Basic Statistical Analysis for a Data Set

1. Data Collection

The data set under consideration is the distance I travelled daily in the first half of 2019, from 01/01/2019 to 26/06/2019. The tripmeter reading from my vehicle was noted down after I returned home from every trip. The readings were noted down as a 3-tuple as (Date, Tripmeter reading, Route) in a pocket diary. They were then digitized and tabulated using MS Excel.

1.1 Format

The data before pre-processing is presented in the file Trip1.pdf.

- The date is in date(dd)-month(mon) format. All dates except the first row are from the year 2019. 31-Dec-2018 is noted as the baseline.
- The tripmeter reading is recorded as 5-digit natural number. This is a variable (numerical value).
- The route column consists of mnemonics that represent the place I visited. These are attributes (non-numerical). YCM stands for Yuvaraja's College, Mysuru. UoM represents University of Mysore. All the other place markers are self-explanatory.

1.2 Pre-processing

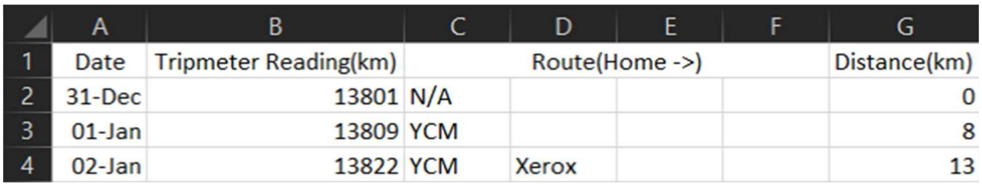
The cleaned data is presented in the file Trip2.pdf.

- Since on some days I did not go out, there are missing values in Trip1.pdf. These dates were later filled in, and the tripmeter reading was imputed using the previous row value, so that the distance travelled on that day is zero. To represent such days, filler attribute 'Nil' is used.
- On some days I've taken my car out more than once. Since the analysis is about the distance travelled without concerning the places I visited, multiple trips on a day have been combined to one, and route has been concatenated.

2. Data description and visualization

A portion of the cleaned table is shown below for representation purpose. Complete tables are available in the Moodle submission. The data points of interest are row3 to row140. row2 and row141 are only for reference and using them in computation leads to faulty conclusions. Microsoft Word, Excel are used primarily. The results were verified using python, the code and plot are attached in Code.pdf.

Table 1. A sample of time series of tripmeter readings along with route.



	A	B	C	D	E	F	G
1	Date	Tripmeter Reading(km)		Route(Home ->)			Distance(km)
2	31-Dec	13801	N/A				0
3	01-Jan	13809	YCM				8
4	02-Jan	13822	YCM	Xerox			13

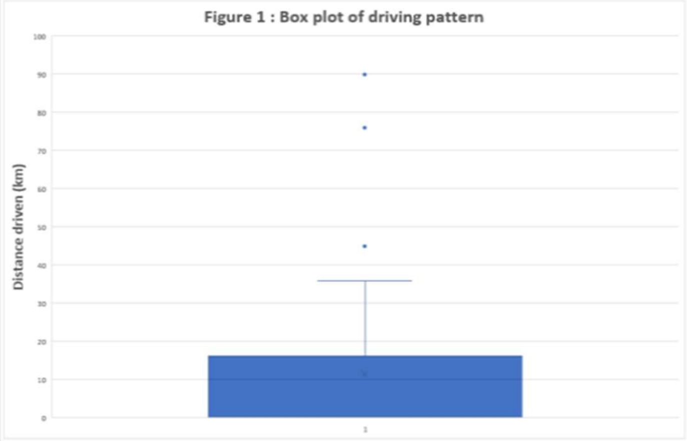


Figure 1 : Box plot of driving pattern

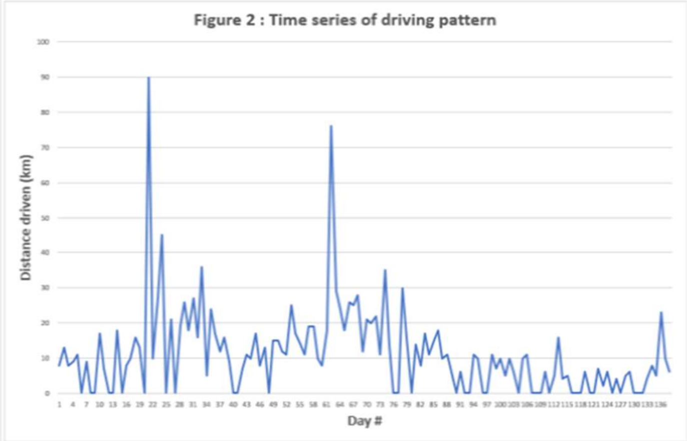


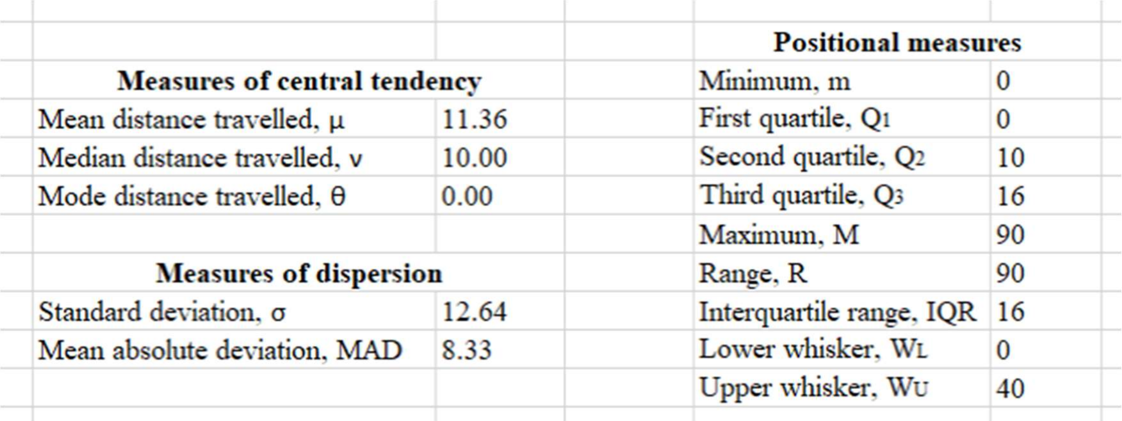
Figure 2 : Time series of driving pattern

3. Data summary

The summary statistics are tabulated and presented in the file Trip3.pdf.

Table 2. Summary statistics of distance driven

(All numerical values are in km)



			Positional measures	
Measures of central tendency			Minimum, m	0
Mean distance travelled, μ	11.36		First quartile, Q_1	0
Median distance travelled, ν	10.00		Second quartile, Q_2	10
Mode distance travelled, θ	0.00		Third quartile, Q_3	16
			Maximum, M	90
Measures of dispersion			Range, R	90
Standard deviation, σ	12.64		Interquartile range, IQR	16
Mean absolute deviation, MAD	8.33		Lower whisker, W_L	0
			Upper whisker, W_U	40

- IQR = $Q_3 – Q_1$; $W_L = Q_1 – 1.5$*IQR; $W_U = Q_3 + 1.5$*IQR. W_L is truncated to 0 as it is negative.

4. Inferences

Not every measure is meaningful for this data, so I've explained those of significance.

- Mean - On an average I've driven 11.36 km daily. This is somewhat ambiguous. This is correct due to the fact that on the days I've driven, 10 – 11 km is the most frequent trip length. This is faulty as there are large number of outliers - 36 days of no driving and some very long trips.
- Mode – I haven't taken my car out on "many" days - 36 to be exact. The next most frequent trips are 10 km and 11 km long. This is correct, as my college about 4.5 km away from my home. Also, during this period, I was enrolled in Spanish classes at University of Mysore and went there directly from college (YCM). The trip length is approximately 11 km.
- Over dispersion - There is huge variation in the data, which is evident from deviations which are comparable to mean. The index of dispersion or the Variance-to-Mean Ratio is VMR = 14.06.
- Quartiles – Minimum = Q_1, and Maximum >> Q_3, implying huge number of undriven days and extremities on the upper end.
- Outliers - From whiskers, values higher than W_U = 40, namely 45, 76, and 90 are outliers, with frequency one each. $W_L < 0$ and is truncated to 0 to validate positiveness of distance.
- Total – I've driven 1567 km in a span of 138 days.

5. Conclusion

The above analysis shows my daily driving pattern in 2019.

Future scope:

- Discarding data - It's better to use a trimmed data set removing the days on which I did not drive since we are interested in summary statistics of distance driven.
- Grouping - The data can be grouped on the basis of months and days to find monthly summary statistics.
- Dates can be matched with days of the week to find correlation between distance driven and day. For e.g. to check if I stay home on weekends and holidays or drive more than normal.

6. References

- Wikipedia
  - Central tendency - https://en.wikipedia.org/wiki/Central_tendency
  - Quartile - https://en.wikipedia.org/wiki/Quartile
- Pandas library - https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
- Numpy library - https://numpy.org/doc/

7. Appendix (Report end, the data is attached for reference/glance, not to be considered for evaluation)

Date	Tripmeter Reading(km)	Route(Home ->)		
31-Dec	13801	N/A		
01-Jan	13809	YCM		
02-Jan	13822	YCM	Xerox	
03-Jan	13830	YCM		
04-Jan	13839	YCM		
05-Jan	13850	YCM	City	
07-Jan	13859	YCM		
10-Jan	13874	YCM	UoM	
	13876	Petrol		
11-Jan	13883	UoM		
14-Jan	13901	YCM		
16-Jan	13909	YCM		
17-Jan	13919	YCM	Hostel	Xerox
18-Jan	13928	YCM		
	13935	UoM		
19-Jan	13948	YCM	Xerox	
21-Jan	14038	YCM	Roaming	
22-Jan	14048	Roaming		
23-Jan	14050	Petrol		
	14075	Roaming		
24-Jan	14110	YCM	UoM	
	14120	Roaming		
26-Jan	14141	Roaming		
28-Jan	14154	Roaming	YCM	
	14160	Roaming		
29-Jan	14186	YCM	Roaming	
30-Jan	14204	YCM		
31-Jan	14206	Petrol		
	14226	YCM		
	14231	Roaming		
01-Feb	14247	YCM		
02-Feb	14283	YCM	YCM	
03-Feb	14288	Roaming		
04-Feb	14300	Roaming		
	14312	UoM		
05-Feb	14329	YCM	UoM	
06-Feb	14341	YCM	UoM	
07-Feb	14350	YCM		
	14356	UoM		
	14357	Petrol		
08-Feb	14366	YCM		
11-Feb	14373	UoM		
12-Feb	14384	YCM	UoM	
13-Feb	14394	YCM	UoM	
14-Feb	14411	YCM	UoM	
15-Feb	14419	YCM		
16-Feb	14423	Roaming		
	14432	YCM		
18-Jan	14441	YCM		

	14447	Roaming		
19-Feb	14462	YCM		
20-Feb	14474	YCM	UoM	
21-Feb	14478	YCM		
	14485	YCM	UoM	
22-Feb	14498	YCM		
	14502	UoM		
	14508	Petrol		
	14510	UoM		
23-Feb	14522	YCM	UoM	
	14527	YCM		
24-Feb	14541	Roaming		
25-Feb	14552	YCM		
26-Feb	14564	YCM	Xerox	
	14571	UoM		
27-Feb	14580	YCM		
	14590	YCM	UoM	
28-Feb	14600	YCM	UoM	
01-Mar	14608	UoM		
02-Mar	14617	YCM		
	14626	YCM		
03-Mar	14662	Zoo		
	14700	Roaming		
	14702	Roaming		
04-Mar	14731	Roaming		
05-Mar	14748	YCM	UoM	
	14755	Petrol		
06-Mar	14773	YCM	UoM	
07-Mar	14799	YCM	UoM	
08-Mar	815	YCM		
	824	UoM		
09-Mar	836	YCM		
	852	Roaming		
10-Mar	860	petrol		
	864	Roaming		
11-Mar	885	UoM		
12-Mar	905	YCM	UoM	
13-Mar	927	YCM	UoM	
14-Mar	938	YCM	UoM	
15-Mar	961	YCM	UoM	
	973	UoM		
16-Mar	987	YCM		
19-Mar	15017	YCM	UoM	
20-Mar	20	Petrol		
	26	YCM		
	33	UoM		
22-Mar	47	YCM		
23-Mar	55	YCM		
24-Mar	65	YCM	Roaming	
	72	YCM	Roaming	

25-Mar	83	YCM	UoM	
26-Mar	88	YCM	Xerox	
	98	YCM	UoM	
27-Mar	107	YCM		
	114	Roaming	Roaming	
	116	Roaming		
28-Mar	126	YCM		
29-Mar	137	YCM		
30-Mar	15143	YCM		
01-Apr	15149	YCM		
04-Apr	15150	Petrol		
	15155	YCM		
	15160	UoM		
05-Apr	15166	YCM		
	15170	UoM		
08-Apr	15176	YCM		
	15181	UoM		
09-Apr	15186	UoM		
	15188	Roaming		
10-Apr	15194	YCM		
	15198	UoM		
11-Apr	15203	UoM		
12-Apr	15209	YCM		
	15213	UoM		
13-Apr	15219	YCM		
15-Apr	15225	YCM		
	15229	UoM		
16-Apr	15236	YCM		
	15240	UoM		
20-Apr	15246	YCM		
22-Apr	15251	UoM		
23-Apr	15257	Dentist		
	15262	Hospital		
	15267	UoM		
24-Apr	15271	UoM		
25-Apr	15276	UoM		
29-Apr	15282	YCM		
02-May	15289	YCM	UoM	
03-May	15291	Petrol		
04-May	15297	YCM		
06-May	15301	UoM		
08-May	15306	UoM		
09-May	15312	YCM		
13-May	15316	UoM		
14-May	15324	YCM		
15-May	15329	UoM		
16-May	15346	Roaming		
	15352	UoM		
17-May	15362	UoM		
18-May	15368	YCM	Crash	

26-Jun	15399	UoM		

Date	Tripmeter Reading(km)	Route(Home ->)				Distance(km)
31-Dec	13801	N/A				0
01-Jan	13809	YCM				8
02-Jan	13822	YCM	Xerox			13
03-Jan	13830	YCM				8
04-Jan	13839	YCM				9
05-Jan	13850	YCM	City			11
06-Jan	13850	Nil				0
07-Jan	13859	YCM				9
08-Jan	13859	Nil				0
09-Jan	13859	Nil				0
10-Jan	13876	YCM	UoM	Petrol		17
11-Jan	13883	UoM				7
12-Jan	13883	Nil				0
13-Jan	13883	Nil				0
14-Jan	13901	YCM				18
15-Jan	13901	Nil				0
16-Jan	13909	YCM				8
17-Jan	13919	YCM	Hostel	Xerox		10
18-Jan	13935	YCM	UoM			16
19-Jan	13948	YCM	Xerox			13
20-Jan	13948	Nil				0
21-Jan	14038	YCM	Roaming			90
22-Jan	14048	Roaming				10
23-Jan	14075	Petrol	Roaming			27
24-Jan	14120	YCM	UoM	Roaming		45
25-Jan	14120	Nil				0
26-Jan	14141	Roaming				21
27-Jan	14141	Nil				0
28-Jan	14160	Roaming	YCM	Roaming		19
29-Jan	14186	YCM	Roaming			26
30-Jan	14204	YCM				18
31-Jan	14231	Petrol	YCM	Roaming		27
01-Feb	14247	YCM				16
02-Feb	14283	YCM	YCM			36
03-Feb	14288	Roaming				5
04-Feb	14312	Roaming	UoM			24
05-Feb	14329	YCM	UoM			17
06-Feb	14341	YCM	UoM			12
07-Feb	14357	YCM	UoM	Petrol		16
08-Feb	14366	YCM				9
09-Feb	14366	Nil				0
10-Feb	14366	Nil				0
11-Feb	14373	UoM				7
12-Feb	14384	YCM	UoM			11
13-Feb	14394	YCM	UoM			10
14-Feb	14411	YCM	UoM			17
15-Feb	14419	YCM				8
16-Feb	14432	Roaming	YCM			13
17-Feb	14432	Nil				0

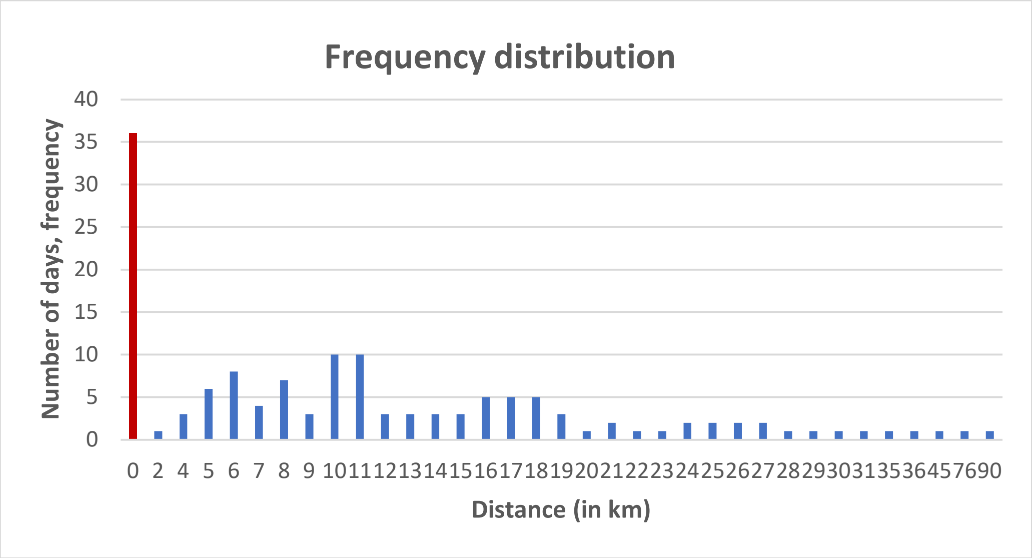
18-Jan	14447	YCM	Roaming			15
19-Feb	14462	YCM				15
20-Feb	14474	YCM	UoM			12
21-Feb	14485	YCM	YCM	UoM		11
22-Feb	14510	YCM	UoM	Petrol	UoM	25
23-Feb	14527	YCM	UoM	YCM		17
24-Feb	14541	Roaming				14
25-Feb	14552	YCM				11
26-Feb	14571	YCM	Xerox	UoM		19
27-Feb	14590	YCM	YCM	UoM		19
28-Feb	14600	YCM	UoM			10
01-Mar	14608	UoM				8
02-Mar	14626	YCM	YCM			18
03-Mar	14702	Zoo	Roaming	Roaming		76
04-Mar	14731	Roaming				29
05-Mar	14755	YCM	UoM	Petrol		24
06-Mar	14773	YCM	UoM			18
07-Mar	14799	YCM	UoM			26
08-Mar	14824	YCM	UoM			25
09-Mar	14852	YCM	Roaming			28
10-Mar	14864	petrol	Roaming			12
11-Mar	14885	UoM				21
12-Mar	14905	YCM	UoM			20
13-Mar	14927	YCM	UoM			22
14-Mar	14938	YCM	UoM			11
15-Mar	14973	YCM	UoM	UoM		35
16-Mar	14987	YCM				14
17-Mar	14987	Nil				0
18-Mar	14987	Nil				0
19-Mar	15017	YCM	UoM			30
20-Mar	15033	Petrol	YCM	UoM		16
21-Mar	15033	Nil				0
22-Mar	15047	YCM				14
23-Mar	15055	YCM				8
24-Mar	15072	YCM	Roaming	YCM	Roaming	17
25-Mar	15083	YCM	UoM			11
26-Mar	15098	YCM	Xerox	YCM	UoM	15
27-Mar	15116	YCM	Roaming	Roaming	Roaming	18
28-Mar	15126	YCM				10
29-Mar	15137	YCM				11
30-Mar	15143	YCM				6
31-Mar	15143	Nil				0
01-Apr	15149	YCM				6
02-Apr	15149	Nil				0
03-Apr	15149	Nil				0
04-Apr	15160	Petrol	YCM	UoM		11
05-Apr	15170	YCM	UoM			10
06-Apr	15170	Nil				0
07-Apr	15170	Nil				0
08-Apr	15181	YCM	UoM			11

09-Apr	15188	UoM	Roaming			7
10-Apr	15198	YCM	UoM			10
11-Apr	15203	UoM				5
12-Apr	15213	YCM	UoM			10
13-Apr	15219	YCM				6
14-Apr	15219	Nil				0
15-Apr	15229	YCM	UoM			10
16-Apr	15240	YCM	UoM			11
17-Apr	15240	Nil				0
18-Apr	15240	Nil				0
19-Apr	15240	Nil				0
20-Apr	15246	YCM				6
21-Apr	15246	Nil				0
22-Apr	15251	UoM				5
23-Apr	15267	Dentist	Hospital	UoM		16
24-Apr	15271	UoM				4
25-Apr	15276	UoM				5
26-Apr	15276	Nil				0
27-Apr	15276	Nil				0
28-Apr	15276	Nil				0
29-Apr	15282	YCM				6
30-Apr	15282	Nil				0
01-May	15282	Nil				0
02-May	15289	YCM	UoM			7
03-May	15291	Petrol				2
04-May	15297	YCM				6
05-May	15297	Nil				0
06-May	15301	UoM				4
07-May	15301	Nil				0
08-May	15306	UoM				5
09-May	15312	YCM				6
10-May	15312	Nil				0
11-May	15312	Nil				0
12-May	15312	Nil				0
13-May	15316	UoM				4
14-May	15324	YCM				8
15-May	15329	UoM				5
16-May	15352	Roaming	UoM			23
17-May	15362	UoM				10
18-May	15368	YCM	Crash			6
26-Jun	15399	UoM				31

Date	Distance(km)		Frequency distribution				
31-Dec	0		Distance	Frequency		**Measures of central tendency**	
01-Jan	8		0	36		Mean distance travelled, μ	11.36
02-Jan	13		2	1		Median distance travelled, ν	10.00
03-Jan	8		4	3		Mode distance travelled, θ	0.00
04-Jan	9		5	6			
05-Jan	11		6	8		**Measures of dispersion**	
06-Jan	0		7	4		Standard deviation, σ	12.64
07-Jan	9		8	7		Mean absolute deviation, MAD	8.33
08-Jan	0		9	3			
09-Jan	0		10	10		**Positional measures**	
10-Jan	17		11	10		Minimum, m	0
11-Jan	7		12	3		First quartile, Q_1	0
12-Jan	0		13	3		Second quartile, Q_2	10
13-Jan	0		14	3		Third quartile, Q_3	16
14-Jan	18		15	3		Maximum, M	90
15-Jan	0		16	5		Range, R	90
16-Jan	8		17	5		Interquartile range, IQR	16
17-Jan	10		18	5		Lower whisker, W_L	0
18-Jan	16		19	3		Upper whisker, W_U	40
19-Jan	13		20	1			
20-Jan	0		21	2			
21-Jan	90		22	1			
22-Jan	10		23	1			
23-Jan	27		24	2			
24-Jan	45		25	2			
25-Jan	0		26	2			
26-Jan	21		27	2			
27-Jan	0		28	1			
28-Jan	19		29	1			
29-Jan	26		30	1			
30-Jan	18		31	1			
31-Jan	27		35	1			

01-Feb	16		36	1			
02-Feb	36		45	1			
03-Feb	5		76	1			
04-Feb	24		90	1			
05-Feb	17						
06-Feb	12		**Trimmed frequency distribution and computation table**				
07-Feb	16		**Distance, d**	**Frequency, f**	**d*f**	**Cumulative frequency**	
08-Feb	9		2	1	2	1	
09-Feb	0		4	3	12	4	
10-Feb	0		5	6	30	10	
11-Feb	7		6	8	48	18	
12-Feb	11		7	4	28	22	
13-Feb	10		8	7	56	29	
14-Feb	17		9	3	27	32	
15-Feb	8		10	10	100	42	
16-Feb	13		11	10	110	52	**N/2**
17-Feb	0		12	3	36	55	
18-Jan	15		13	3	39	58	
19-Feb	15		14	3	42	61	
20-Feb	12		15	3	45	64	
21-Feb	11		16	5	80	69	
22-Feb	25		17	5	85	74	
23-Feb	17		18	5	90	79	
24-Feb	14		19	3	57	82	
25-Feb	11		20	1	20	83	
26-Feb	19		21	2	42	85	
27-Feb	19		22	1	22	86	
28-Feb	10		23	1	23	87	
01-Mar	8		24	2	48	89	
02-Mar	18		25	2	50	91	
03-Mar	76		26	2	52	93	
04-Mar	29		27	2	54	95	
05-Mar	24		28	1	28	96	

06-Mar	18		29	1	29	97	
07-Mar	26		30	1	30	98	
08-Mar	25		31	1	31	99	
09-Mar	28		35	1	35	100	
10-Mar	12		36	1	36	101	
11-Mar	21		45	1	45	102	
12-Mar	20		76	1	76	103	
13-Mar	22		90	1	90	104	**N**
14-Mar	11		**Total**	104	1598		
15-Mar	35		* This table contains data of days I've driven, neglecting days on which I din't drive				
16-Mar	14						
17-Mar	0						
18-Mar	0						
19-Mar	30						
20-Mar	16						
21-Mar	0						
22-Mar	14						
23-Mar	8						
24-Mar	17						
25-Mar	11						
26-Mar	15						
27-Mar	18						
28-Mar	10						
29-Mar	11						
30-Mar	6						
31-Mar	0						
01-Apr	6						
02-Apr	0						
03-Apr	0						
04-Apr	11						
05-Apr	10						
06-Apr	0						
07-Apr	0						

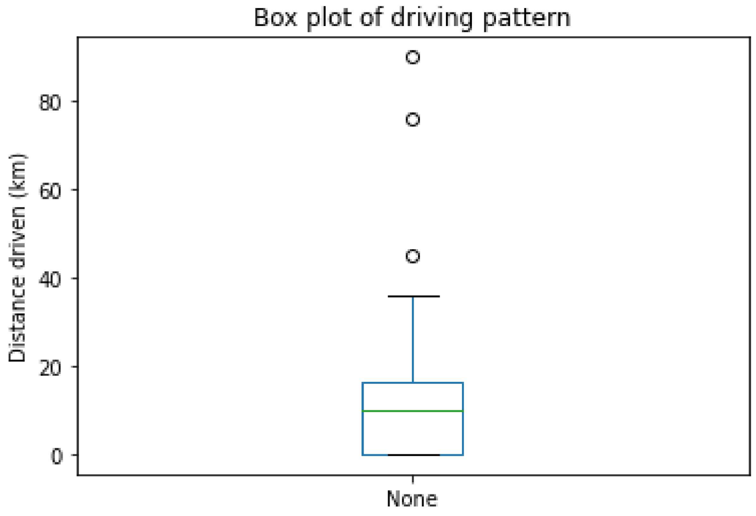


08-Apr	11						
09-Apr	7						
10-Apr	10		**Summary statistics for trimmed data**				
11-Apr	5		**Measures of central tendency**				
12-Apr	10		Mean distance travelled, μ	15.37			
13-Apr	6		Median distance travelled, ν	11.50			
14-Apr	0		Mode distance travelled, θ	10, 11	**Bimodal**		
15-Apr	10						
16-Apr	11		**Measures of dispersion**				
17-Apr	0		Standard deviation, σ	12.49			
18-Apr	0		Mean absolute deviation, MAD	8.33			
19-Apr	0						
20-Apr	6		**Positional measures**				
21-Apr	0		Minimum, m	2			
22-Apr	5		First quartile, Q_1	8			
23-Apr	16		Second quartile, Q_2	11.5			
24-Apr	4		Third quartile, Q_3	18			
25-Apr	5		Maximum, M	90			
26-Apr	0		Range, R	88			
27-Apr	0		Interquartile range, IQR	10			
28-Apr	0		Lower whisker, W_L	8			
29-Apr	6		Upper whisker, W_U	33			
30-Apr	0						
01-May	0						
02-May	7						
03-May	2						
04-May	6						
05-May	0						
06-May	4						
07-May	0						
08-May	5						
09-May	6						
10-May	0						

11-May	0						
12-May	0						
13-May	4						
14-May	8						
15-May	5						
16-May	23						
17-May	10						
18-May	6						
26-Jun	31						

```
import numpy as np
import pandas as pd
from google.colab import drive
import matplotlib.pyplot as plt

plt.close('all')

filename = '/content/drive/My Drive/Trip.csv'
drive.mount('/content/drive', force_remount=True)

date = np.array([])
distance = np.array([])

filer = open(filename, 'r')
lines = filer.readlines()
for line in lines:
  line = line.rstrip().split(',')
  date = np.append(date, line[0])
  distance = np.append(distance, line[1])

titles = [date[0], distance[0]]
date = pd.Series(date[1:])
distance = pd.Series(int(distance[i]) for i in range(1, len(distance)))
data = pd.concat([date, distance], axis=1)
#data = pd.read_csv(filename)
distance.plot.box(ylabel="Distance driven (km)", title = 'Box plot of driving pattern')
print(distance[1:-1].describe())
print(distance[1:-1].mad())
print(distance.value_counts().transpose())
```

```
Mounted at /content/drive
<matplotlib.axes._subplots.AxesSubplot at 0x7fea81b20ef0>
```


Box plot of driving pattern
Distance driven (km)
None