Exploratory Data Analysis

Setup and Imports

```
knitr::opts_chunk$set(echo = TRUE)
library(readr)
library(dplyr)
library(Hmisc)
library(ggplot2)
library(lubridate)
library(parsedate)
setwd('/home/xuanthu/Dropbox/W210/IDEA_dataset/')
viber_raw <- read_delim('./android/viber/total_info.txt',
                      delim='******',
                      col_names=F)

viber <- data.frame(rating=viber_raw$X1, review=viber_raw$X7, date=viber_raw$X13, version=viber_raw$X19)
viber$date <- parse_date(viber$date)
```

Summary and cleaning

```
summary(viber)
```

```
##      rating
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :4.000
##  Mean   :3.326
##  3rd Qu.:5.000
##  Max.   :5.000
##
##
##  it 39 . s very good .
##  recently i find application earn click on view publicity globe mobayl . chesny site really withdrawa
##  it do n't show real afk time !
##  if you wan na clash of clan advantage go to http . eu you receive free clash of clan gem instantly
##  unfortunately look ugly . there 39 . s always on the phone while wifi excellent work on the call pho
##  when registration be require to write to invite you login druga privet84 and get a bonus in the euro
##  (Other)
##       date                          version
##  Min.   :2014-12-17 16:00:00   5.3.0.2339:7635
##  1st Qu.:2015-03-05 16:00:00   5.2.2.478 :3872
##  Median :2015-04-03 17:00:00   5.2.1.36  :2389
##  Mean   :2015-03-21 00:06:17   5.2.1.26  :1455
##  3rd Qu.:2015-04-18 17:00:00   5.3.0.2331: 675
##  Max.   :2015-06-02 17:00:00   5.3.0.2274: 655
##                                (Other)   : 445
```

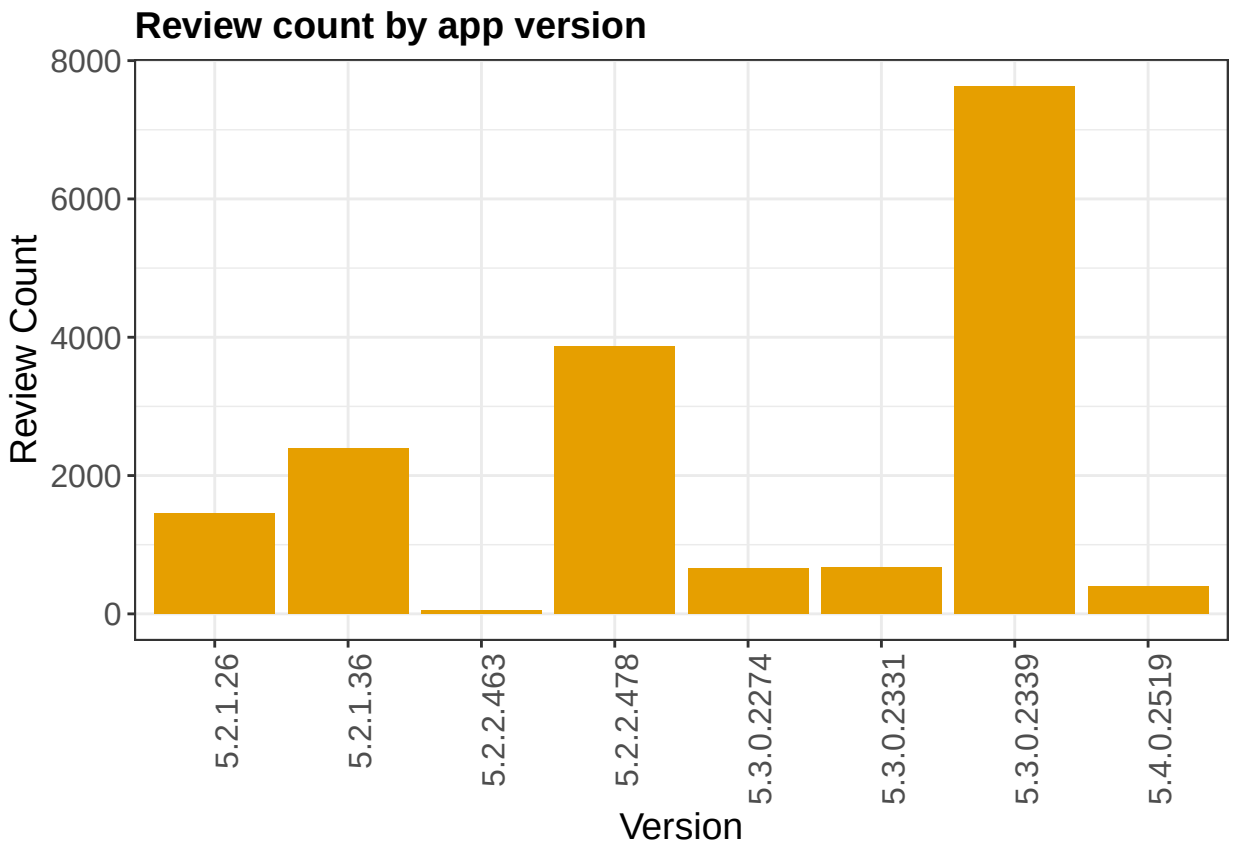
```
glimpse(viber)
```

```
## Observations: 17,126
## Variables: 4
## $ rating  <dbl> 1, 5, 2, 5, 5, 4, 3, 4, 1, 4, 5, 4, 1, 4, 5, 4, 1, 1, ...
## $ review  <fct> very good , more i want to know why me and my friend w...
## $ date    <dttm> 2015-05-01 17:00:00, 2015-04-22 17:00:00, 2015-04-22 ...
## $ version <fct> 5.2.1.26, 5.2.1.26, 5.2.1.26, 5.2.1.26, 5.2.1.26, 5.2....
```

```
describe(viber)
```

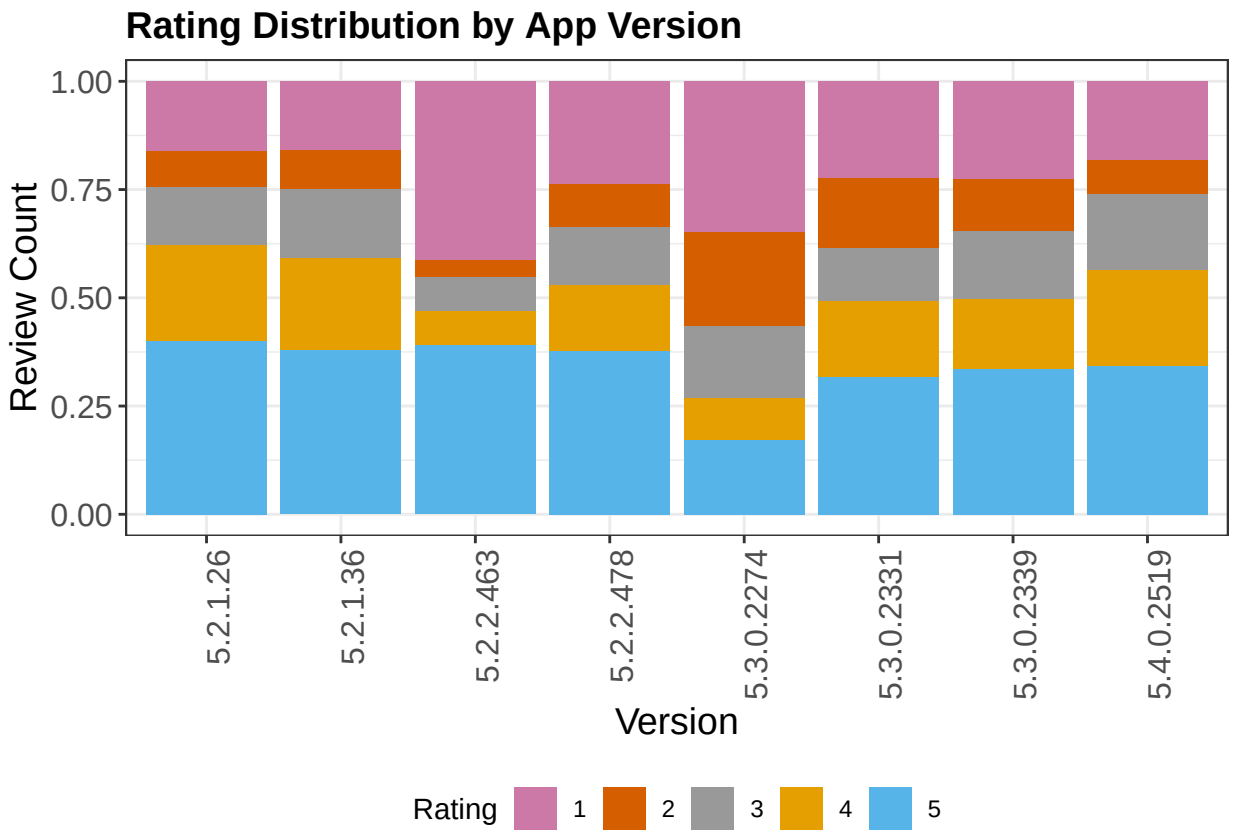
```
## viber
##
##  4  Variables      17126  Observations
## ---------------------------------------------------------------------------
## rating
##        n  missing distinct     Info     Mean      Gmd
##    17126        0        5    0.937    3.326    1.736
##
## Value          1     2     3     4     5
## Frequency   3715  1923  2567  2906  6015
## Proportion 0.217 0.112 0.150 0.170 0.351
## ---------------------------------------------------------------------------
## review
##        n  missing distinct
##    17126        0    16436
##
## lowest : 0 star for your video call im use with wifi but it buffring hahah .
## highest: zee asphalt possible tgela message or contact me mesh byrne and mesh bhovha be to open it .
## ---------------------------------------------------------------------------
## date
##                    n              missing             distinct
##                17126                    0                  155
##                 Info                 Mean                  Gmd
##                    1 2015-03-21 00:06:17 1970-02-15 04:35:53
##                  .05                  .10                  .25
## 2015-01-02 16:00:00 2015-01-07 16:00:00 2015-03-05 16:00:00
##                  .50                  .75                  .90
## 2015-04-03 17:00:00 2015-04-18 17:00:00 2015-05-01 17:00:00
##                  .95
## 2015-05-07 17:00:00
##
## lowest : 2014-12-17 16:00:00 2014-12-18 16:00:00 2014-12-19 16:00:00 2014-12-20 16:00:00 2014-12-21 1
## highest: 2015-05-24 17:00:00 2015-05-30 17:00:00 2015-05-31 17:00:00 2015-06-01 17:00:00 2015-06-02 1
## ---------------------------------------------------------------------------
## version
##        n  missing distinct
##    17126        0        8
##
## Value        5.2.1.26   5.2.1.36  5.2.2.463  5.2.2.478 5.3.0.2274
## Frequency        1455       2389         51       3872        655
## Proportion      0.085      0.139      0.003      0.226      0.038
##
## Value      5.3.0.2331 5.3.0.2339 5.4.0.2519
## Frequency         675       7635        394
## Proportion      0.039      0.446      0.023
## ---------------------------------------------------------------------------
```

```r
viber$day <- floor_date(viber$date, "day")

# color blind palette
cbPalette <- c("#CC79A7", "#D55E00", "#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2" )

# count review by version
ggplot(data = viber, aes(x = version)) +
  geom_bar(fill = "#E69F00") +
  theme_bw() +
  ylab("Review Count") +
  xlab("Version") +
  ggtitle("Review count by app version") +
  theme(axis.text.x = element_text(size = 12, angle = 90, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14),
        plot.title = element_text(lineheight=1, face="bold", size = 14))
```

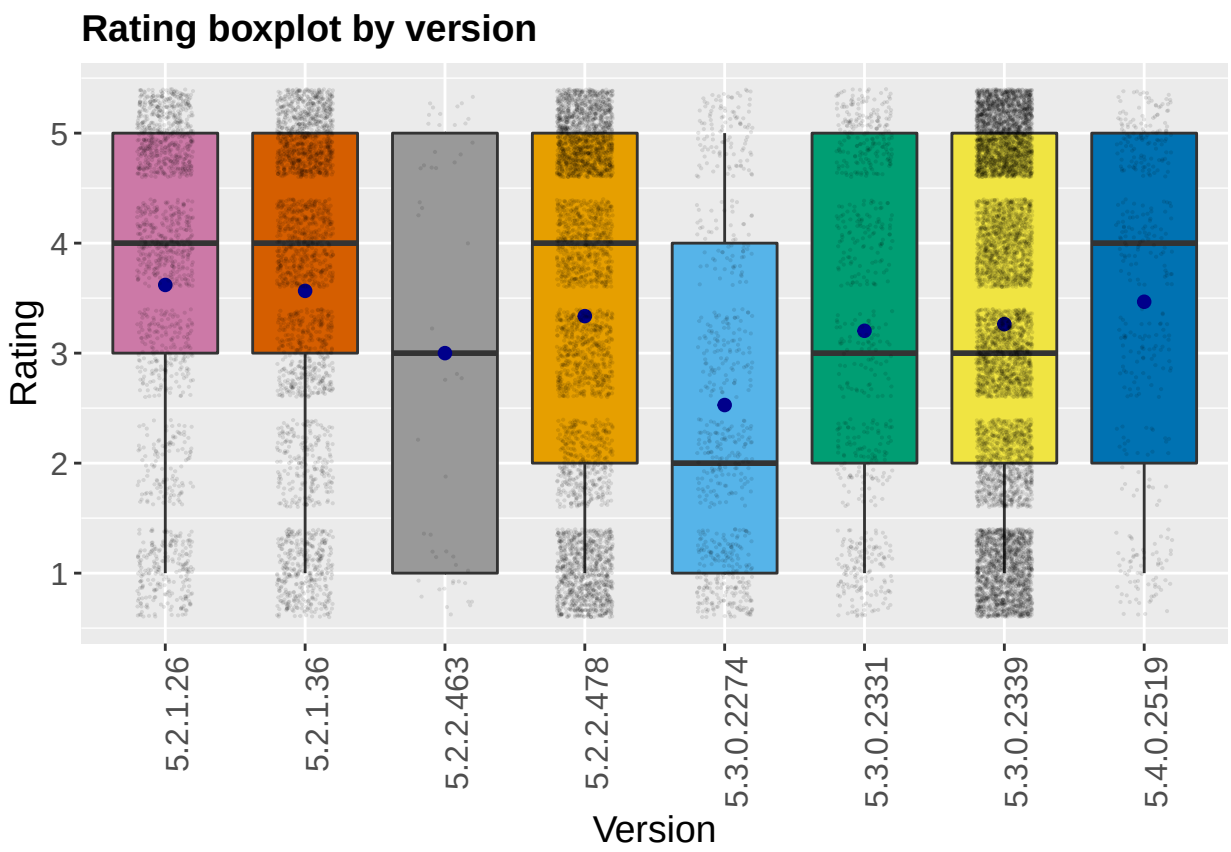



```r
ggsave(filename = 'Viber_review_count_by_version.png', width=10, height=5)

# Ratings by app versions
ggplot(viber, aes(x = version, fill = factor(rating))) +
  geom_bar(position = "fill") +
  theme_bw() +
  ylab("Review Count") +
  xlab("Version") +
  labs(fill='Rating') +
  scale_fill_manual(values=cbPalette) +
```

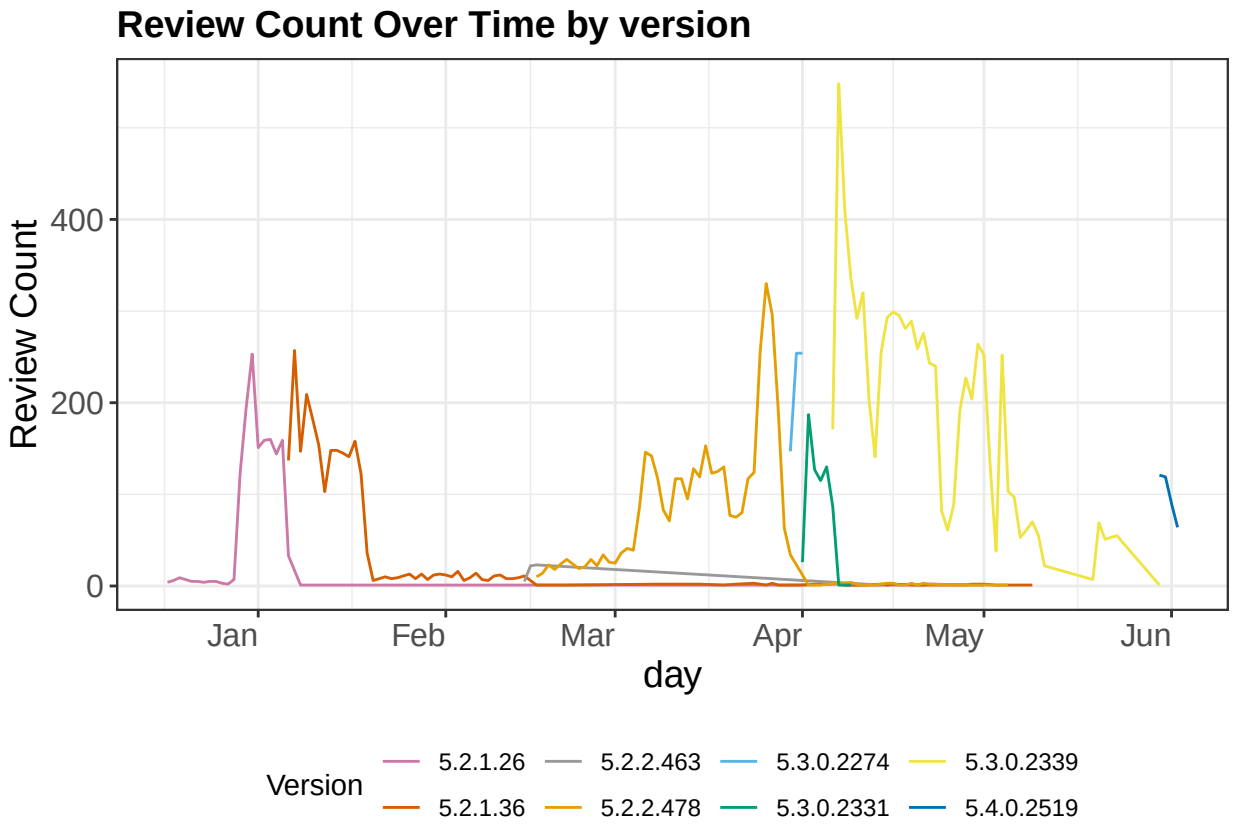
```
  ggtitle("Rating Distribution by App Version") +
  theme(axis.text.x = element_text(size = 12, angle = 90, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14),
        plot.title = element_text(lineheight=1, face="bold", size = 14),
        legend.position="bottom")
```




```
ggsave(filename = 'Viber_review_distribution_by_version_stacked_bar.png', width=10, height=5)

# boxplot, ratings by app version
ggplot(viber, aes(x = version, y = rating, fill = as.factor(version))) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=20, size=3, color="darkblue", fill="darkblue") +
  geom_jitter(width = 0.2, alpha = 0.1, size = 0.05) +
  scale_fill_manual(values=cbPalette) +
  ylab('Rating') +
  xlab('Version') +
  labs(fill = "Version") +
  ggtitle('Rating boxplot by version') +
  theme(axis.text.x = element_text(size = 12, angle = 90, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14),
        plot.title = element_text(lineheight=1, face="bold", size = 14),
        legend.position="none")
```

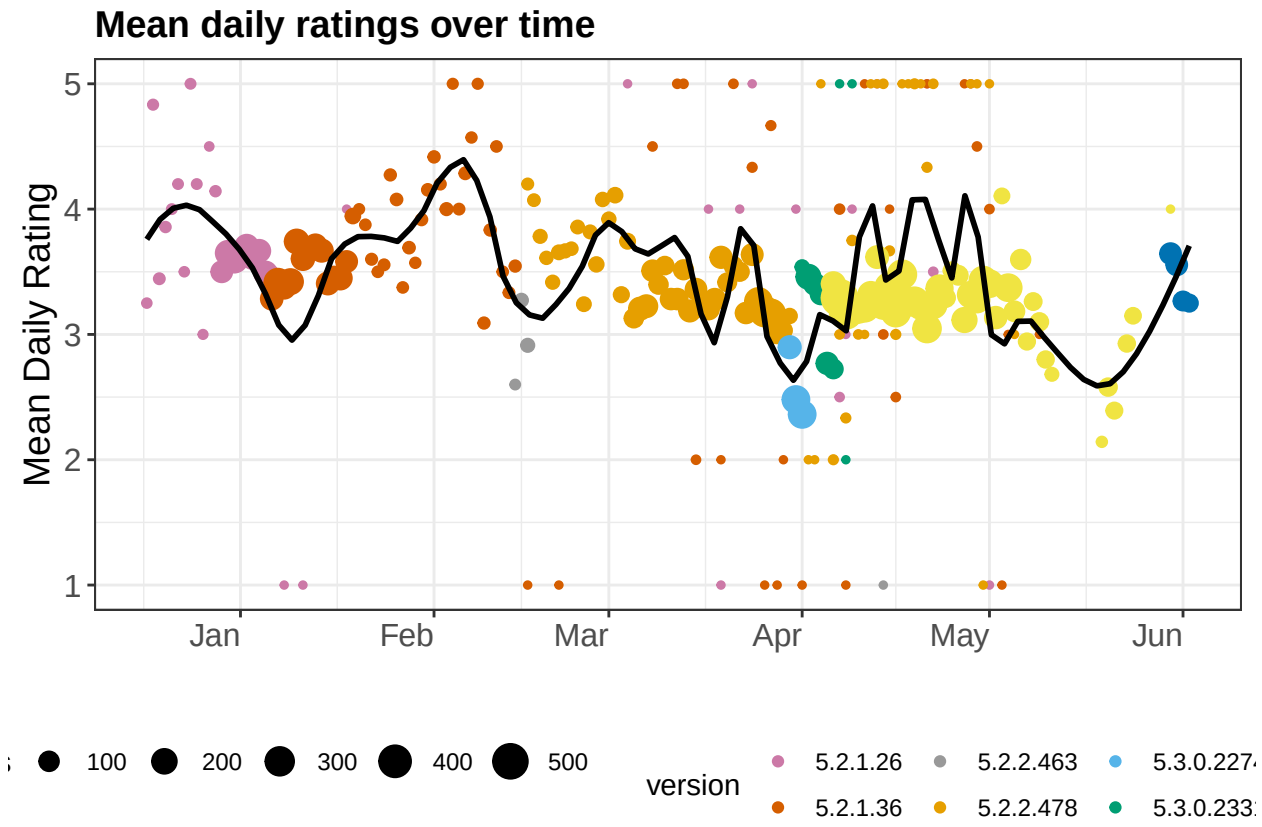



```
ggsave(filename = 'Viber_boxplot_by_version.png', width=10, height=5)

# how often do apps update?
viber %>% group_by(day, version) %>%
  summarise(review_count=n()) %>%
  ggplot(aes(x = day, y = review_count, color = version)) +
    geom_line() +
    scale_colour_manual(values=cbPalette) +
    theme_bw() +
    ylab("Review Count") +
    labs(color ='Version') +
    ggtitle("Review Count Over Time by version") +
    theme(axis.text.x = element_text(size = 12, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14),
        plot.title = element_text(lineheight=1, face="bold", size = 14),
        legend.position="bottom")
```




```
ggsave(filename = 'Viber_app_reviews_over_time_by_version.png', width=10, height=5)

viber_agg_day <- with(viber, aggregate(rating, by=list(day=day, version=version), mean))
viber_agg_day$n_reviews <- with(viber, aggregate(rating, by=list(day=day, version=version), length))[,3]

#date vs. rating
ggp <- ggplot(viber_agg_day,
              aes(x=day, y=x, color=version,size=n_reviews))
ggp + geom_point() +
  geom_smooth(inherit.aes=F, aes(x=day, y=x), span=.1, se=F, color='black') +
  scale_color_manual(values=cbPalette) +
  theme_bw() +
  ylab('Mean Daily Rating') + xlab('') +
  labs(fill = "Version") +
  ggtitle('Mean daily ratings over time') +
  theme(axis.text.x = element_text(size = 12, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14),
        plot.title = element_text(lineheight=1, face="bold", size = 14),
        legend.position="bottom")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggsave(filename = 'Viber_ratings_over_time.png', width=10, height=5)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.