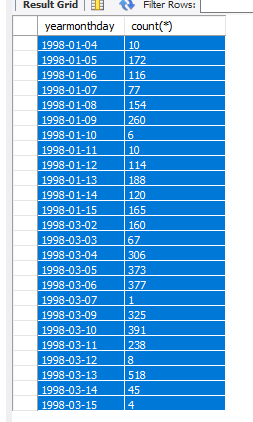
***How sensibly you were able to interpret the results you have obtained;***

From this table I could formulate a couple of questions:

I am not going to lie, the main issue since I saw the data was to answer what does a -1 in occupancy really mean?

1. An outlier? Well the block exit = 1 is an outlier, because it has no consistency and happens on different patterns randomly and most of the time, the intervals of which is blocked can’t be any of the situations that I tried to describe.
2. Complete Standstill? Cannot determine with this data the value of -1 in occupancy.
3. Faulty Sensor?

N60442A the first 15 days of January marked the road blocked every 10-20 minutes at least one or 20 intervals (8 seconds to 2 minutes maximum). Here the explanation is that must be faulty falling on our third category, because the same sensor did not mark a blocked road in March. Also, did not help us in the analysis of the -1 in occupancy, because sometimes it marked occupancy and sometimes it did not.



N60311G almost every day there is a road blocked = 1 reported by this sensor, that is on the image, the problem is that are more outliers. Every sensor has 21600 intervals or unique records every day. The maximum intervals that are covered by one day is 1998 - 03-13 with 518 intervals but they are intermittent and not sequential,

Other reasons for an exit blocked could be?

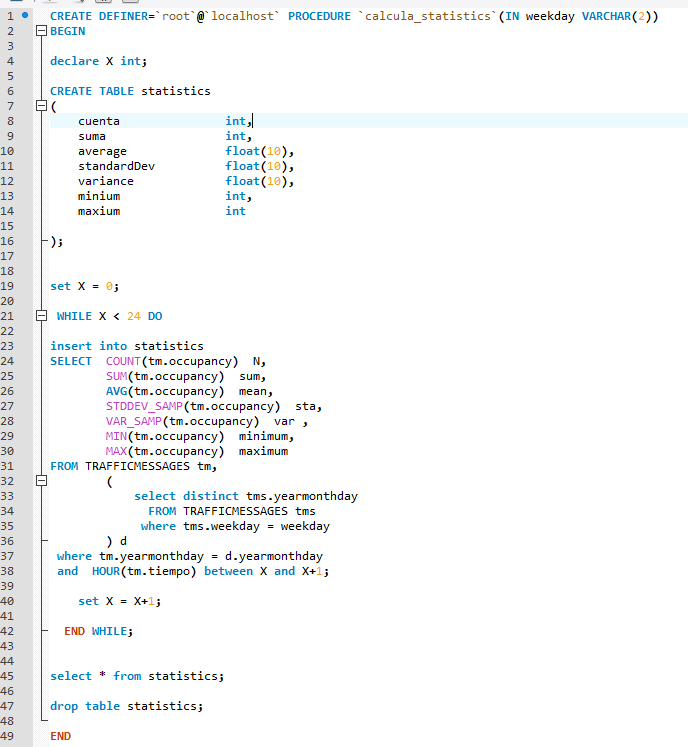
1. Road maintenance? Can’t determine with this data, but sounds very weird maintenance that last 2-3 consecutive minutes at best.
2. Accidents? Accidents that close a road for 2-3 consecutive minutes must be bicycles colliding at best.

Based that intervals of lectures are listed from 0 to 59 and they last 4 seconds each, so they cover ranges of 4 minutes of road.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SensorId | Monday’s | SensorId | Tuesday’s | SensorId | Wednesday’s | SensorId | Thursday’s | SensorId | Friday’s | SensorId | Saturday’s | SensorId | Sunday’s |
| N60442A | 923 | N60431E | 30 | N60442X | 32 | N60442A | 1758 | N60442A | 1403 | N60321J | 54 | N60442B | 28 |
| N60442B | 105 | N60331H | 21 | N60311H | 176 | N60442X | 22 | N60311H | 436 | N60311H | 80 | N60442C | 28 |
| N60331H | 9 | N60431F | 30 | N60331H | 2 | N60311H | 208 | N60321J | 331 | N60442A | 2286 | N60442X | 28 |
| N60442C | 105 | N60442A | 1414 | N60311G | 741 | N60431E | 68 | N60421B | 8 | N60311G | 52 | N60442A | 1028 |
| N60421C | 1 | N60321J | 380 | N60442A | 1670 | N60311G | 700 | N60442X | 21 | N60442B | 22 | N60311G | 24 |
| N60442X | 105 | N60442B | 10 | N60442B | 32 | N60431F | 68 | N60331H | 7 | N60442C | 22 | N60311H | 20 |
| N60421B | 1 | N60311G | 762 | N60442C | 32 | N60321J | 396 | N60321B | 8 | N60442X | 22 | N60321J | 4 |
| N60311G | 771 | N60442C | 10 | N60321J | 201 | N60431G | 68 | N60421A | 8 |  |  |  |  |
| N60421A | 2 | N60311H | 244 | N60421R | 4 | N60442B | 22 | N60331F | 8 |  |  |  |  |
| N60311H | 324 | N60442X | 10 | N60321B | 15 | N60431H | 69 | N60311G | 1155 |  |  |  |  |
| N60321J | 399 |  |  | N60321A | 15 | N60442C | 22 | N60321A | 8 |  |  |  |  |
| N60321B | 1 |  |  | N60331F | 4 | N60331H | 1 | N60421C | 8 |  |  |  |  |
| N60321A | 1 |  |  |  |  | N60321B | 40 | N60442B | 21 |  |  |  |  |
|  |  |  |  |  |  | N60321A | 40 | N60421R | 8 |  |  |  |  |
|  |  |  |  |  |  |  |  | N60442C | 21 |  |  |  |  |

***How much metadata (statistics) you have obtained;***

Created a stored procedure to obtain as much statistics as they could be, trying to find a good pattern to show my case.



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Monday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 298793 | 50125 | 0.167758 | 2.0324 | 4.13063 | -1 | 52 |
| 1:00 | 2:00 | 296763 | 11982 | 0.0403757 | 1.52772 | 2.33394 | -1 | 45 |
| 2:00 | 3:00 | 269776 | 9446 | 0.0350142 | 1.54076 | 2.37395 | -1 | 45 |
| 3:00 | 4:00 | 244802 | 35936 | 0.146796 | 2.05287 | 4.21428 | -1 | 60 |
| 4:00 | 5:00 | 244787 | 129826 | 0.530363 | 3.23369 | 10.4568 | -1 | 60 |
| 5:00 | 6:00 | 244785 | 332164 | 1.35696 | 5.10797 | 26.0914 | -1 | 66 |
| 6:00 | 7:00 | 244809 | 828908 | 3.38594 | 8.59007 | 73.7892 | -1 | 67 |
| 7:00 | 8:00 | 244809 | 1568069 | 6.40528 | 12.1136 | 146.74 | -1 | 89 |
| 8:00 | 9:00 | 244798 | 1845221 | 7.53773 | 13.0925 | 171.413 | -1 | 89 |
| 9:00 | 10:00 | 244794 | 1688257 | 6.89664 | 12.3674 | 152.951 | -1 | 86 |
| 10:00 | 11:00 | 244798 | 1687269 | 6.8925 | 12.2348 | 149.69 | -1 | 87 |
| 11:00 | 12:00 | 244802 | 1765050 | 7.21011 | 12.4536 | 155.092 | -1 | 87 |
| 12:00 | 13:00 | 244790 | 1773420 | 7.24466 | 12.457 | 155.176 | -1 | 82 |
| 13:00 | 14:00 | 244695 | 1722846 | 7.04079 | 12.2537 | 150.152 | -1 | 89 |
| 14:00 | 15:00 | 244702 | 1751911 | 7.15937 | 12.3455 | 152.412 | -1 | 89 |
| 15:00 | 16:00 | 226368 | 1668439 | 7.37047 | 12.5676 | 157.946 | -1 | 87 |
| 16:00 | 17:00 | 205994 | 1518582 | 7.37197 | 12.6303 | 159.524 | -1 | 82 |
| 17:00 | 18:00 | 224339 | 1391755 | 6.2038 | 11.5368 | 133.098 | -1 | 81 |
| 18:00 | 19:00 | 244711 | 1061602 | 4.33819 | 9.34295 | 87.2908 | -1 | 72 |
| 19:00 | 20:00 | 244799 | 782430 | 3.19621 | 7.7395 | 59.8998 | -1 | 79 |
| 20:00 | 21:00 | 244798 | 651319 | 2.66064 | 7.01455 | 49.2039 | -1 | 79 |
| 21:00 | 22:00 | 244784 | 488027 | 1.9937 | 6.1135 | 37.3749 | -1 | 72 |
| 22:00 | 23:00 | 244797 | 291958 | 1.19265 | 4.7488 | 22.5511 | -1 | 68 |
| 23:00 | 0:00:00 | 122407 | 102051 | 0.833702 | 3.92677 | 15.4196 | -1 | 56 |
|  |  | Tuesday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 244796 | 37461 | 0.153029 | 2.02675 | 4.10773 | -1 | 57 |
| 1:00 | 2:00 | 243089 | 13178 | 0.0542106 | 1.63656 | 2.67833 | -1 | 57 |
| 2:00 | 3:00 | 216089 | 9533 | 0.0441161 | 1.58709 | 2.51885 | -1 | 41 |
| 3:00 | 4:00 | 190796 | 20071 | 0.105196 | 1.87318 | 3.5088 | -1 | 52 |
| 4:00 | 5:00 | 190793 | 90496 | 0.474315 | 3.08141 | 9.4951 | -1 | 52 |
| 5:00 | 6:00 | 190794 | 256354 | 1.34362 | 5.09516 | 25.9607 | -1 | 58 |
| 6:00 | 7:00 | 190805 | 630870 | 3.30636 | 8.53725 | 72.8846 | -1 | 70 |
| 7:00 | 8:00 | 190801 | 1180658 | 6.1879 | 11.9233 | 142.166 | -1 | 90 |
| 8:00 | 9:00 | 190792 | 1377215 | 7.21841 | 12.7889 | 163.556 | -1 | 90 |
| 9:00 | 10:00 | 190793 | 1257595 | 6.59141 | 12.1058 | 146.551 | -1 | 89 |
| 10:00 | 11:00 | 190794 | 1246237 | 6.53185 | 12.0024 | 144.057 | -1 | 87 |
| 11:00 | 12:00 | 190800 | 1287915 | 6.75008 | 12.1727 | 148.174 | -1 | 82 |
| 12:00 | 13:00 | 190782 | 1307633 | 6.85407 | 12.2193 | 149.312 | -1 | 90 |
| 13:00 | 14:00 | 190781 | 1272246 | 6.66862 | 11.9992 | 143.98 | -1 | 90 |
| 14:00 | 15:00 | 190792 | 1295381 | 6.78949 | 12.0938 | 146.259 | -1 | 87 |
| 15:00 | 16:00 | 205475 | 1479481 | 7.2003 | 12.5068 | 156.42 | -1 | 86 |
| 16:00 | 17:00 | 232478 | 1743469 | 7.4995 | 12.8069 | 164.017 | -1 | 81 |
| 17:00 | 18:00 | 244796 | 1585878 | 6.47837 | 11.8882 | 141.328 | -1 | 77 |
| 18:00 | 19:00 | 244785 | 1087928 | 4.44442 | 9.53897 | 90.992 | -1 | 68 |
| 19:00 | 20:00 | 244787 | 811360 | 3.31456 | 7.92524 | 62.8095 | -1 | 68 |
| 20:00 | 21:00 | 244794 | 694629 | 2.83761 | 7.30069 | 53.3001 | -1 | 69 |
| 21:00 | 22:00 | 244797 | 519095 | 2.12051 | 6.28849 | 39.5451 | -1 | 69 |
| 22:00 | 23:00 | 244805 | 307318 | 1.25536 | 4.76441 | 22.6996 | -1 | 68 |
| 23:00 | 0:00:00 | 122397 | 114629 | 0.936534 | 4.1495 | 17.2184 | -1 | 68 |
|  |  | Wednesday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 244799 | 36625 | 0.149613 | 2.0032 | 4.01282 | -1 | 49 |
| 1:00 | 2:00 | 244792 | 5817 | 0.023763 | 1.50776 | 2.27333 | -1 | 49 |
| 2:00 | 3:00 | 244809 | 4345 | 0.0177485 | 1.49052 | 2.22165 | -1 | 49 |
| 3:00 | 4:00 | 244802 | 21333 | 0.0871439 | 1.81667 | 3.3003 | -1 | 50 |
| 4:00 | 5:00 | 244783 | 112792 | 0.460784 | 3.0592 | 9.35868 | -1 | 63 |
| 5:00 | 6:00 | 244799 | 343713 | 1.40406 | 5.30106 | 28.1013 | -1 | 65 |
| 6:00 | 7:00 | 244808 | 856874 | 3.50019 | 8.84594 | 78.2507 | -1 | 70 |
| 7:00 | 8:00 | 244796 | 1572895 | 6.42533 | 12.1253 | 147.023 | -1 | 85 |
| 8:00 | 9:00 | 244804 | 1771190 | 7.23514 | 12.6695 | 160.516 | -1 | 85 |
| 9:00 | 10:00 | 244805 | 1565779 | 6.39603 | 11.7035 | 136.973 | -1 | 84 |
| 10:00 | 11:00 | 244782 | 1581900 | 6.46248 | 11.7598 | 138.294 | -1 | 88 |
| 11:00 | 12:00 | 244780 | 1663050 | 6.79406 | 12.0878 | 146.116 | -1 | 88 |
| 12:00 | 13:00 | 221598 | 1521074 | 6.86411 | 12.1883 | 148.555 | -1 | 90 |
| 13:00 | 14:00 | 194598 | 1312388 | 6.7441 | 12.0675 | 145.626 | -1 | 90 |
| 14:00 | 15:00 | 190800 | 1302380 | 6.82589 | 12.128 | 147.087 | -1 | 86 |
| 15:00 | 16:00 | 190797 | 1354103 | 7.09709 | 12.3989 | 153.733 | -1 | 86 |
| 16:00 | 17:00 | 190789 | 1387592 | 7.27291 | 12.5986 | 158.725 | -1 | 78 |
| 17:00 | 18:00 | 190801 | 1202471 | 6.30223 | 11.7513 | 138.094 | -1 | 85 |
| 18:00 | 19:00 | 190799 | 863745 | 4.52699 | 9.71351 | 94.3523 | -1 | 85 |
| 19:00 | 20:00 | 190793 | 644829 | 3.37973 | 8.12093 | 65.9495 | -1 | 74 |
| 20:00 | 21:00 | 190806 | 549665 | 2.88075 | 7.44853 | 55.4806 | -1 | 74 |
| 21:00 | 22:00 | 190802 | 439816 | 2.30509 | 6.70596 | 44.9698 | -1 | 69 |
| 22:00 | 23:00 | 190796 | 261696 | 1.3716 | 5.10204 | 26.0308 | -1 | 68 |
| 23:00 | 0:00:00 | 95399 | 90847 | 0.952285 | 4.19424 | 17.5917 | -1 | 63 |
|  |  | Thursday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 190793 | 36003 | 0.188702 | 2.2009 | 4.84394 | -1 | 48 |
| 1:00 | 2:00 | 192996 | 17882 | 0.0926548 | 1.85601 | 3.44478 | -1 | 49 |
| 2:00 | 3:00 | 227193 | 34361 | 0.151241 | 2.07079 | 4.28818 | -1 | 51 |
| 3:00 | 4:00 | 259197 | 46965 | 0.181194 | 2.16408 | 4.68326 | -1 | 51 |
| 4:00 | 5:00 | 259207 | 102644 | 0.395992 | 2.90644 | 8.44737 | -1 | 62 |
| 5:00 | 6:00 | 259197 | 270303 | 1.04285 | 4.63197 | 21.4551 | -1 | 70 |
| 6:00 | 7:00 | 259205 | 644371 | 2.48595 | 7.58202 | 57.4871 | -1 | 70 |
| 7:00 | 8:00 | 258382 | 1193847 | 4.62047 | 10.7692 | 115.976 | -1 | 89 |
| 8:00 | 9:00 | 277537 | 1495440 | 5.38825 | 11.6604 | 135.965 | -1 | 89 |
| 9:00 | 10:00 | 305365 | 1526755 | 4.99977 | 11.0105 | 121.232 | -1 | 89 |
| 10:00 | 11:00 | 313199 | 1664272 | 5.31378 | 11.0631 | 122.392 | -1 | 90 |
| 11:00 | 12:00 | 313192 | 1806405 | 5.76772 | 11.3383 | 128.557 | -1 | 90 |
| 12:00 | 13:00 | 313203 | 1834695 | 5.85785 | 11.3977 | 129.907 | -1 | 88 |
| 13:00 | 14:00 | 311292 | 1832683 | 5.88734 | 11.4192 | 130.397 | -1 | 89 |
| 14:00 | 15:00 | 311216 | 1885660 | 6.05901 | 11.609 | 134.768 | -1 | 89 |
| 15:00 | 16:00 | 313129 | 1939612 | 6.19429 | 11.7796 | 138.759 | -1 | 88 |
| 16:00 | 17:00 | 313199 | 1941663 | 6.19945 | 11.8525 | 140.482 | -1 | 86 |
| 17:00 | 18:00 | 310768 | 1684930 | 5.42183 | 11.0306 | 121.674 | -1 | 78 |
| 18:00 | 19:00 | 310743 | 1275112 | 4.10343 | 9.21629 | 84.9399 | -1 | 78 |
| 19:00 | 20:00 | 287572 | 956946 | 3.32767 | 8.01779 | 64.285 | -1 | 81 |
| 20:00 | 21:00 | 253401 | 723105 | 2.8536 | 7.33849 | 53.8535 | -1 | 81 |
| 21:00 | 22:00 | 244795 | 534512 | 2.18351 | 6.3655 | 40.5196 | -1 | 69 |
| 22:00 | 23:00 | 244794 | 348882 | 1.42521 | 5.03173 | 25.3183 | -1 | 69 |
| 23:00 | 0:00:00 | 122401 | 147762 | 1.2072 | 4.61642 | 21.3114 | -1 | 62 |
|  |  | Friday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 244794 | 101705 | 0.415472 | 2.83416 | 8.03248 | -1 | 55 |
| 1:00 | 2:00 | 244810 | 71332 | 0.291377 | 2.46909 | 6.09642 | -1 | 51 |
| 2:00 | 3:00 | 244797 | 44868 | 0.183287 | 2.11766 | 4.48449 | -1 | 49 |
| 3:00 | 4:00 | 244793 | 36095 | 0.147451 | 1.99679 | 3.98716 | -1 | 43 |
| 4:00 | 5:00 | 244792 | 117836 | 0.481372 | 3.09494 | 9.57867 | -1 | 54 |
| 5:00 | 6:00 | 244790 | 291037 | 1.18893 | 4.76813 | 22.7351 | -1 | 63 |
| 6:00 | 7:00 | 244808 | 697370 | 2.84864 | 7.82525 | 61.2345 | -1 | 88 |
| 7:00 | 8:00 | 244007 | 1369862 | 5.61403 | 11.3566 | 128.973 | -1 | 88 |
| 8:00 | 9:00 | 243984 | 1689154 | 6.92322 | 12.5469 | 157.424 | -1 | 88 |
| 9:00 | 10:00 | 244786 | 1662666 | 6.79232 | 12.2723 | 150.61 | -1 | 88 |
| 10:00 | 11:00 | 244800 | 1783973 | 7.28747 | 12.6705 | 160.54 | -1 | 89 |
| 11:00 | 12:00 | 244791 | 1925631 | 7.86643 | 13.1804 | 173.722 | -1 | 89 |
| 12:00 | 13:00 | 244786 | 1973428 | 8.06185 | 13.3253 | 177.563 | -1 | 90 |
| 13:00 | 14:00 | 244784 | 1964480 | 8.02536 | 13.2511 | 175.591 | -1 | 90 |
| 14:00 | 15:00 | 244788 | 2013308 | 8.2247 | 13.3844 | 179.142 | -1 | 90 |
| 15:00 | 16:00 | 244798 | 2044280 | 8.35089 | 13.4784 | 181.668 | -1 | 88 |
| 16:00 | 17:00 | 244801 | 1919492 | 7.84103 | 13.0436 | 170.134 | -1 | 84 |
| 17:00 | 18:00 | 244646 | 1544011 | 6.3112 | 11.6539 | 135.813 | -1 | 84 |
| 18:00 | 19:00 | 244617 | 1176633 | 4.8101 | 9.87727 | 97.5604 | -1 | 72 |
| 19:00 | 20:00 | 244774 | 1063958 | 4.3467 | 9.24147 | 85.4048 | -1 | 75 |
| 20:00 | 21:00 | 244811 | 917596 | 3.74818 | 8.61449 | 74.2094 | -1 | 86 |
| 21:00 | 22:00 | 244798 | 656703 | 2.68263 | 7.20292 | 51.8821 | -1 | 86 |
| 22:00 | 23:00 | 244798 | 549911 | 2.24639 | 6.55036 | 42.9072 | -1 | 87 |
| 23:00 | 0:00:00 | 122409 | 290293 | 2.3715 | 6.81521 | 46.4471 | -1 | 87 |
|  |  | Saturday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 244789 | 266754 | 1.08973 | 4.47351 | 20.0123 | -1 | 72 |
| 1:00 | 2:00 | 244797 | 201588 | 0.823491 | 3.87477 | 15.0138 | -1 | 88 |
| 2:00 | 3:00 | 244817 | 133776 | 0.546433 | 3.21884 | 10.3609 | -1 | 88 |
| 3:00 | 4:00 | 244813 | 57412 | 0.234514 | 2.24807 | 5.05384 | -1 | 56 |
| 4:00 | 5:00 | 244797 | 87463 | 0.357288 | 2.65137 | 7.02974 | -1 | 56 |
| 5:00 | 6:00 | 244787 | 170887 | 0.698105 | 3.54775 | 12.5865 | -1 | 60 |
| 6:00 | 7:00 | 244798 | 278520 | 1.13775 | 4.91066 | 24.1146 | -1 | 67 |
| 7:00 | 8:00 | 244807 | 601912 | 2.45872 | 7.68711 | 59.0916 | -1 | 87 |
| 8:00 | 9:00 | 244797 | 1080373 | 4.41334 | 10.1637 | 103.301 | -1 | 87 |
| 9:00 | 10:00 | 244787 | 1501824 | 6.13523 | 11.8051 | 139.36 | -1 | 88 |
| 10:00 | 11:00 | 244776 | 1833645 | 7.49111 | 13.0041 | 169.107 | -1 | 90 |
| 11:00 | 12:00 | 244784 | 1951477 | 7.97224 | 13.2938 | 176.725 | -1 | 90 |
| 12:00 | 13:00 | 244798 | 1901232 | 7.76653 | 12.9955 | 168.883 | -1 | 90 |
| 13:00 | 14:00 | 244787 | 1838791 | 7.5118 | 12.8073 | 164.027 | -1 | 90 |
| 14:00 | 15:00 | 244797 | 1773341 | 7.24413 | 12.5999 | 158.756 | -1 | 89 |
| 15:00 | 16:00 | 244799 | 1637742 | 6.69015 | 12.0419 | 145.006 | -1 | 88 |
| 16:00 | 17:00 | 244793 | 1433077 | 5.85424 | 11.1271 | 123.813 | -1 | 77 |
| 17:00 | 18:00 | 244804 | 1163949 | 4.75462 | 9.99254 | 99.8509 | -1 | 72 |
| 18:00 | 19:00 | 244787 | 969122 | 3.95904 | 8.96406 | 80.3544 | -1 | 86 |
| 19:00 | 20:00 | 244803 | 934610 | 3.8178 | 8.65715 | 74.9462 | -1 | 86 |
| 20:00 | 21:00 | 244796 | 795734 | 3.2506 | 8.00044 | 64.0071 | -1 | 84 |
| 21:00 | 22:00 | 244781 | 584675 | 2.38856 | 6.82156 | 46.5337 | -1 | 84 |
| 22:00 | 23:00 | 244821 | 535952 | 2.18916 | 6.55226 | 42.9321 | -1 | 89 |
| 23:00 | 0:00:00 | 122416 | 280598 | 2.29217 | 6.74235 | 45.4593 | -1 | 89 |
|  |  | Sunday | | | | | | |
|  |  | Count | Sum | Average | StandardDev | Variance | Minium | Maxium |
| 0:00 | 1:00 | 244778 | 290016 | 1.18481 | 4.73218 | 22.3935 | -1 | 75 |
| 1:00 | 2:00 | 246832 | 245318 | 0.993866 | 4.42285 | 19.5616 | -1 | 79 |
| 2:00 | 3:00 | 273848 | 190146 | 0.694349 | 3.69495 | 13.6526 | -1 | 79 |
| 3:00 | 4:00 | 298801 | 80859 | 0.270612 | 2.36571 | 5.59657 | -1 | 55 |
| 4:00 | 5:00 | 298801 | 44429 | 0.148691 | 1.92449 | 3.70365 | -1 | 53 |
| 5:00 | 6:00 | 298803 | 76051 | 0.254519 | 2.27943 | 5.19579 | -1 | 53 |
| 6:00 | 7:00 | 298771 | 32671 | 0.109351 | 2.62423 | 6.8866 | -1 | 62 |
| 7:00 | 8:00 | 298781 | 8340 | 0.0279134 | 3.2482 | 10.5508 | -1 | 62 |
| 8:00 | 9:00 | 298801 | 202570 | 0.677943 | 4.8051 | 23.0889 | -1 | 66 |
| 9:00 | 10:00 | 298796 | 679774 | 2.27504 | 7.40761 | 54.8727 | -1 | 72 |
| 10:00 | 11:00 | 298794 | 1136701 | 3.8043 | 9.13287 | 83.4093 | -1 | 72 |
| 11:00 | 12:00 | 298687 | 1328049 | 4.44629 | 9.78242 | 95.6957 | -1 | 72 |
| 12:00 | 13:00 | 298700 | 1344364 | 4.50072 | 9.83341 | 96.696 | -1 | 72 |
| 13:00 | 14:00 | 298791 | 1312727 | 4.39346 | 9.70325 | 94.1532 | -1 | 72 |
| 14:00 | 15:00 | 298774 | 1277536 | 4.27593 | 9.57304 | 91.6431 | -1 | 71 |
| 15:00 | 16:00 | 298809 | 1168174 | 3.90943 | 9.15716 | 83.8536 | -1 | 71 |
| 16:00 | 17:00 | 298806 | 1059868 | 3.54701 | 8.69818 | 75.6584 | -1 | 71 |
| 17:00 | 18:00 | 298787 | 1018205 | 3.4078 | 8.45265 | 71.4473 | -1 | 71 |
| 18:00 | 19:00 | 298792 | 917293 | 3.07001 | 7.80327 | 60.8909 | -1 | 69 |
| 19:00 | 20:00 | 298797 | 783397 | 2.62184 | 6.92677 | 47.9802 | -1 | 69 |
| 20:00 | 21:00 | 298785 | 749749 | 2.50933 | 6.84022 | 46.7886 | -1 | 79 |
| 21:00 | 22:00 | 298778 | 593433 | 1.9862 | 6.09543 | 37.1543 | -1 | 79 |
| 22:00 | 23:00 | 298805 | 366547 | 1.22671 | 4.66943 | 21.8036 | -1 | 63 |
| 23:00 | 0:00:00 | 149404 | 147515 | 0.987356 | 4.21957 | 17.8048 | -1 | 63 |

As far as I can tell with this data, the only statistic that could somehow be useful in my honest point of view is sum, remember I’m matching all the sensors and aggregating all Sundays in my data (2 from January and 3 from March) into Sunday field. I am sure if I somehow mix the statistics operators I can find relevant data, but as it is right now with the basic operations not being interconnected is hard to find anything relevant or that tell us clear pattern.

***How well you were able to associate your results and interpretations with issues in data***

***quality discussed during the lectures;***

Some of the issues we discussed during the lectures in class about outliers and incomplete data, could had applied if I did not cut those records before getting in my main table, regardless I do not think that If I included that data with any of the given techniques my results would had change. Sadly, I couldn’t find good patterns In this data regardless.

***How well you have combined attributes to build the profile of the data;***

I have combined the attributes that were given as primary key that exemplifies an interval in time for a given sensor id with their respective attributes of occupancy, LQ, BQ, EB and SS trying to create sense of the given data, as well to match the patterns that lead me to findings and deep analysis.

***How well you were able to describe your findings, in terms of depth of analysis, sufficient***

***detail and clarity of description.***

Describe your findings: I could find some traffic patterns with the median of the data, where it shows that some hours are busier than others taking into consideration that 17 vehicles are 1 PDU and that traffic is something that is somewhat predictable in some time intervals.

Depth of Analysis: When I was trying to stablish hypothesis, and get a pattern the flag that I selected, because created a suitable scenario to find out the value of -1 in occupancy. That could describe if it was an outlier or a traffic jam where no car could move, the data was too noisy and couldn’t find anything interesting.

Sufficient Detail: I think some of the flags given lacks sufficient detail, and some data were just added to create cross roads and confusion, like exit block really reflects nothing because if a road is closed for a maximum of 2 minutes in the 30 days we analyze it means there was not a full maintenance and no accidents at all on it.

***Task 6: Describe your Experiences with the Tools You Used (30 marks)***

Read about the suggested tools, play with a subset of them, so that you can comment on the

advantages and disadvantages of each.

***Which of the tools is best suited to cope with the amount of data you are working with?***

MySQL can cope with the amount of data we are working with, as long as you optimize your table and index the columns that you are going to perform your queries and is easy to back up and create dumps if needed. That is one of its strongest features and could be considered as main advantage over the other tools.

***Which is the least suited?***

When I was testing the tools, I found Open Refined as suitable tool to clean data and transform data, the transformations that I made work on Excel for these tasks were done easier on OR. You can filter, there are dedicated interfaces to find inconsistencies, you can trim trailing spaces with one click for instance. Based on heuristics OR can help you merge the same text value that is wrote in 3-4 different ways into a single group value. You can upload your cleaned data to a central database making it relational, so you can apply normal queries on it and functions that can help you while clearing data.

But the biggest disadvantage of this tool is that it struggles with volume, it cannot cope with couple of hundred thousand of records. And is web based, making it less flexible than desired, is a very good tool for cleaning data on the web and it could be very useful once working in the order of thousands of records not much more than ten thousand. So that is a red flag for this task, since it needs to interact with millions.

***Which tool allows you to profile your data with the least level of interation / effort from you so the profiling task is almost fully automated?***

I find MySQL having the edge over the rest of the offered tools, his biggest drawback is that it requires you to have knowledge and experience thinking on data sets, but if you have the programming skills and knowledge. It is going to be easy for you to automate the results and perform data cleans that are identical thus you can reverse engineer in case of an error, it can handle millions of records and if you optimize your tables and query plans, queries should take 15-30 seconds.

***Which one requires more effort from you to obtain a good profile of the data?***

Microsoft Excel main disadvantage is that requires a lot of effort to obtain a good profile of data, since everything is manual and must be done based on clicks, there are macros that can help you with some data cleaning but they are not designed for processing millions of registers, based on this task experience since the data files were fragmented, used it to quick scan the data with filters to recognize the cut or double records and once I found those to create the .csv to export the data to MySQL.

***What are the main steps for obtaining a good data profile using each of the tools?***

1. The first step was changing the extension of the files from .data to .csv to prepare them for some analysis in excel file by file.
2. Once you open them in excel, add filters on top of each column and find the cut and double records that could be in the data and filter by M14 content type.
3. Once you remove or complete the cut and double records you save the .csv and import it to a temporal table in MySQL.
4. Once in MySQL you apply the script that you prepared to solve the different inconsistencies cases and errors.
5. You add it to your main table, that is optimized for profile data by queries.

All things considered, all the tools that were offered might work but depend on the context of the work that we want to perform, some must be new developments oriented to perform data analysis on the web and others to perform on great chunks of data like ours, all of them have their advantages and disadvantages. So, in order to succeed you need to pick what is best suited to perform the task.