Ridership Across North American Transit Systems

Alex Pan

February 2, 2018

Contents

Transit Ridership

Set up

```
library(tidyverse)
library(jsonlite)
library(lubridate)
library(ggplot2)
theme_set(theme_bw())
```

Ridership by Region

Edmonton

From https://dashboard.edmonton.ca/Dashboard/Transit-Ridership/q4c4-5fu4 (updated January 9, 2018)

This file contains the annual total transit system ridership based on the last 12 months of fare revenue.

```
edmonton <- read_csv('Data/edmonton_ridership.csv')
```

```
head(edmonton, 3)
```

```
## # A tibble: 3 x 7
##      ID              DateTime  YEAR REPORT_PERIOD MONTH_RIDERSHIP
##   <int>                 <chr> <int>         <chr>           <int>
## 1     1 01/01/2005 12:00:00 AM  2005        05-Jan         4762992
## 2     2 02/01/2005 12:00:00 AM  2005        05-Feb         4741322
## 3     3 03/01/2005 12:00:00 AM  2005        05-Mar         5081793
## # ... with 2 more variables: `LAST 12 MONTHS` <int>, `CHANGE_%` <chr>
```

`DateTime` is being read as a factor, which will have to be converted to dates

`Change_.` is also being read as a factor, but I'm not immediately interested in that variable.

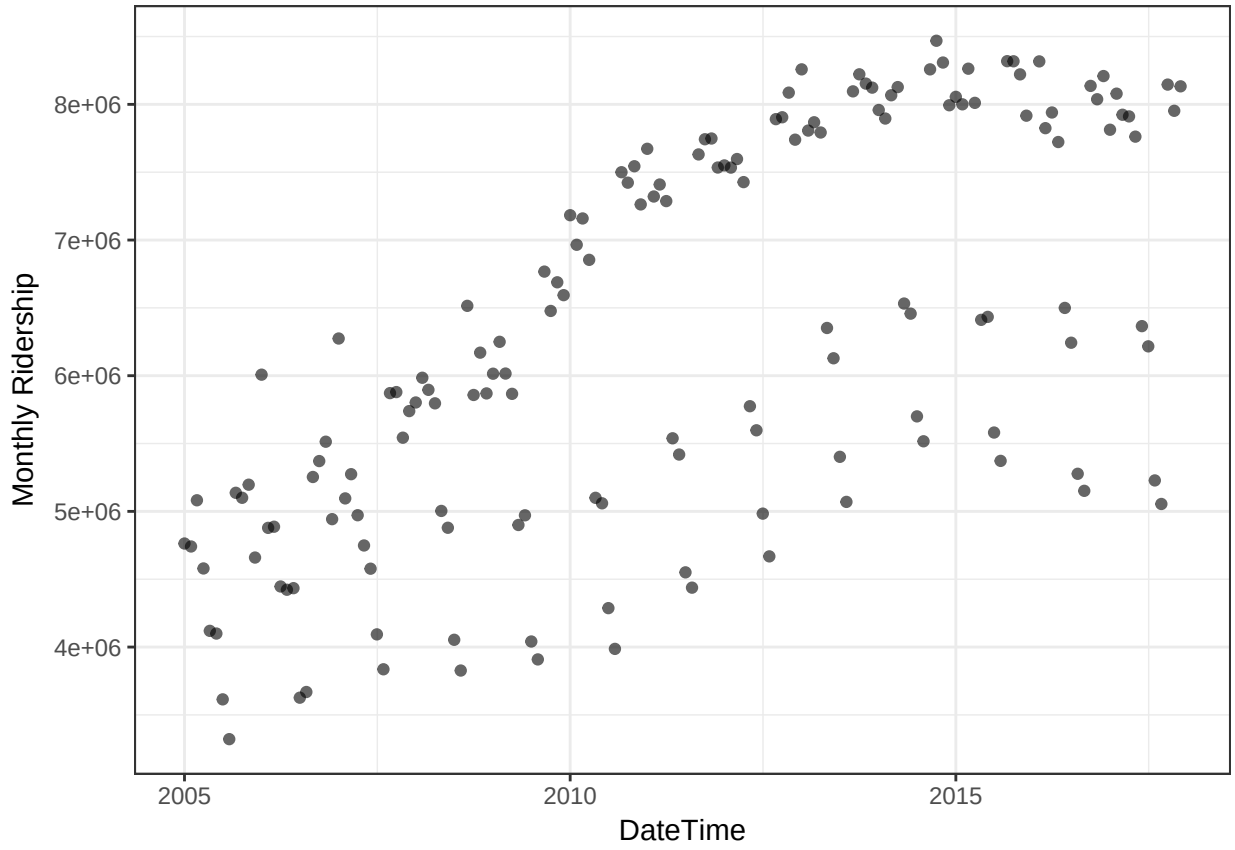
```
# In addition to converting `DateTime` into a date object, I'm also going to
# pull the month from that date object into a new column.

edmonton <- edmonton %>%
    mutate(DateTime = mdy_hms(as.character(DateTime)),
```

```
            Month = month(DateTime)) %>%
    rename(Year = YEAR)
```

Exploratory Plots

A simple ridership vs. time scatterplot

```
ggplot(edmonton) +
    aes(x = DateTime, y = MONTH_RIDERSHIP) +
    geom_point(alpha = 0.6) +
    ylab('Monthly Ridership')
```



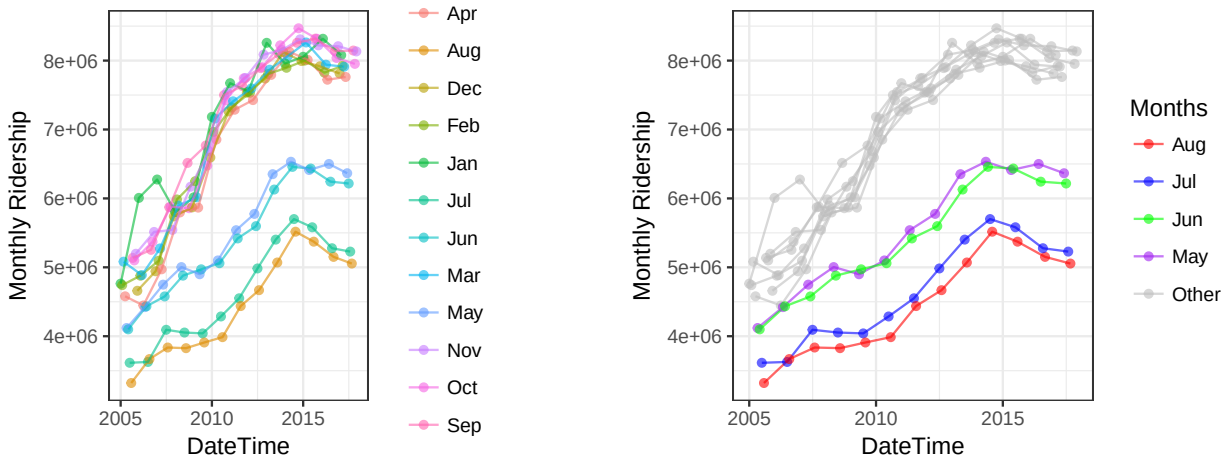
This is an interesting plot. There seem to be three distinct 'groups'

Split the plot by month

```
summer_months <- as.factor(
    ifelse(!(edmonton$Month %in% seq(5,8)), 'Other', month.abb[edmonton$Month]))
palette <- c('red', 'blue', 'green', 'purple', 'grey')

# All months
p1 <- ggplot(edmonton) +
    aes(x = DateTime, y = MONTH_RIDERSHIP, col = month.abb[Month]) +
    geom_point(alpha = 0.6) +
```

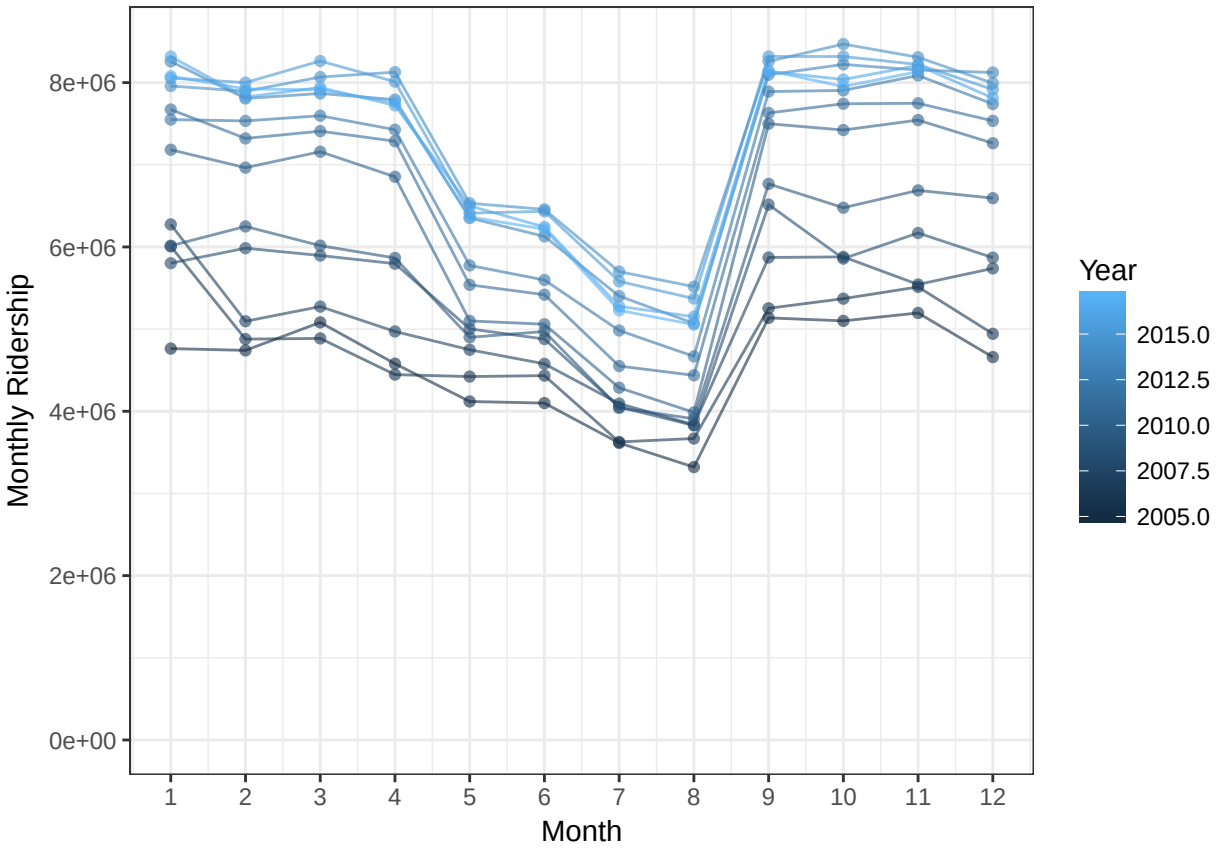
```
    geom_line(aes(group = month.abb[Month]), alpha = 0.6) +
    ylab('Monthly Ridership')

# Highlight summer months
p2 <- ggplot(edmonton) +
    aes(x = DateTime, y = MONTH_RIDERSHIP) +
    geom_line(aes(group = Month, col = summer_months), alpha = 0.6) +
    scale_colour_manual('Months', values = palette) +
    geom_point(aes(col = summer_months), alpha = 0.6) +
    ylab('Monthly Ridership')
```

```
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



Here is another way to look at this:

```
ggplot(edmonton) +
    geom_line(aes(x = Month, y = MONTH_RIDERSHIP, group = Year, col = Year),
              alpha = 0.6) +
    geom_point(aes(x = Month, y = MONTH_RIDERSHIP, group = Year, col = Year),
               alpha = 0.6) +
    scale_x_continuous(breaks = seq(1, 12)) +
    ylab('Monthly Ridership') +
    ylim(c(0, 8500000))
```



We can see that the summer months have much lower transit usage that the oher months. This shoots way back up in September. Blame the university students?

How big is this difference?

```
temp <- edmonton %>%
    group_by(summer = ifelse(Month %in% seq(5,8), 'summer', 'other')) %>%
    summarize(mean = mean(MONTH_RIDERSHIP))

print(temp)
```

```
## # A tibble: 2 x 2
##   summer    mean
##    <chr>   <dbl>
## 1  other 6972246
## 2 summer 5010399
```

```
temp$mean[1] - temp$mean[2]
```

```
## [1] 1961847
```

On average, summer months have 1,961,847 fewer rides monthly than other times of the year.

Back to top

Chicago

Back to top

Maryland

From: https://catalog.data.gov/dataset/mta-average-weekday-ridership-by-month

These are data for the entire state, rather than for a single city.

The data are given as the average weekday ridership across each medium of transit:

- Intra-city transit: bus, metro, light rail
- Accessibility services: MobilityLink and Taxi Access
- MARC commuter train system, which connects the Baltimore-Washington DC area.
- Commuter Bus system, which serves to connect suburban areas to urban Baltimore/Washington DC.

Also note that these data are daily ridership numbers, whereas Edmonton's data were supplied as Monthly numbers. We can easily calculate that by multiplying by days in the month, but I won't be doing anything that requires that.

Data Exploration

```
json_maryland <- fromJSON('Data/maryland_ridership.json', flatten = TRUE)

maryland <- data.frame(json_maryland$data)
names(maryland) <- json_maryland$meta$view$columns$fieldName
```

```
names(maryland)
```

```
##  [1] ":sid"                           ":id"
##  [3] ":position"                      ":created_at"
##  [5] ":created_meta"                  ":updated_at"
##  [7] ":updated_meta"                  ":meta"
##  [9] "average_weekday_ridership"      "bus"
## [11] "metro"                          "light_rail"
## [13] "mobility"                       "taxi_access"
## [15] "marc_average"                   "marc_brunswick"
## [17] "marc_camden"                    "marc_penn"
## [19] "commuter_bus_total"             "baltimore"
## [21] "washington"                     "icc"
## [23] "total_average_weekday_ridership"
```

`total_average_weekday_ridership` is the total average ridership across all transit media, calculated by the folks at the MTA. In general, we'll prefer to use our own calculations rather than ones provided to us.

```
maryland[1:3, 9:23]    # Display first 3 rows/The first 8 columns are meta data
```

```
##   average_weekday_ridership    bus metro light_rail mobility taxi_access
## 1                   Jul-06 205015 43358      22997     2838        1072
## 2                   Aug-06 215455 44427      22708     2860        1162
## 3                   Sep-06 251719 44062      24085     2985        1211
##   marc_average marc_brunswick marc_camden marc_penn commuter_bus_total
## 1         <NA>           <NA>        <NA>      <NA>               <NA>
```

```
## 2         <NA>           <NA>        <NA>      <NA>                <NA>
## 3         <NA>           <NA>        <NA>      <NA>                <NA>
##   baltimore washington  icc total_average_weekday_ridership
## 1      <NA>       <NA> <NA>                          275280
## 2      <NA>       <NA> <NA>                          286612
## 3      <NA>       <NA> <NA>                          324062
```

Some problems with the data:

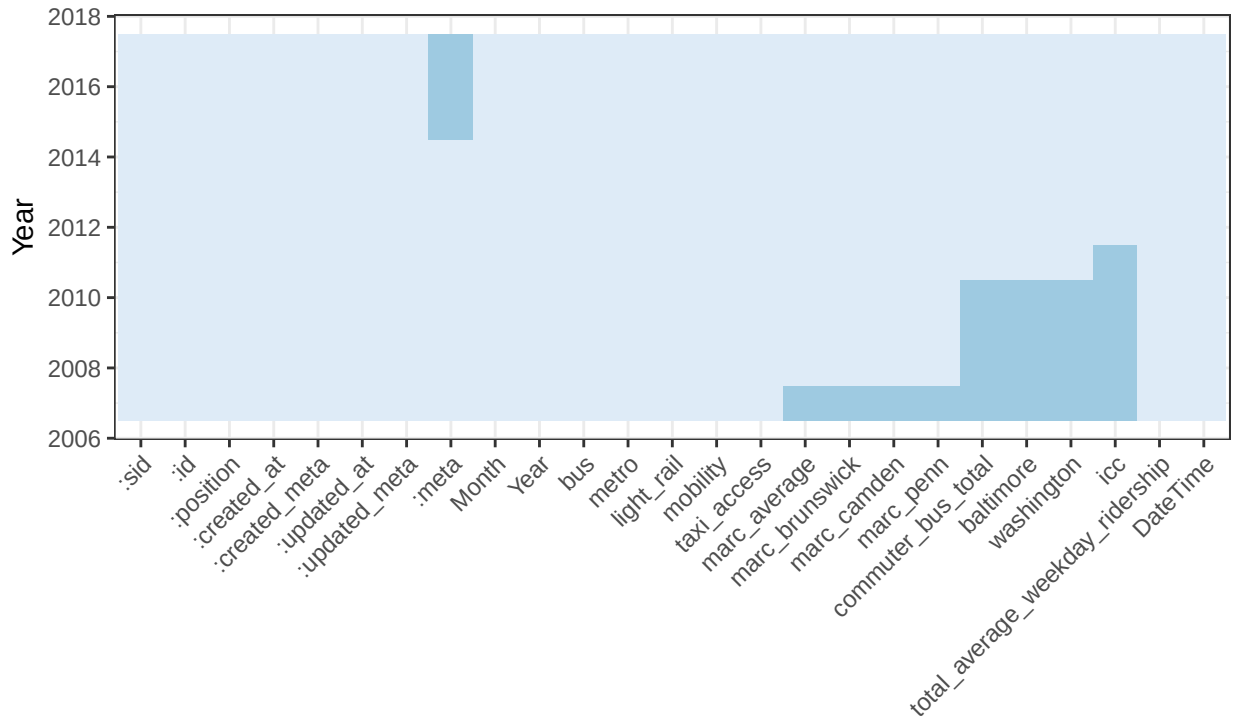
- Everything has been read in as a string/factor
- The MARC and Commuter Bus data are missing (at least for these rows)

Fix the variable typing:

```
# Convert columns to correct type
# Also add usable date columns
maryland[] <- lapply(maryland[], function(x) type.convert(as.character(x)))

maryland <- maryland %>%
    separate(average_weekday_ridership, into = c('Month', 'Year'), sep = '-') %>%
    mutate(Month = match(Month, month.abb),
           Year = as.numeric(paste('20', Year, sep = '')),
           DateTime = Year + Month / 12)
```

Investigate the missingness:

```
# Plot each NA as a bar, sorted by date.

# Sort by date
maryland <- maryland %>% arrange(DateTime)

# Find if entry is missing
x <- as.matrix(is.na(maryland))

# Name columns to match year. To prevent congestion of the axis, only name rows
# that correspond to January (i.e. Month == 1)
dimnames(x)[[1]] <- ifelse(maryland$Month == 1, maryland$Year, NA)

# Reshape matrix x into a single column of data
x <- reshape2::melt(x)

ggplot(x, aes(x = Var2, y = Var1, fill = value)) +
    geom_tile() +
    scale_fill_brewer() +
    ylab('Year') + xlab('') +
    theme(legend.position = "none",
          axis.text.x = element_text(angle=45, hjust = 1))
```



We see that we don't have data on the MARC until ~2007, and we don't have data on the Commuter Bus or Intracounty Commuter Bus (ICC) until 2011. These will bias the `total_average_weekday_ridership` variable, so we should probably calculate totals ourselves when making comparisons across years.

We can also see that our data do not extend to the end of 2017.

```
max(maryland$DateTime, na.rm = T)
```

```
## [1] 2017.5
```

They only extend until June

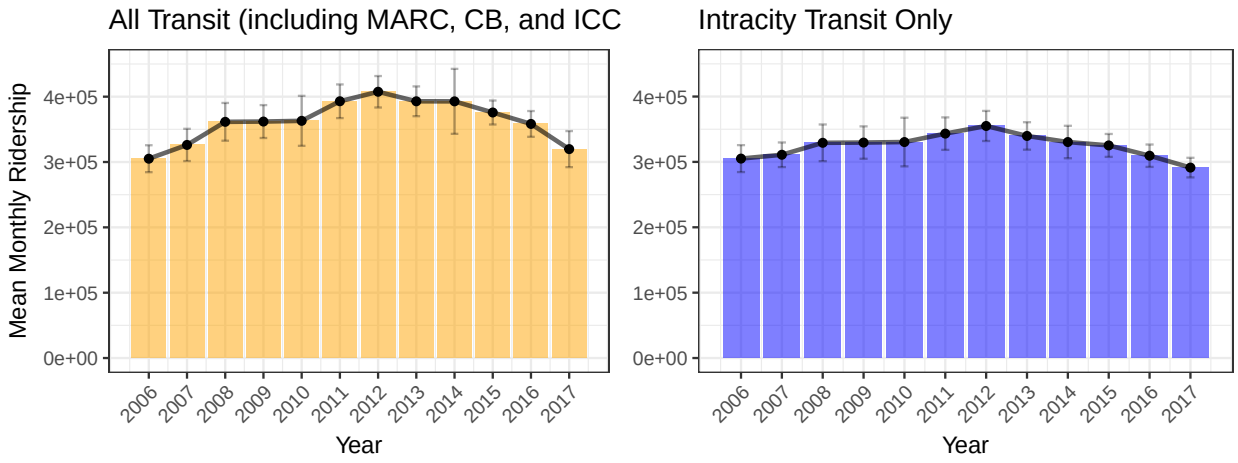
Compare total ridership including MARC, commuter bus, and ICC to intracity ridership

```
# New calculation of total ridership without MARC, Commuter bus, and ICC

maryland <- maryland %>%
    mutate(ridership_intracity = bus + metro + light_rail + mobility + taxi_access)

maryland_by_year <- maryland %>% group_by(Year) %>%
    summarize(mean = mean(total_average_weekday_ridership),
              sd = sd(total_average_weekday_ridership),
              mean_intracity = mean(ridership_intracity),
              sd_intracity = sd(ridership_intracity))
```

```
# Compare all trips
p1 <- ggplot(maryland_by_year) +
    aes(x = Year, y = mean) +
    geom_col(fill = 'orange', alpha = 0.5) +
    geom_point() +
    geom_line(alpha = 0.6, size = 1) +
```

```
    geom_errorbar(aes(ymin = mean - sd,
                      ymax = mean + sd),
                  width = 0.2, alpha = 0.3) +
    ggtitle('All Transit (including MARC, CB, and ICC)') +
    ylab('Mean Monthly Ridership') +
    scale_x_continuous(breaks = seq(2006, 2017)) +
    ylim(c(0, 450000)) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

p2 <- ggplot(maryland_by_year) +
    aes(x = Year, y = mean_intracity) +
    geom_col(fill = 'blue', alpha = 0.5) +
    geom_point() +
    geom_line(alpha = 0.6, size = 1) +
    geom_errorbar(aes(ymin = mean_intracity - sd_intracity,
                      ymax = mean_intracity + sd_intracity),
                  width = 0.2, alpha = 0.3) +
    ggtitle('Intracity Transit Only') +
    scale_x_continuous(breaks = seq(2006, 2017)) +
    ylim(c(0, 450000)) +
    theme(axis.title.y = element_blank(),
          axis.text.x = element_text(angle = 45, hjust = 1))
```

```
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



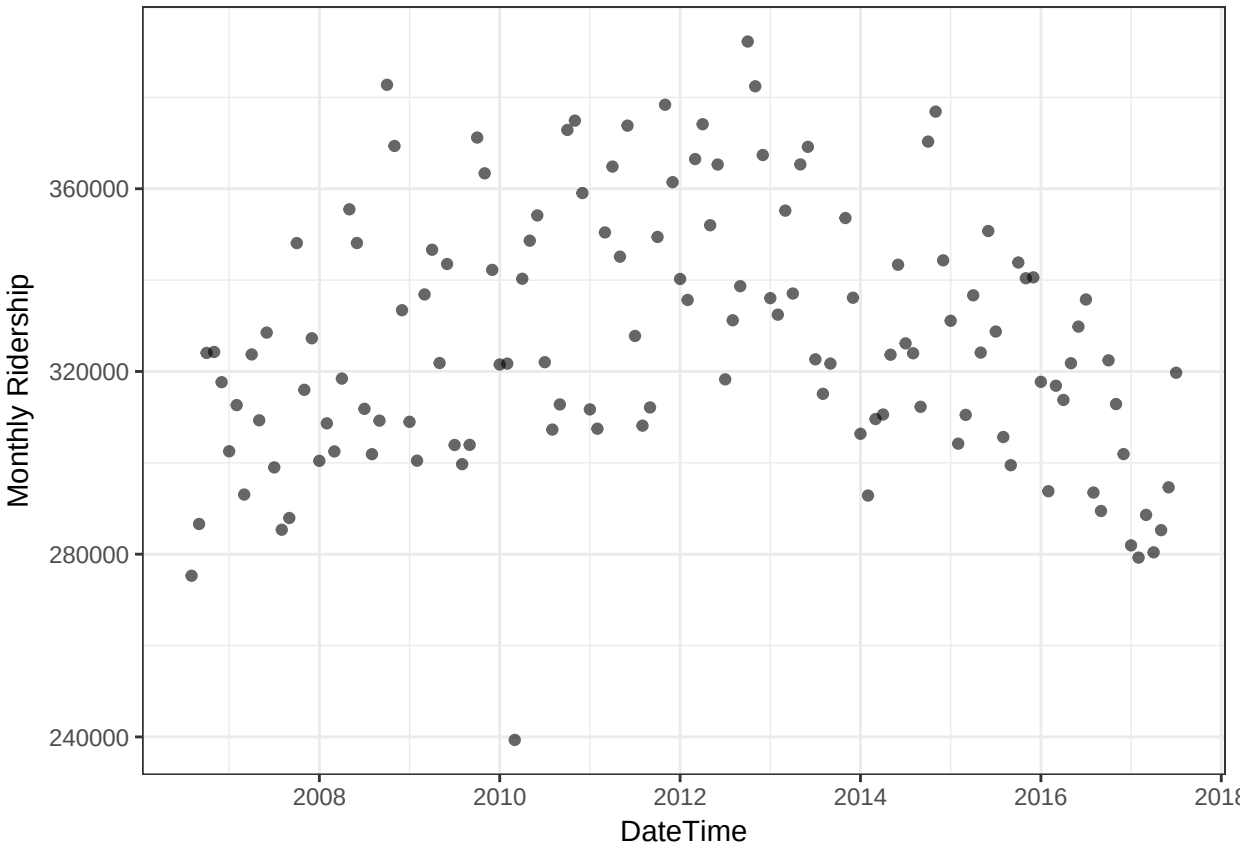
We can see that the 'jumps' in ridership in 2007/2008 and 2011/2012 mostly disappear when we only look at intracity transit.

Also recall that the data for 2017 only go up to June.

Exploratory Plots

Let's just look at ridership over time for the urban transit system

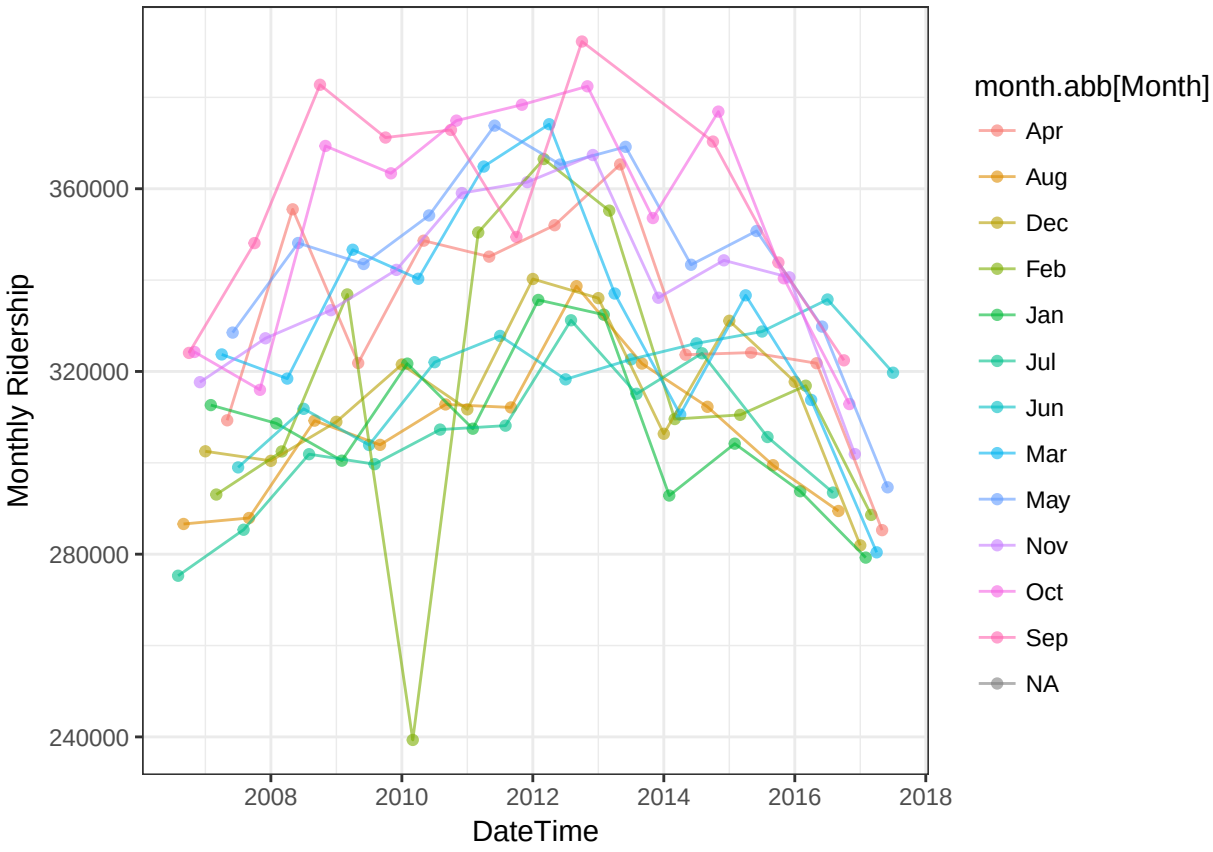
```
ggplot(maryland) +
    aes(x = DateTime, y = maryland$ridership_intracity) +
    geom_point(alpha = 0.6) +
    ylab('Monthly Ridership')
```



Unlike the Edmonton transit system, the obvious trend is that the transit system is being used less over time. We also don't see any sort of obvious monthly trends.

Looking at Monthly trends

```
ggplot(maryland) +
    aes(x = DateTime, y = ridership_intracity, col = month.abb[Month]) +
    geom_point(alpha = 0.6) +
    geom_line(aes(group = month.abb[Month]), alpha = 0.6) +
    ylab('Monthly Ridership')
```



The dip in February 2010 is interesting. A quick internet search reveals that there was a bad blizzard in February 2010 that covered the roads with 2 feet of snow.

There are no notable seasonal trends. The four least busy months appear to be December, January, July, and August.

June was one of the least busy months for transit ridership, but for what ever reason has become one of the busiest months in recent years.

Back to top