Singapore Grab Data Analysis

Hour distribution

```
data1 <- alldata[[1]]
hourdistr <- data1 %>%
  mutate(time = as_datetime(pingtimestamp)) %>%
  mutate(time = with_tz(time, tzone = "Singapore")) %>%
  mutate(hour = hour(time), day_of_week = weekdays(time), day = day(time)) %>%
  distinct(trj_id, hour, day_of_week, day)

for (v in 2:10) {
  temp <- alldata[[v]]
  temp <- temp %>%
    mutate(time = as_datetime(pingtimestamp)) %>%
    mutate(time = with_tz(time, tzone = "Singapore")) %>%
    mutate(hour = hour(time), day_of_week = weekdays(time), day = day(time)) %>%
    distinct(trj_id, hour, day_of_week, day)

  hourdistr <- rbind(hourdistr, temp)
}

hourdistr <- hourdistr %>%
  mutate(day_of_week = factor(day_of_week,
                              levels = c("Monday", "Tuesday", "Wednesday", "Thursday",
                                         "Friday", "Saturday", "Sunday"))) %>%
  distinct() %>%
  arrange(day)
```
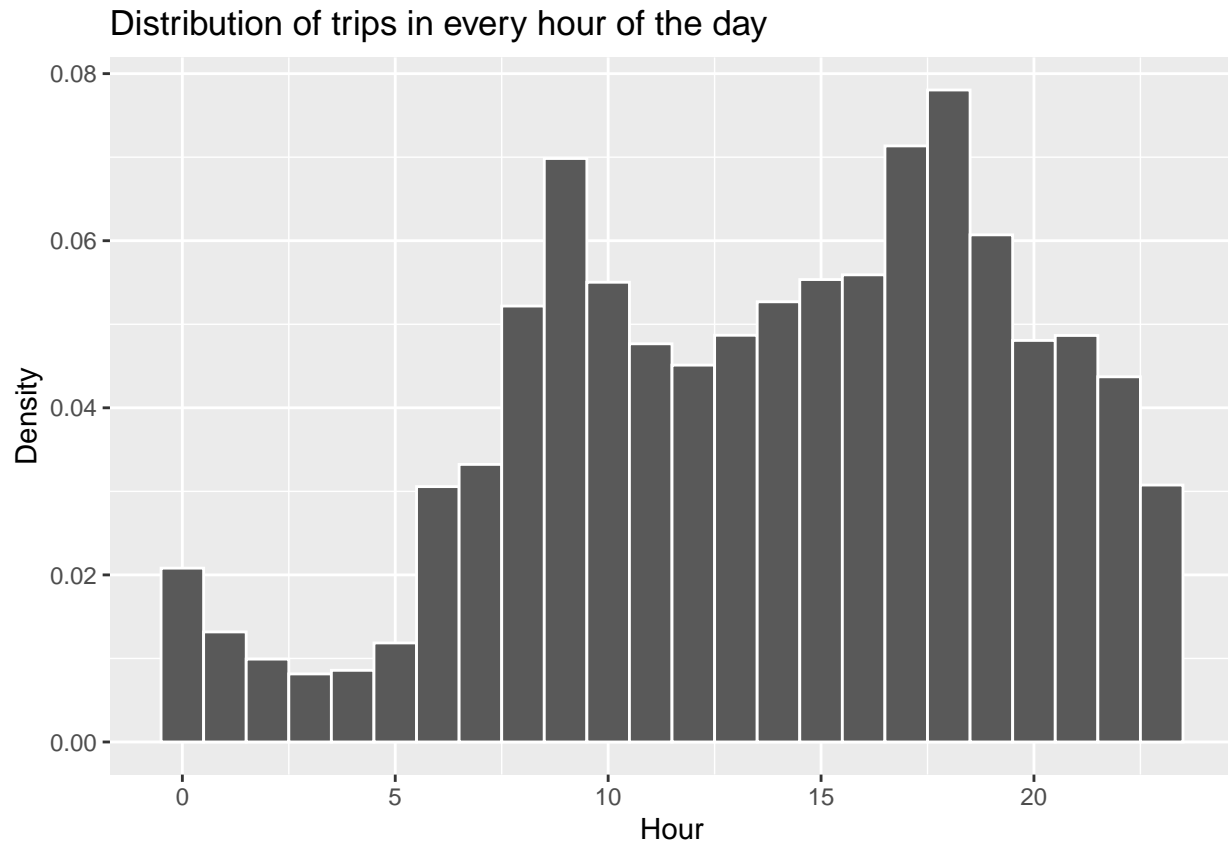
In the above R code we store the date, day of the week and hours of the day where each `trj_id` occurs into `hourdistr`.

Overall distribution of trips by hour

```
ggplot(hourdistr) +
  geom_histogram(aes(hour, y = ..density..), binwidth = 1, color = "white") +
  labs(title = "Distribution of trips in every hour of the day", y = "Density", x = "Hour")
```

## Distribution of trips in every hour of the day

By looking at the distribution of trips in every hour, we can see that there is a peak around 9am in the morning and 6pm in the evening, which suggest that there is a higher demand for Grab rides during those hours.

Distribution of trips by hour, by day of the week

```
ggplot(hourdistr) +
  geom_histogram(aes(hour), binwidth = 1, color = "white") +
  labs(title = "Distribution of trips in every hour of the day, by day of the week", y = "Number of trip
  facet_wrap(~day_of_week)
```

## Distribution of trips in every hour of the day, by day of the week
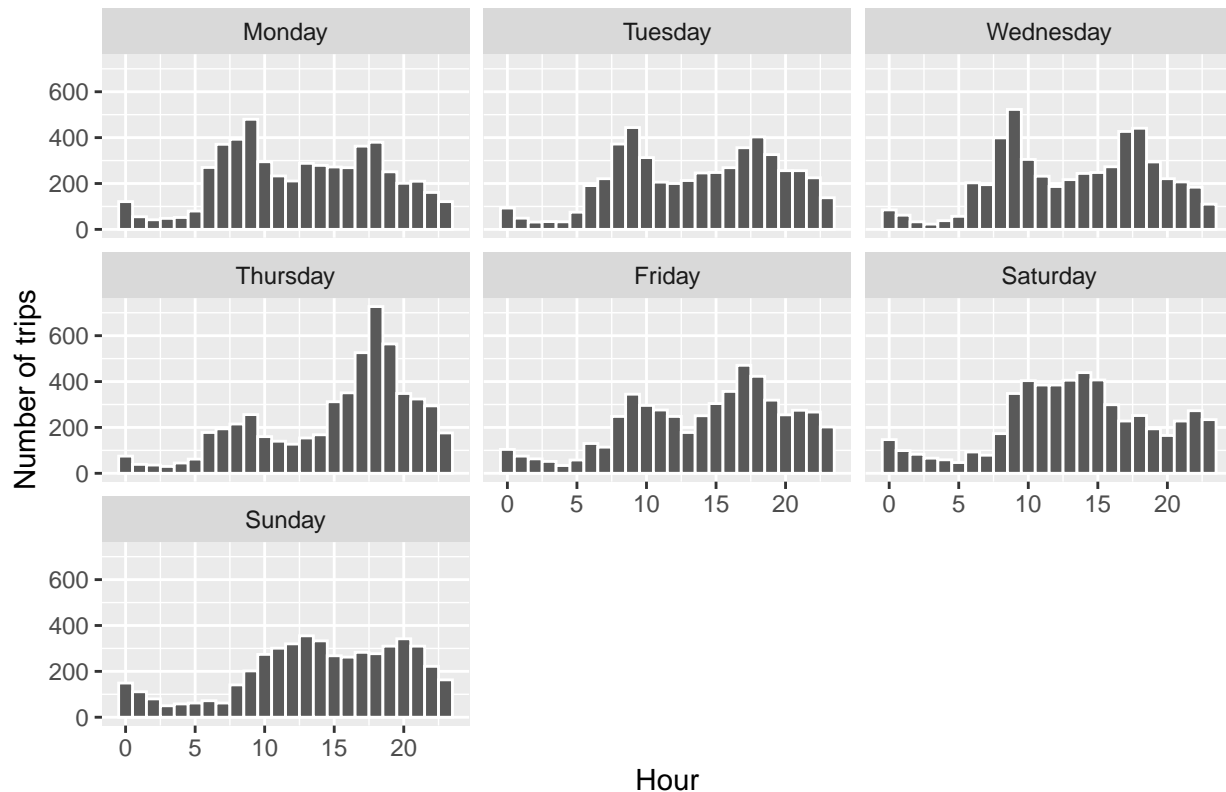
When we look at the distribution of trips by day of the week, we can see a few trends. Firstly, there is a large number of trips specifically at 6pm on Thursdays. There also seems to be a larger spike in demand on weekdays (especially Monday, Tuesday and Wednesday) around 9am compared to on weekends (Saturday and Sunday) around that time.

Possible impact of Good Friday

```
friday_names <- c("12" = "12 April 2019", "19" = "19 April 2019 (Good Friday)")
hourdistr %>%
  filter(day_of_week == "Friday") %>%
  ggplot() +
  geom_histogram(aes(hour), color = "white", binwidth = 1) +
  labs(title = "Distribution of trips by hour on 2 different Fridays", y = "Count", x = "Hour") +
  facet_wrap(~day, labeller = as_labeller(friday_names))
```

## Distribution of trips by hour on 2 different Fridays
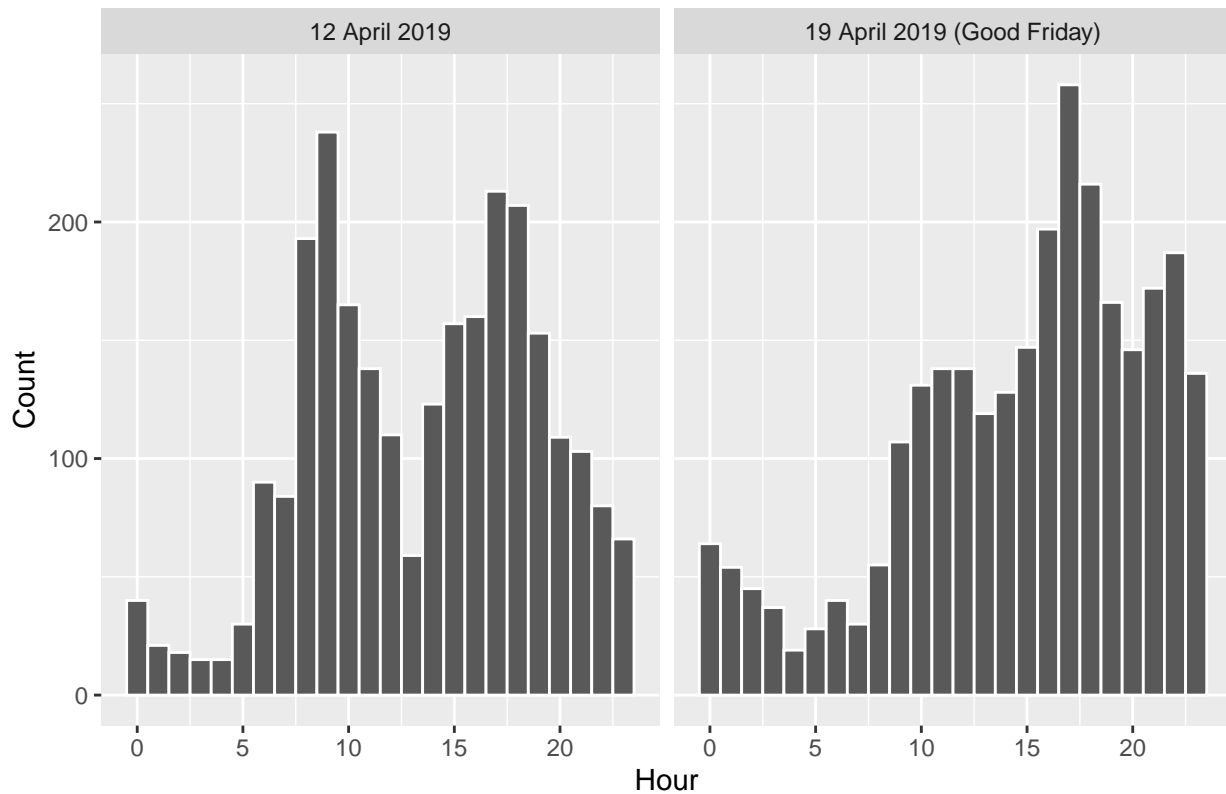
Since out of the 2 Fridays in the data, one of them is a public holiday (19 April 2019, Good Friday), we also try to compare the distribution by hour on those days. We can see that compared to 12 April, 19 April sees no spike in demand at 9am, but there is greater demand from 8pm onwards.

Possible impact of weather

```
weather_labels <- c("13" = "13 April 2019", "20" = "20 April 2019 (Rainy)")
hourdistr %>%
  filter(day == 20 | day == 13) %>%
  ggplot() +
  geom_histogram(aes(hour), color = "white", binwidth = 1) +
  labs(title = "Distribution of trips by hour on 2 different Saturdays", y = "Count", x = "Hour") +
  facet_wrap(~day, labeller = as_labeller(weather_labels))
```

Distribution of trips by hour on 2 different Saturdays
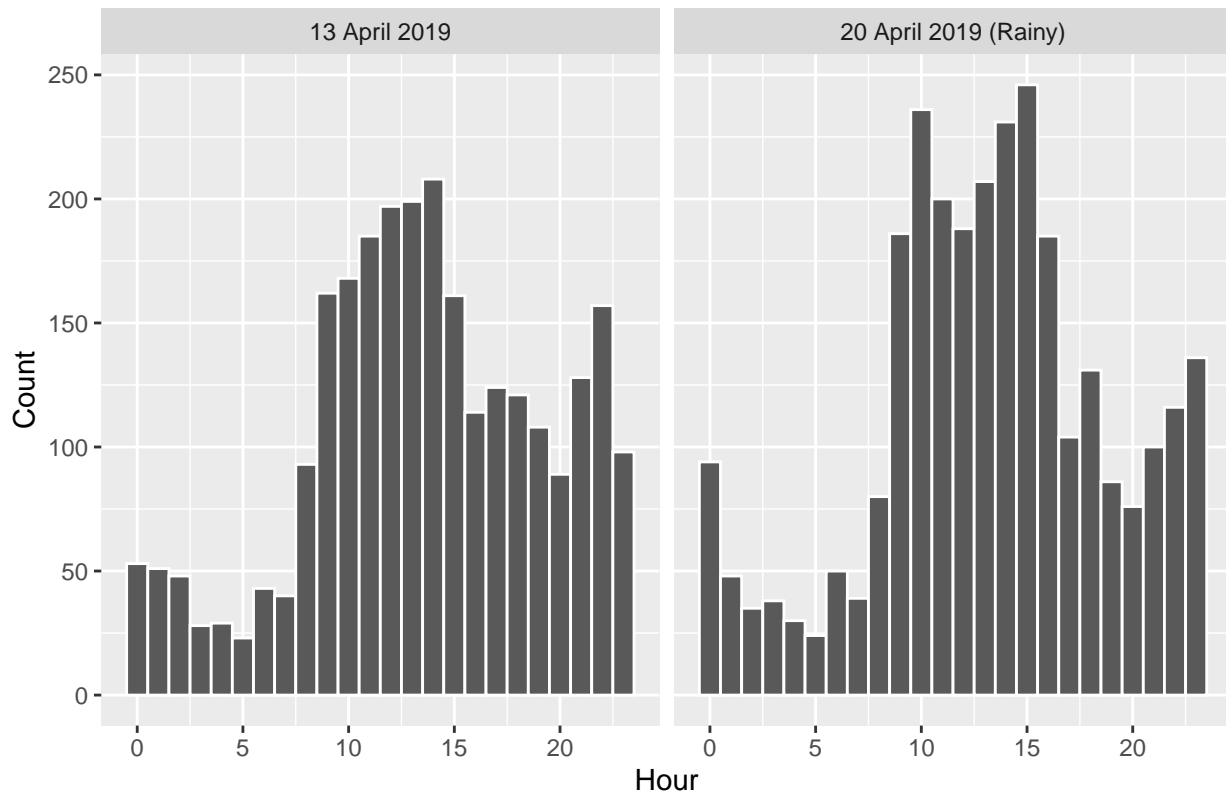
Between the 2 Saturdays in our data, there was a thunderstorm on 20 April 2019 between 10am and 4pm. We can see that between 10am and 4pm, there was higher demand on 20 April compared to 13 April.

Final thoughts

As we are only using 2 weeks worth of data, our observations are very sensitive to outliers and it may be difficult to draw any concrete conclusions. However, we can use similar data analysis on more data that we gather to see if the trends we identified hold over a larger set of data, which would mean that they are more likely to be statistically significant. For example, we can look to see if the trend of greater demand on Thursday 6pm holds over more weeks, or if the average demand during periods with thunderstorm is greater than when the weather is fine.