RedEye Supply-Demand Gap Analysis & Possible Solutions

Group 13 - DS 5110 - Spring 2022

Shagun Saboo | Dhruvilsinh Jhala | Amritanj Ayush | Sanjana Gadalay

1. Summary

RedEye is a one-of-a-kind off-campus shuttle service run by the Northeastern University Police Department (NUPD) that helps Northeastern University students staying off-campus, get home safely at night. Students who live within two miles of the centre of campus can use the RedEye shuttle services by downloading the Redeye app and reserving a ride through it.

NUPD started offering this service in 2017, and between Fall 2021 and Spring 2021, they received over 90,000 requests, which is the highest number ever in such a short period of time. However, due to rising demand, there has been a significant supply deficit for the past two years. Most students deal with long wait times, ride cancellations, and a shortage of available seats on a regular basis. Since graduate students who live off-campus, and students who stay in college until late rely on this service, we chose to focus our efforts on finding a solution to this problem, or at the very least easing the current situation, because the problem we are seeking to address is important and would have a significant impact on the Northeastern University student community. As a result, through this project, we will aim to bridge the demand-supply gap by first identifying operational bottlenecks, then reducing student wait times, and ultimately providing recommendations to boost the chances of students receiving a confirmed seat on the RedEye.

NUPD provided us with a large dataset of all requests made on the RedEye app over the course of roughly three years, from 2019 to 2022. There are 314,344 rows/observations and 31 columns/variables in the dataset. The following are key variables for each dataset observation:

- ➔ Date and time of the request created (eg: 2022-01-11 17:28:09)
- ➔ Request Status (Completed, Cancelled, Seat Unavailable, No-showed etc.)
- ➔ Address & coordinates of Origin/Pickup location
- ➔ Address & coordinates of Destination/Drop location
- ➔ Time from request creation to planned pickup i.e. estimated waiting time
- ➔ Ride Distance (in miles)
- ➔ Ride Duration (in minutes)

Request Status	Count	Percent
Completed	166,141	52.85%
Seat Unavailable	70,747	22.51%
Not Accepted	33,961	10.80%
Cancelled	21,332	6.79%
Out of Service Hours	9,263	2.95%
Other Error	7,259	2.31%
No Show	5,641	1.79%
Total	**314,344**	**100%**

Table 1: Breakdown of Request status of RedEye

According to our findings from RedEye data, there were a total of 314,344 requests, of which only 166,141 were successfully completed and the remaining 148,203 were not completed for a variety of reasons (Seat Unavailable, Out of Service hours, No Show, Not Accepted, Cancelled). This indicates that approximately 47.15 percent of users are unable to use the service owing to a variety of factors and the rate of completed rides is 52.85 percent, indicating that the service needs improvement.

2. Methods

2.1. Programming language: R

2.2. Libraries used - Ggplot2, dplyr, tidyr, ggmap, lubridate, geosphere, RColorBrewer, patchwork, maps, osmdata, here, movis, read_xlxs, readr

2.3. Pre Processing

Data preprocessing is the first step towards transforming raw data into a clean data set. This is about turning data into a base form that makes it easy to use. A feature of a tidy data set is one observation per row and one variable per column. This part is vital and must be done properly and systematically. Otherwise, we will construct models that are not specific. Below are the pre-processing steps that are applied to our dataset.

- **Split Date & Time Column**: Initially the dataset contained date and time merged as a single feature. So, we have split this into a separate column containing date and time separately. Further, Date has been divided into year and month columns and time has been split to hours.
- **Removed NA Values:** Next, we have deleted all the observations that contain missing values.
- **Encoded Categorical Data**: Then we have encoded the categorical data. Here, Encoding refers to transforming text data into numeric data: all the ride status which are Completed is taken as 0 and those which are not completed are stored as 1
- **Group into 4 Areas**: Using the Destination Latitude and Longitude, we have divided the Area into 4 parts as northwest, northeast, southeast and southwest where the centre point is the redeye pickup point near Snell library.
- **Group Data to seasons:** We also grouped the data according to the seasons like fall, spring and summer.
- **Convert Date to Weekday:** Lastly, we have converted the date into the weekday like Monday, Tuesday, etc.

2.4. EDA

Exploratory Data Analysis (EDA) refers to describing data using statistical techniques and visualisation to develop significant aspects of this data for more in-depth analysis.

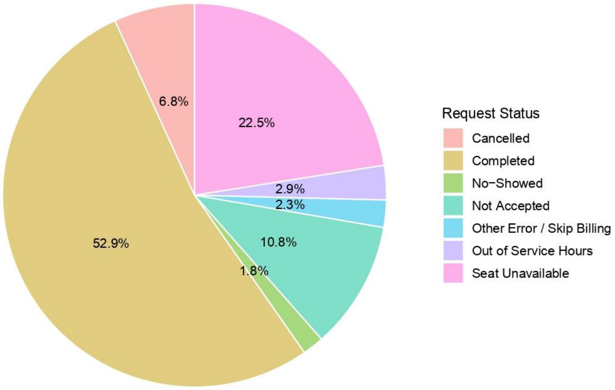
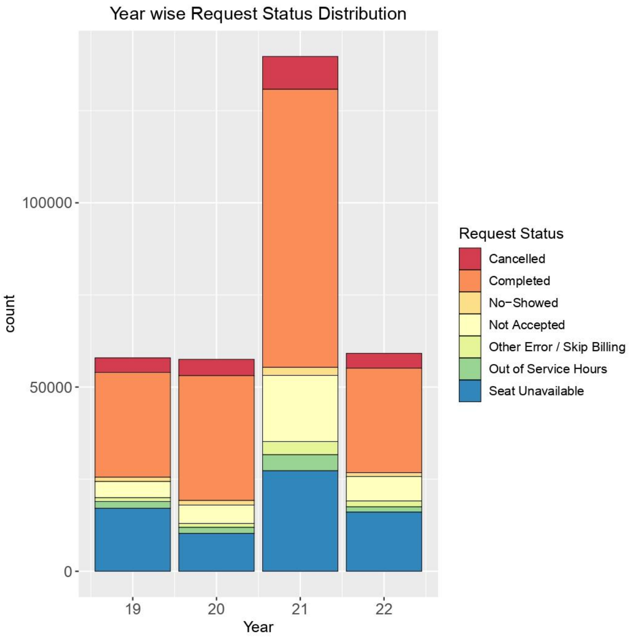


Figure 1: Request Status on Daily basis

As seen in the above Pie Chart, out of total data, only 52.9% rides were Completed, 22.5% requests had Seat Unavailable, 10.8% were Not Accepted, 6.8% were Cancelled, 2.9% times Out of Service Hours, 2.3% Other error and 1.8% were No Showed.



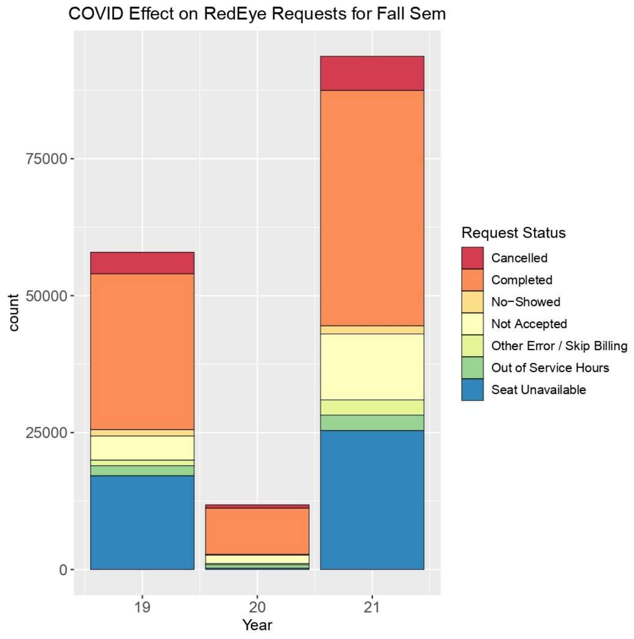
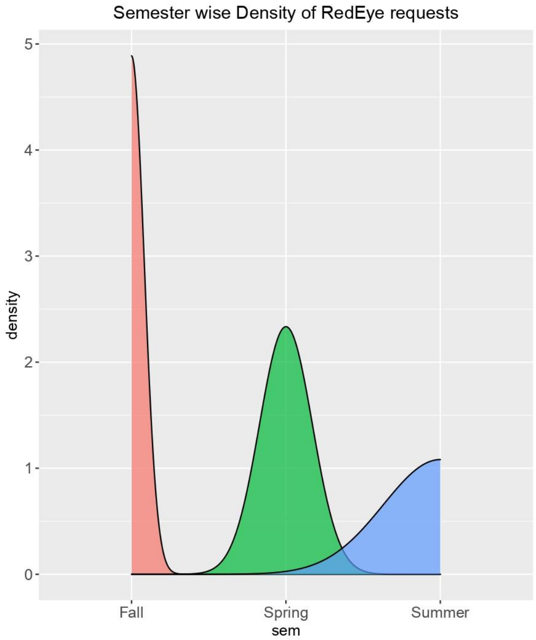


Figure 2: Request Status on Yearly basis

Figure 3: COVID19 effect on RedEye Requests for Fall Sem

The above stacked bar charts represent the distribution of Request Status given with respect to Year. From the first plot, it can be noted that maximum requests occur during 2021. This is because our data only has data from Sep 2019 till March 2022. This is the reason 2019 and 2022 show such less values. As we have data for all Fall semesters from 2019-2021, we can go ahead and check for that period. According to the second plot, 2021 has the highest ride requests, followed by 2019. Also 2020 has the least number of ride requests most likely because the university was closed due to COVID Pandemic and the classes were taken online for that duration.



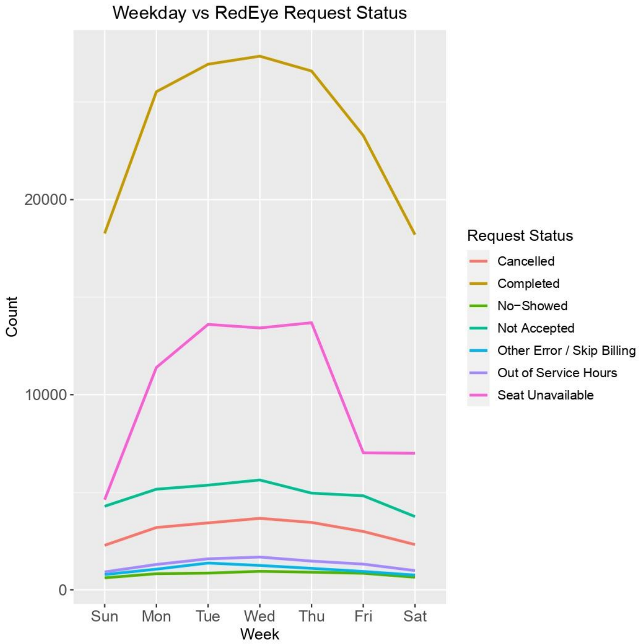


Figure 4: Request Status on Semester Basis

Figure 5:Request status on Weekly Basis

The above density plot visualises the amount of requests made for the RedEye, broken down by semester. The line chart shows the trend of requests made for the Redeye during the days of the week. As per the first plot, Fall semester seems to have maximum ride requests followed by Spring semester and Summer with minimum requests. This observation makes sense as less courses are available during Summer and students usually go for Internships during this period. Request Status was also compared with weekdays, where it was found that weekdays have a higher number of requests compared to weekends.

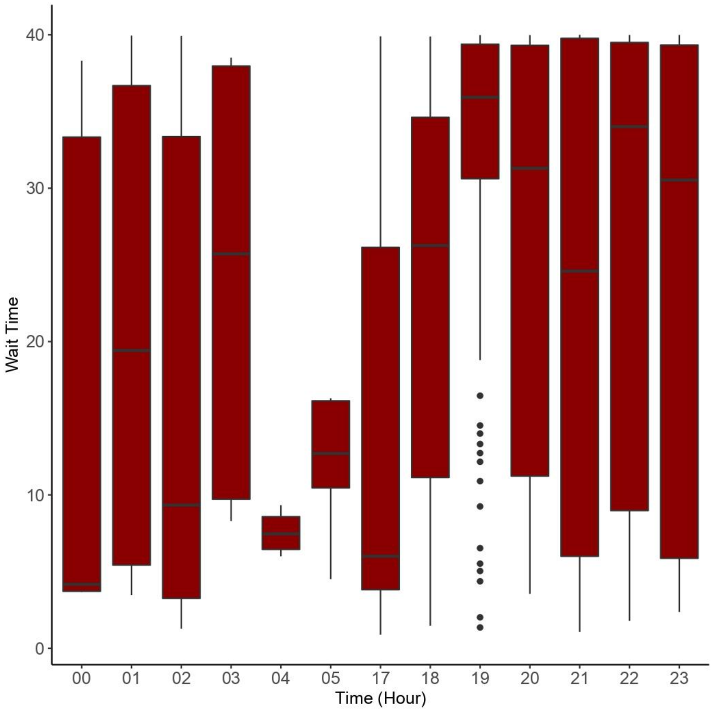
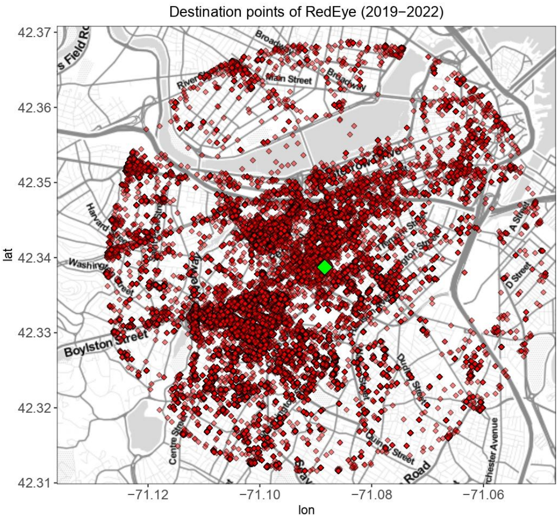


Figure 6:Wait-time Graph on hourly basis

The box-plot chart demonstrates the waiting time distribution (time from request creation to planned pickup) during specific hours of the day when the Redeye service is operational. As seen in the above box-plot we can observe the minimum wait-time at midnight which increases in the next hour and again it's dropped at 2:00 am with a significant increase at 3:00 am, where as the maximum Wait time is at 7:00 pm which keeps on decreasing as we approach the midnight. The increase and decrease pattern observed after 12:00 am is because of the turn around time of 40 mins, which increases because of the decrease in RedEye during midnight.



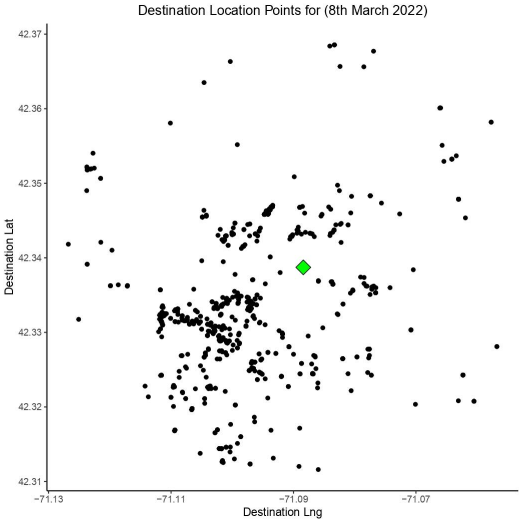
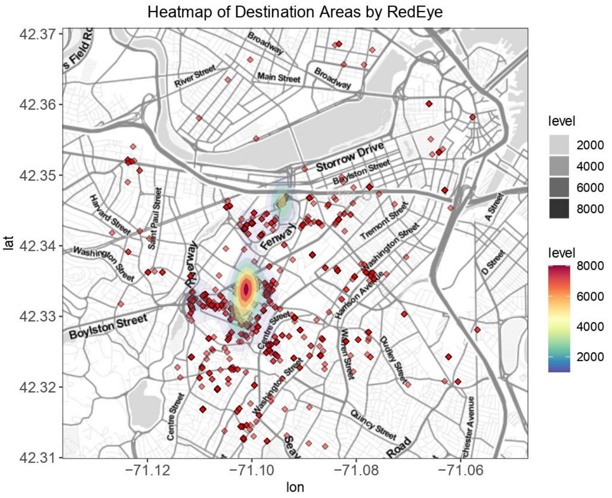


Figure 7:Destination Points of RedEye

Figure 8:Destination Points of RedEye for 8th March 2022.

In the visualisations shown above, we have plotted all the destination coordinates on the Boston map for 3 years (2019-2022) and our test day - 8th March, 2022, respectively. The radius of RedEye service is 2 miles from Snell Library, and so a circular boundary is formed from the centre green point which in this case is Snell Library. We have also plotted a single day (22 March, 2022) coordinates with green point being the origin point.



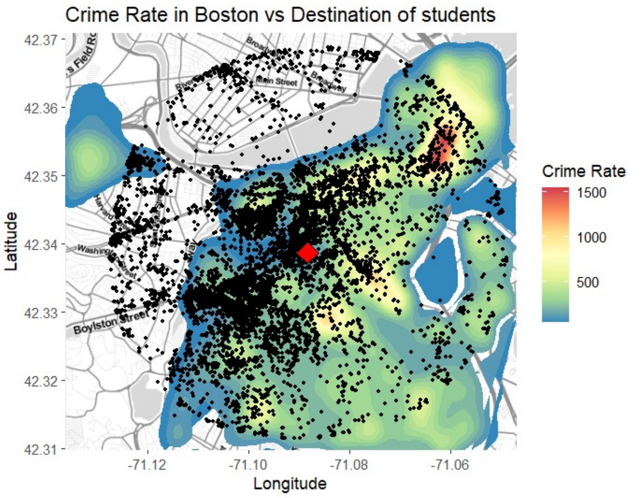


Figure 9: Heatmap of Destination Areas

Figure 10: Crime rate in Boston vs Destination of students

Using the 22 March, 2022 day's coordinates, we plotted a heatmap, which shows that there is a high demand in the Longwood Medical Area. This may be because several northeastern graduate students live in Jvue, Longwood and Mission Main apartments in the same area. There also seem to be a little density near Bolyston area. Our analysis also included checking if RedEye is able to serve areas with high crime rate, so we scattered our destination coordinates over the crime rate heatmap. It was noted that more crimes took place in Downtown boston and less requests were made for that location.

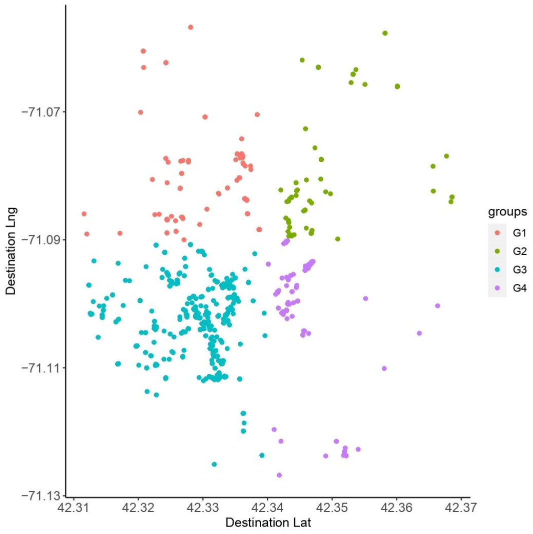


Figure 11: Division of Destination Locations into groups

We grouped the coordinates into four groups to better comprehend the different regions on the map. Each group in the March 22, 2022 data has been colour coded, as seen in the above visualisation. Using these categories, we hope to assign a certain number of RedEye cars to each region, which will drop students off at a common drop-off site or serve students who stay in the region and return to college, minimising turn-around time and distance travelled per trip by each van.

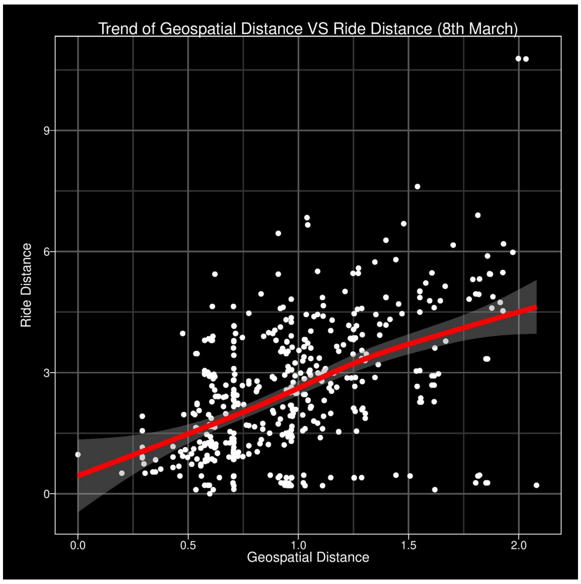
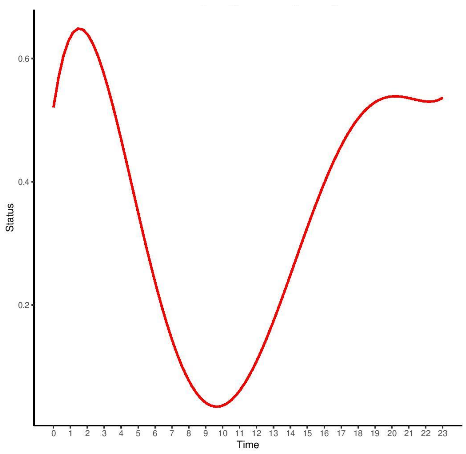


Figure 12: Trend of geospatial distance vs ride distance on 8th march 2022

We also used the ‘geosphere’ library to get the Geospatial distance of each ride from origin coordinate to destination coordinate and compared it with how much distance the rider actually travelled. We found that for every 1 mile of geospatial distance, the rider travelled approximately 3 miles.

2.5. Modelling



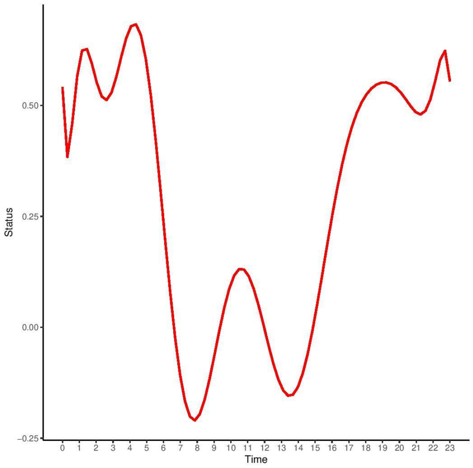


Figure 13: Polynomial Regression Model (n=5)

Figure 14: Polynomial Regression Model (n=14)

A Polynomial Regression model was built with ‘Time’ as predictor variable and ‘Status of RedEye’ as the response variable. This model can predict the chance of successfully booking a RedEye at any given time. Initially, we have considered n=5 in our model but the model performed to be underfitting on any new data as seen in the first graph. Hence, we updated n value to 14 and the model performed better as shown in the second plot. Here status close to 1 indicated that a seat will be available in the RedEye and status close to 0 shows that a seat can’t be booked. N, here stands for regression degree, which gives the relation between independent variable x and dependent variable y as nth degree polynomial in x.

3. Results

Using sample data from March 22, 2022, we compared the effectiveness of the current working approach to our proposed method. We used data from 7-8pm to gain a better analysis because it is the peak time of day.

Total 190 Requests were made out of which 87 were Completed, and 103 Not Completed (45.78%)

Proposed Method

1) Divide regions served into groups
2) Evaluate the number of requests in each region
3) Allocate number of cars accordingly

	Requests
Group 1	12
Group 2	15
Group 3	131
Group 4	32

Table 2 - Requests made per region

	Cars
Group 1	1
Group 2	1
Group 3	5
Group 4	1

Table 3 - No. of cars allocated to each group

	Group 1	Group 2	Group 3	Group 4	Total
Completed	13	13	65	13	104

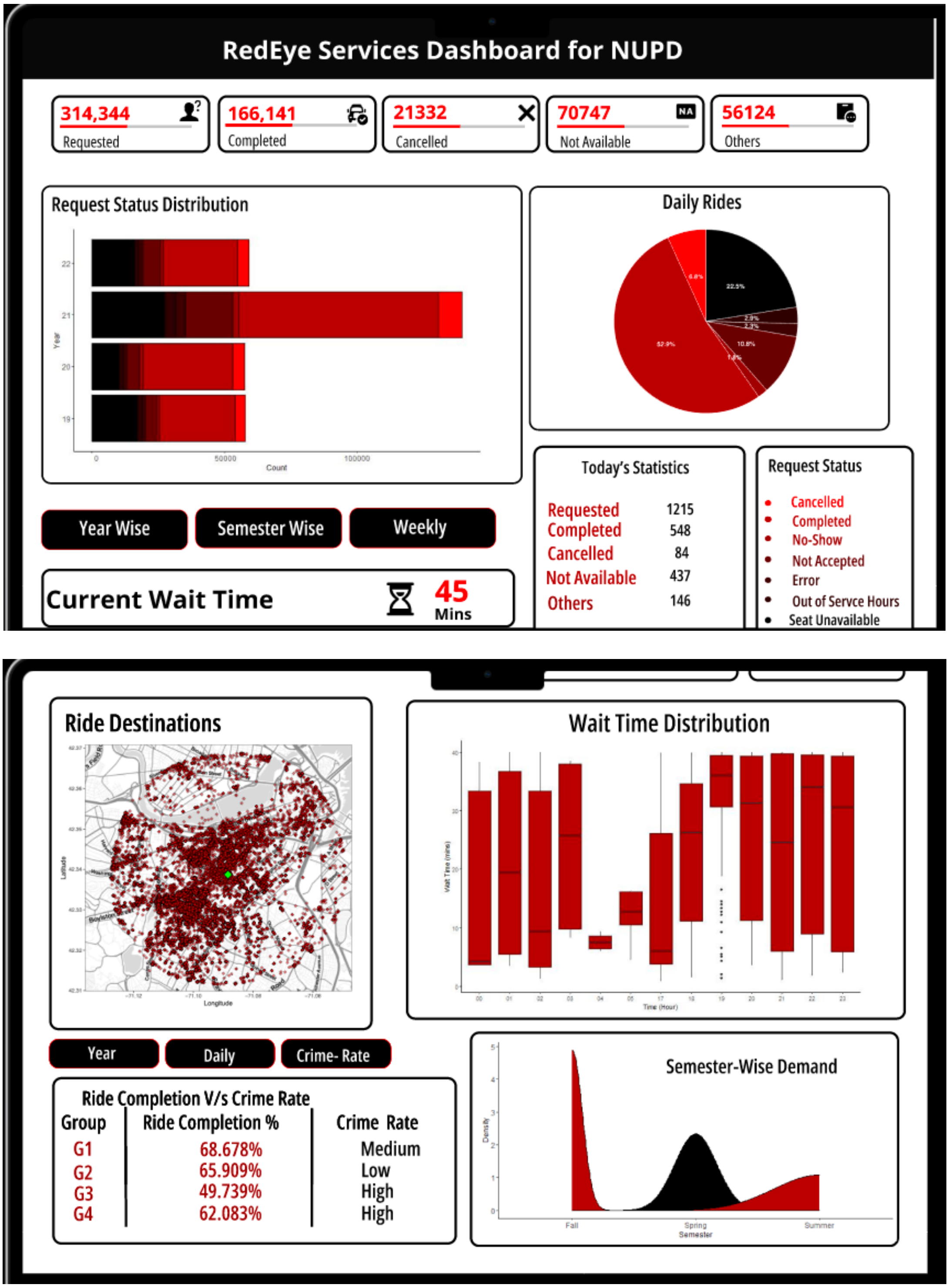
Table 4 - Number of students served per group

Table 2 shows the number of requests on 22nd March, 2022 from 7-8 pm and Table 3 shows the number of cars allocated to each group of regions based on the number of requests. Now statistically implementing the number of requests that can be completed using our method in Table 5, the total comes up to 104. This is significantly higher than 84 which was before. Hence using our proposed method, there is a 19.54% increase in the number of rides completed.

Cars	Completed	Percent Completed
+1	117	61.5%
+2	130	68.4%
+3	143	75.2%
+4	156	82.1%
+5	169	88.9%
+6	182	95.78%
+7	195	100%

Table 5 - No. of students served if additional cars are added to existing inventory

We went forward and got the number of RedEye cars needed to make all requests completed. Here each car added makes approximately 6.8% difference, and hence 7 more cars are required so that each request can be served, and 100% of the requests can be completed.



RedEye Services Dashboard for NUPD

314,344	Requested
166,141	Completed
21332	Cancelled
70747	Not Available
56124	Others

Current Wait Time 45 Mins

Today's Statistics

Requested	1215
Completed	548
Cancelled	84
Not Available	437
Others	146

Request Status

- Cancelled
- Completed
- No-Show
- Not Accepted
- Error
- Out of Servce Hours
- Seat Unavailable

Ride Completion V/s Crime Rate

Group	Ride Completion %	Crime Rate
G1	68.678%	Medium
G2	65.909%	Low
G3	49.739%	High
G4	62.083%	High

Figure 15: RedEye Services Dashboard for NUPD

Additionally, using all our analysis, a dashboard has been created for NUPD which contains all the yearly, semester-wise data for the usage statistics of RedEye. A snapshot of the dashboard we created for NUPD has been displayed above in Figure 15.

4. Discussion

The current student population served by the RedEye shuttle services is fairly small, according to our findings. The majority of students are unable to get a seat on the shuttle since demand far outnumbers supply. We, as students, chose to seek a solution to this problem or, at the very least, to alleviate the current situation because the ultimate benefactor of our analysis is the student population. The project's findings and analysis will undoubtedly benefit the student population, but it will also help the NUPD create successful measures to make the Redeye more accessible and expand the number of students served. Our findings will aid NUPD in determining the hours of the day and months of the year when they can expect strong demand and effectively deploy more or fewer cars accordingly. With that stated, we recognize that NUPD places utmost priority on student safety, which is why the service was created in the first place, whereas our aim is to expand RedEye's accessibility. As a result, we defer to the university and supporting authorities to determine where the intersection will be placed and how to strike a balance between making the RedEye more accessible while maintaining safety.

To summarise, grouping rides according to their destinations would undoubtedly improve operations and efficiency. This is a low-investment strategy for increasing efficiency. Increasing the number of shuttles running during peak hours is the most obvious and capital-intensive approach. If we add additional cars, we'll need to hire more drivers, which means more people on the university payroll. It is entirely up to the university and supporting authorities to determine if such an expenditure is compatible with their goals.

Finally, if we implement all the above recommendations we will be able to reduce the turnaround time and that would result in more trips per van and hence more students served. We requested NUPD for more data on the student to help us determine whether the student is a graduate or undergraduate; the number of cars running on a single day; and which students/observations were grouped in one ride on a particular day, all of which were missing from our dataset due to confidentiality concerns. These missing data points would have allowed us to look into more interesting links and improve our recommendations, but we respect the need for anonymity and did our best with the dataset we had.

5. Statement of contribution

We chose the RedEye data after debating three alternatives presented by Shagun and Dhruvilsinh, as well as their candidates for obtaining the data. The group together approached the respective administrations for the same. Apart from ideation Dhruvilsinh worked upon Data preprocessing, EDA, Animation, Modelling and the documentation. Shagun contributed to EDA, Animation, Proposed Solution and development of Dashboard. Amritanj Ayush worked upon EDA and further turned them attractive using various ggplot elements, he also worked on getting the animation up and running, and extensively worked on all documentation of the project. Sanjana was always ready to pick any work and was responsible to submit it as soon as possible. She contributed majorly in Data preprocessing along with EDA and documentation.

6. References

- https://nupd.northeastern.edu/
- https://cfss.uchicago.edu/notes/raster-maps-with-ggmap/
- https://www.geeksforgeeks.org/adding-data-points-to-world-map-in-r/
- http://movevis.org/
- https://statisticsglobe.com/geospatial-distance-between-two-points-in-r
- https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston

7. Appendix

Code Repository - https://github.com/dhruviljhala/RedEye-Data-Analysis