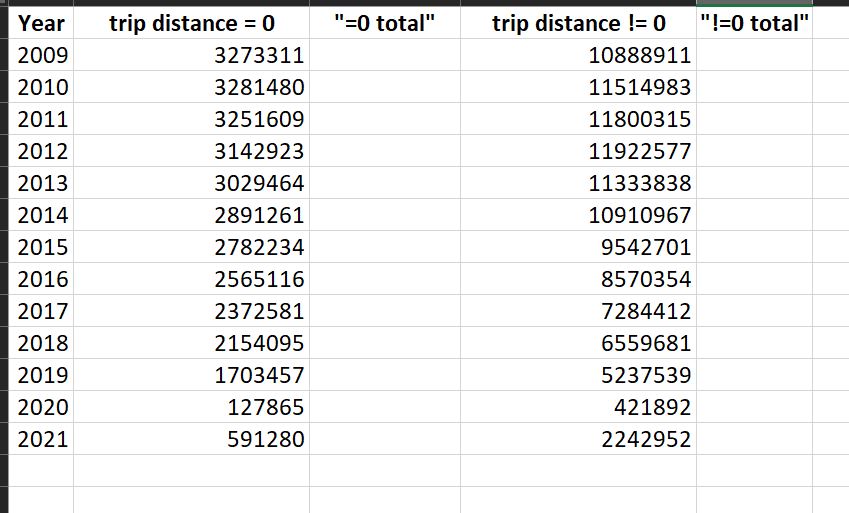
**Goals**

At the beginning of this project, we had to decide whether we’d be taking a descriptive or prescriptive approach. I thought it made more sense for this data set to take a descriptive approach. For this final deliverable, my main goals were to calculate the average rate of dollars/trip\_distance for June of each year between 2009 to 2021 and also describe the rate of change in the amount of cab trips for June of each year between 2009 and 2021. I thought it would be interesting to see if this rate of change increased or decrease each year and also predict what 2020 and 2021 would have looked like for yellow cabs if they had been normal years. In the business sense, this would give the cab company a goal of where they should be back at some time this year since covid case number have dropped so much in the last few months.

**Average rate of dollars/trip\_distance for June of each year between 2009 to 2021**

One issue with calculating this value was if the cab trip was less than a mile then the value entered was zero. The permissions for the sandbox prevented me from changing this column in Athena. So instead of changing the data, I first took a count of how many entries each year were equivalent to zero and how many weren’t. This is shown in the table below:



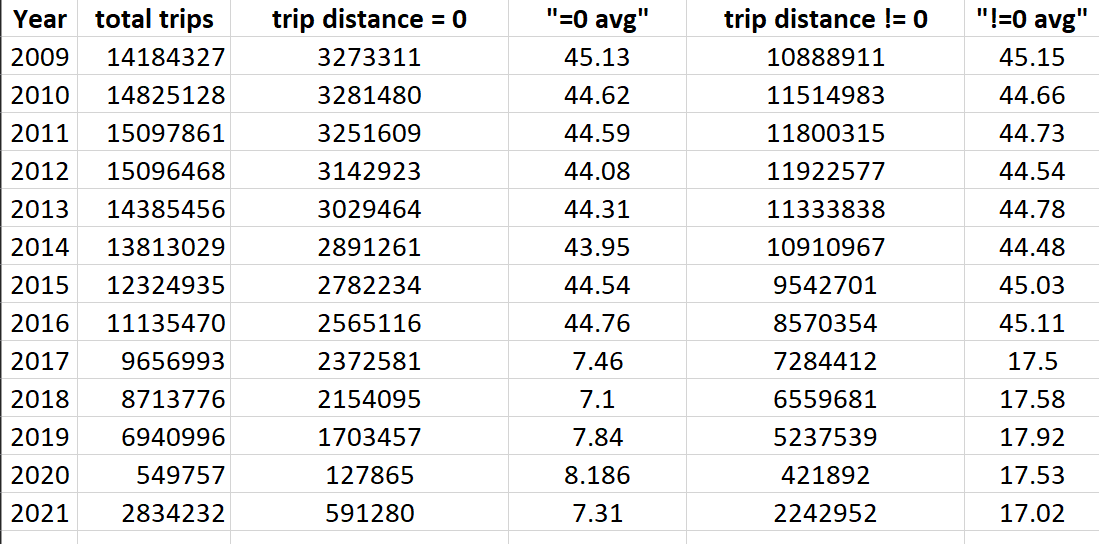
Below is the query I edited each time to get the values shown above:

SELECT COUNT(\*)

FROM "yellowtaxidata"."junedata"

WHERE EXTRACT(YEAR FROM tpep\_pickup\_datetime) = 2010 AND trip\_distance = 0;

Then I calculated the AVG(fare\_amount/trip\_distance) where the trip\_distance != 0 and AVG(fare\_amount/.9) where the trip\_distance = 0. I used .9 because I think it’s more likely if someone has hailed a cab that the distance was closer to 1 mile than 0 miles. These values are shown below:



Below is the query I edited each time to get the values shown above:

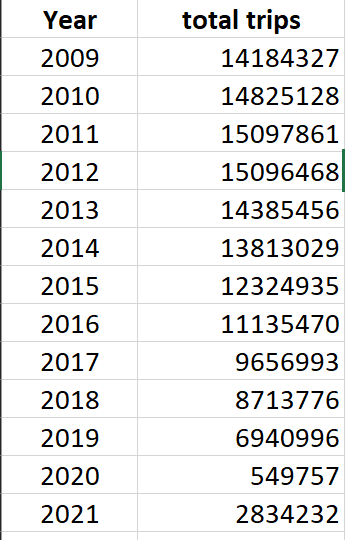
SELECT AVG(fare\_amount/.9) as dollars\_per\_mile

FROM "yellowtaxidata"."junedata"

WHERE EXTRACT(YEAR FROM tpep\_pickup\_datetime) = 2012 AND trip\_distance = 0;

**Rate of change in the number of trips for June of each year between 2009 to 2021**

For this task, I wanted to see what the percent decrease or increase in rides was for June of each year and also see the rate that this changed each year. To begin I shown below the total trips each year in June.



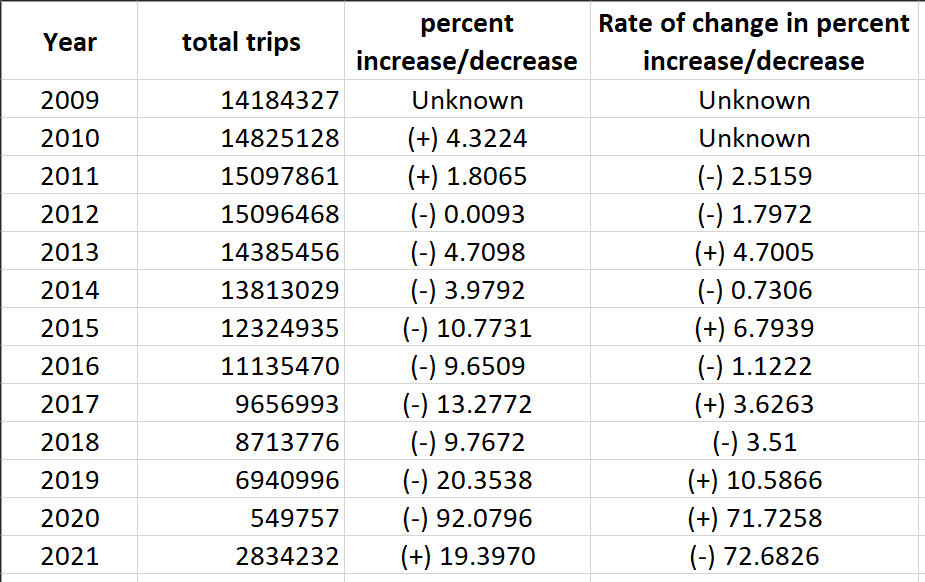
For this I just used the simple query below for each year:

SELECT COUNT(\*)

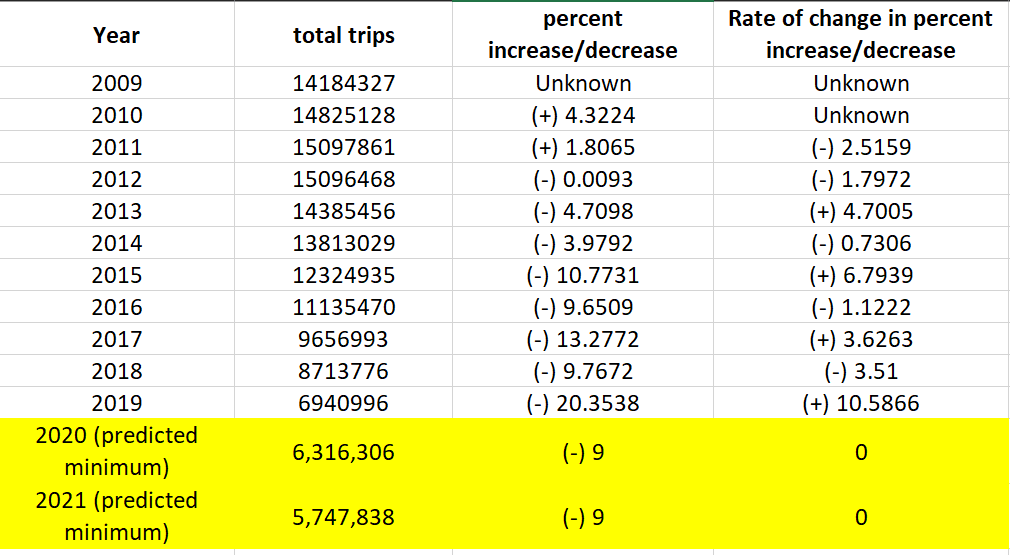
FROM "yellowtaxidata"."junedata"

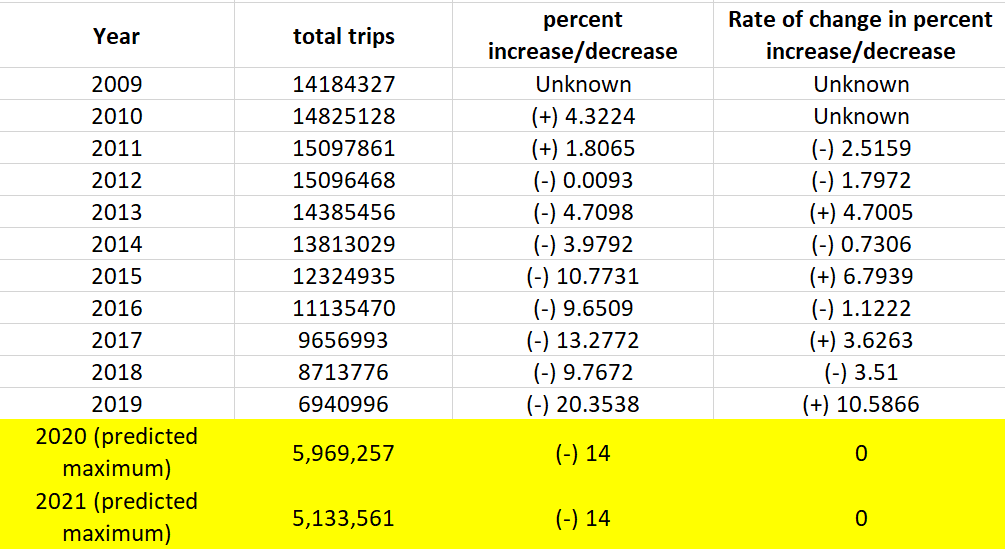
WHERE EXTRACT(YEAR FROM tpep\_pickup\_datetime) = 2009;

Then I calculated the percent change from the last year and how this change increased or decreased from the previous year, these values are shown below:



The predicted minimum and maximum percent change in rides for 2020 and 2021 if the pandemic was not a factor are shown below:





*The minimum predicted percent change is 9% and the maximum is 14%*

**Results of Average rate of dollars/trip\_distance for June of each year between 2009 to 2021**

I think it’s interesting that with occurrences of both trip distances equal and not equal to 0, there is a dramatic drop in the average fare\_amount/trip\_distance. This could be an error in the data set or it could be that the cab company took note of the toll that Uber was taking on their business and decided to closer match their pricing with a drastic price drop.

**Results of rate of change in the number of trips for June of each year between 2009 to 2021**

When thinking about creating this chart I had hypothesized that the drop in rides would be a steady rate decrease, however the rate of change goes up and down sporadically. Because I don’t think there’s an accurate and informed way to average this data, I decided to not use column 4 as a predictor for the change in rides for 2020. Similar to things we learned in chapter 7, because prediction and optimization can be very tricky, it’s best to find ways to simplify the process.

Obviously because of the pandemic we see very dramatic increases and decreases in 2020 and 2021. The percent change for the year 2019 I also believe is an outlier. It could be that at this point Lyft and Uber were really cutting into the New York cab companies but I think it’s more likely that the change coincidentally was much higher that year. For most other years the change ranges between much lower numbers and specifically in the years leading up to 2019/2020 the percent change ranges from 9 to 14. I decided to look at what the minimum (9%) and maximum (14%) of this change in 2020 and 2021 would look like.

As shown in the tables above, the predicted minimum change results in 6,316,306 rides for 2020 and 5,747,838 rides for 2021, while the predicted maximum change results in 5,969,257 for 2020 and 5,133,561 for 2021. This means as things get back to normal after the pandemic rides for each month should be between 5,133,561-6,316,306. As we can see from the actual data collected in 2021 this number is still much lower at 2,834,232. As we know, covid case numbers were decreasing in summer 2021 but then picked up heavily in the winter. It may not be until later this year or summer at the least that they reach this goal range. Without further data it’s hard to say whether this slow increase in rides is specific to the yellow cab company or if it has also been a slow increase for Uber and Lyft. This would be an interesting case for future work.

**Comments and Obstacles**

As previously mentioned, there were several times, like with the average (fare\_amount/trip\_distance), where the data would dramatically change within one year and then remain steady again around that new value. This could be an error in data collection or could be true to a change that actually occurred.

It was really inconvenient that several values, like the trip\_distance, fare\_amount, tip\_amount and other monetary values were entered as whole numbers. Depending on how these values were rounded, this could’ve had a huge effect on the data. This is especially true for the trip distance since mileage under 1 was entered as 0. Because of this it’s hard to gauge how customers were charged based on trip distance, but in these cases the value was replaced with .9.

Some of the querying I did, I would’ve liked to use to create a new column and enter the calculated values but the sandbox we used had strict permissions that prevented a lot of actions. When trying to alter tables in Athena, there was a permissions error that stated the user wasn’t allowed to perform this action. When I tried to create a new user, it stated that the sandbox didn’t allow for new user creation. This also prevented me from connecting an S3 bucket to Quicksight or using Quicksight at all, so in the end I used Tableau to visualize the data I collected from queries. It makes sense that there would be limits on what you can do in AWS using this sandbox since this is likely used a lot for student access to learn, but it definitely created some obstacles.

**Future Work**

If such data is published somewhere, I think it would be interesting to directly compare yellow cab data to Lyft and Uber data for each year. It would be interesting to note when they increase or decrease in the same time periods or when it’s clear their businesses directly affect one another.

Around 2016/2017 the CEO of Uber faced a lot of criticism and was eventually replaced. During this time, millions of Uber users deleted their accounts. It would be interesting to see how this affected the business of yellow cabs and Lyft.

Another future consideration is to see if between this spring and summer (2022) whether the yellow cab ride numbers return somewhere between the predicted minimum and maximum values shown previously. Currently the yellow cab data is only available through June of 2021, which is why I chose that month for each year so the most recent numbers could be analyzed.

If someone were to hypothetically use my work, I have provided details and documentation on what queries were used and how calculations were done throughout Deliverable 2 and 3.