NYC Taxi Data Analysis (The B Team)

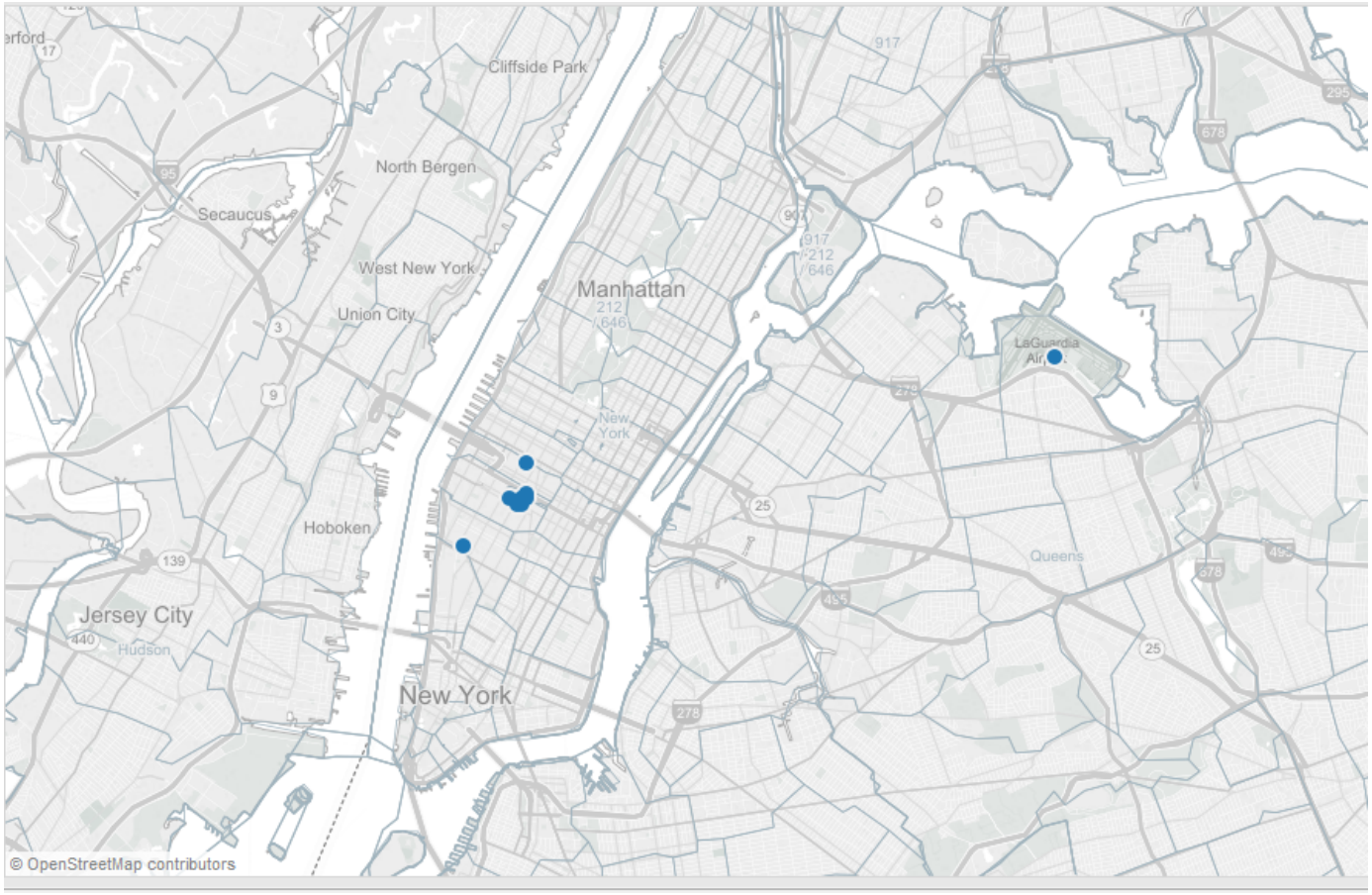
Objective: Thoroughly analyze the Taxi trip and fare datasets and provide useful.

Link for Final findings:

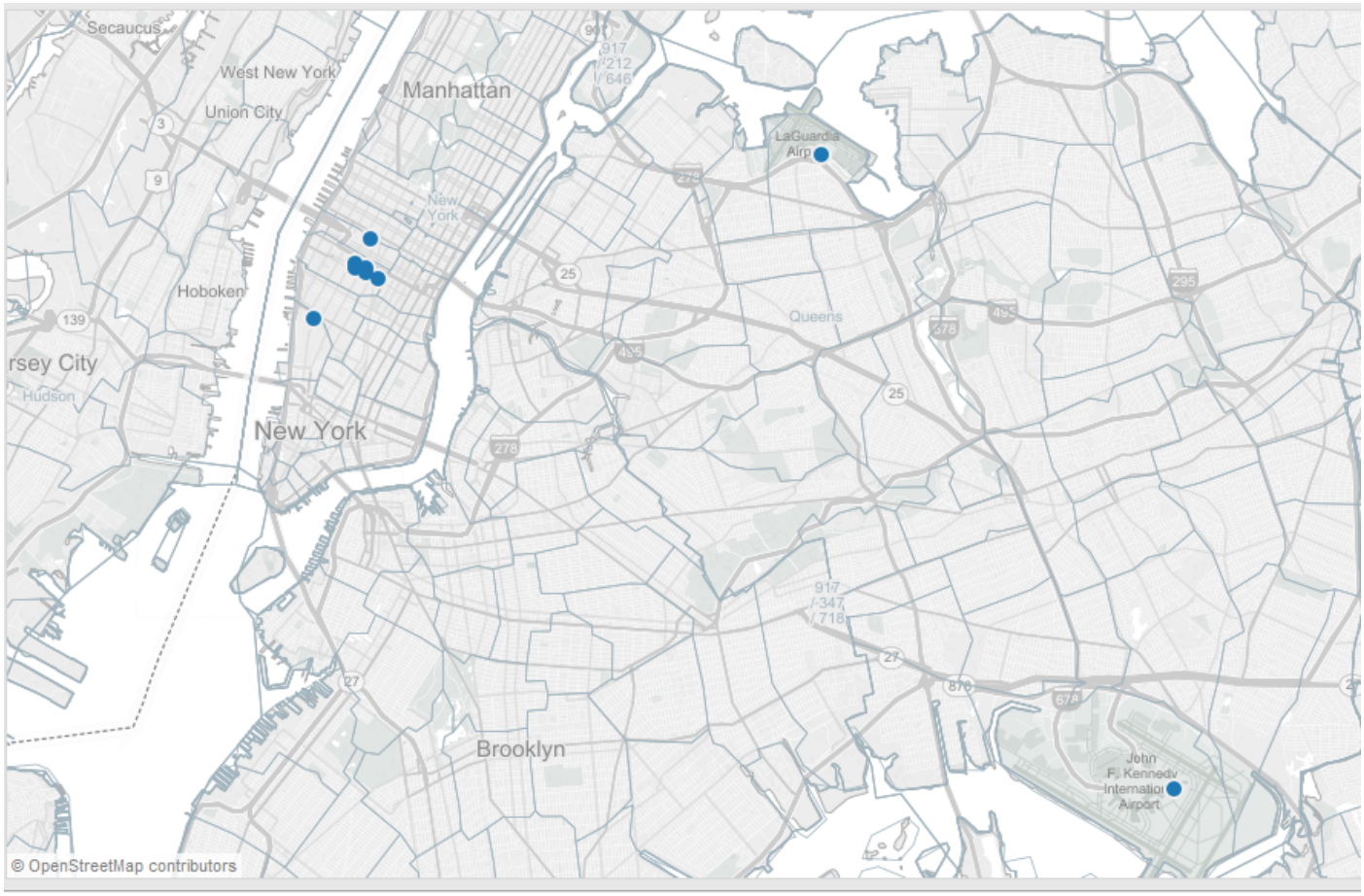
https://docs.google.com/presentation/d/1enU2gm-xBdUvwuEPPNpzowiAoxSy-vLmDbYduORs3A8/edit?usp=sharing

Preliminary Findings.

1. The top 10 busiest spots of dropoff.
2. On an average, 34 st Penn Station leads.
3. Also, we can see that **JFK does not make it to this list because people do not prefer an airport at such long distance, also most of the domestic airlines shed at LaGuardia.**



4. Top 10 busiest spots of pickup



5. On an average, again 34 st Penn Station leads.
6. Anomaly in the data wherein the the co-ordinates for pickup and dropoff is recorded as (0.0,0.0). Surprisingly there are plenty of medallions with this defect. We are assuming that this is due to faulty operation of GPS. Faulty because there are numerous records which has a pickup location but the dropoff location is recorded as 0.0 and vice versa.

Final Findings:

1. **GPS Anomaly: Presence of tall buildings that scatter the GPS signals**

- Further analysing on the GPS data we found that about 589 taxis do NOT have GPS working throughout the year and just these taxis contribute to about 92545 rides per year
- An additional 610 taxis did not have GPS readings for majority of the year
- 224 taxis did not have GPS working at some month throughout the year
- Thus, this incomplete data does not help determine the busiest hub in manhattan wherein the difference in number of rides for Top 10 are fairly close.

Further we found that around **101,183** had **no entry** for **pickup** locations and **220,460** did **not have entry** for **drop-off** locations. To better understand this we queried to find multiple taxis with such records and tracked its movement in ascending order of time and with little difference in time we picked the next pickup location to the record with drop off as '0.0'. **Shockingly most of these locations were in the Manhattan**, predominantly **lower manhattan and east midtown**. On further research we learnt that it is **caused by the presence of tall buildings that scatter the GPS signals**.

2. **Tip amount: To analyze human behaviour**

- We decided to delve further into tip amount in the fare data set. Some interesting findings were:

  Total rides in 2013 = 173,179,771

  Total rides paid by Card = 93,334,004

  Total tip data for rides paid by card = 90,450,929

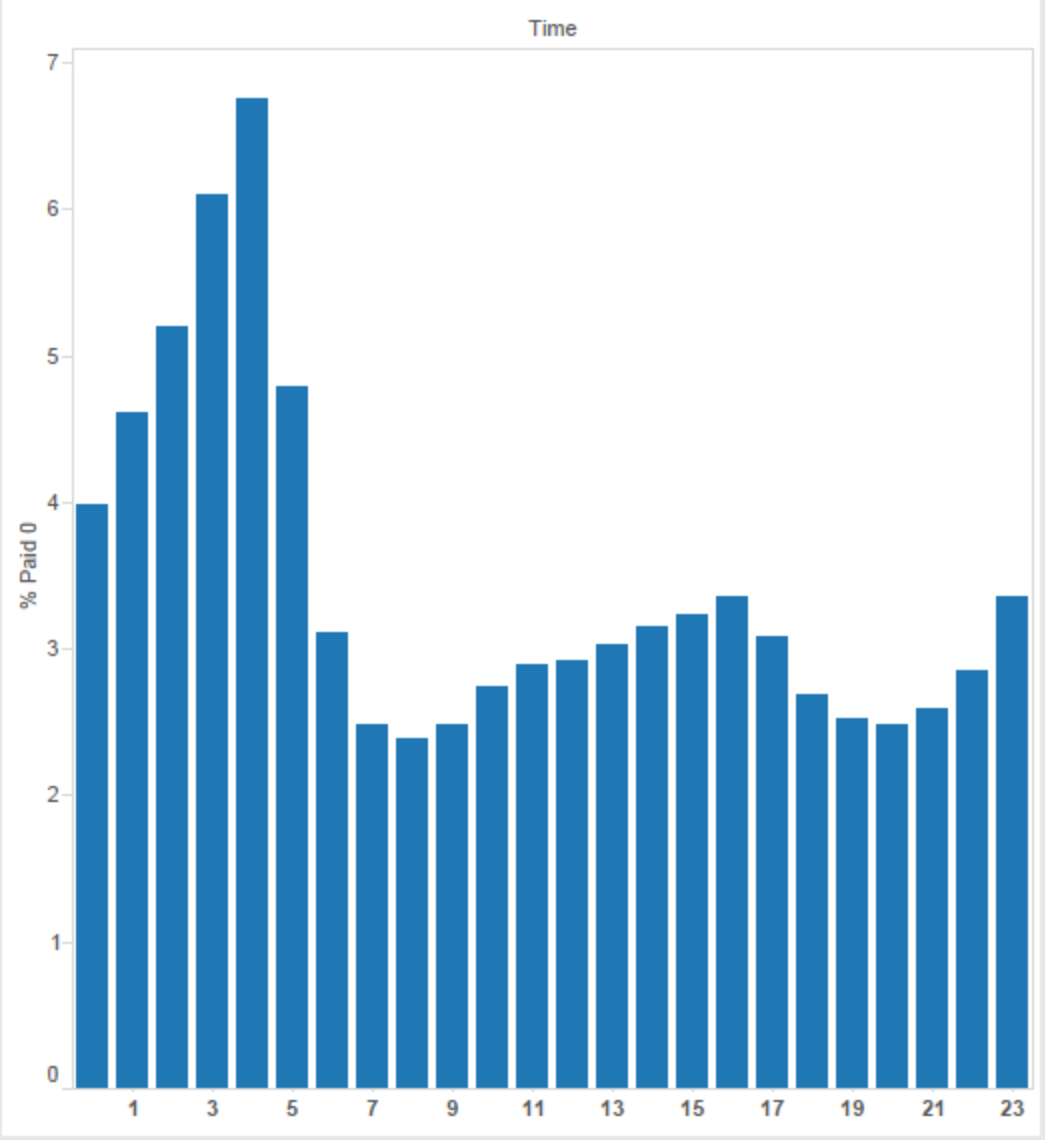
  Total rides paid in Cash = 79,110,096

  Total tip data for rides paid in cash = 6,708

  Because there is not enough data for tip paid in cash we decided to perform further analysis only with Card payment records.

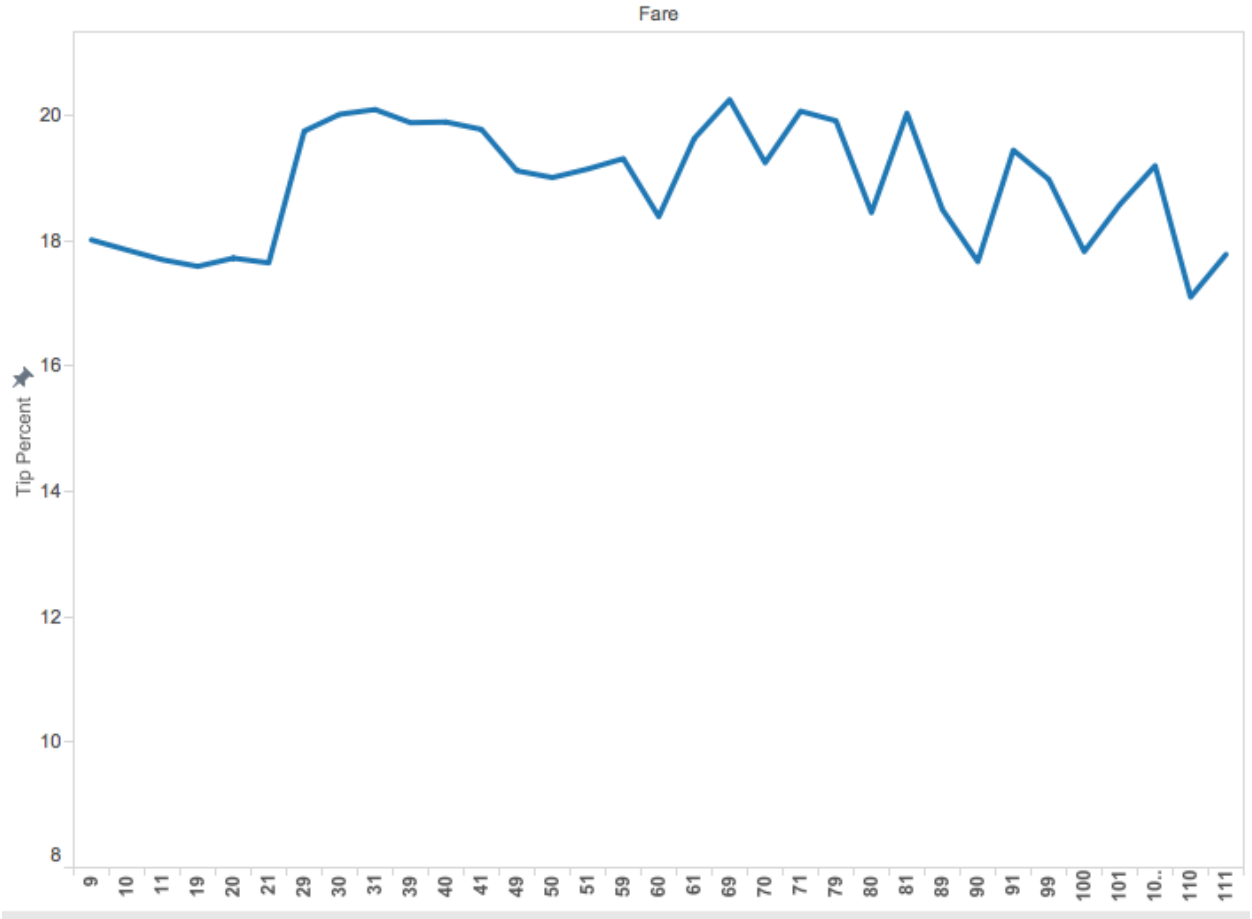
- **Drinking and tipping don’t go together**

  From tip amount analysis for payment with credit card we found that the percentage of people who do not tip vary greatly between midnight and 5:00 AM which is otherwise constant for the rest of the day. Our assumption for this pattern is; longer the person stayed at a bar lesser the tip. Apart from this major factor other reason that could justify this data is probably frustrated people working late shifts.



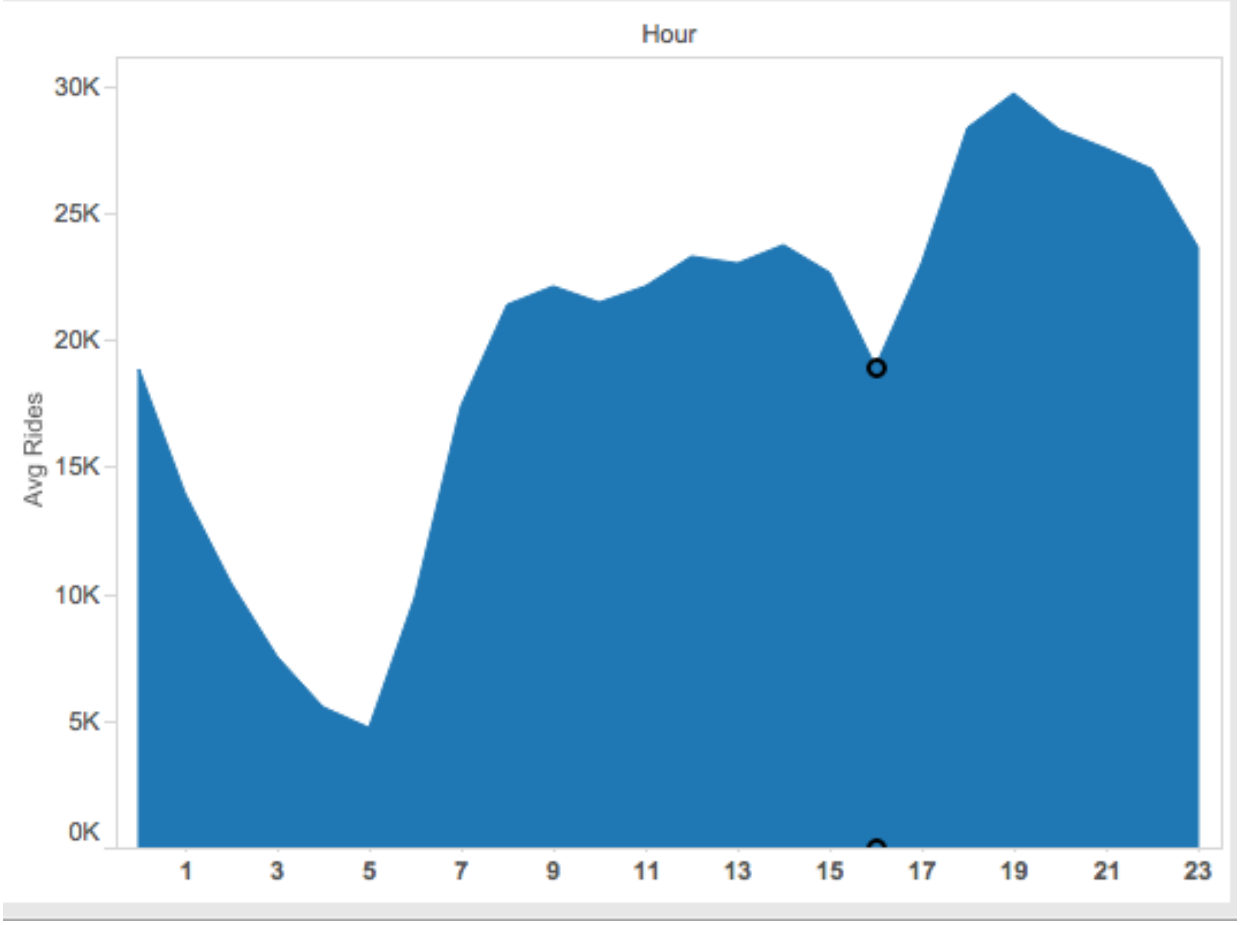
- **Tip amount decreases when trip fare reaches tens place and then increases gradually**

  Further performing boundary analysis around fare amounts that has 10 as multiple (e.g., 68,69,70,71,72) as the average of tip steadily increases with fare amount, we find that people tend to tip less when fare is exactly a multiple of 10 in this case 70. This could be for two reasons, One is the person's tendency to round off to nearest multiple of 5 or 10 to do an easier math and ignore a standard 18% tip, other reason could be that person's thinking that spending $70 or $71 is lot more than $69 or $68.



3. **Drivers change shift around 4 pm and its difficult to find taxi during that time**

In trip data; on querying to find the average number of trips in a month for each hour of the day, we noticed an evident drop between midnight and 5AM as expected and also noticed a sudden drop in number of rides between 4:00 PM and 5:00 PM which was not expected. To understand this, we further analysed number of trips and total number of taxis on daily basis. And this fact was still evident that there was a drop in total rides just between 4 to 5 PM and it rises again after that. To get a better reasoning we pulled out the total number of taxis(medallions) and the drivers(hack license) for that time interval and found that there were more drivers than taxis (an average of 235). This could mean that there was a **shift of drivers.**



hour	medallion	hack lic.	diff.	rides
12	10321	10328	7	24170
13	10360	10361	1	23495

14	10398	10400	2	24834
15	9563	9598	35	22711
16	8072	8277	205	17646
17	9324	9388	64	23258
18	10746	10752	6	30720
19	11309	11313	4	33034

Failed Attempts:

1. We were able to find a rich dataset (NYC Open Data) on the Traffic collisions. That had relatable fields like latitude and latitude of collision and the type of vehicle involved as 'TAXI'. An unbelievable number of around 15,000 major and minor TAXI collisions have occurred. We wanted to find that if any of these taxi's had a pickup entry and if the dropoff was at the collision location also if the passenger paid for the trip. We failed to find and match because most of the location data in the collision data set are street and junction names. And we also cannot relate dropoff time with collision time because it's not exact. Another hurdle was being unable to relate collisions without passengers to the trip data provided.
2. With the medallion, hack licence and taxi running times we set out to find number of drivers working day and night shift, also how long the taxi was in motion without any passenger. We could not obtain accurate results because the taxi could have been without any passenger for really long hours between the actual shift end and last drop off.

Technologies Used:

- Initially we struggled with the performance of NYU HPC with uploading large files from local machine so moved the game to AWS EMR with hive.

- Exhausting the credits shortly after preliminary findings, after a prolonged period of time we uploaded all the data sets into NYU HPC. The performance was much better than the initial period.
- For analysis we build complex and efficient Hive Query Language queries, such that the have would run as few mapper functions as possible.
- The resulting tables were exported and used to plot graphs for better visualizations.

Team Contribution:

All tasks including discussions, planning, execution and documentation were always carried out in group ensuring equal contribution from all.

The following link is a ppt for the prelimnary finidings

https://docs.google.com/a/nyu.edu/presentation/d/1bgv0si4ddk0ybhQLapPngREQIxSXB3-wVP7drhaLmMw/edit?usp=sharing