CSCI 0111 : Computing Foundations, Data
Project 1: Transit Report

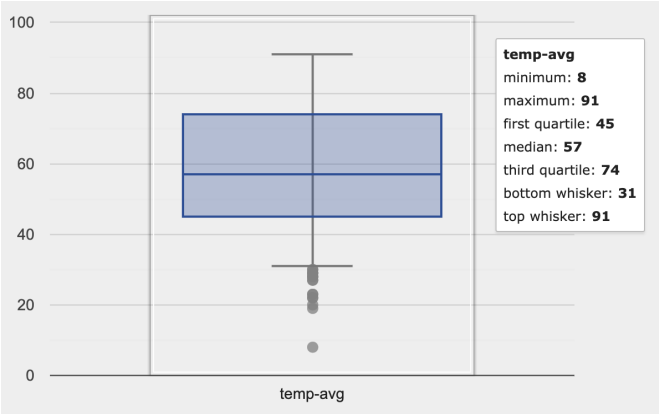
1. How Bad Weather Influences Taxi Rides

I. Temperature

Definitions: We defined three different temperature ranges:

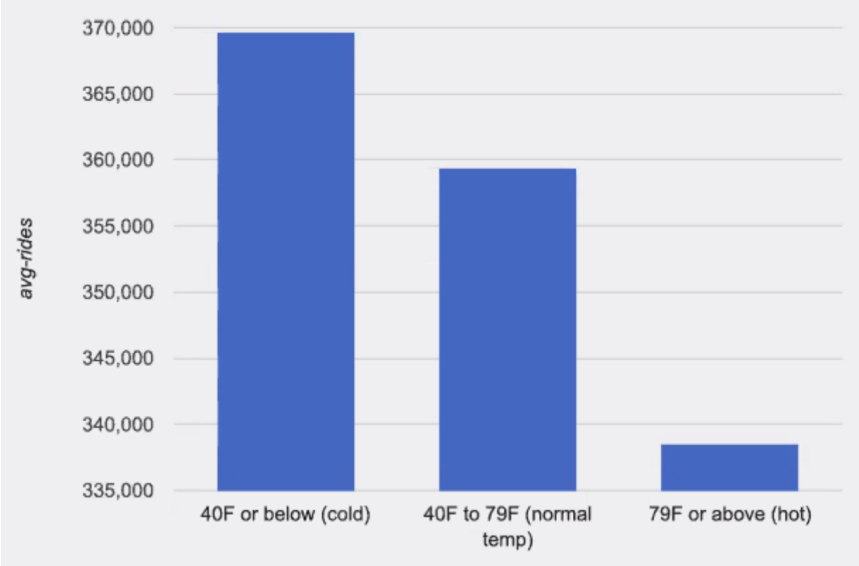
- Cold days: days in which the average temperature was **40 F or lower**.
- Hot days: days in which the average temperature was **79 F or higher**.
- Normal days: days when the average temperature was **between 40 and 79 F**

We determined these temperature bounds by calculating the first and third quartile of yearly temperatures in New York from the LaGuardia weather data, and then used those temperatures as a metric to define "extremely cold" and "extremely warm" days in the context of New York City. We chose our cold bound to be 5 degrees below the first quartile (45 - 5 = **40 F**) and our hot bound to be 5 degrees above the third quartile (74 + 5 = **79 F**).



***The above box-plot is called temp-box-plot in our code.

Results & Methodology:
We generated a bar chart with three bars measuring the average number of rides across all four periods of a given day on dates we classified as cold, normal, and hot respectively.



***The above bar chart is called temp-rides-bar-chart

In the chart above…
The average taxi ridership when temperatures are 40F or below is 369,713.
The average taxi ridership when temperatures are 40F to 79F is 359,388.
The average taxi ridership when temperatures are 79F or above is 338,548 .

We determined that temperature does have a measurable effect on the number of taxi rides that occur on a particular day. Generally, as average temperature decreases, taxi ridership rises. When temperatures are lower than "normal" (i.e., they are 40 F or lower), average taxi rides on a given day are higher than the number of rides on a "normal" day (*369,713 > 359,388.*). This means that **lower temperatures increase taxi ridership**. Moreover, when temperatures are higher than "normal" (i.e., they are 79 F or higher), average taxi rides on a given day are lower than the number of rides on a "normal" day (*338,548 < 359,388)*. This means that **higher temperatures decrease taxi ridership**. For reference, when we refer to the number of rides on a normal day, we are referring to the *359,388* average number of rides.

To summarize, on cold days, there are a higher average number of rides compared to normal days (**~10325 rides more than normal on cold days**), indicating that colder weather conditions encourage people to take taxis. In contrast, on hot days, there is a lower average number of rides compared to normal days (**~20840 rides less than normal on warm days**), indicating that hotter weather discourages people from taking taxis.

However, there are also several other variables that could influence the average number of taxi rides on days with cold, hot, and normal temperatures. Seasonal factors, like tourism, could impact the average frequency of taxi rides on a given day. Additionally, we should also consider that during the summer months, when days are likely to be hot (79 F or above), many New York City residents leave the city for second homes in The Hamptons and in other areas, which could also be contributing to the lower average taxi rides on hotter days.

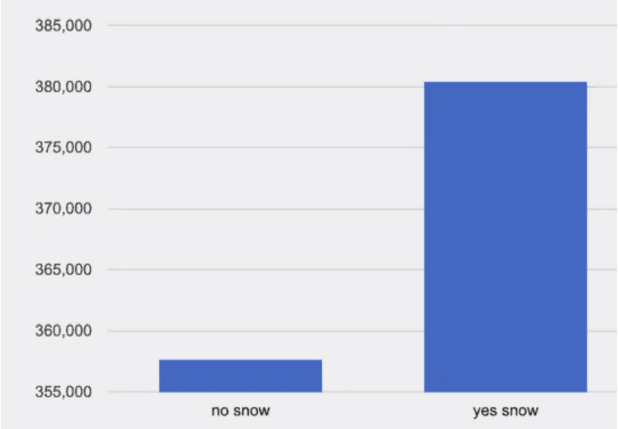
We must also consider the implications of how we defined "cold", "hot" and "normal" temperature days. Changing the temperature range for which we defined these "extreme

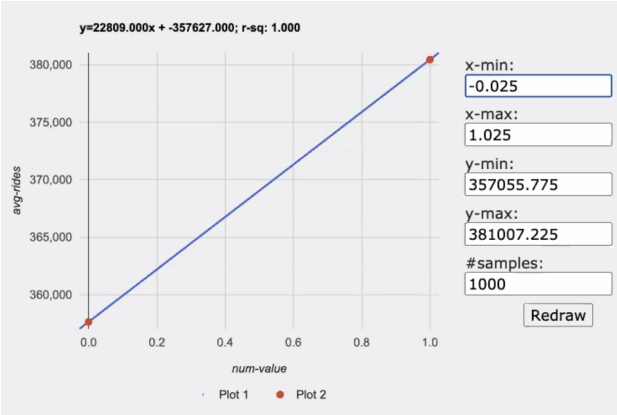
weather" days would also affect our findings for how cold or hot temperature impacts taxi ridership. Lastly, we must consider that the number of cold days and the number of hot days is less than the number of "normal days" — inherently affecting the sample size from which we generate our average taxi ridership. To fix this issue, we analyzed the mean ridership on hot, cold, and normal days.

II. Precipitation

A. Snow

We created a bar chart to display the average taxi ridership when there is no snow (i.e., snow == 0) and the average ridership whenever this is any snow (i.e., snow > 0). This graph shows us that whenever there is any amount of snow, taxi ridership would increase from an average of 357,627 riders when there is no snow to 380,436 riders when there is snow. We also created a linear regression model which further breaks down this math, showing that the actual increase in ridership (y) is equal to 22809(x)+ (~357627). The x variable, in this scenario, is our boolean variable that indicates if there is snow when someone is deciding whether or not to ride in a taxi. From this model, it is clear that **if it is snowing on a given day, an additional 22,809 people on average will take a taxi ride** .



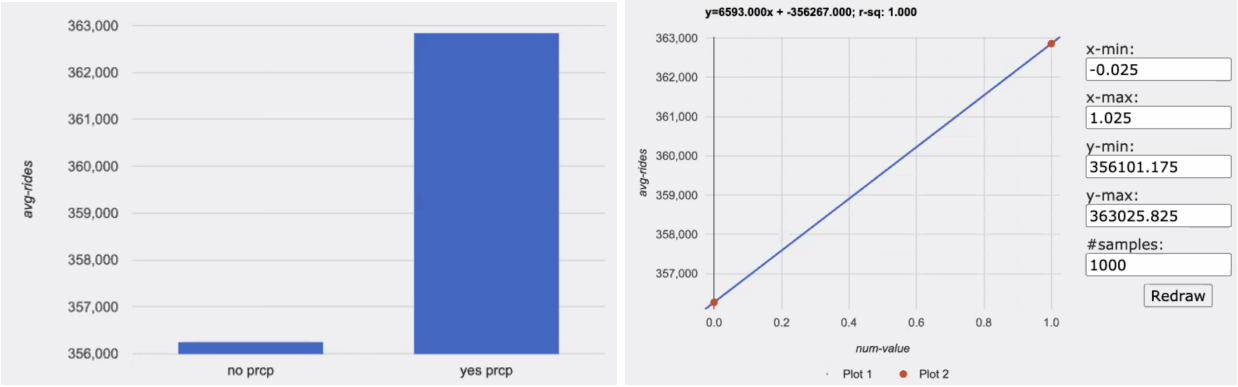


***The graphs above are called snow-bar-chart and snow-regression, respectively

B. Rain

We created a bar chart to display the average taxi ridership when there is no rain (i.e., prcp == 0) and the average ridership whenever there is any rain (i.e., prcp > 0). We decided to only compare two different buckets — "no rain" and "yes rain." This graph shows us that whenever there is any amount of rain (i.e., prcp > 0), taxi ridership would increase from an average of 356,267 riders when conditions are clear to 362,860 riders when there is rain. We also created a linear regression model which further breaks down this math, showing that the actual increase in ridership (y) is equal to 6593(x)+ (~356,267). The x variable, in this scenario, is our boolean variable that indicates if there is rain when someone is deciding whether or not to

ride in a taxi. From this model, it is clear that if **it is raining on a given day, an additional 6,593 people on average will take a taxi ride**.



***The graphs above are called rain-bar-chart and rain-regression, respectively

III. Other Considerations

A. Wind

We chose not to include a wind variable from the LaGuardia Weather Data table because winds less than 25 mph are not considered severe. We visually scanned the table and did not see any winds we thought would be classified as bad weather, so we elected not to include it in our final combined table and analysis. We also spoke with our TA during our design check, and they approved of this reasoning.

2. Patterns in Taxi Rides and Taxi Fares Across Days of the Week

Results & Methodology:

To analyze patterns across the same given weekday throughout the year (e.g., ridership and fares on all Mondays), we chose to generate the relative standard deviation (RSD) table shown below. The RSD is expressed as a percentage and was calculated by multiplying the standard deviation by 100 and dividing by the mean. We chose to use the RSD as a metric to analyze patterns in taxi rides and taxi fares across a day of the week instead of an non-standardized standard deviation because the RSD "normalizes" the deviation, which allows us to compare the spread in data between different days. (Standard deviation alone wouldn't have allowed us to compare the spread in data between different days, since different days of the week have different daily average total rides and daily average total fares.)

We created a function to build a RSD table for rides and fares — by using a helper function to calculate RSD given a row and column name. We then input table-days-rides-fares, a table which included one row for each date and had columns representing the total daily rides and total daily fares (by summing rides and fares across all four time periods in the day). The standard deviations determined in the following table represent how much daily total rides or

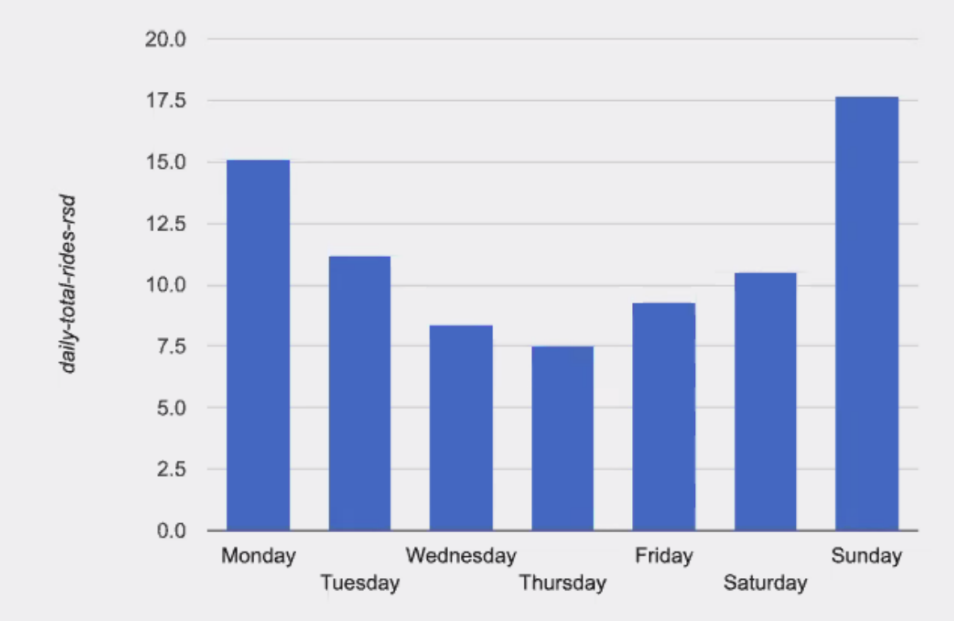
daily total fares deviate from the average daily total rides or average daily total fares across all Mondays, for example.

The following table represents the calculated relative standard deviation (RSD), which is a ratio of the standard deviation of a weekday's value divided by its mean value, obtained using the methods outlined above:

weekday	daily-total-rides-rel-stdev	daily-total-fares-rel-stdev
"Sunday"	17.7	17.3
"Monday"	15.1	15
"Tuesday"	11.2	10.6
"Saturday"	10.5	11.3
"Friday"	9.3	10.6
"Wednesday"	8.4	8.4
"Thursday"	7.5	8.1

***The table above is called sorted-daily-rsd-table

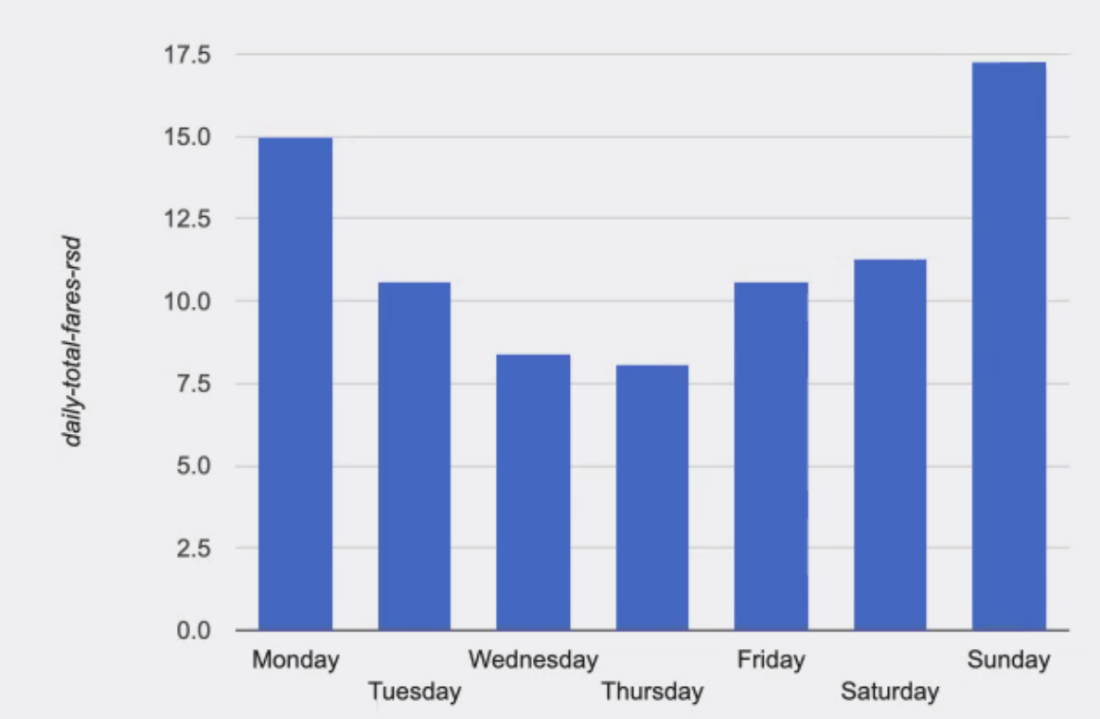
We then graphed our table to facilitate visual analysis of our findings. When answering the question about how consistent are ridership numbers and total fares on any given day of the week, one can look at our generated relative standard deviation table to find answers. For example, given that standard deviation is a measure "how spread out values are," we can tell from our table that Sunday is the least consistent day with a standard deviation of 17.7% from its mean whereas Thursday is the most consistent day with a standard deviation of just 7.5% from its mean.



***The chart above is called daily-rides-rsd-chart

In the table above…
The highest standard deviation in ridership is **17.7% *deviation from the mean on Sundays***.
The lowest standard deviation in ridership is **7.5% deviation from the mean rides *on Thursdays***.

Moreover, in terms of daily total fares, we can see that again Sunday is the least consistent day (largest standard deviation of 17.3% from the mean) whereas Thursday is our most consistent day (smallest standard deviation of 8.1% from the mean).



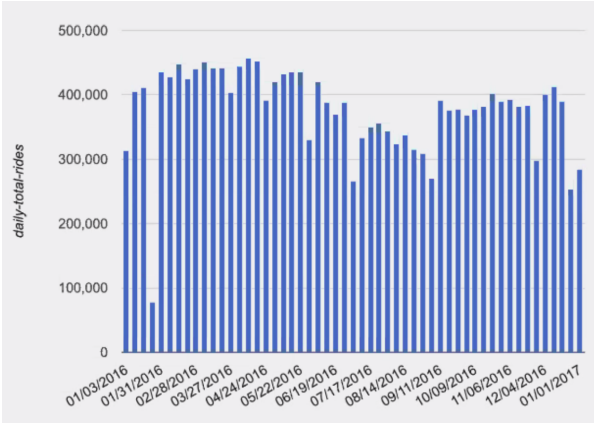
***The chart above is called daily-rides-rsd-chart

In the table above…
The highest standard deviation in fares is ***17.3% deviation from the mean on Sundays****.*
The lowest standard deviation in fares is ***8.1% deviation from the mean on Thursdays****.*

Again, to summarize our findings, this means that Sunday is the least consistent day (largest standard deviation) in terms of both ridership and collected fares whereas Thursdays are our most consistent days (smallest standard deviation). It is not surprising that the day that deviates most in ridership is also the day that deviates the most in fares given that ridership generates fares (i.e., a strong positive correlation exists between these two variables).

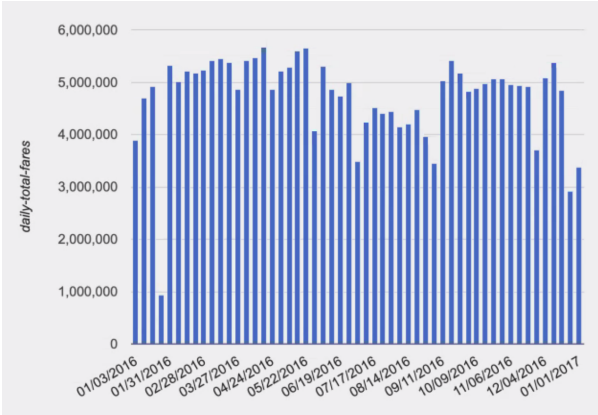
The following table places charts of the ridership and fare data for both Sundays and Thursdays next to one another to facilitate a quick visual analysis.

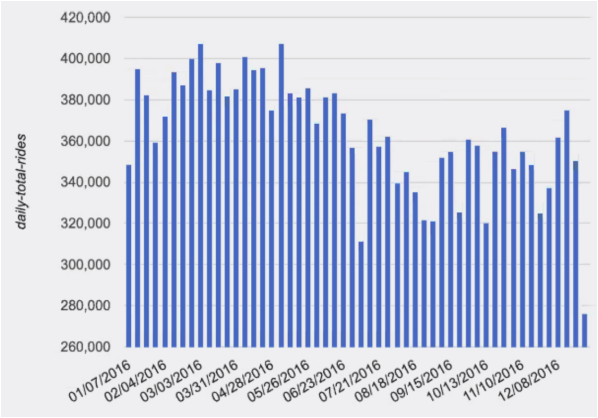
Daily Ridership on all Sundays (the least consistent day according to our standard deviations table)	**Daily Ridership on all Thursdays (the most consistent day according to our standard deviations table)**
Daily ridership on Sundays:	Daily ridership on Thursdays:



***Charts above and below are called sunday-total-rides-chart, and sunday-total-fares-chart respectively

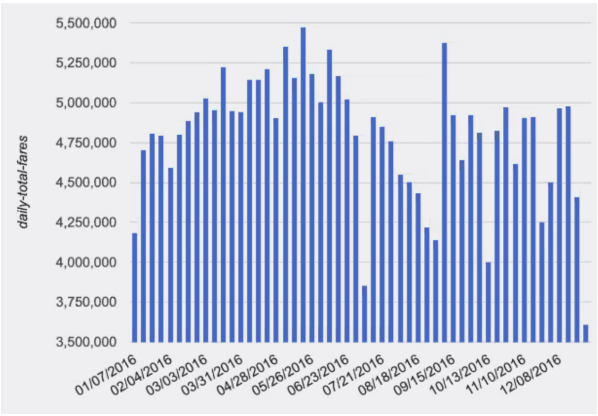
Daily Total Fares:





***Charts above and below are called thursday-total-rides-chart, and thursday-total-fares-chart respectively

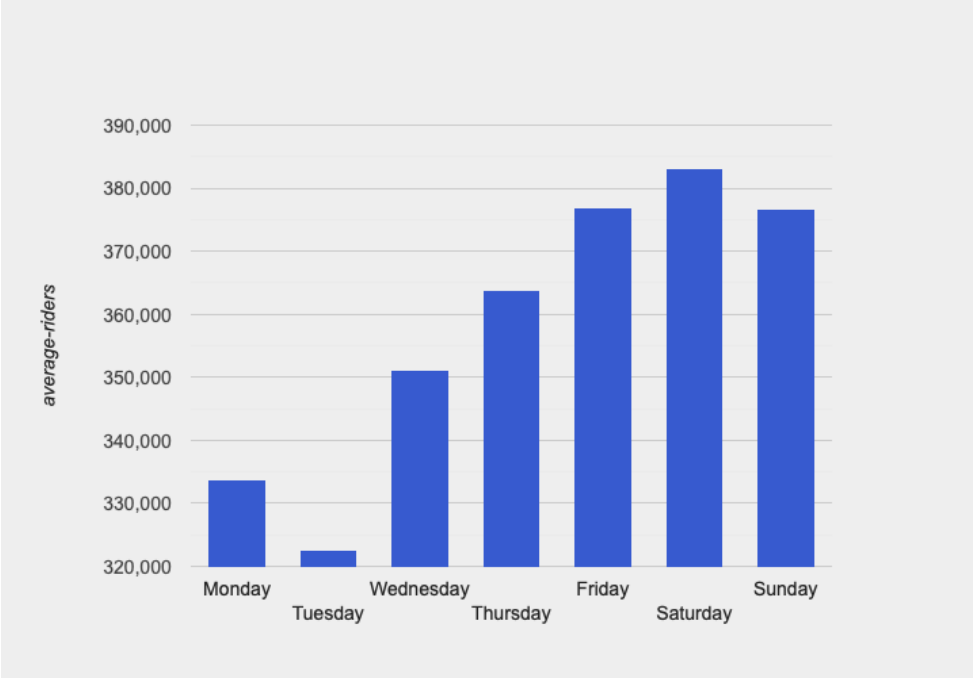
Daily Total Fares:



3. How Day of the Week Influences Taxi Rides

Results & Methodology:

We used filter-with to create smaller tables representing all data from a certain day of the week (i.e. every Monday) and then calculated the average taxi ridership using a mean function and displayed this data in the following bar chart:



***The chart above is called which-day-is-popular-bar

In the table above…
*The highest average taxi ridership is on Saturdays with an average of **383,024**.*
*The second highest average taxi ridership is on Fridays with an average of **376,845**.*
*The third highest average tax ridership is on Sundays with an average of **376,707**.*
*Also, from this graph it is clear that Tuesdays are the least popular days to take a cab (i.e., lowest average taxi ridership) with an average of **322,575** riders.*

To summarize our findings, weekends tend to be more popular times to take a cab (i.e., higher average tax ridership) as compared to weekdays. Moreover, Saturdays are the most popular days to take a cab at **383,024 riders** and Tuesdays are the least popular at **322,575 riders**.

A Summary Table of Rides and Weather Conditions

Definitions:

A table that displays the mean number of annual rides for each time frame given the weather condition.

Results & Methodology:

We created a new table called sum-table with the one column “time” which contains rows Morning, Afternoon, Evening and Night.

We then created helper function sum-table-builder to help create the additional 3 columns (Rain, Snow and Clear) with the cells containing the number of rides if row-header and column-header are true. Sum-table-builder has a nested builder function within it that filters the data table to only have rows where time and weather conditions are met, then runs the given function f to generate a number (mean, sum, or stdev) of the values.

The function is summary-table which takes in a table (like table-with-all-data) and uses the helper function to build 3 columns for prcp, snow, and clear

The following table represents the average number of rides during a given timeframe and weather condition (e.g., the average ridership on rainy mornings (where time == "Morning" **and** it is raining (i.e., prcp > 0)) is 87,461.88).

time	prcp	snow	clear
"Morning"	87461.88	82677.58	87543.94
"Afternoon"	108158.56	113173.67	106045.95
"Evening"	122392.03	124379.83	121013.89
"Night"	44847.39	60204.92	41663.05

***The table above is called mean-summary-table

Reflection questions to be included in our report:

1. **One key insight that we each gained about programming or data analysis**
    a. **Finn:** I learned quite a bit about nesting helper functions within other functions. Prior to this project, I preferred to write my helper functions outside of my larger functions. For whatever reason, that separation made it easier for me to process and understand my code. However, during this project, I watched my partners code using a series of nested functions and saw they build them (i.e., watching the process occur before me), which made their construction make more sense.
    b. **Alex:** I learned about making functions more modular so that it can be applied to multiple things (i.e., grabbing snow and rain data because they are so similar). When we wrote the code, we noticed that a lot of our functions were just copied

and pasted versions of each other. Restructuring them so they could just become one function was interesting.

c. **Eugenia:** I learned more about choosing charts to visualize data and to help set boundaries. At first, we thought of using a pie-chart to show which day had the most rides on average. However, that shows parts to a whole unlike bar charts which clearly indicate the towering data. Also, I first thought of using forums/articles about New York City to set the boundary between cold and hot as bad weather. Unfortunately, those articles were very subjective and vague. After thinking about it, we decided to use box plots as that shows a five-figure summary using the provided data and shows outliers. I was more assured of this being correct when the 1st and 3rd quartile values were significantly close to the values provided by a more-detailed article. In addition, we had to work our way around the regression line plot to make correlation more readable since Pyret limits samples.

d. **Spencer:** Generalizing functions so that they could be used in multiple different ways was a major takeaway for me. Thinking about new ways to simplify and reduce code was really cool and helpful in learning how to write functions that can be applied to multiple different contexts.

e. **Travess:** When in doubt break each problem down into smaller solvable problems. Make simple functions with tables that I understand and then re-define those tables when needed. This sort of building block functionality scales quickly. Working on the later sections of code I can see the growing complexity of the program, all based on smaller and simpler units.

2. One key misconception that each partner had to work through

a. **Finn:** At first, I was not sure how we would be able to compare standard deviations from Sundays to Thursdays. Although the SD is a good indicator of the spread of data within each of those day "datapoints," it is not a good mathematical choice to compare the SD from day to day without standardizing their magnitudes. However, Spencer found out how to convert SD into relative standard deviation (RSD), which measures a data's spread from its own mean in percentage terms. The RSD takes into account differences in taxi ridership from day to day and allows us to compare the SD of Sunday to Thursday, for example, in a mathematically sound way.

b. **Alex:** Having coded individually before this project, I did not realize how important doc strings, putting the data type after variables (t :: Table), and testing were. When we tried combining parts of code we did separately, having clear doc strings was super helpful in figuring out exactly what a function was supposed to do. When I created a table in pyret, I initially was lazy and did not put a type for each column and in the end there was an issue where I incorrectly assumed that a variable was a boolean and not a string, a problem easily solved had I put the types.

c. **Eugenia:** Prior to coding as a group, I assumed that for the summary function to work, we had to write an input function specifically for it. At the time, I forgot that any function of the format (Table, String ->Number) should be able to work

with being able to be used as an input. Working with the group and going to TA hours helped me understand that concept.

d. **Spencer:** When working through the function to generate the the standard deviation of the number of rides on a given day, we realized that we were not calculating the standard deviation in ridership on a specific day (i.e., all Mondays), but rather the standard deviation in ridership in the different time frames on a specific day (i.e., rides in morning, afternoon, evening, night across all Mondays). This was because our input table did not combine total fares across all time periods on a given day. Nothing was wrong with our code and it was functional, but we were using the wrong input table. I learned that even if your code is correct, you can still get erroneous results. It's important to ensure that you are using the right inputs and always "sanity checking" the results.

e. **Travess:** One misconception I had was how sums and means can misinterpret data depending on the type of analysis I'm doing. In part three of the analysis, I started by adding up the total number of fares for each day of the week across the entire year. The issue with adding as opposed to using the mean is that there are certain days of the week that can have one more appearance than others. I know it's basic, but if Alex hadn't called out the issue I would have kept going as if the data was correct.

3. **How confident are we in the quality of our results? What other information or skills could have improved the accuracy and precision of our analysis?**

To improve the accuracy and precision of our analysis, we could run a more sophisticated regression of our data. If we ran a multivariate regression — using all weather variables at once ($y = c1x1 + c2x2 + c3x3 + b$, for example) — we would obtain a more sophisticated model to help us understand the relative influence of different weather variables on daily ridership and reduce omitted variable bias.

If we had quarterly weather data from LaGuardia, we would not have had to reduce our dataset to contain fewer data points (daily data as compared to the four data points per day) for parts 2-4.

We could have separated rain and snow data so that our analysis of each was independent of the other va riable.

To improve the readability of our graphs, we wanted to scale the axes so that they were consistent when comparing multiple graphs.

4. **One or two follow up questions we have about programming or data analysis after working on this project:**

    a. How can we visually show and compare three different variables on the same graph? For example, how would we present data for every day of a week for both fares and rides?

b. When we have many data points (like in Part 1 of this lab), how do we run a meaningful regression analysis? In other words, how do we prevent additional/useful information from being “capped” and not represented in the visual presentation of the regression analysis?