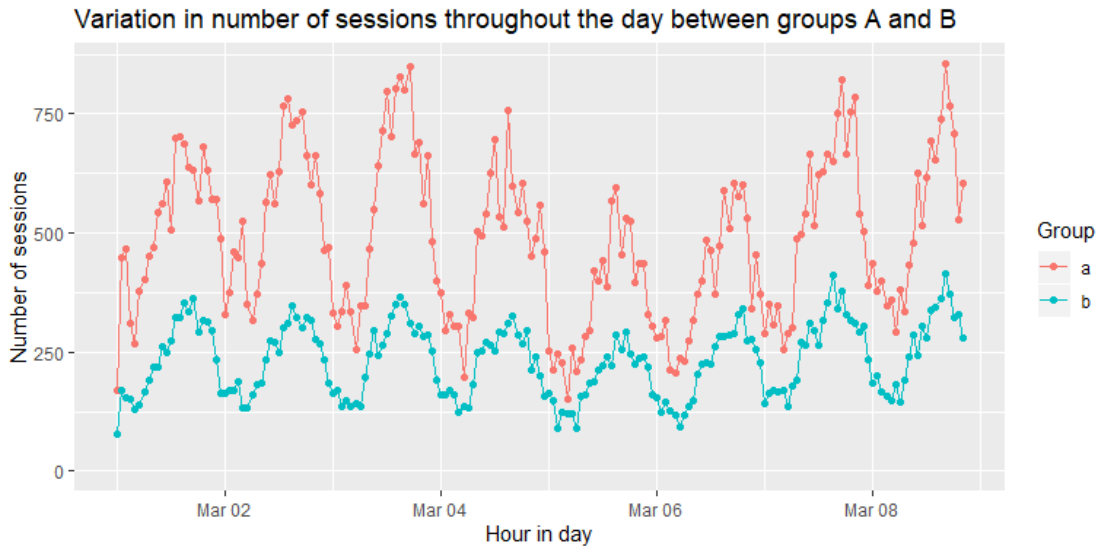
Lab_1-Wikimedia_Interview_Task-STATS_369

Pavan Mani

2/08/2019

TASK 1

```
# Create dataframe for logged events
events_log <- read.csv("C:/Users/pavan/Downloads/events_log.csv")

# In order to carry out our analysis, we will parse and format the timestamp
data. This will make it easier to make any groupings on our original
dataframe that requires the timestamp.
events_log <- transform(events_log, timestamp=
parse_date_time(as.character(timestamp),orders = "ymdHMS"))

# Round the timestamp data too the nearest whole hour for easier analysis.
Group logged events by timestamp and group (a or b). Calculate the total
number of sessions for each timestamp for each group.
sessions_by_day_df <- events_log %>%
  mutate(new_timestamp=round_date(timestamp, unit = "hour"))%>%
  group_by(new_timestamp, group) %>%
  summarise(number_of_sessions = sum(action == 'searchResultPage'))


# Visually see how session numbers vary throughout each day between groups A
and B.
ggplot(data = sessions_by_day_df, aes(x = new_timestamp, y =
number_of_sessions, color = group)) + geom_line() + geom_point() +
labs(title="Variation in number of sessions throughout the day between groups
A and B", x = "Hour in day", y = "Number of sessions", colour = "Group")
```

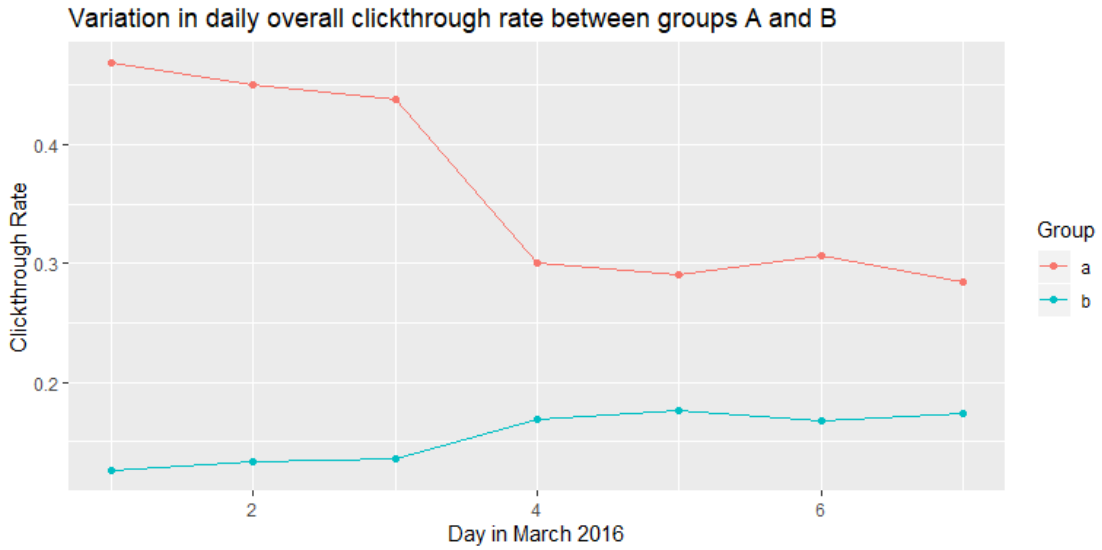



```
# From the side by side line plot generated above, we can see that we do not
have logged events for the entirety of 2016-03-08 (or day 8). Therefore, we
will filter logged events that occur on 2016-03-08 and only consider the
first 7 days (start of 2016-03-01 to end of 2016-03-07) when calculating the
daily clickthrough rate for both groups.
sessions <- events_log %>%
  mutate(new_timestamp=round_date(timestamp, unit = "hour")) %>%
  filter(day(new_timestamp) != 8)

#Group logged events by day and group. Calculate the daily clickthrough rate
by dividing total number of clicks by the total number of sessions each day
for each group.
daily_clickrate_df <- sessions %>%
  mutate(day = day(new_timestamp)) %>%
  group_by(day, group) %>%
  summarise(number_of_sessions = sum(action ==
'searchResultPage'),number_of_clicks = sum(action ==
'visitPage'),daily_clickthrough_rate = number_of_clicks/number_of_sessions)

# Show event details by group in a summary table.
print.data.frame(daily_clickrate_df)
```

```
##    day group number_of_sessions number_of_clicks daily_clickthrough_rate
## 1    1     a              12442             5828               0.4684134
## 2    1     b               5708              719               0.1259636
## 3    2     a              13180             5938               0.4505311
## 4    2     b               5692              756               0.1328180
## 5    3     a              13259             5809               0.4381175
## 6    3     b               5874              797               0.1356827
## 7    4     a              11336             3403               0.3001941
## 8    4     b               5382              912               0.1694537
## 9    5     a               8584             2496               0.2907735
## 10   5     b               4592              809               0.1761760
```

```
## 11    6      a                9491                2908                0.3063955
## 12    6      b                5142                 862                0.1676391
## 13    7      a               12618                3590                0.2845142
## 14    7      b                6304                1099                0.1743338
```

```
# Plot variation in daily clickthrough rate between groups A and B.
ggplot(data = daily_clickrate_df,aes(x= day, y=daily_clickthrough_rate, color = group)) + geom_point() + geom_line()+labs(title = "Variation in daily overall clickthrough rate between groups A and B", x = "Day in March 2016", y = "Clickthrough Rate", colour = "Group")
```




```
# From the plot of daily clickthrough rates, we can observe that the clickthrough rate is larger in group A than in group B for all the 7 days. Furthermore, we can observe that the clickthrough rate decreases after day 3 for group A while it increases for group B. This suggests the effectiveness of the Search Engine in returing relevant results to users in group A by the Search Engine was lower, whereas the opposite occurred for those in group B.
```