Weather and Rideshare Ridership

John Cruz

2023-04-25

Introduction

The National Oceanic and Atmospheric Administration ('NOAA') defines the heat index as the apparent temperature of what the temperature feels like to the human body when relative humidity is combined with the air temperature. This has important considerations for the human body's comfort. When the body gets too hot, it begins to perspire or sweat to cool itself off.

As for the New York City subway system during the summer, it is notoriously known to have unbearable temperatures where the platform can be 104 degrees, compared to 86 degrees outside ('Curbed NY').

Given the health risks, and general discomfort during high heat days, this project will look into alternative modes of transportation, particularly ridesharing companies such as Uber and Lyft.

Research question

Does high heat index days (>=90 degrees) increase the number of trips taken with Uber or Lyft compared to non-high heat index days?

Data Source

Weather (Oikolab)

Data was collected using Oikolab API historical data API service. It collects its data from the ECWMF and NOAA. Each case represents hourly weather measurements in August 2022.

Uber & Lyft Trips (NYC Taxi and Limousine Commission)

Data was collected using the available 'parquet files'. The agency collects the data from Uber and Lyft. Each case represents a trip taken either via Uber or Lyft in the month of August 2022.

Type of study

This is an observational study.

Variables

Dependent The response variable is total trips and is numerical

Independent Variable(s) The independent variables are:

- type_of_day: categorical
- precipitation: numerical
- day_of_week: categorical

Note: Other potential factors that are important but not included: special events (i.e. sporting event), major delays with public transportation (MTA Subway) or alternative transportation such as Citi bikes.

Required Libraries

```
library(tidyverse)
library(arrow)
library(lubridate)
library(infer)
library(psych)
library(GGally)
```

Data Preparation

Load Historical Weather Data

Note: The data has been cleaned and filtered using *weather_filter.Rmd* that is within the same GitHub repo. Here are the changes:

Calculate Heat Index

The measurements for the United States is generally in Fahrenheit. The weather data will be converted from Celsius to Fahrenheit using the *weathermetrics* library.

- Relative humidity is calculated using the temperature and dewpoint temperature.
- Heat index is calculated using the temperature and relative humidity.

Day of Week

Using the *lubridate* library, we will determine the day of the week and transform the data type with factor levels. The datetime_utc will also be updated to New York's local time to match the trip records.

```
weather <- read_parquet('weather.parquet') |>
  janitor::clean_names()

knitr::kable(head(weather))
```

datetime_ny	temp_deg_f	rel_humidity	heat_idx	total_precip
2022-07-31 20:00:00	79.34	69.89607	82	0.0007874
2022-07-31 21:00:00	78.69	73.22915	81	0.0003937
2022-07-31 22:00:00	78.17	75.51439	80	0.0055118
2022-07-31 23:00:00	77.13	77.92819	78	0.0000000
2022-08-01 00:00:00	76.59	77.55404	78	0.0000000
2022-08-01 01:00:00	75.56	79.02121	77	0.0000000

Load Uber and Lyft Trips

The NYC Taxi and Limousine Commission provides a data dictionary 'here'. The rideshare app companies such as Uber is coded as (HV0003) and Lyft (HV0005).

Note: The data has been cleaned and filtered using *tlc_data_filter.Rmd* that is within the same GitHub repo. Here are the changes:

- All the trips performed in August were found to be within two separate files for both August and September.
- Trips were filtered because of huge outliers that were present such as:
  - Trip time had to be >0 seconds and <= 5 hours.
  - Trip miles had to be >= 0.
  - Driver pay > $0.01.
  - Base passenger fare > $0.01.
  - Pickup locations had to be within the NYC region and not unknown/outside of it.

```
tlc_trips <- read_parquet('tlc_trips.parquet')

knitr::kable(head(tlc_trips))
```

pickup_datetime	total_trips	total_trip_dist	total_trip_time	total_base_fare
2022-08-01 00:00:00	16367	95228.86	16290733	392988.8
2022-08-01 01:00:00	9386	49653.46	8660832	191243.2
2022-08-01 02:00:00	6842	36202.17	6285157	140393.1
2022-08-01 03:00:00	5736	33645.40	5453926	131228.5
2022-08-01 04:00:00	6362	46159.03	6863760	206663.8
2022-08-01 05:00:00	10228	74357.79	11680789	317810.9

Merge Datasets

```
trip_weather_data <-
  tlc_trips |>
  left_join(weather, by = join_by(pickup_datetime == datetime_ny)) |>
  mutate(pickup_date = date(pickup_datetime)) |>
  select(!pickup_datetime) |>
  group_by(pickup_date) |>
  mutate(daily_trips = sum(total_trips),
         daily_trip_dist = sum(total_trip_dist),
         daily_trip_time = sum(total_trip_time),
```

```
        daily_base_fare = sum(total_base_fare),
        max_temp_deg_f = max(temp_deg_f),
        max_heat_idx = max(heat_idx),
        daily_precip = sum(total_precip),
        .keep = "none") |>
  distinct() |>
  mutate(type_of_day = ifelse(max_heat_idx >= 90, 'hot', 'not_hot'),
        day_of_week = factor(wday(pickup_date, label = TRUE, week_start = 1), ordered = FALSE))

trip_weather_data$day_of_week = relevel(trip_weather_data$day_of_week, ref='Mon')
trip_weather_data$type_of_day = relevel(factor(trip_weather_data$type_of_day, ordered = FALSE), ref='no

write_parquet(trip_weather_data, "trip_weather_data.parquet")
```

Summary Statistics

Maximum Daily Heat Index Highest Daily Heat Index throughout August 2022

```
trip_weather_data |>
  ggplot(aes(x = pickup_date, y = max_heat_idx, color = type_of_day)) +
  geom_point(stat = 'identity') +
  theme_classic() +
  labs(x = '', y ='', title = "Maximum Daily Heat Index", subtitle = "August 2022", caption = "Figure 1
  theme(legend.position = "top",
        legend.justification = c(0, 1)) +
  scale_color_manual(values = c('#4E79A7', '#F28E2B'), name = '')
```

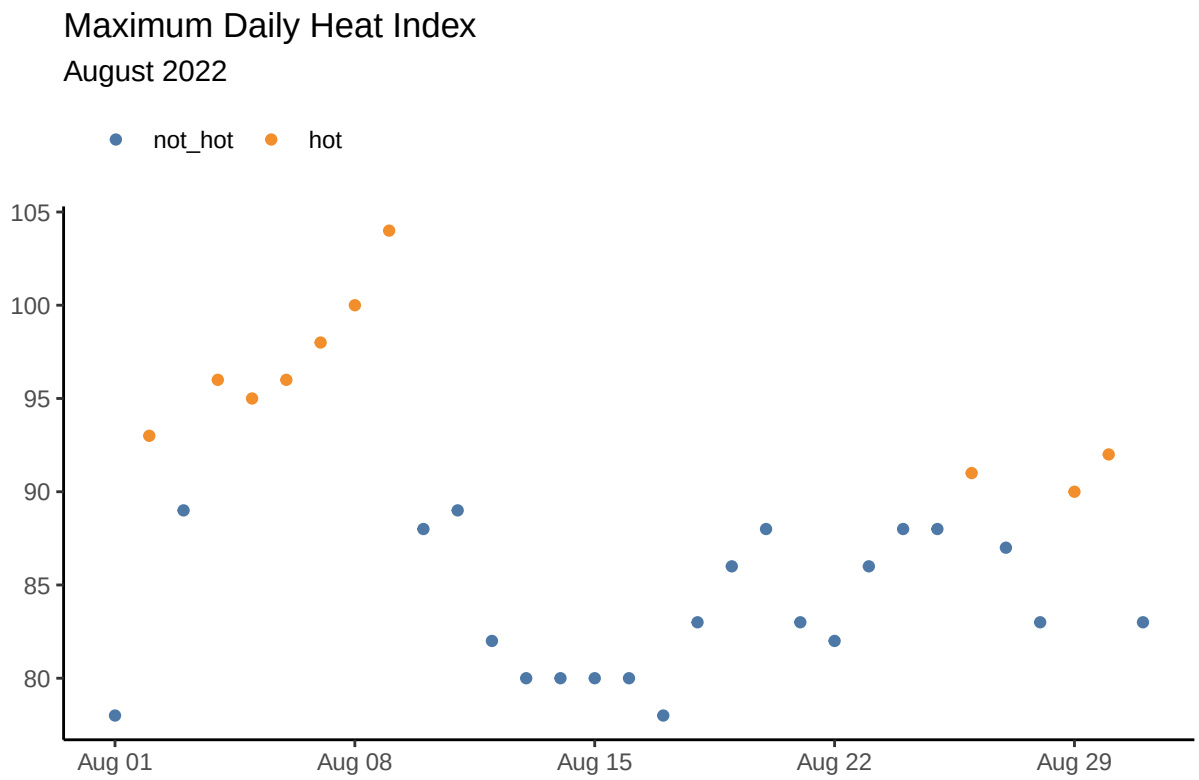


Figure 1

Total Daily Trips Total daily trips in August.

```
trip_weather_data |>
  ggplot(aes(x = pickup_date, y = daily_trips)) +
  geom_line(stat = 'identity') +
  scale_y_continuous(labels = scales::comma) +
  theme_bw() +
  labs(title = "Total Daily Trips",
       subtitle = "August 2022",
       caption = "Figure 2")
```

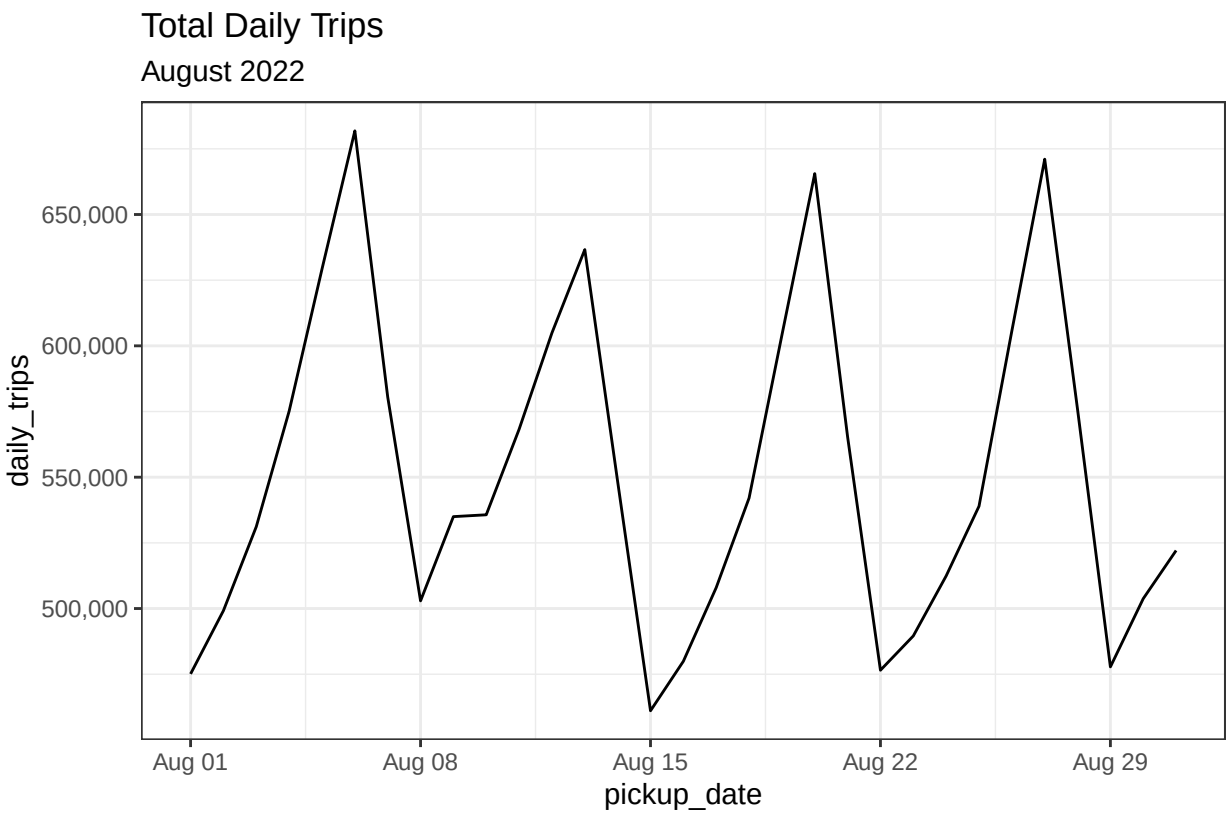


Figure 2

Heat Index vs. Daily Trips

```
trip_weather_data |>
  ggplot(aes(x = max_heat_idx, y = daily_trips)) +
  geom_point(stat = 'identity') +
  labs(title = 'Max Heat Index vs. Daily Trips', subtitle = 'August 2022', caption = "Figure 3", x = 'h
  theme_classic()
```

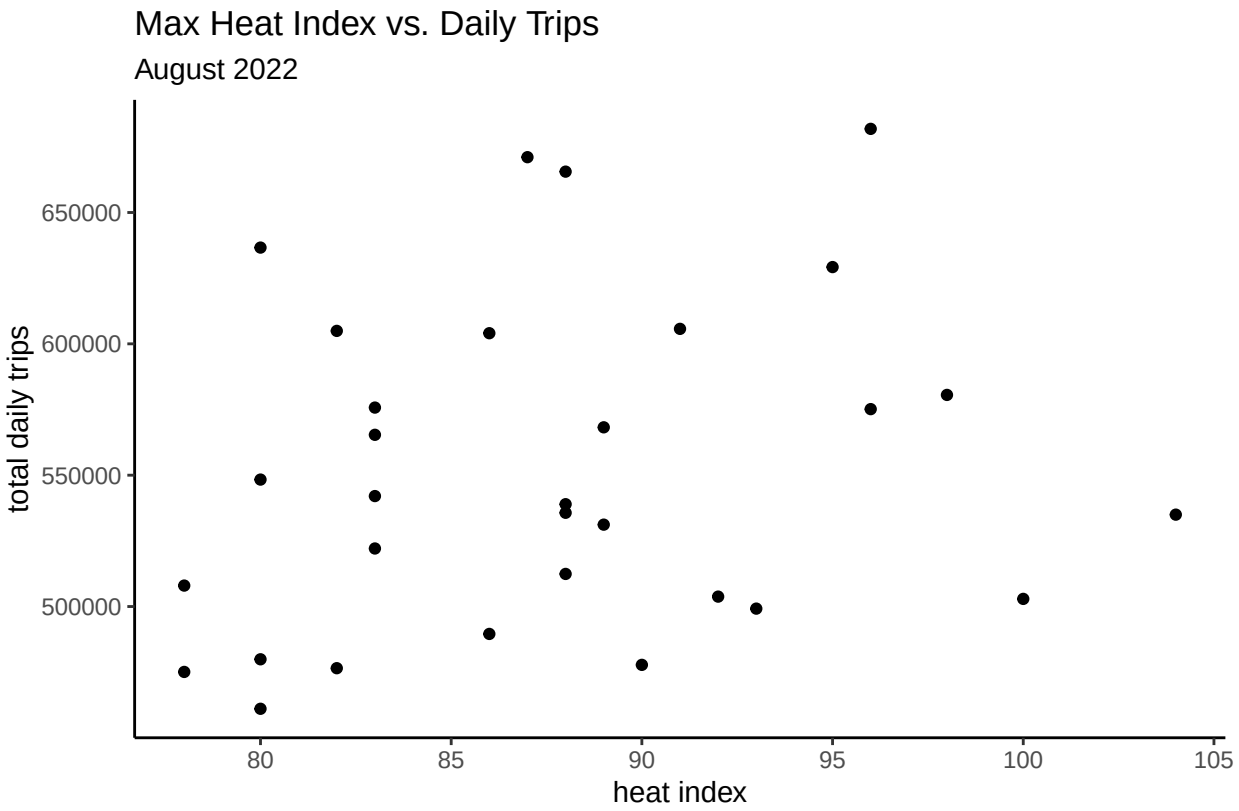


Figure 3

Observed Difference

```
comparison <-
  trip_weather_data |>
  group_by(type_of_day) |>
  summarise(avg_trips = mean(daily_trips),
            n = n())

knitr::kable(comparison, caption = "Observed Differences")
```

Table 3: Observed Differences

type_of_day	avg_trips	n
not_hot	548213.7	21
hot	559114.9	10

Here we can see a scatterplot that shows no general trend where as the heat index increases, the number of trips increase as well. However, viewing the observed difference averages, we do see that there is a small increase of average trips taken on high heat index days compared to non-high heat index days.

Hypothesis Testing

To test if the small increase of ridership is statistically significant, a hypothesis test will be used to confirm this. The test will utilize bootstrapping to create a normal distribution of the one-month sample.

```
set.seed(2023)
```

```
obs_diff <- trip_weather_data %>%
  specify(daily_trips ~ type_of_day) %>%
  calculate(stat = "diff in means", order = c("hot", "not_hot"))
```

```
null_dist <- trip_weather_data %>%
  specify(daily_trips ~ type_of_day) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("hot", "not_hot"))
```

```
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram(bins = 30) +
  labs(caption = "Figure 4")
```

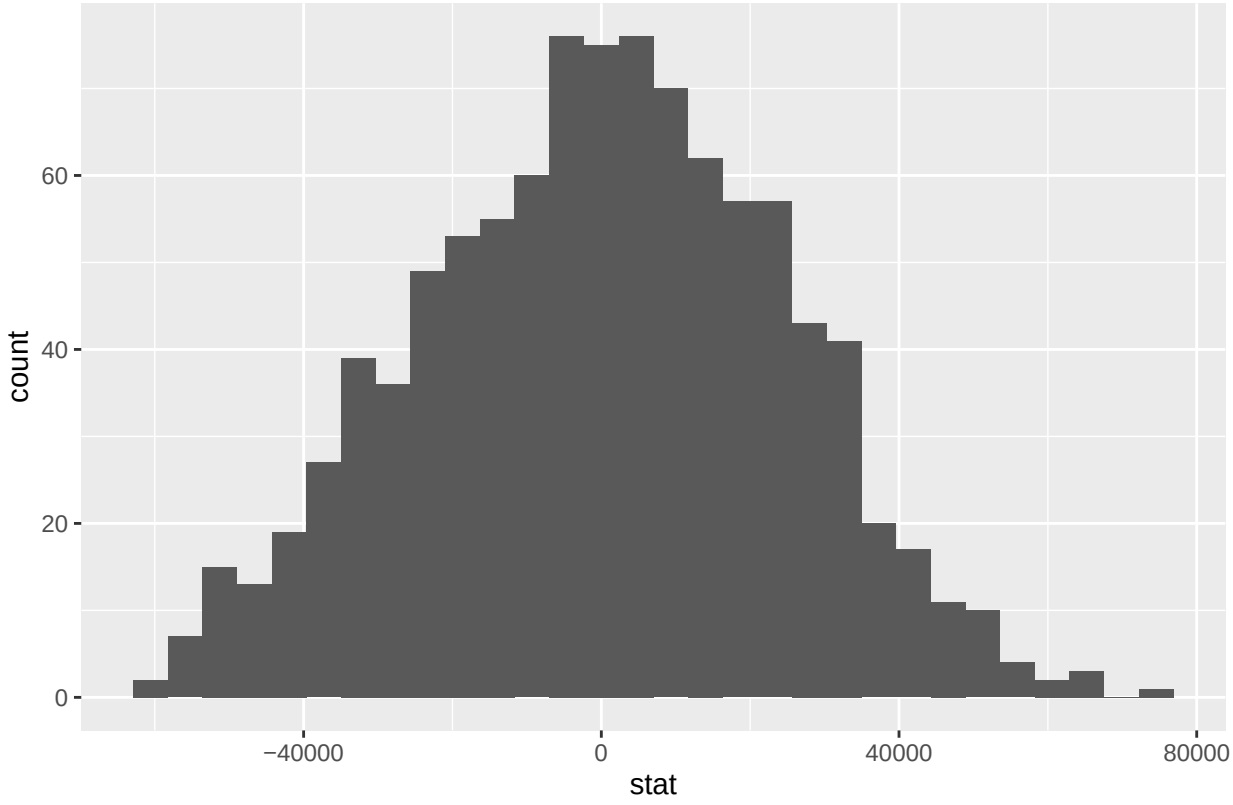


Figure 4

After confirming the normality of the distribution a two-tailed t-test will be performed.

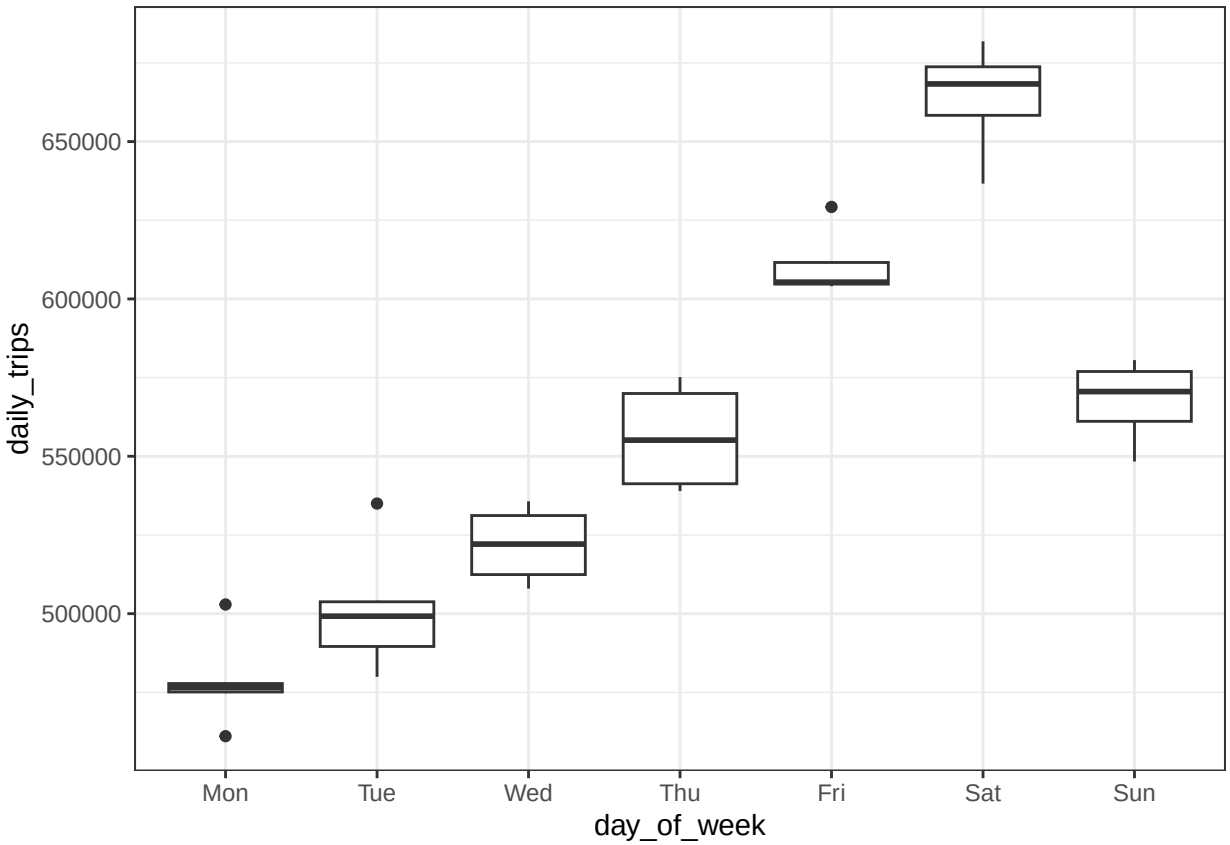
```
t.test(daily_trips ~ type_of_day, data = trip_weather_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  daily_trips by type_of_day
## t = -0.43742, df = 16.491, p-value = 0.6675
## alternative hypothesis: true difference in means between group not_hot and group hot is not equal to
## 95 percent confidence interval:
##  -63605.41  41802.94
## sample estimates:
## mean in group not_hot     mean in group hot
##              548213.7              559114.9
```

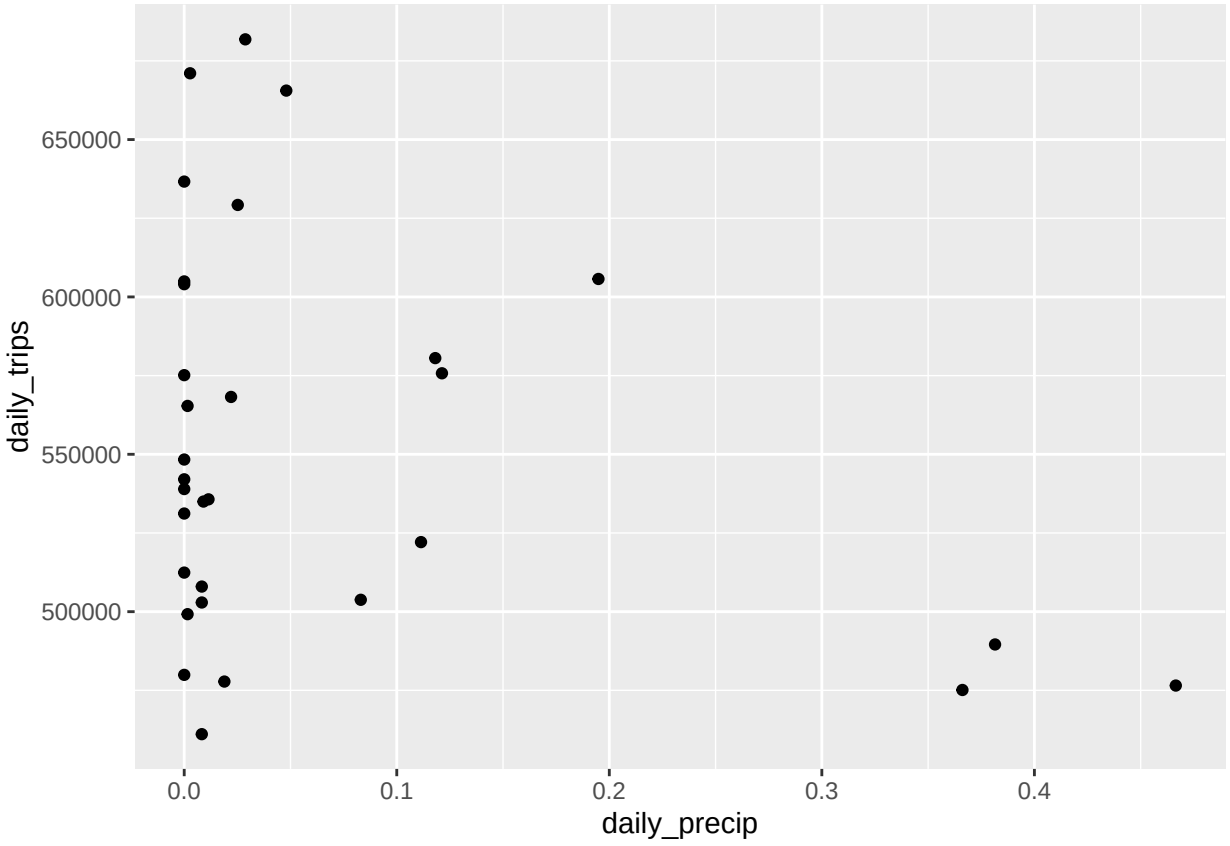
Using a 95% confidence t-test, it tells us that we are unable to reject the null hypothesis and claim that a high heat index increases the total daily trips taken using Uber or Lyft. We see in the confidence interval, it includes zero, where as the p-value is also 0.6675 which is />0.05 alpha threshold.

Even with the failure to reject the null hypothesis, there are other controlling variables to account for that may influence how a heat index may be prove to be useful. A multiple linear regression model will be created to account for the previous seasonal variability accounted for in *Figure 2.* This will factor out each day of the week, having Monday as the reference variable. Also included, will be the daily total precipitation.

```
ggplot(data = trip_weather_data, aes(x = day_of_week, y = daily_trips)) +
  geom_boxplot() +
  theme_bw()
```




```
ggplot(data = trip_weather_data, aes(x = daily_precip, y = daily_trips)) +
  geom_point()
```



Linear Regression Model

```
lm_mod <- lm(daily_trips ~ type_of_day + daily_precip + day_of_week, data = trip_weather_data)
summary(lm_mod)
```

```
##
## Call:
## lm(formula = daily_trips ~ type_of_day + daily_precip + day_of_week,
##     data = trip_weather_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21427  -8832   1532   7288  25691
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      467970       7699  60.782  < 2e-16 ***
## type_of_dayhot    21912       5693   3.849 0.000871 ***
## daily_precip      11237      22534   0.499 0.622962
## day_of_weekTue    19306       8518   2.267 0.033586 *
## day_of_weekWed    53607       9429   5.685 1.02e-05 ***
## day_of_weekThu    82585       9712   8.503 2.11e-08 ***
## day_of_weekFri   131433       9190  14.302 1.28e-12 ***
```

```
## day_of_weekSat   190104       9584  19.836 1.58e-15 ***
## day_of_weekSun    93376       9274  10.068 1.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13140 on 22 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9551
## F-statistic: 80.83 on 8 and 22 DF,  p-value: 1.624e-14
```

Results

The linear regression model is: $\hat{y} = 467970 + 21912 \times type\_of\_dayhot + 11237 \times daily\_precip + 19306 \times day\_of\_weekTue + 53607 \times day\_of\_weekWed + 82585 \times day\_of\_weekThu + 131433 \times day\_of\_weekFri + 190104 \times day\_of\_weekSat + 93376 \times day\_of\_weekSun + \epsilon$

When controlling for the seasonality of the days of the week, the heat index (*type_of_dayhot*) became statistically significant. Holding all other variables constant, when the heat index is />90 we can estimate an increase of 21,912 trips than non high heat index days. As for the practical significance of the variable, it is quite surprisingly a significant predictor as the adjusted Rˆ2 of the model is 0.9551.

Checking Assumptions

Normality The Q-Q plot shows there is some "S" curvature within the band of residuals, but overall is straight.

```
ggplot(data = lm_mod, aes(sample = .resid)) +
  stat_qq() +
  labs(caption = "Figure 5")
```

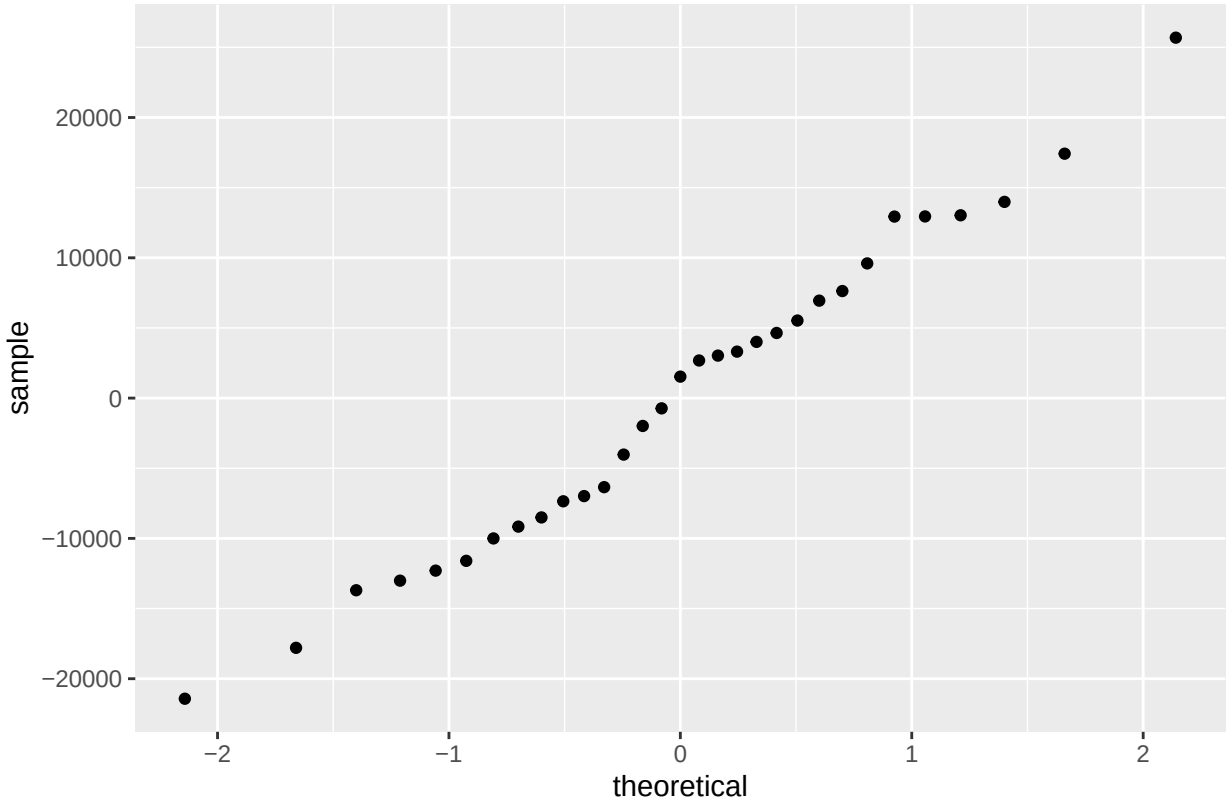


Figure 5

Constant variability The spread around zero does appear to have some heteroskedasticity as it is cone-shaped from the middle but overall nothing alarming.

```
ggplot(data = lm_mod, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals") +
  labs(caption = "Figure 6")
```

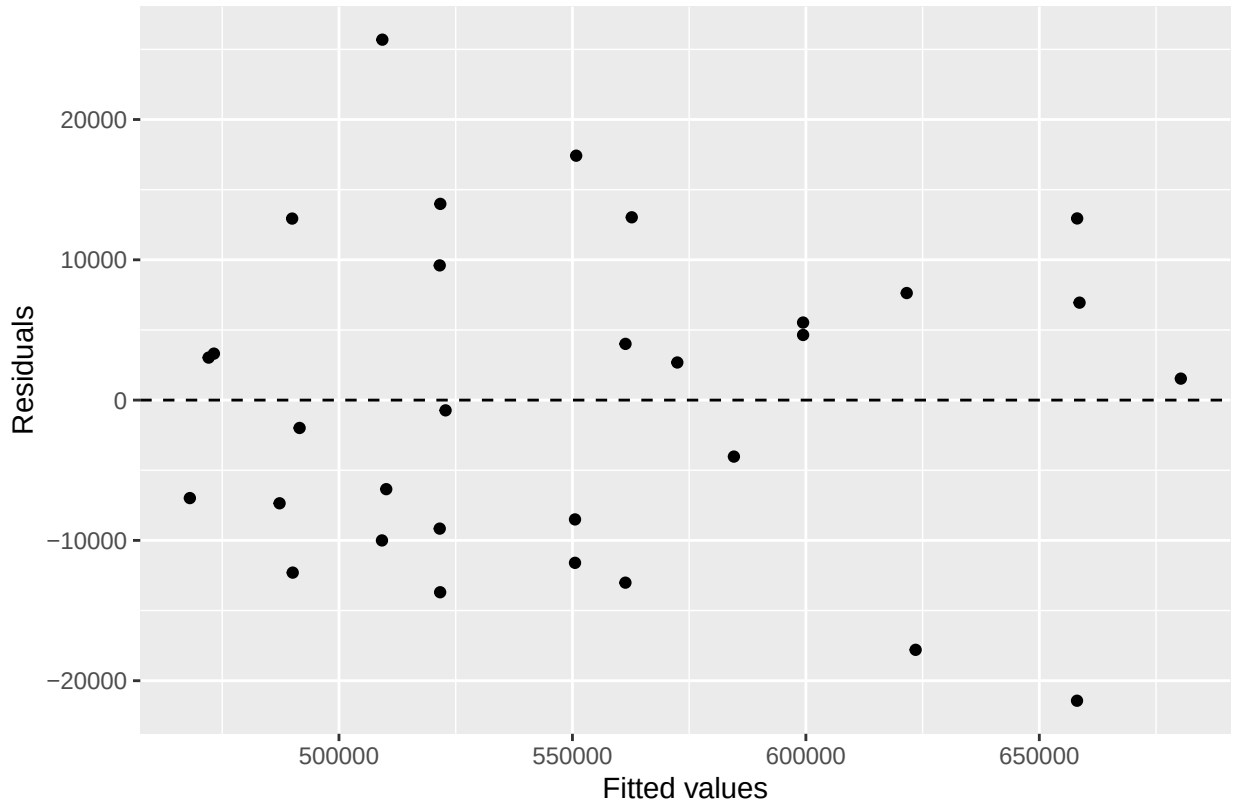


Figure 6

Linearity It does not pass the linearity test as from Figure 3, there is not a linear shape for heat index and trips taken.

Conclusions

When using the heat index as a measure of trips taken, it alone does not seem to be an indicator of higher trips being taken via Uber and Lyft. However, when controlling for the seasonal effects on the day of the week, it does become a good indicator of trip counts. However, there are some concerns about the high results of the adjusted R^2 and small scale of data used. For future research and without limitations to processing capacity, I would want to increase the data from being just using August 2022 and study the trends for a few years worth of summer data. It is also important to note that this analysis focuses on the unique environment of NYC but can the same be said for other hot climates such as Miami or Los Angeles.