Case Study 2

Sonia Mazzi

18/12/2018

Data Science Campus

CASE STUDY Massachusetts Bay Transport Authority (MBTA) data from an excel file

Data on transportation data in Boston, USA: monthly averages of weekday number of passangers (in thousands) by mode of transportation.

Data Science Campus

A snapshot of part of the data (not all columns are included), in excel format, is below.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**MBTA Avg Weekday Unlinked Passenger Trips (thousands)**												
2		mode	2007-01	2007-02	2007-03	2007-04	2007-05	2007-06	2007-07	2007-08	2007-09	2007-10	2007-1
3	1	All Modes by Qtr	NA	NA	1187.653	NA	NA	1245.959	NA	NA	1256.571	NA	NA
4	2	Boat	4	3.6	40	4.3	4.9	5.8	6.521	6.572	5.469	5.145	3.
5	3	Bus	335.819	338.675	339.867	352.162	354.367	350.543	357.519	355.479	372.598	368.847	330.
6	4	Commuter Rail	142.2	138.5	137.7	139.5	139	143	142.391	142.364	143.051	146.542	145.
7	5	Heavy Rail	435.294	448.271	458.583	472.201	474.579	477.032	471.735	461.605	499.566	457.741	488.
8	6	Light Rail	227.231	240.262	241.444	255.557	248.262	246.108	243.286	234.907	265.748	241.434	250.
9	7	Pct Chg / Yr	0.02	-0.04	0.114	-0.002	0.049	0.096	-0.037	0.004	-0.007	-0.064	-0.
10	8	Private Bus	4.772	4.417	4.574	4.542	4.768	4.722	3.936	3.946	4.329	4.315	4.
11	9	RIDE	4.9	5	5.5	5.4	5.4	5.6	5.253	5.308	5.609	5.806	5.
12	11	Trackless Trolley	12.757	12.913	13.057	13.444	13.479	13.323	13.311	13.142	14.393	14.622	13.
13	10	TOTAL	1166.974	1191.639	1204.725	1247.105	1244.755	1246.129	1243.952	1223.323	1310.764	1244.453	1241.
14													
15													
16													
17													
18													

- 4 variables: transportation mode, year, month, and monthly weekday average number of trips.
- The first row in the excel sheet is a title, so we skip this row when reading the data in.

```r
#skip the first row.
#NA character is "NA". Blank space is the default value
mbta = read_excel("DataFiles/mbta.xlsx", na = "NA", skip = 1)
```

```r
mbta
```

```
## # A tibble: 11 x 60
##     X__1 mode  `2007-01` `2007-02` `2007-03` `2007-04` `2007-05` `2007-06`
##    <dbl> <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
##  1     1 All ~     NA        NA     1188.       NA        NA      1246.
##  2     2 Boat       4         3.6     40         4.3       4.9       5.8
##  3     3 Bus      336.      339.     340.      352.      354.      351.
##  4     4 Comm~    142.      138.     138.      140.      139       143
##  5     5 Heav~    435.      448.     459.      472.      475.      477.
##  6     6 Ligh~    227.      240.     241.      256.      248.      246.
##  7     7 Pct ~      0.02     -0.04     0.114    -0.002     0.049     0.096
##  8     8 Priv~      4.77      4.42     4.57      4.54      4.77      4.72
##  9     9 RIDE       4.9       5        5.5       5.4       5.4       5.6
## 10    10 Trac~     12.8      12.9     13.1      13.4      13.5      13.3
## 11    11 TOTAL   1167.     1192.    1205.     1247.     1245.     1246.
## # ... with 52 more variables: `2007-07` <dbl>, `2007-08` <dbl>,
## #   `2007-09` <dbl>, `2007-10` <dbl>, `2007-11` <dbl>, `2007-12` <dbl>,
## #   `2008-01` <dbl>, `2008-02` <dbl>, `2008-03` <dbl>, `2008-04` <dbl>,
## #   `2008-05` <dbl>, `2008-06` <dbl>, `2008-07` <dbl>, `2008-08` <dbl>,
## #   `2008-09` <dbl>, `2008-10` <dbl>, `2008-11` <dbl>, `2008-12` <dbl>,
## #   `2009-01` <dbl>, `2009-02` <dbl>, `2009-03` <dbl>, `2009-04` <dbl>,
## #   `2009-05` <dbl>, `2009-06` <dbl>, `2009-07` <dbl>, `2009-08` <dbl>,
## #   `2009-09` <dbl>, `2009-10` <dbl>, `2009-11` <dbl>, `2009-12` <dbl>,
## #   `2010-01` <dbl>, `2010-02` <dbl>, `2010-03` <dbl>, `2010-04` <dbl>,
## #   `2010-05` <dbl>, `2010-06` <dbl>, `2010-07` <dbl>, `2010-08` <dbl>,
## #   `2010-09` <dbl>, `2010-10` <dbl>, `2010-11` <dbl>, `2010-12` <dbl>,
## #   `2011-01` <dbl>, `2011-02` <dbl>, `2011-03` <dbl>, `2011-04` <dbl>,
## #   `2011-05` <dbl>, `2011-06` <dbl>, `2011-07` <dbl>, `2011-08` <dbl>,
## #   `2011-09` <dbl>, `2011-10` <dbl>
```

```r
#display all the column names
names(mbta)
```

```
##  [1] "X__1"    "mode"    "2007-01" "2007-02" "2007-03" "2007-04" "2007-05"
##  [8] "2007-06" "2007-07" "2007-08" "2007-09" "2007-10" "2007-11" "2007-12"
## [15] "2008-01" "2008-02" "2008-03" "2008-04" "2008-05" "2008-06" "2008-07"
## [22] "2008-08" "2008-09" "2008-10" "2008-11" "2008-12" "2009-01" "2009-02"
## [29] "2009-03" "2009-04" "2009-05" "2009-06" "2009-07" "2009-08" "2009-09"
## [36] "2009-10" "2009-11" "2009-12" "2010-01" "2010-02" "2010-03" "2010-04"
## [43] "2010-05" "2010-06" "2010-07" "2010-08" "2010-09" "2010-10" "2010-11"
## [50] "2010-12" "2011-01" "2011-02" "2011-03" "2011-04" "2011-05" "2011-06"
## [57] "2011-07" "2011-08" "2011-09" "2011-10"
```

```
mbta
```

```
## # A tibble: 11 x 60
##     X__1 mode  `2007-01` `2007-02` `2007-03` `2007-04` `2007-05` `2007-06`
##    <dbl> <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
##  1     1 All ~     NA        NA      1188.       NA        NA       1246.
##  2     2 Boat       4         3.6      40         4.3       4.9        5.8
##  3     3 Bus      336.      339.      340.      352.      354.       351.
##  4     4 Comm~    142.      138.      138.      140.      139        143
##  5     5 Heav~    435.      448.      459.      472.      475.       477.
##  6     6 Ligh~    227.      240.      241.      256.      248.       246.
##  7     7 Pct ~      0.02     -0.04      0.114    -0.002     0.049      0.096
##  8     8 Priv~      4.77      4.42      4.57      4.54      4.77       4.72
##  9     9 RIDE       4.9       5         5.5       5.4       5.4        5.6
## 10    10 Trac~     12.8      12.9      13.1      13.4      13.5       13.3
## 11    11 TOTAL   1167.     1192.     1205.     1247.     1245.      1246.
## # ... with 52 more variables: `2007-07` <dbl>, `2007-08` <dbl>,
...
```

- 1st column enumerates rows. Rows are identified by mode of transporatation. 1st column is unnecessary.
- 1st row is a quarterly aggregation. Not needed.
- The last row (11th) has totals. Not needed.
- 7th row has % change in the year. Not needed.

```r
#Leave out the first column
mbta2 = select(mbta, -1)
```

```r
#Eliminate rows 1, 7, 11
mbta2 = slice(mbta2, -c(1, 7, 11))
```

```r
mbta2
```

```
## # A tibble: 8 x 59
##   mode  `2007-01` `2007-02` `2007-03` `2007-04` `2007-05` `2007-06`
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Boat       4         3.6      40         4.3       4.9       5.8
## 2 Bus      336.      339.      340.      352.      354.      351.
## 3 Comm~    142.      138.      138.      140.      139       143
## 4 Heav~    435.      448.      459.      472.      475.      477.
## 5 Ligh~    227.      240.      241.      256.      248.      246.
## 6 Priv~      4.77      4.42      4.57      4.54      4.77      4.72
## 7 RIDE       4.9       5         5.5       5.4       5.4       5.6
## 8 Trac~     12.8      12.9      13.1      13.4      13.5      13.3
## # ... with 52 more variables: `2007-07` <dbl>, `2007-08` <dbl>,
...
```

- Variables are mode of transportation, year, month and monthly average number of passengers.
- All column names, except for the first one, `mode`, are values of year and month combined.
- To correct this we use the `gather()` function

```
#gather column names, except first one, into "year_month"
#with the corresponding value in the column "NrPassengers"
#
mbta3 = gather(mbta2, "year_month", "NrPassengers", -1)
```

```
glimpse(mbta3)
```

```
## Observations: 464
## Variables: 3
## $ mode         <chr> "Boat", "Bus", "Commuter Rail", "Heavy Rail", "Li...
## $ year_month   <chr> "2007-01", "2007-01", "2007-01", "2007-01", "2007...
## $ NrPassengers <dbl> 4.000, 335.819, 142.200, 435.294, 227.231, 4.772,...
```

- `year_month` has values of 2 variables. Keep the year in one column and month in another column.
- We separate them using the `separate()` function.

Data Science Campus

```
mbta4 =
  separate(mbta3, "year_month", c("year", "month"), sep = "-")
```

```
mbta4$year = as.numeric(mbta4$year)
```

```
mbta4$month = as.numeric(mbta4$month)
```

```
glimpse(mbta4)
```

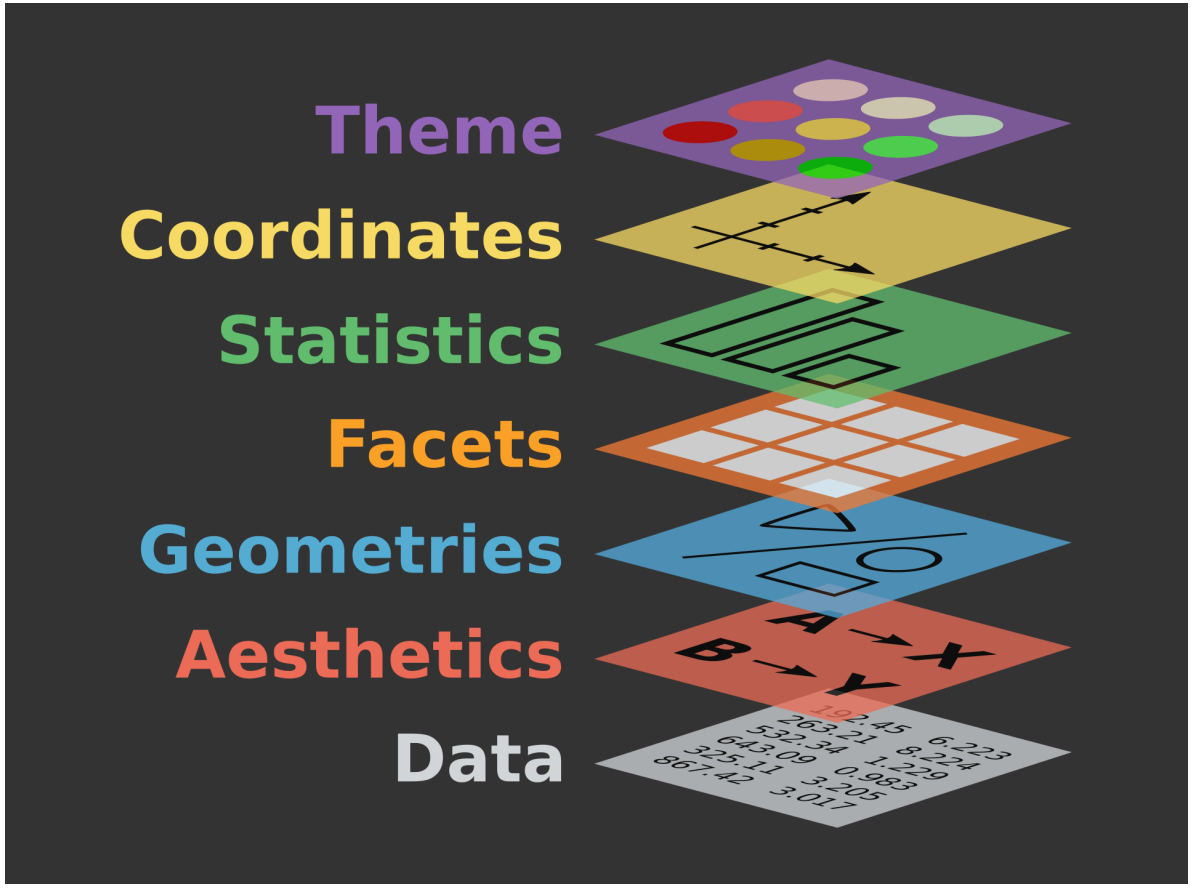
```
## Observations: 464
## Variables: 4
## $ mode          <chr> "Boat", "Bus", "Commuter Rail", "Heavy Rail", "Li...
## $ year          <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
## $ month         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3...
## $ NrPassengers <dbl> 4.000, 335.819, 142.200, 435.294, 227.231, 4.772,...
```

- The data is tidy and ready to be explored.

Using ggplot2 to visualise data

Data Science Campus

`ggplot2` is based on "The Grammar of Graphics".



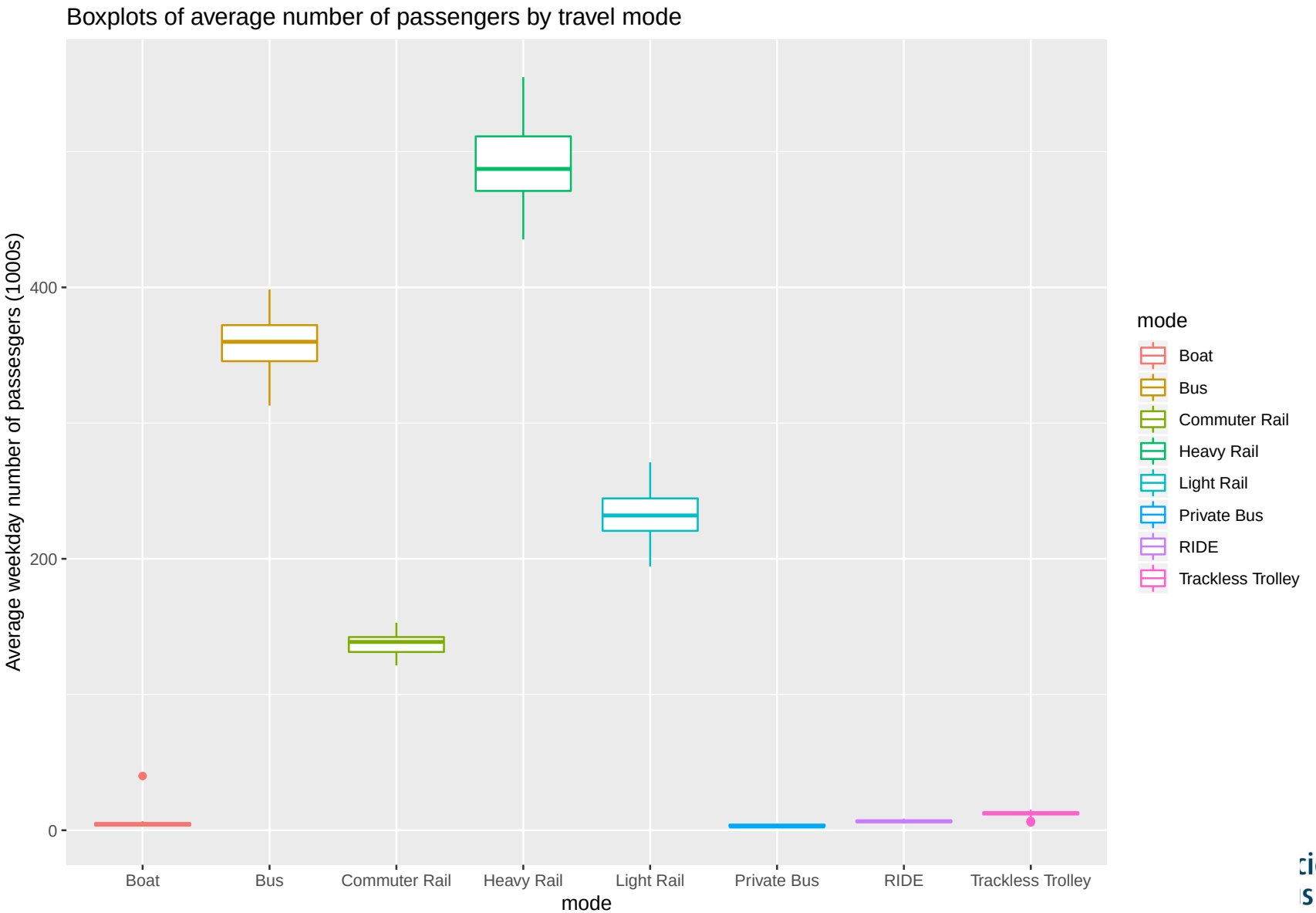
Data Science Campus

- The function `ggplot()`, in the package `ggplot2`, is used to visualise data.
- The basic use is

```
ggplot(myData, aes = (myMapping)) + myGeometryLayer
```

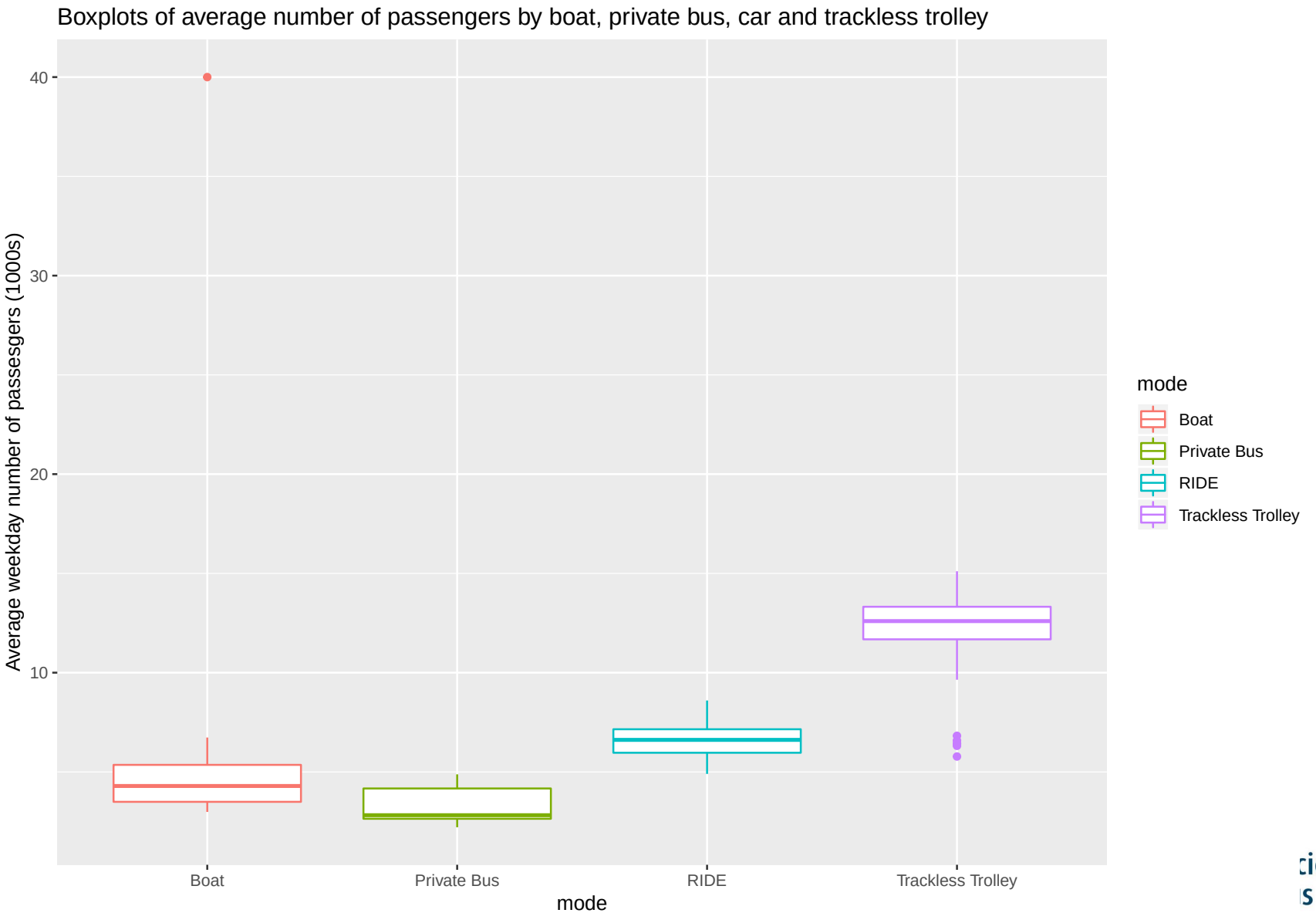
- mydata: data frame with variables to use in plot.
- myMapping: mapping from the data to the aesthetics (visual dimension) in the graph. For example, the mapping can be `x = Varx, y = Vary` for a scatter plot of `Vary` vs. `Varx`.
- myGeometryLayer: specify what you want, points, lines, boxes, etc. e.g.: `geom_point()` for a scatter plot, `geom_line` for a line plot, etc.
- One can add many layers to the basic ggplot object created with the `ggplot()` function.

```
ggplot(mbta4, aes(x = mode , y = NrPassengers, color = mode)) +
  geom_boxplot() +
  labs(y = "Average weekday number of passesgers (1000s)") +
  ggtitle("Boxplots of average number of passengers by travel mode")
```



- Most trips were made by heavy rail, bus, light rail and commuter rail, in descending order.
- Number of passengers are on a different scale for boat, private bus and trackless trolley. Plot them separately

```r
aux = filter(mbta4, mode %in% c("Boat", "Private Bus", "RIDE", "Trackless Trolley"))
ggplot(aux, aes(x = mode , y = NrPassengers, color = mode)) +
  geom_boxplot() +
  labs(y = "Average weekday number of passesgers (1000s)") +
  ggtitle("Boxplots of average number of passengers by boat, private bus, car and trackless trolley")
```



- There is a very large observation for Boat. Let us find out when it was observed.

```
#pb contains NrPassengers by boat only
pb = with(mbta4, NrPassengers[mode == "Boat"])
```

```
#gives the index of the maximum value of pb
aux = which.max(pb)
```

```
with(mbta4, year[mode == "Boat"][aux])
```

```
## [1] 2007
```

```
with(mbta4, month[mode == "Boat"][aux])
```

```
## [1] 3
```

- The unusual observation occurred in March 2007.

Data Science Campus

- Look at the distribution of the other values of number of passengers who travelled by boat

```
summary(pb[-aux])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.985   3.488   4.285   4.455   5.189   6.733
```

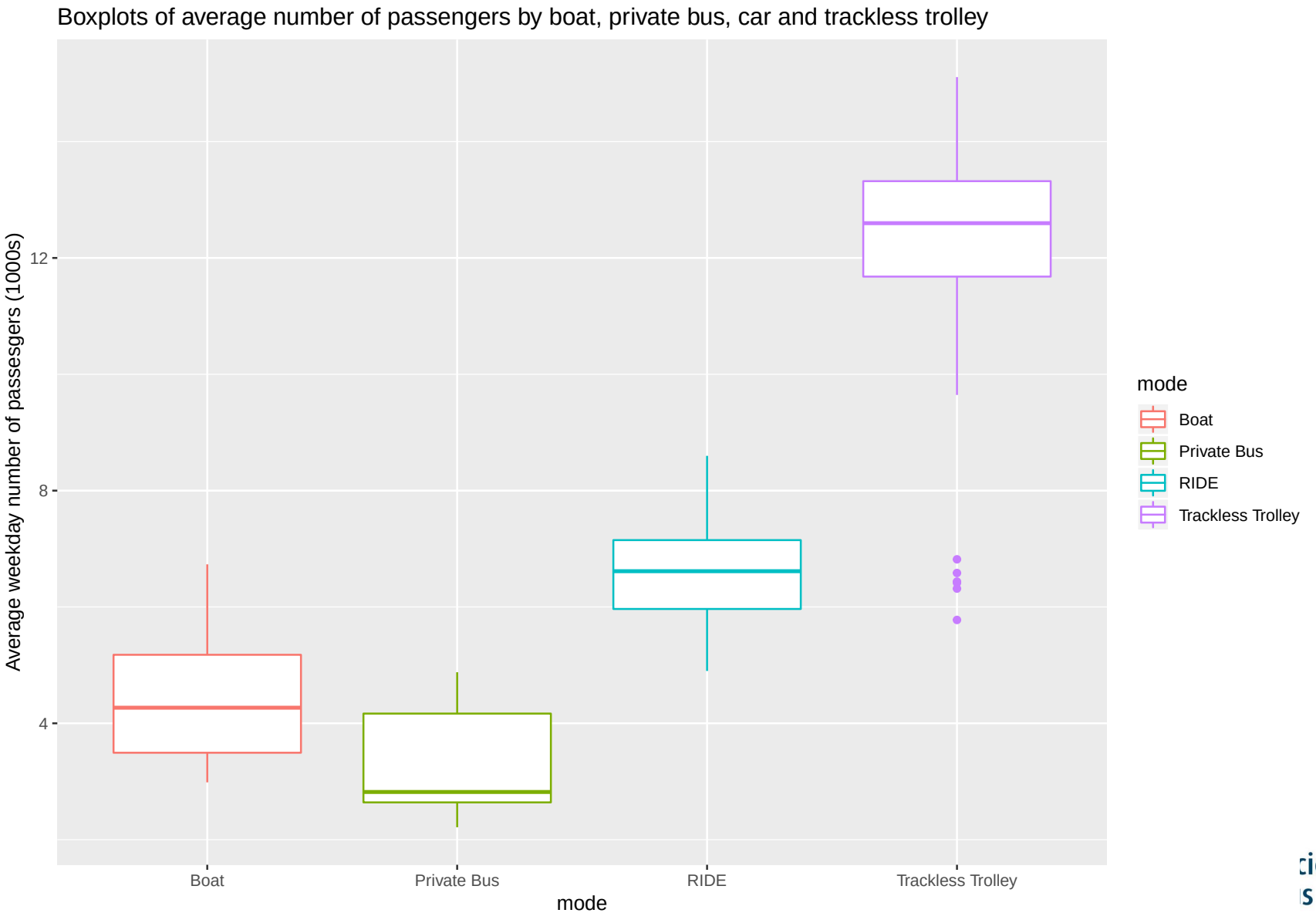
- No big event happenend in Boston in March 2007.
- We conclude that it's quite likely the person who entered the data added an extra zero.
- Change this value to 4. Check with data originators and report.

```
mbta4$NrPassengers[mbta4$mode == "Boat"][aux] = 4
```

- Let us see the boxplots again

Data Science Campus

```r
aux = filter(mbta4, mode %in% c("RIDE", "Boat", "Private Bus", "Trackless Trolley"))
ggplot(aux, aes(x = mode , y = NrPassengers, color = mode)) +
  geom_boxplot() +
  labs(y = "Average weekday number of passesgers (1000s)") +
  ggtitle("Boxplots of average number of passengers by boat, private bus, car and trackless trolley")
```



- There are some unusually low values for the number of passengers travelling by trackless trolley.
- Find out when they occurred.

```r
#this is all the data for trackless trolley only
trtr = filter(mbta4, mode == "Trackless Trolley")
```

```r
aux = arrange(trtr, NrPassengers)
```

```r
head(aux, n=8)
```

```
## # A tibble: 8 x 4
##   mode                year month NrPassengers
##   <chr>              <dbl> <dbl>        <dbl>
## 1 Trackless Trolley  2010    12         5.78
## 2 Trackless Trolley  2010     8         6.32
## 3 Trackless Trolley  2010    11         6.42
## 4 Trackless Trolley  2010     9         6.44
## 5 Trackless Trolley  2010     7         6.58
## 6 Trackless Trolley  2010    10         6.82
## 7 Trackless Trolley  2009     2         9.64
## 8 Trackless Trolley  2011     7        11.1
```

- The unusually low observations for trackless trolley occurred in the second semester of 2010.
- Don't change or delete, but be aware.
- Let us visualise.

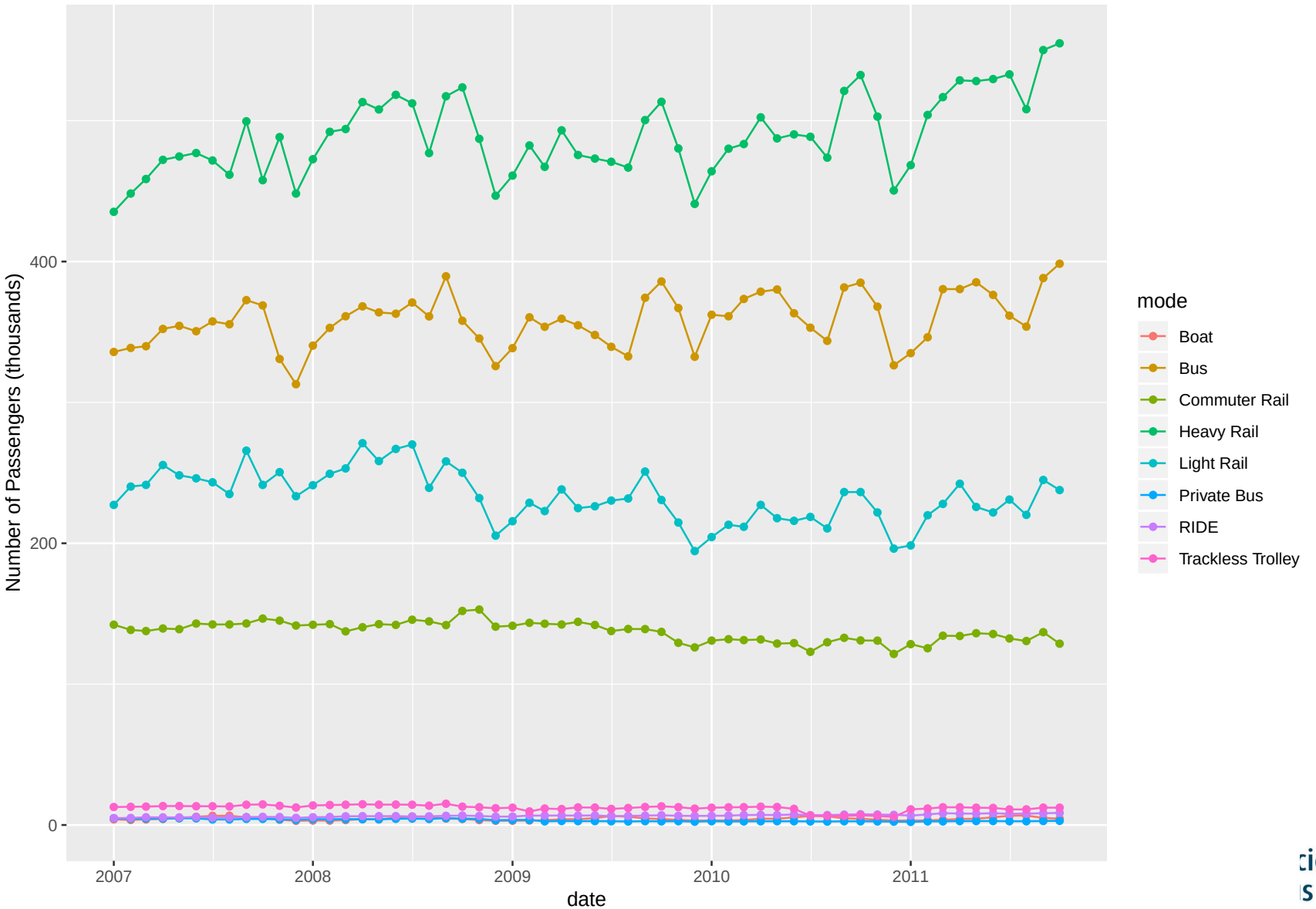
- We will plot the data on numbers of passengers against time.
- Create a new variable, `date`, with the `lubridate` package function `make_date()`

```
#create and append the new variable date
mbta4 = mutate(mbta4, date = make_date(year, month) )
```

```
head(mbta4)
```

```
## # A tibble: 6 x 5
##   mode           year month NrPassengers date
##   <chr>         <dbl> <dbl>        <dbl> <date>
## 1 Boat           2007     1         4    2007-01-01
## 2 Bus            2007     1       336.   2007-01-01
## 3 Commuter Rail  2007     1       142.   2007-01-01
## 4 Heavy Rail     2007     1       435.   2007-01-01
## 5 Light Rail     2007     1       227.   2007-01-01
## 6 Private Bus    2007     1         4.77 2007-01-01
```

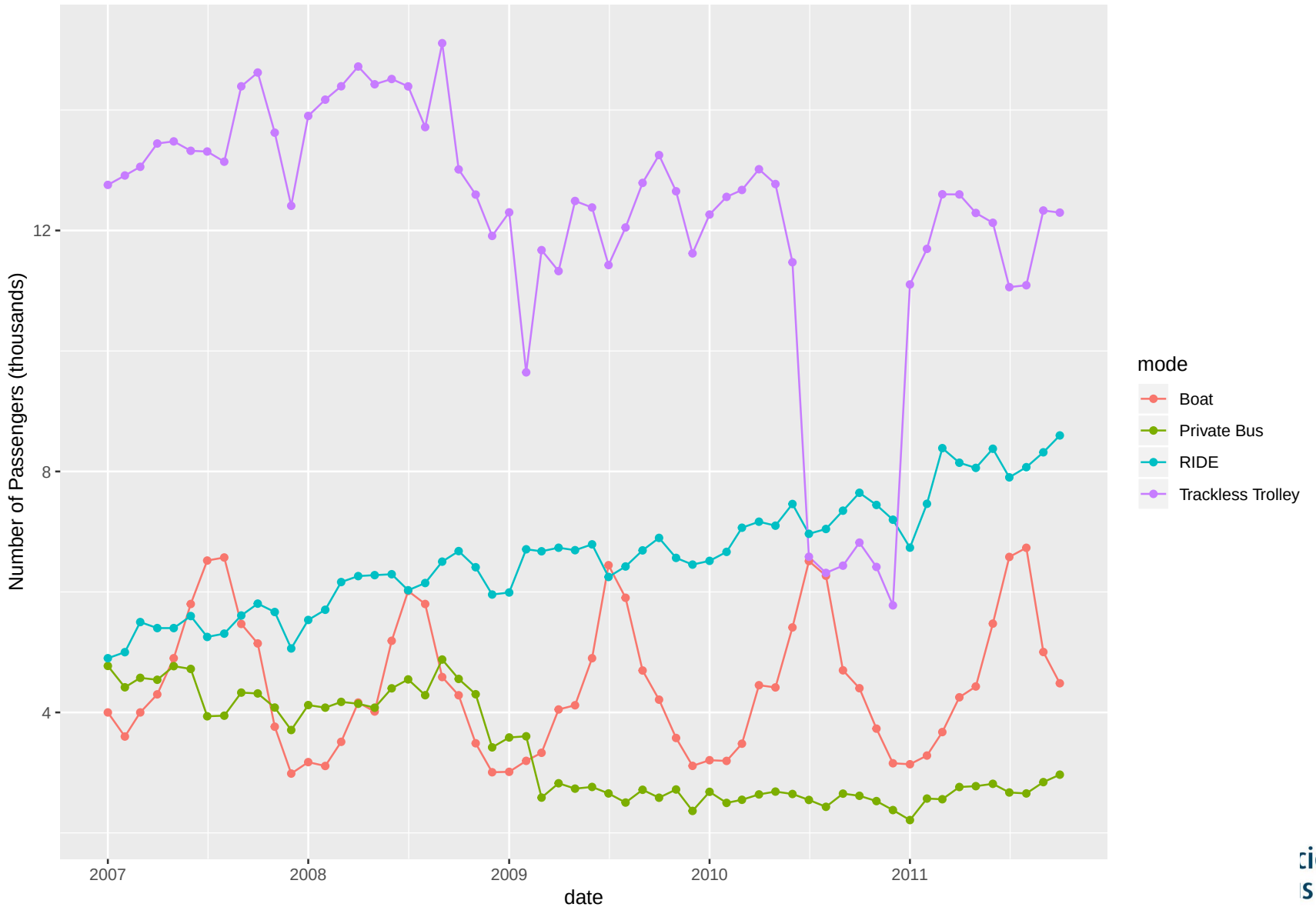
```
ggplot(mbta4, aes(x = date, y = NrPassengers, col = mode)) +
 geom_line() +
 geom_point()+
 labs(y = "Number of Passengers (thousands)")
```



- Different scales for the data corresponding to the number of passengers travelling by boat, car, trackless trolley and private bus.

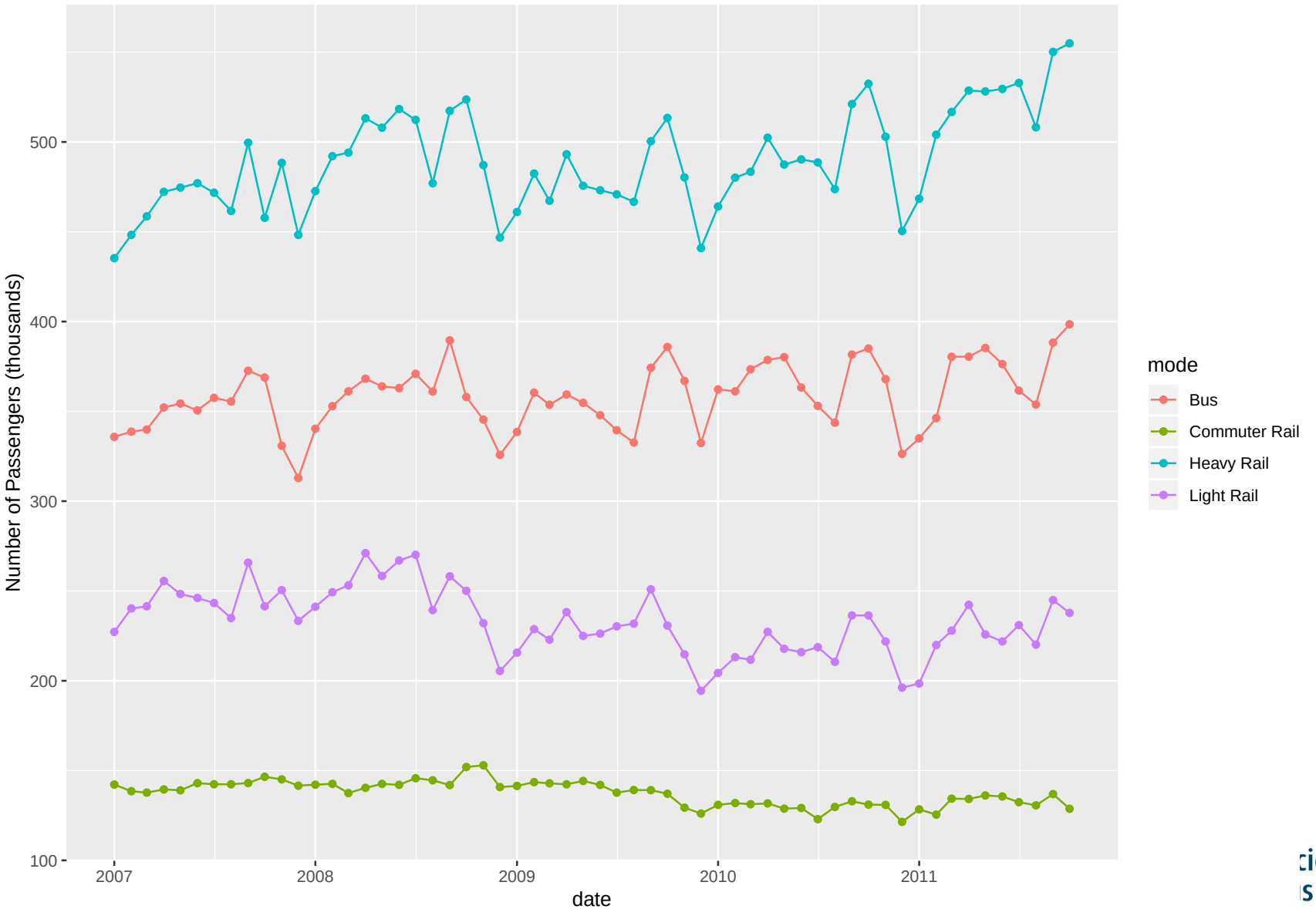
```
#select rows for boat, RIDE, trackless trolley, private bus
#
mbta5 = filter(mbta4, mode %in% c("Boat", "Private Bus","RIDE",
                                  "Trackless Trolley"))
```

```
ggplot(mbta5, aes(x = date, y = NrPassengers, col = mode)) +
  geom_line() +
  geom_point() +
  labs(y = "Number of Passengers (thousands)")
```



- Strong seasonal component in number of passengers travelling by boat.
- The use of RIDE seems to be steadily increasing during time.
- The use of private bus and trackless trolley had a sharp decrease since 2009 and something unusual made the use of trackless trolley dramatically decrease in the second half of 2010.

```
mbta6 = filter(mbta4, mode %in% c("Bus", "Commuter Rail", "Heavy Rail", "Light Rail"))
ggplot(mbta6, aes(x = date, y = NrPassengers, col = mode)) +
  geom_line() +
  geom_point() +
  labs(y = "Number of Passengers (thousands)")
```



- The number of passengers travelling by light and commuter rail has decreased since 2009.
- There seems to be an upwards trend on number of passengers travelling by heavy rail and perhaps by bus as well.