R_hw9

Sihyuan_Han

Capital Bikeshare Data

- 1-1 Use a readr function to load in the trips data and the station data from the data folder

```
Sys.setlocale("LC_TIME","English")
```

```
## [1] "English_United States.1252"
```

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v dplyr   1.0.2
## v tibble  3.0.4     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
capital_trips_2016_df <- read_csv("../R_data/capital_trips_2016.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `Duration (ms)` = col_double(),
##   `Start date` = col_character(),
##   `End date` = col_character(),
```

```
##   `Start station number` = col_double(),
##   `Start station` = col_character(),
##   `End station number` = col_double(),
##   `End station` = col_character(),
##   `Bike number` = col_character(),
##   `Member Type` = col_character()
## )
```

```
capital_stations <- read_csv("../R_data/capital_stations.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   id = col_double(),
##   name = col_character(),
##   terminalName = col_double(),
##   lastCommWithServer = col_double(),
##   lat = col_double(),
##   long = col_double(),
##   installed = col_logical(),
##   installDate = col_double(),
##   removalDate = col_double(),
##   temporary = col_logical(),
##   public = col_logical(),
##   capacity = col_double()
## )
```

```
# Review and rename variables that have spaces in the names
# glimpse(capital_trips_2016_df)
capital_trips_2016_df %>%
  rename(Duration_ms = `Duration (ms)`,
         Start_date = `Start date`,
         End_date = `End date`,
         Start_station_number = `Start station number`,
         Start_station = `Start station`,
         End_station_number = `End station number`,
         End_station = `End station`,
         Bike_number = `Bike number`,
         Member_Type = `Member Type`) ->
  capital_trips_2016
glimpse(capital_trips_2016)
```

```
## Rows: 552,399
## Columns: 9
## $ Duration_ms          <dbl> 301295, 557887, 555944, 766916, 139656, 967713...
## $ Start_date           <chr> "3/31/2016 23:59", "3/31/2016 23:59", "3/31/20...
## $ End_date             <chr> "4/1/2016 0:04", "4/1/2016 0:08", "4/1/2016 0:...
## $ Start_station_number <dbl> 31280, 31275, 31101, 31226, 31011, 31266, 3122...
## $ Start_station        <chr> "11th & S St NW", "New Hampshire Ave & 24th St...
## $ End_station_number   <dbl> 31506, 31114, 31221, 31214, 31009, 31600, 3127...
## $ End_station          <chr> "1st & Rhode Island Ave NW", "18th St & Wyomin...
## $ Bike_number          <chr> "W00022", "W01294", "W01416", "W01090", "W2193...
## $ Member_Type          <chr> "Registered", "Registered", "Registered", "Reg...
```

```
glimpse(capital_stations)
```

```
## Rows: 221
## Columns: 12
## $ id                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...
## $ name               <chr> "20th & Bell St", "18th & Eads St.", "20th & Cry...
## $ terminalName       <dbl> 31000, 31001, 31002, 31003, 31004, 31005, 31006,...
## $ lastCommWithServer <dbl> 1.36837e+12, 1.36837e+12, 1.36837e+12, 1.36837e+...
## $ lat                <dbl> 38.85610, 38.85725, 38.85640, 38.86024, 38.85730...
## $ long               <dbl> -77.05120, -77.05332, -77.04920, -77.05028, -77....
## $ installed          <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ...
## $ installDate        <dbl> 1.31606e+12, 1.28499e+12, 1.28446e+12, 1.28446e+...
## $ removalDate        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ temporary          <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,...
## $ public             <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ...
## $ capacity           <dbl> 11, 19, 15, 11, 10, 11, 11, 19, 11, 11, 11, 13, ...
```

- 1-2 Use lubridate functions to convert the date-time information in the trip data to a date time variable

```
capital_trips_2016 %>%
  mutate(Start_date = force_tz(mdy_hm(Start_date), tzone = "America/New_York"),
         End_date = force_tz(mdy_hm(End_date), tzone = "America/New_York")) ->
  capital_trips_2016
glimpse(capital_trips_2016)
```

```
## Rows: 552,399
## Columns: 9
## $ Duration_ms          <dbl> 301295, 557887, 555944, 766916, 139656, 967713...
## $ Start_date           <dttm> 2016-03-31 23:59:00, 2016-03-31 23:59:00, 201...
## $ End_date             <dttm> 2016-04-01 00:04:00, 2016-04-01 00:08:00, 201...
## $ Start_station_number <dbl> 31280, 31275, 31101, 31226, 31011, 31266, 3122...
## $ Start_station        <chr> "11th & S St NW", "New Hampshire Ave & 24th St...
## $ End_station_number   <dbl> 31506, 31114, 31221, 31214, 31009, 31600, 3127...
## $ End_station          <chr> "1st & Rhode Island Ave NW", "18th St & Wyomin...
## $ Bike_number          <chr> "W00022", "W01294", "W01416", "W01090", "W2193...
## $ Member_Type          <chr> "Registered", "Registered", "Registered", "Reg...
```

- 1-3 Calculate the average number of trips for each weekday, given the day has trips. There are several days with no trips

```
capital_trips_2016 %>%
  separate(Start_date, into = c("Start_date", "Start_time"), sep = " ") %>%
  group_by(Start_date)%>%
  summarise(total_trips = n()) %>%
  mutate(wday = wday(Start_date, label = TRUE)) %>%
  group_by(wday) %>%
  summarise(mean_num_trips = mean(total_trips)) ->
  sumdf
```

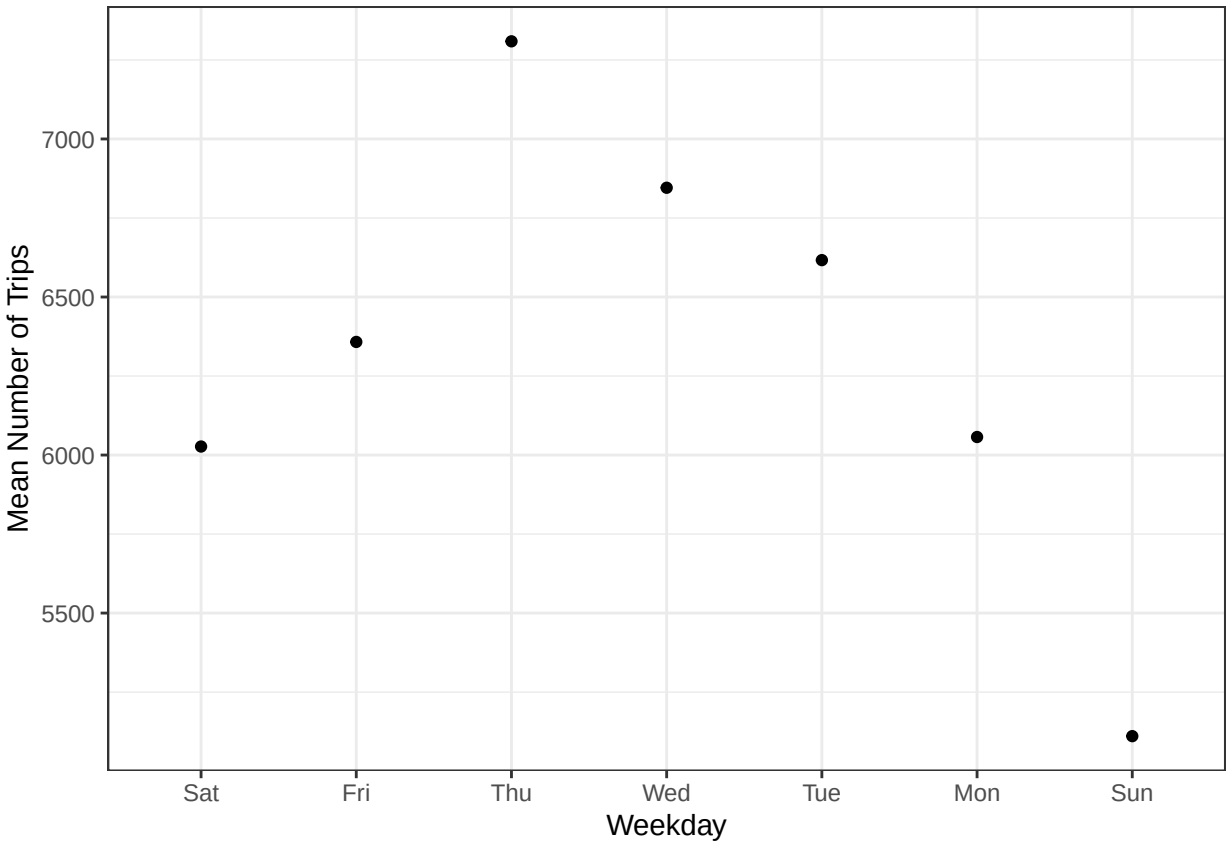
```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
glimpse(sumdf)
```

```
## Rows: 7
## Columns: 2
## $ wday           <ord> Sun, Mon, Tue, Wed, Thu, Fri, Sat
## $ mean_num_trips <dbl> 5110.667, 6057.083, 6616.667, 6845.692, 7308.923, 63...
```

- 1-4 Reproduce this plot in R

```
sumdf %>%
  ggplot(aes(x = fct_rev(wday), y = mean_num_trips)) +
  geom_point() +
  theme_bw() +
  xlab("Weekday") +
  ylab("Mean Number of Trips")
```



- 1-5 In a stunning show of contempt, the IEEE Computer Society decided to add a new weekday called "Fooday" with abbreviation "Foo"

```
Fooday_df <- tribble(~wday, ~mean_num_trips,
                     "Foo", 12567)
wday_abb <- c("Foo","Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat" )
bind_rows(Fooday_df, sumdf) %>%
```

```
  mutate(wday = parse_factor(wday, levels = wday_abb)) %>%
  mutate(mean_num_trips = round(mean_num_trips)) ->
  sumdf_w_foo
sumdf_w_foo
```

```
## # A tibble: 8 x 2
##   wday  mean_num_trips
##   <fct>          <dbl>
## 1 Foo            12567
## 2 Sun             5111
## 3 Mon             6057
## 4 Tue             6617
## 5 Wed             6846
## 6 Thu             7309
## 7 Fri             6358
## 8 Sat             6027
```

- 1-6 In another stunning show of contempt, the IEEE Computer Society decided to change the abbreviations from three letters to two letters

```
sumdf_w_foo %>%
  mutate(wday = fct_recode(wday,
                           "Fo" = "Foo",
                           "Su" = "Sun",
                           "Mo" = "Mon",
                           "Tu" = "Tue",
                           "We" = "Wed",
                           "Th" = "Thu",
                           "Fr" = "Fri",
                           "Sa" = "Sat")) ->
  sumdf_shortabb
sumdf_shortabb
```

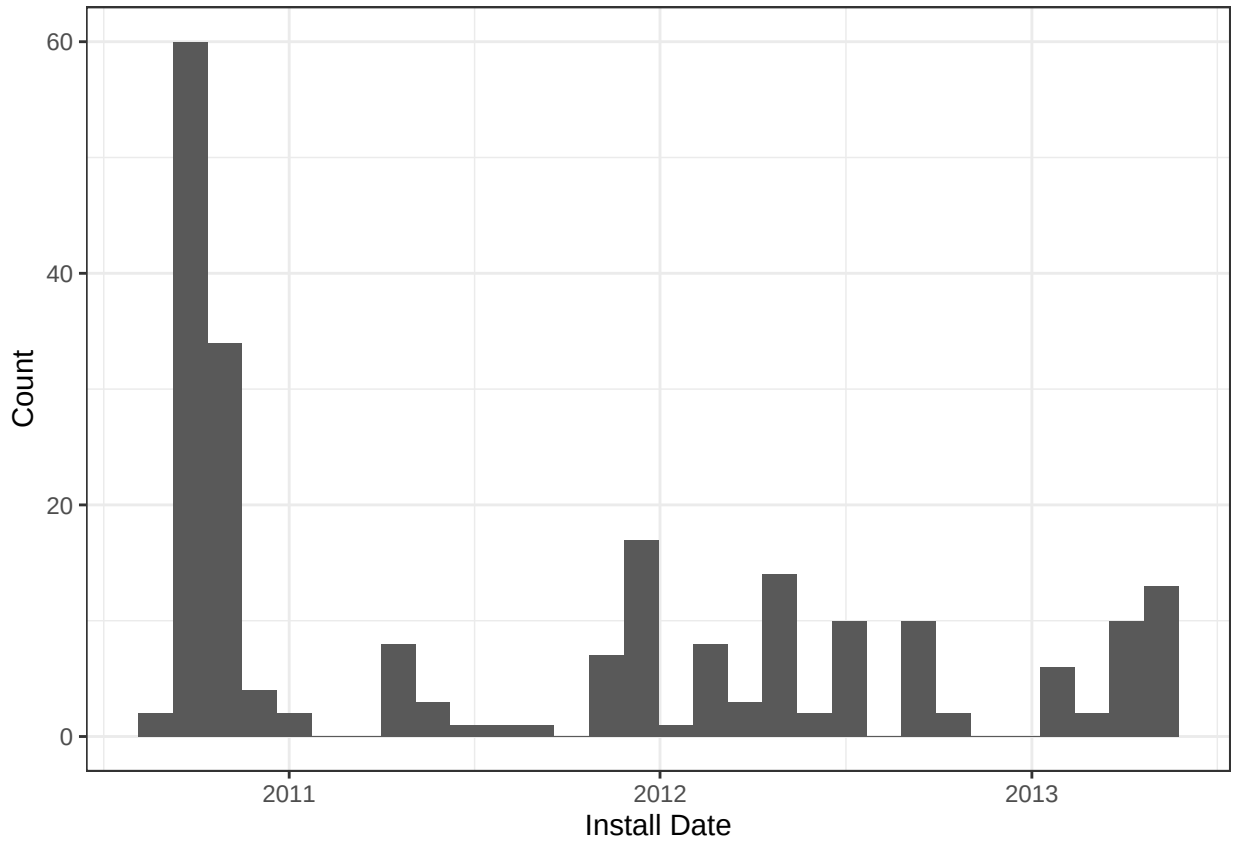
```
## # A tibble: 8 x 2
##   wday  mean_num_trips
##   <fct>          <dbl>
## 1 Fo             12567
## 2 Su              5111
## 3 Mo              6057
## 4 Tu              6617
## 5 We              6846
## 6 Th              7309
## 7 Fr              6358
## 8 Sa              6027
```

- 1-7

```
capital_stations %>%
  mutate(start_time = ymd_hms("1970-01-01 00:00:00", tz = "America/New_York"),
         installDate = start_time + dmilliseconds(installDate)) %>%
  ggplot(aes(x = installDate)) +
```

```
  geom_histogram() +
  xlab("Install Date") +
  ylab("Count") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Reddit Data

- 2-1 Use a readr function to read in the data from the all_comments.csv in the data folder

```
all_comments_df <- read_csv("../R_data/all_comments.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   id = col_double(),
##   structure = col_character(),
##   post_date = col_character(),
##   comm_date = col_character(),
##   num_comments = col_double(),
##   subreddit = col_character(),
##   upvote_prop = col_double(),
```

```
##   post_score = col_double(),
##   author = col_character(),
##   user = col_character(),
##   comment_score = col_double(),
##   controversiality = col_double(),
##   comment = col_character(),
##   title = col_character(),
##   post_text = col_character(),
##   link = col_character(),
##   domain = col_character(),
##   URL = col_character(),
##   flair = col_character()
## )
```

```
glimpse(all_comments_df)
```

```
## Rows: 37,106
## Columns: 19
## $ id               <dbl> 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1,...
## $ structure        <chr> "1", "1_1", "1", "1", "1", "1", "1", "2", "1", "1"...
## $ post_date        <chr> "01-04-20", "01-04-20", "01-04-20", "01-04-20", "0...
## $ comm_date        <chr> "01-04-20", "01-04-20", "01-04-20", "01-04-20", "0...
## $ num_comments     <dbl> 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1,...
## $ subreddit        <chr> "anaesthesia", "anaesthesia", "anaesthesia", "anae...
## $ upvote_prop      <dbl> 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1....
## $ post_score       <dbl> 6, 6, 1, 2, 1, 1, 1, 1, 2, 3, 2, 2, 3, 6, 1, 5, 2,...
## $ author           <chr> "PA1GR", "PA1GR", "alfentazolam", "alfentazolam", ...
## $ user             <chr> "alfentazolam", "PA1GR", "alfentazolam", "alfentaz...
## $ comment_score    <dbl> 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1,...
## $ controversiality <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ comment          <chr> "Are you an HMO or registrar? \n\nTry this link\nh...
## $ title            <chr> "Resources on ventilators? Would be good if its fo...
## $ post_text        <chr> NA, NA, NA, NA, NA, NA, NA, NA, "# International A...
## $ link             <chr> "https://www.reddit.com/r/anaesthesia/comments/fsq...
## $ domain           <chr> "self.anaesthesia", "self.anaesthesia", "drmarksus...
## $ URL              <chr> "http://www.reddit.com/r/anaesthesia/comments/fsqm...
## $ flair            <chr> "none", "none", "none", "none", "none", "none", "n...
```

- 2-2 Use a lubridate function to convert the character dates into date variables so there are no parsing errors

```
all_comments_df %>%
  mutate(post_date = dmy(post_date),
         comm_date = dmy(comm_date),
         subreddit = as.factor(subreddit)) ->
  all_comments
glimpse(all_comments)
```

```
## Rows: 37,106
## Columns: 19
## $ id               <dbl> 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1,...
## $ structure        <chr> "1", "1_1", "1", "1", "1", "1", "1", "2", "1", "1"...
```

```
## $ post_date        <date> 2020-04-01, 2020-04-01, 2020-04-01, 2020-04-01, 2...
## $ comm_date        <date> 2020-04-01, 2020-04-01, 2020-04-01, 2020-04-01, 2...
## $ num_comments     <dbl> 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1,...
## $ subreddit        <fct> anaesthesia, anaesthesia, anaesthesia, anaesthesia...
## $ upvote_prop      <dbl> 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1....
## $ post_score       <dbl> 6, 6, 1, 2, 1, 1, 1, 1, 2, 3, 2, 2, 3, 6, 1, 5, 2,...
## $ author           <chr> "PA1GR", "PA1GR", "alfentazolam", "alfentazolam", ...
## $ user             <chr> "alfentazolam", "PA1GR", "alfentazolam", "alfentaz...
## $ comment_score    <dbl> 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1,...
## $ controversiality <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ comment          <chr> "Are you an HMO or registrar? \n\nTry this link\nh...
## $ title            <chr> "Resources on ventilators? Would be good if its fo...
## $ post_text        <chr> NA, NA, NA, NA, NA, NA, NA, NA, "# International A...
## $ link             <chr> "https://www.reddit.com/r/anaesthesia/comments/fsq...
## $ domain           <chr> "self.anaesthesia", "self.anaesthesia", "drmarksus...
## $ URL              <chr> "http://www.reddit.com/r/anaesthesia/comments/fsqm...
## $ flair            <chr> "none", "none", "none", "none", "none", "none", "n...
```

- 1-3 Compute the difference between the post date and the comment date as a period and remove all records where the difference is 0 and save the data frame

```
all_comments %>%
  mutate(period = as.duration(comm_date-post_date)) %>%
  filter(period != 0) ->
  all_comments_new
glimpse(all_comments_new)
```

```
## Rows: 15,261
## Columns: 20
## $ id               <dbl> 2, 1, 1, 2, 1, 3, 2, 2, 3, 4, 5, 5, 6, 7, 8, 9, 13...
## $ structure        <chr> "1_1", "1", "1", "1_1", "1", "3", "1_1", "2", "2_1...
## $ post_date        <date> 2020-04-17, 2020-05-02, 2020-03-07, 2020-03-07, 2...
## $ comm_date        <date> 2020-04-18, 2020-05-10, 2020-03-14, 2020-03-14, 2...
## $ num_comments     <dbl> 2, 1, 2, 2, 1, 3, 2, 5, 5, 5, 5, 24, 24, 24, 24, 2...
## $ subreddit        <fct> anaesthesia, anaesthesia, Cardiology, Cardiology, ...
## $ upvote_prop      <dbl> 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 0.60, 0....
## $ post_score       <dbl> 2, 2, 2, 2, 2, 11, 3, 1, 1, 1, 1, 24, 24, 24, 24, ...
## $ author           <chr> "alfentazolam", "jianfa-ben-tsai", "Super_tachy", ...
## $ user             <chr> "alfentazolam", "alfentazolam", "[deleted]", "Supe...
## $ comment_score    <dbl> 2, 1, 1, 1, 1, 0, 2, 2, 1, 2, 1, 3, 1, 1, 1, 1, 3,...
## $ controversiality <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ comment          <chr> "https://www.jaybro.com.au/about/choosing-the-righ...
## $ title            <chr> "Use of coveralls as an alternative option for non...
## $ post_text        <chr> NA, NA, "Hi everyone - quick question for the sub ...
## $ link             <chr> "https://icmanaesthesiacovid-19.org/news/use-of-co...
## $ domain           <chr> "icmanaesthesiacovid-19.org", "self.ideas_jianfa",...
## $ URL              <chr> "http://www.reddit.com/r/anaesthesia/comments/g31m...
## $ flair            <chr> "none", "none", "none", "none", "PhD, FAHA", "none...
## $ period           <Duration> 86400s (~1 days), 691200s (~1.14 weeks), 6048...
```

- 1-4 Reproduce the following plot

```r
all_comments_new %>%
  group_by(subreddit, period) %>%
  summarise(median_num_com = median(num_comments)) %>%
  mutate(subreddit = fct_reorder(subreddit, median_num_com)) %>%
  mutate(subreddit = fct_relevel(subreddit, "anaesthesia",
                                 after = length(levels(all_comments_new$subreddit))-1)) %>%
  mutate(subreddit = fct_relevel(subreddit, "COVID19",
                                 after = length(levels(all_comments_new$subreddit))-3)) %>%
  ggplot(aes(x = period, y = median_num_com, color = subreddit)) +
  scale_y_log10() +
  geom_point() +
  geom_smooth(se = F, method = "lm") +
  xlab("Time After Posting of Comments") +
  ylab("Medain Number of Comments") +
  ggtitle("Comments by Subreddit Over Time")
```

```
## `summarise()` regrouping output by 'subreddit' (override with `.groups` argument)
```

```
## `geom_smooth()` using formula 'y ~ x'
```