Final Project

Taxi 838

Introduction

If you live in a big city and want to get somewhere without your car, the first thing you think is: "can I get there using the subway?"

In a big city the subway is usually the most reliable and efficient way of traveling. Trains come around every 5 minutes, never get stuck in traffic and get you to the destination as fast as possible without giving you a ridiculous paycheck.

But what if subway isn't an option? What if the destination is too far from the nearest subway station, or, perhaps, there's some maintenance duty going on? While underground means of transportation might be the best kind of public transport, there still are reliable ways of getting to your destination above the ground.

Maybe there could be cases or certain circumstances that make underground transportation the second best option, or, even third? Our team decided to dig deeper into this topic. Let's see what we uncovered!

Data Import and Cleaning

We decided to use the NYC Taxi data and MTA Daily Rider Data to compare two means of transportation: taxi and subway.

First of all, let's download the data related to taxis and clean the names of columns.

Now let's join the spatial data with the numbers and clean column names.

```
glimpse(nyc_taxi_tbl)
```

```
Rows: ??
Columns: 25
Database: DuckDB 1.4.0 [User@Windows 10 x64:R 4.5.1/D:\KSE\Year2-Term1\R\FinalProject\nyc.du
$ vendor_id             <int> 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 1, 2, 1, 2, 2, 2, ~
```

```
$ tpep_pickup_datetime  <dttm> 2024-01-24 15:17:12, 2024-01-24 15:52:24, 2024-~
$ tpep_dropoff_datetime <dttm> 2024-01-24 15:34:53, 2024-01-24 16:01:39, 2024-~
$ passenger_count       <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, ~
$ trip_distance         <dbl> 3.33, 1.61, 4.38, 0.95, 2.58, 15.80, 7.69, 8.31,~
$ ratecode_id           <dbl> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, ~
$ store_and_fwd_flag    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"~
$ pu_location_id        <int> 239, 234, 88, 211, 68, 164, 231, 138, 132, 226, ~
$ do_location_id        <int> 246, 249, 211, 234, 144, 132, 161, 262, 192, 7, ~
$ payment_type          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, ~
$ fare_amount           <dbl> 20.50, 10.70, 25.40, 9.30, 18.40, 70.00, 40.80, ~
$ extra                 <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 2.5, 0.0, 5.0, 0.0, 0.0~
$ mta_tax               <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5~
$ tip_amount            <dbl> 3.00, 3.67, 5.88, 2.66, 4.48, 10.00, 6.50, 10.00~
$ tolls_amount          <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 6.94, 0.00, 6.94, ~
$ improvement_surcharge <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ total_amount          <dbl> 27.50, 18.37, 35.28, 15.96, 26.88, 90.94, 51.30,~
$ congestion_surcharge  <dbl> 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 0.0, 0.0~
$ airport_fee           <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 1.75, ~
$ month                 <chr> "01", "01", "01", "01", "01", "01", "01", "01", ~
$ year                  <dbl> 2024, 2024, 2024, 2024, 2024, 2024, 2024, 2024, ~
$ pu_borough            <chr> "Manhattan", "Manhattan", "Manhattan", "Manhatta~
$ pu_zone               <chr> "Upper West Side South", "Union Sq", "Financial ~
$ do_borough            <chr> "Manhattan", "Manhattan", "Manhattan", "Manhatta~
$ do_zone               <chr> "West Chelsea/Hudson Yards", "West Village", "So~
```

Nice and clean!

Let's move on to the MTA dataset. There's not much work to do here, really:

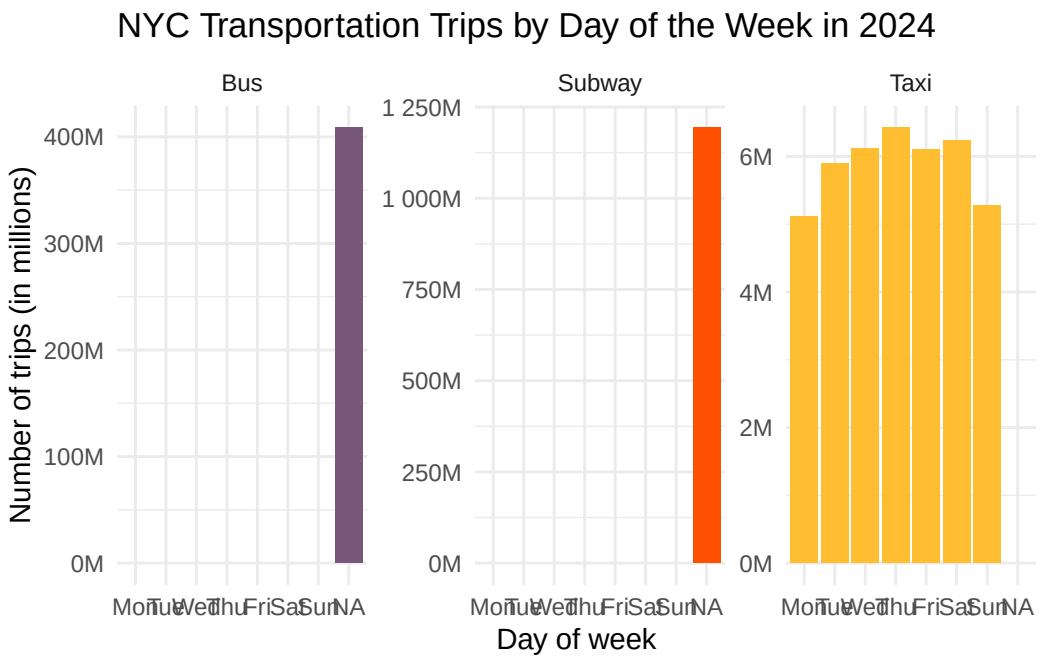
And that's it for data collection and cleaning. Now let's try and unveil the intrigue.

Analysis

First of all, let's see what the most popular means of public transportation are for New-Yorkers:

```
taxi_day_week <- nyc_taxi_tbl |>
  mutate(day_of_week = wday(tpep_pickup_datetime, label = TRUE)) |>
  group_by(day_of_week) |>
  summarise(total_count = n()) |>
  collect() |>
  mutate(transport_type = "Taxi")
```

```r
subway_day_week <- metro_tbl |>
  mutate(day_of_week = wday(date, label = TRUE)) |>
  group_by(day_of_week) |>
  summarise(total_count = sum(subways_total_estimated_ridership)) |>
  mutate(transport_type = "Subway")

busses_day_week <- metro_tbl |>
  mutate(day_of_week = wday(date, label = TRUE)) |>
  group_by(day_of_week) |>
  summarise(total_count = sum(buses_total_estimated_ridership)) |>
  mutate(transport_type = "Bus")


combined_data <- bind_rows(taxi_day_week, subway_day_week, busses_day_week)

day_order <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")

combined_data <- combined_data |>
  mutate(day_of_week = factor(day_of_week, levels = day_order))

ggplot(combined_data, aes(x = day_of_week, y = total_count, fill = transport_type)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ transport_type, scales = "free_y") +
  scale_fill_manual(values = c("Taxi" = "#FFBD31", "Subway" = "#FF5000",
                               "Bus" = "#77587B")) +
  scale_y_continuous(labels = scales::label_number(scale = 1e-6, suffix = "M")) +
  labs(
    title = "NYC Transportation Trips by Day of the Week in 2024",
    x = "Day of week",
    y = "Number of trips (in millions)"
  ) +
  theme_minimal()
```



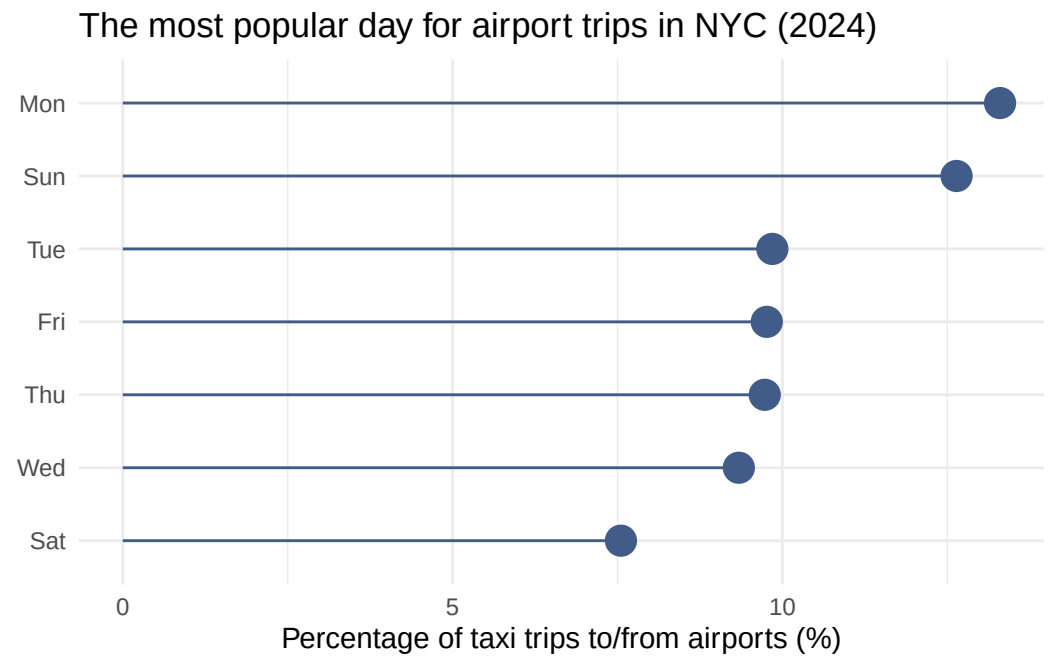
It is clear now that subway is the best way of getting around the city for the most of the citizens. And it's really logical if you think about it. What's the problem with the taxi? It might be too expensive and get stuck in traffic. The bus fixes the expensiveness of taxi, but still has the problem of getting stuck in traffic. On the other hand, subway doesn't have any of those issues. It's cheap, fast, reliable and, let's be honest, has a vibe to it.

Also, the bar chart suggests that on Saturday and Sunday there is a dip in bus and subway usage, but we can't say the same about taxi.

Honestly, I didn't think that taxi is so rarely used. What's the reason behind its low usage rate? Maybe it's too expensive or too unnecessary? Perhaps the use cases for taxi are narrow? What if taxi is only used to go to, for example, the airport? Let's see:

```
airport_location_ids <- zones |>
  filter(str_detect(zone, "Airport")) |>
  pull(location_id)

airport_percentage_by_day <- nyc_taxi_tbl |>
  mutate(
    day_of_week = wday(tpep_pickup_datetime, label = TRUE),
    is_airport_trip = do_location_id %in% airport_location_ids |
      pu_location_id %in% airport_location_ids
  ) |>
  group_by(day_of_week) |>
```

```
  summarise(
    total_trips = n(),
    total_airport_trips = sum(is_airport_trip, na.rm = TRUE)
  ) |>
  mutate(
    airport_trip_percentage = total_airport_trips / total_trips * 100
  ) |>
  collect()

ggplot(airport_percentage_by_day, aes(x = reorder(day_of_week, airport_trip_percentage),
                                      y = airport_trip_percentage)) +
  geom_segment(aes(xend = reorder(day_of_week, airport_trip_percentage), yend = 0),
               color = "#415C89") +
  geom_point(aes(), size = 5, color = "#415C89") +
  coord_flip() +
  labs(
    title = "The most popular day for airport trips in NYC (2024)",
    x = NULL,
    y = "Percentage of taxi trips to/from airports (%)"
  ) +
  theme_minimal()
```



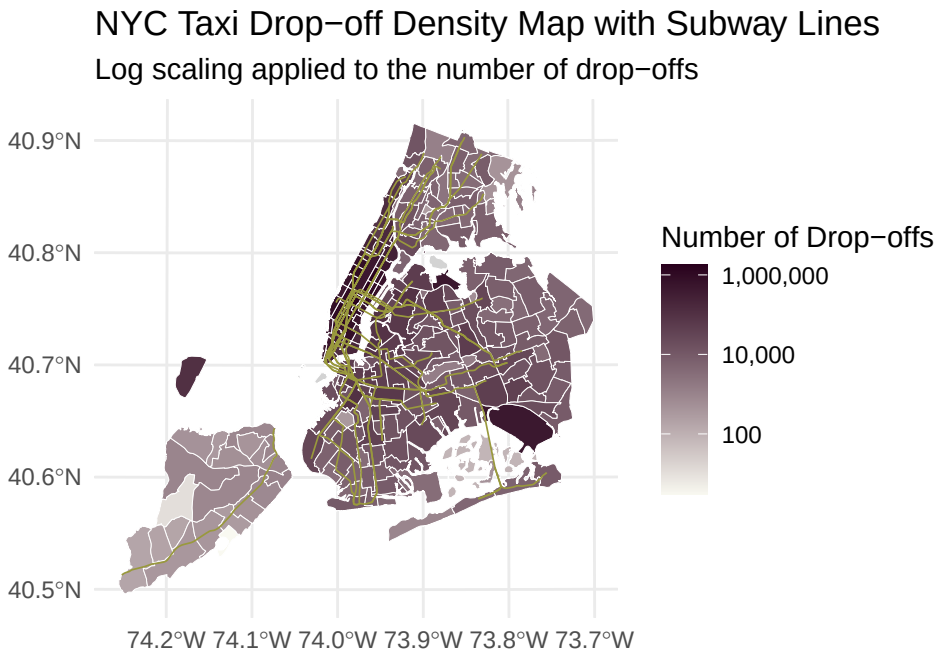
Nope. It seems that taxi is just a luxurious way of getting around (or just that its use case

isn't really getting to the airport).

We didn't come up with another definitive way of using the taxi that also could be analyzed with the data we have. So our plotter, Yevhenia, decided to create a choropleth map of taxi drop offs.

For the sake of accuracy, it would be convenient to also add the subway routes. That way we can clearly see if taxis cover the areas that lack proper subway lines. Let's see what the map has to say:

```
p_combined <- ggplot() +
  geom_sf(
    data = taxi_zones_with_data,
    aes(
      fill = trip_count,
      text = paste("Zone:", zone, "\nNumber of drop-offs:", scales::comma(trip_count))
    ),
    color = "white",
    linewidth = 0.05
  ) +
  geom_sf(
    data = subway_lines,
    aes(),
    color = "#99993D",
    linewidth = 0.3,
    inherit.aes = FALSE
  ) +
  scale_fill_gradient(
    high = "#29001D",
    low = "#F9F9F1",
    trans = "log10",
    labels = scales::label_comma(),
    name = "Number of Drop-offs",
    na.value = "lightgrey"
  ) +
  labs(
    title = "NYC Taxi Drop-off Density Map with Subway Lines",
    subtitle = "Log scaling applied to the number of drop-offs"
  ) +
  theme_minimal()

p_combined
```



As we can see, taxi is really not that popular even in the small areas where the subway is underdeveloped. The biggest number of drop-offs is concentrated around airports of New-York, so getting a cab to the airport is indeed the biggest use case of the taxi services. Also, for some reason, Manhattan happens to be another concentration of taxi drop-offs. I asked ChatGPT about why that is the case and here's what he told me:

"Manhattan stands out as a major drop-off area for taxis largely because it is the economic, cultural, and tourist heart of New York City. Dense clusters of offices, hotels, theaters, and attractions concentrate a high volume of trips into relatively small areas. Even though outer boroughs may have more residents, their trips are spread over larger, less dense neighborhoods, whereas Manhattan blocks can see thousands of drop-offs daily—numbers comparable to major airports. The combination of heavy commuter traffic, tourism, nightlife, and the convenience of taxis for short, cross-town, or luggage-heavy trips makes Manhattan a consistent hotspot for taxi activity."

Conclusion

Even if taxi isn't the best way of transportation, it still has its own use cases where it shines. It was surprising to find that taxis aren't only widely used for trips to airports, but also for trips to certain areas (Manhattan specifically). I thought that using a taxi comes down to areas where other means of transportation are underdeveloped, which can't be said about Manhattan, the central district of New-York. Good thing we ran a research and now know how it works inside this gigantic city.