Indiana University Bloomington

CSCI B 565

Data Mining

IU Bus Route Optimization - Group Name : Kinesis

Author:
Anup Prasad
Shalab
Hemant Pandey

Submitted to:
Dr. Mehmet
Dalkilic

December 18, 2015

Abstract

With the rise in the number of student population enrolling each year at Indiana University, Bloomington, the demand for transportation for the students within the campus also increases. One such transportation service fulfilling this need is the IU Bus Service. But the current frequency as well as the availability of the bus is proving to be insufficient for the increasing ridership. To deal with this situation, the IU Bus Service provided us with their day to day data and required us to propose a system using the data that would meet the current challenges faced by them. One of the major challenge faced by the IU Bus Service was the effective frequency mapping which would be efficient for both the riders as well as the IU Bus Service employees. Other challenges included dynamic routing as per requirement or the removal or addition of a particular bus at a given time period on the basis of past predictions. All these challenges led to the development of this project. The project required us to treat the raw data available to us, enrich it, transform it into a more structured database and finally perform various mathematical operations on the database which allowed us to generate some conclusions on the basis of our understandings.

Contents

1 Introduction

The IU Bus Service Project has been a significant part of CSCI-B 565 Data mining course. The impact of this project is so significant that upon completion not only it will provide the students with a plethora of new skills, it will be affecting the lives of both the IU Bus Service as well as the riders of the bus in a positive fashion. The Project required us to work in teams to analyze the raw data provided by the IU Bus service, make meaning out of the data by linking various segments of the data to one another, and finally propose a system that could effectively optimize the service so that it could be efficient in all possible means for both the IU Bus Service as well as the riders. The project inputs were a set of goals that the IU Bus Services wishes to get fulfilled through this project. Although the goals were very vague, but we managed to redefine them more objectively as per our understandings. Another set of input was the raw data which the IU Bus Service had collected over a period of time. This raw data comprised of Microsoft Access as well as Excel tables having various entities associated with the Bus activity like: route data, schedule data, weather data, ridership data, etc.

2 Problem Description

One of the most crucial stages of any project is the problem identification or the requirement analysis phase.Therefore, we spent our initial meetings in identifying the problem.We went through the requirements and tried identifying the goals of the project. The goals defined were very vague. Hence, we had to define the goals in a more granular manner so that they could make more sense and would allow us to connect our goals with the data that was available to us.We identified the various characters or objects that were the key players in the project and created use case diagrams to understand their impacts on the Bus Service and make sense out of them.These characters/objects were: Bus, weather, route, driver, etc.The objectives broadly given to us are as follows :-

1. to calculate variance from the given schedules .
2. To describe the effects of various factors on stop timings like weather and passenger count .
3. Determine average travel time between any two stops at any time of the day.
4. Maximize the on-route time with sporadic breaks .Optimize for most passengers carried while maintaining a minimum frequency .

2.1 Presentation of case study

The technologies used for this project are: Java, Python, Json, Mongodb, MySQL, MS Access and MS Excel. The data provided by IU Bus Service had many outliers, and many data sets like the weather data were incomplete which were taken care of using legitimate data from sources like NOAA. The hardest challenge was to make sense out of the terms used in the database. Many challenges were faced during this project and it was hard to propose an effective system under strict time constraints. But we are finally able to propose this system that can propose solutions to almost all the major goals set by the IU Bus Service so far by developing graph plots and route instances (visual representation of a route along with factors affecting it). As one progresses through the report, one can read more about the various modules and phases of the project.

2.2 Mathematical formulation

Formula for variance

Average variance of An Instance of a route $= \frac{\sum_{i=1}^{i=n} ArrivalTime_{actual(i)} - ArrivalTime_{schedule(i)}}{n}$
where,
n = number of stops for which schedule arrival time is given.
Above formula can be extended to calculate average variance in arrival time for a given route for following cases .

1. Given week days i.e what is the variance on Monday, Tuesday, Wednesday and Thursday .

2. Time frame of a day i.e. On a given day at what time e.g 8:00 am to 9:00 am, how much is the variance in arrival time .

Formula To calculate Load Time

Average load time of an instance of a route= $\frac{\sum_{i=1}^{i=n} LoadTime_{Stop(i)}}{n}$
Where,
n = Total number of stops present in the route.

Above formula can be extended to calculate Load time under following conditions.

1. Average load time for all instance= $\frac{\sum_{instance=1}^{instance=n} AverageLoadTimeOfAnInstanceofaRoute}{n}$
   .
   n = total number of instances exists in the data base for the particular Route(A,B,E or X)

2. Average Load time under different weather event like Fog, Haze , Snow etc .

3. Average Load Time at different stops at different point of time .

Formula to calculate travel time

Average travel time for an instance of a route= $\frac{\sum_{i=1,j=i+1}^{i=n-1,j=n} TravelTime_{Stop(i)\ to\ Stop(j)}}{n-1}$
Where,

n = Total number of stops present in the route.

Above formula can be extended to calculate Load time under following conditions.

1. Average travel time for all instances= $\frac{\sum_{instance=1}^{instance=n} AverageLoadTimeOfAnInstanceofaRoute}{n}$
   .
   n = total number of instances exists in the data base for the particular Route(A,B,E or X)

2. Average travel time under different weather event like Fog, Haze , Snow etc .

3. Average travel Time at different stops at different point of time .

Mathematical model to predict average travel time between any two given stop at any time of the day.

We can use Naive Bayes to predict Average Travel Time between any two given stop at any time of the day .For this, a training data set has to be created with following structure.

Week Day	Time Frame	Weather	From Stop	To Stop	Avg Travel Time
Monday	9:00 AM	Snow	Stadium	Jordan	5 min
Monday	10:00 AM	Snow	Jordan	Everman	7 min
Tuesday	12:00 PM	Normal	Wells Library	Everman	2 min

Table 1: Sample training data set.

In the above training set,time frame is discrete value ,it indicates start time of a bus from source stop. If a bus starts at 7:25 Am then its time frame would be 7:00 AM - 8:00 AM.In the same way Average travel time has been changed from continuous to discrete value using below table .

Actual Travel Time Range(in minutes)	Discrete Value(in minutes)
0-5	5
6-10	10
11-15	15
16-20	20
21-25	25
22-30	30

Table 2: Table to map continuous travel time to discrete value .

Now, Naive Bayes can be applied as follows :
$P(Traveltime \mid X\{WeekDay, TimeFrame, Weather, FromStop, ToStop\}) = P(Traveltime) \prod_{i=1}^{5} P(X_i \mid Traveltime)$
Where,
$P(Traveltime \mid X\{WeekDay, TimeFrame, Weather, FromStop, ToStop\})$ = posterior probability.
$P(Traveltime)$ = Prior probability.
$P(X_i \mid Traveltime)$= Class conditional probability.

3 Data Description

Note : The intervaldata2014-2015.tsv file contains data for various bus id's and route id's for as well as the load time and travel time for various transactions of the bus.

The IU Bus Service data was the key input for this project and it would not have been possible to even initiate the project without it. For this, Logan provided us with many MS Access and Excel files that contained information about ridership, climatic factors, routes, buses and their designated stops, GPS data, etc. The data helped us to visualize the approach towards the fulfillment of the goals provided by the IU Bus Service.

3.1 Cleaning and Enrichment

As the saying goes in data mining, that about 80For instance, we neglected all the entries having travel time between stops or the dual time at stops more than 2 hours. While using the weather data, we could only find entries for 114 days and that too having limited weather information which was

not sufficient enough to describe the effect of weather conditions on the Bus Ridership. To deal with this problem, we used the annual data available by NOAA (National Oceanic and Atmospheric Administration) to get all the required data like weather events, precipitation, temperature, etc. to understand the effect of weather conditions on the Bus Ridership. Further we neglected the GPS data as it had many outliers like GPS data pointing to locations in Chicago and Florida. Finally, we grouped the bus data on the basis of route id, and for each route id we again grouped the data on the basis of bus id. This way we grouped all instances for a route for a particular bus id.

3.2 Transformation

Once the data was enriched and cleaned, it was ready to be pushed into the database so that it can be used with the various programs that we had written to achieve the goals assigned to us. For this we split the given intervaldata2014-2015.tsv file on the basis of route id and further spitted it on the basis of bus ids. We created a Java code that identified all the present route instances and converted it into JSON representations which contains Meta data for each route like: weather events, total load time, total travel time, route id, schedule id, temperature, precipitation, and the representation of route in the form of travel sequence containing stops. An example of a route instance is as follows:

```
{
        "TotalTravelTime": "711",
        "Temperature": "20",
        "BusID": "638",
        "TotalLoadTime": "747",
        "EndTime": "08:51:51",
        "TravelSequence": "Stadium(A)--->Alumni Center--->Briscoe--->McNutt--->Kelley School--->Well's Library--->Neal Marshall--->3rd & Jo
        "StartTime": "08:30:29",
        "WeatherEvent": ["Snow"],
        "Date": "2015-01-30",
        "ScheduleName": "A3",
        "WeekDay": "Friday",
        "TotalStops": "18",
        "Percipitation": "T",
        "RoutID": "355",
        "StartPoint": "Stadium(A)",
        "RouteName": "A",
        "Route": [{
                                "From": "Stadium(A)",
                                "To": "Stadium(A)",
                                "LoadTime": "181",
                                "When": "08:30:29"
                        }, {
                                "From": "Stadium(A)",
                                "To": "Alumni Center",
                                "Travel_Time": "80",
                                "When": "08:31:49"
                        }, {
                                "From": "Alumni Center",
                                "To": "Alumni Center",
                                "LoadTime": "24",
                                "When": "08:32:13"
                        }, {
                                "From": "Alumni Center",
                                "To": "Briscoe",
```


Json representation of a route instance

Finally, once all the instances of all the routes have been identified, they are pushed into MongoDB where we created a collection named route. In total we identified 30432 instances of route for the year 2015.

3.3 Existing model

The data provided to us was in MS-ACCESS tables and .tsv file . It was a data dump containing the entire schedule of an year for all the buses across all the routes . The data was not sorted and was totally jumbled up and did not make any sense . There was no particular sequence to it or a pattern to it .So we had to write up java code to extract meaningful information from it in terms of routes and bus ids'. The enrichment phase allowed us to remove many outliers present in the data as well.

3.4 Proposed model

In mongoDb we stored all the route instances along with other relevant data like the schedule of the route, BusId, RouteId, starting point, starting time , total load time, total travel sequence, weather, and weekday.This kind of structured of data cannot be maintained in SQL kind of data base . Sql-kind of databases require all data to be structured and well defined. They treat each record in terms of attributes which are well defined in terms of a particular data type. In MongoDB , however for instance, there were no constraints like that of primary and foreign keys. There was no constraint on the kind of data which could be inserted.We did not have to define multiple tables or define the relationship between those tables.Using JSON , we created structures of each route instances, which had several key value pairs.The keys were actors which impacted the routes.These actors were weather, load time, day, time, schedule of the route etc.

4 Computational Implementation

4.1 Calculation Of Variance

One of the main objective was to determine variance in schedule, where variance is defined as :

$$\frac{\sum_{i=1}^{i=n} ArrivalTime_{actual(i)} - ArrivalTime_{schedule(i)}}{n}$$

Each instance of a given route was mapped to a particular schedule of a given route present in Schedule Data table. In order to calculate variance we needed the actual time and the expected arrival time for each stop present in schedule table. We got the actual time, date, load time, schedule and actual arrival time for the year 2015 on the four weekdays, from the route instances (trips) that were identified during data modeling step. There were approximately 30000 such instances. These instances were arranged from the data provided in the intervaldata2014-2015.tsv file.

We have calculated variance for following conditions:

1 Variance for all the routes i.e. A,B,E,X .

2 Variance at particular time of day .

3 Variance on week days i.e. Monday , Tuesday , Wednesday , Thursday

4.2 Load Time And Travel Time under different weather condition

Load time represents the amount of time a bus waits at a particular stop to allow passengers to board and de-board .So the more the load time , greater the number of passengers availing the bus service .Where as, travel time represents the amount of time , a bus spent in traveling to complete a route . When the weather is bad(hazy) , travel time was found to have increased .These pattern can be inferred from the graphs being attached.

We have calculated travel time and load time for following circumstances .

1 Load Time and Travel time for all the routes i.e. A,B,E,X .

2 Load Time and Travel time for different weather event like snow or fog etc .

3 Load Time and Travel time for different weather event and different precipitation value.

4.3 Calculation Load Time at various stops

Relation between load time and passenger count can be described by following formula.

$$\text{Load Time} \propto \mid PassengerCount \mid$$

i.e. More is the number of passenger getting down or getting into the bus , more is the load time . We have used following formula to calculate Load Time.

$$\frac{\sum_{i=1}^{i=n} LoadTime_{Stop(i)}}{n}$$

where n = number of instances of a route in which $Stop_i$ is present

5 Visualizations

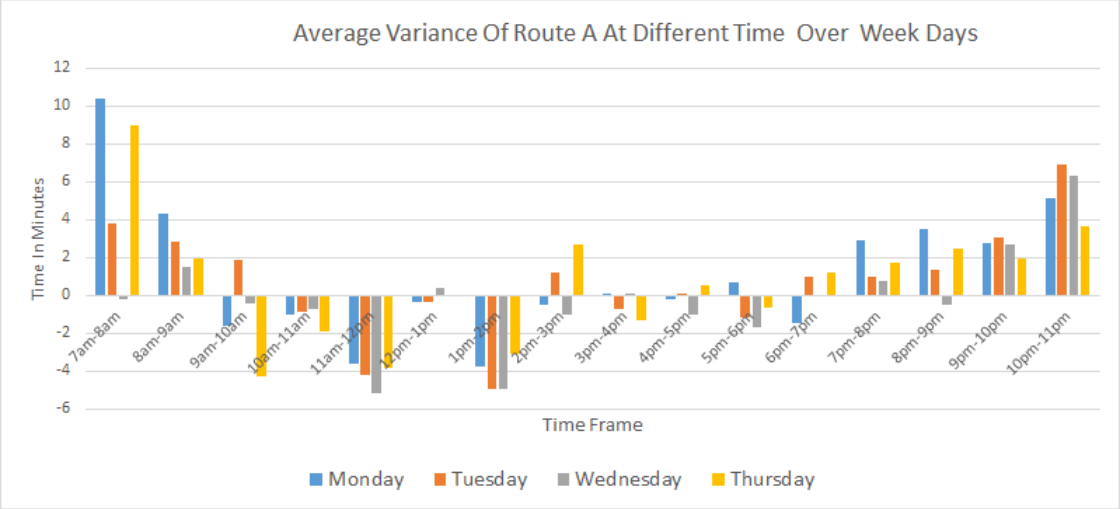
Based on our calculations we have represented our results using different graphs. These visual representations of result can help us in taking a better decision and making good recommendations.

5.1 Variance in arrival time

Two factors contributes to high variance.

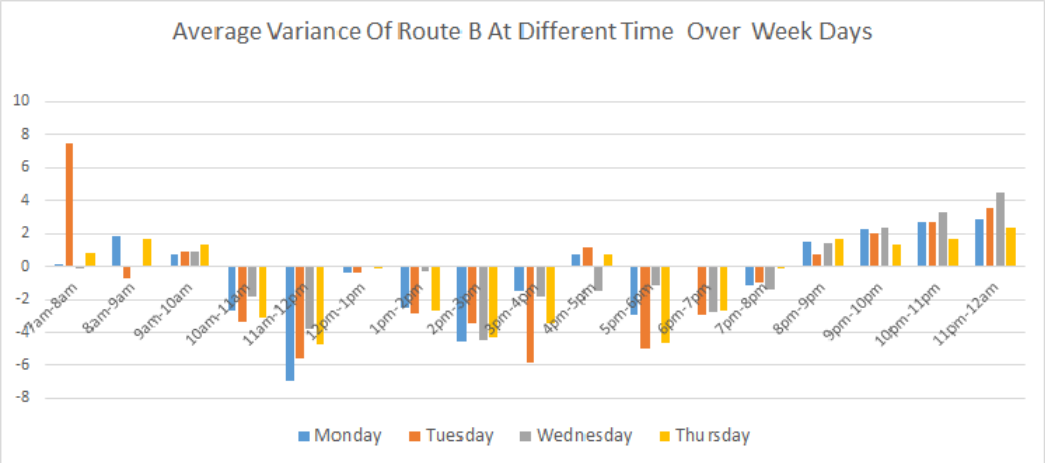
1 Traffic, high traffic leads to increase in travel time and hence more deviation will be there from scheduled arrival time.

2 Passengers count , a high passenger count leads to high load time at stops and thus leads to increase in deviation from schedule arrival time.



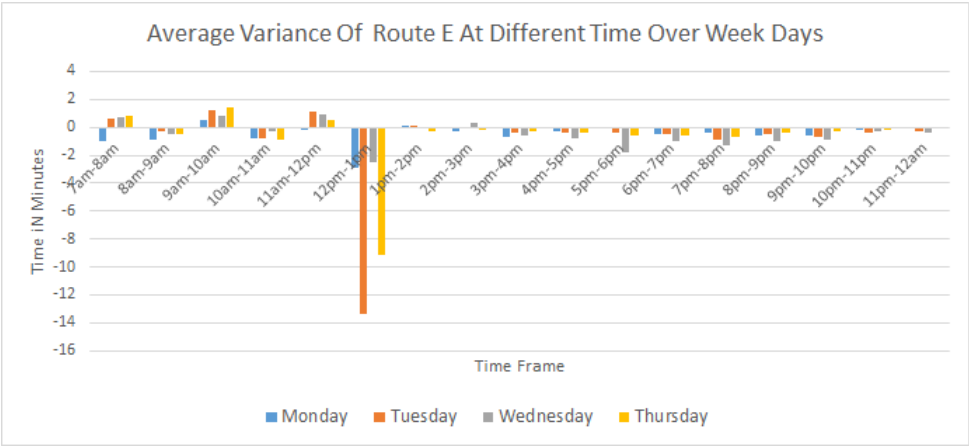
Graph 1: The variance of route A, at different time over weekdays.

From above graph, it can be concluded that for route A , Monday is the busiest day and time interval between 7:00 Am to 8:00 Am is the busiest time of the day, as the variance is highest at that point of time. A negative variance indicates that bus has arrived at stops before the scheduled arrival time.



Graph 2: The variance of route B, at different time over weekdays.

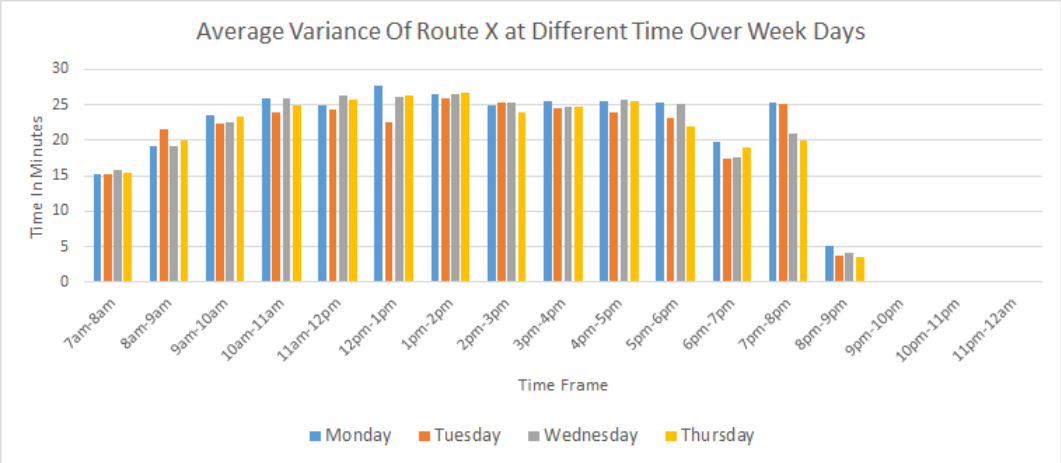
From above graph, it can be concluded that for route B , Tuesday is the busiest day and time interval between 7:00 Am to 8:00 Am is the busiest time of the day, as the variance is highest at that point of time. A negative variance indicates that bus has arrived at stops before the scheduled arrival time.



Graph 3: The variance of route E, at different time over weekdays.

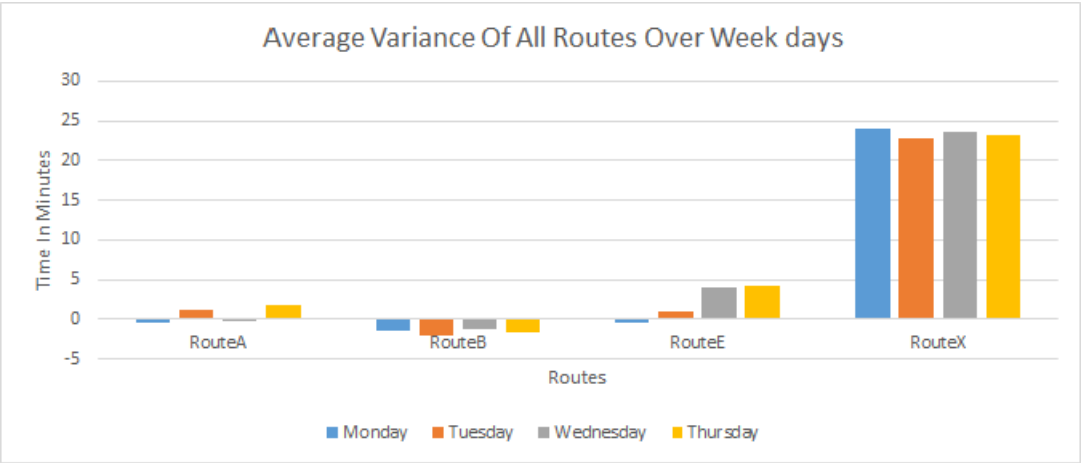
From above graph, it can be concluded that for route E , Thursday is the busiest day and time interval between 7:00 Am to 8:00 Am is the busiest time

of the day, as the variance is highest at that point of time.It is also evident that compared to route A,B,and X route E experiences lest deviation from schedule arrival time.



Graph 4: The variance of route X, at different time over weekdays.

From above graph, it can be concluded that for route X , Monday is the busiest day and time interval between 1:00 Pm to 2:00 Pm is the busiest time of the day, as the variance is highest at that point of time.While Tuesday is comparatively less busy day at all time instances .It is also evident that compared to route A,B,and E,route X experiences highest deviation from schedule arrival time.



Graph 5: The variance of all routes, on different week days.

Above graph shows consolidated result of average variance of all routes on 4 week days . We can infer that .

1 For route A consolidated variance is less than 5 minutes over 4 week days .

2 For route B consolidated variance is (-)ve indicating, route B is not as busy compared to route A, E or X.

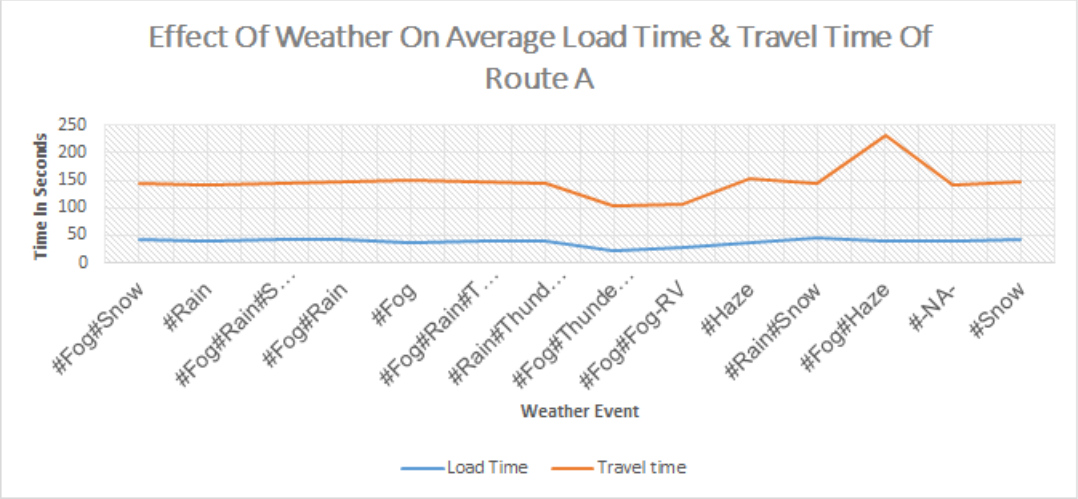
3 For route E , Wednesday and Thursday are busiest day .

4 For route X all 4 days are equally busy.

5.2 Effect Of weather on load time and travel time

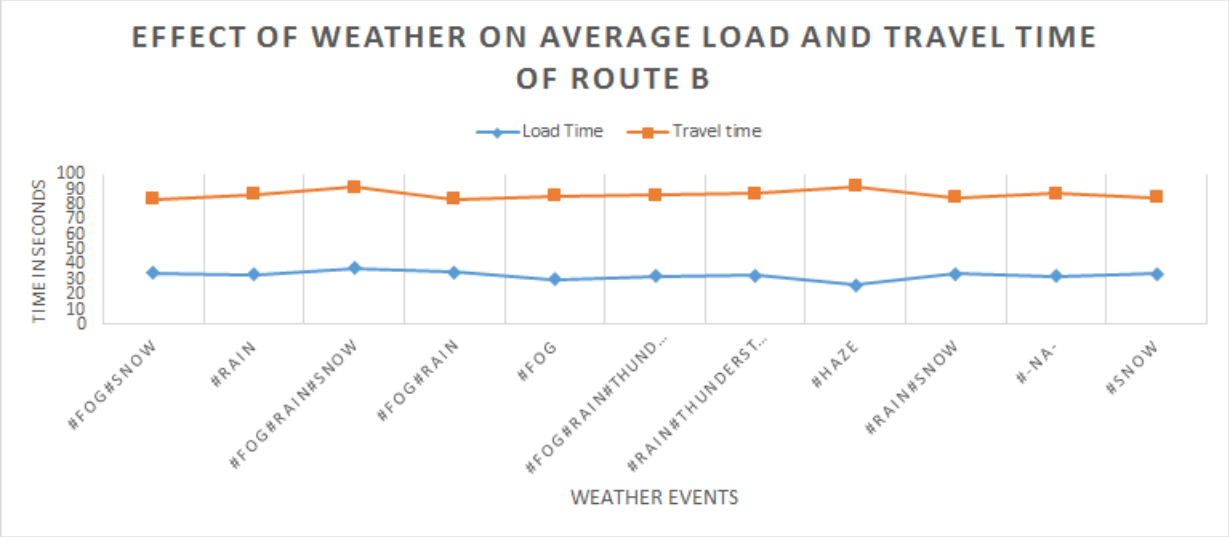
Load time and travel time are affected by weather, inclement weather tends to increase travel time while it's effect on load time is not known. With our results and visualization we have tried to evaluate the effect of inclement weather on load time .

Travel time show in all the graphs is average travel time between two stops .



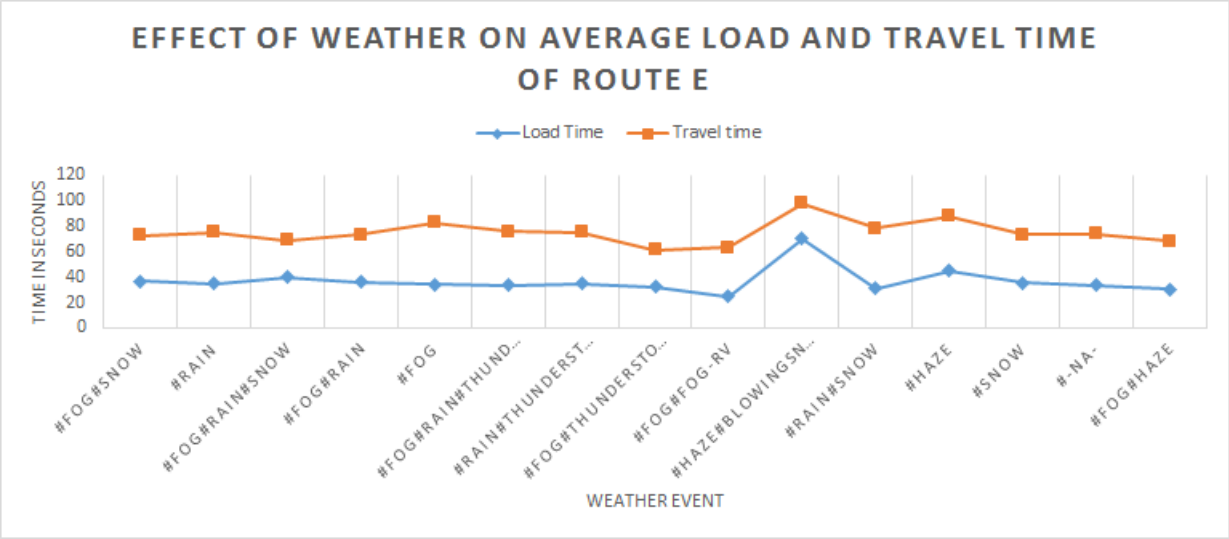
Graph 6: Effect of weather on route A.

There are three local maxima for travel time ,corresponding weather events for these maxima are rain and thunder, Haze , Fog and Haze .It is evident that highest increase in travel time is caused by Fog and Haze . Load time is highest for the weather event rain and snow. This leads to the conclusion that when it rains and snow then more people prefers to travel by bus than by car or other means .



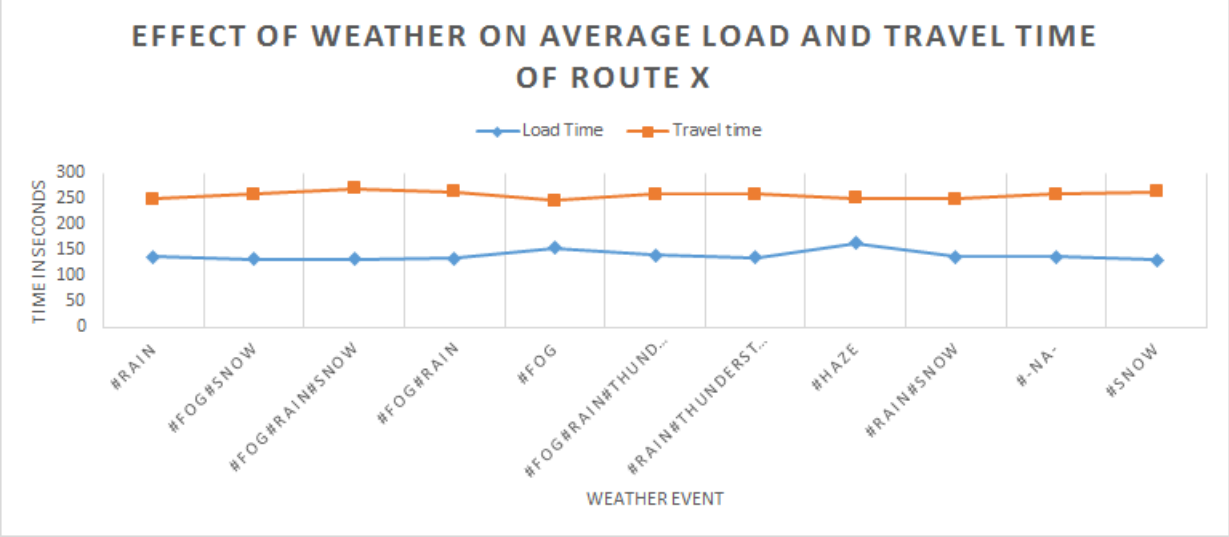
Graph 7: Effect of weather on route B.

Highest travel time is for haze and fog,rain and snow weather event . While there is slight increase in load time when weather event was fog,rain,snow and rain, snow.



Graph 8: Effect of weather on route E.

Highest travel time is for haze and haze, blowing snow weather event . While there is large increase in load time when weather event was haze, blowing snow

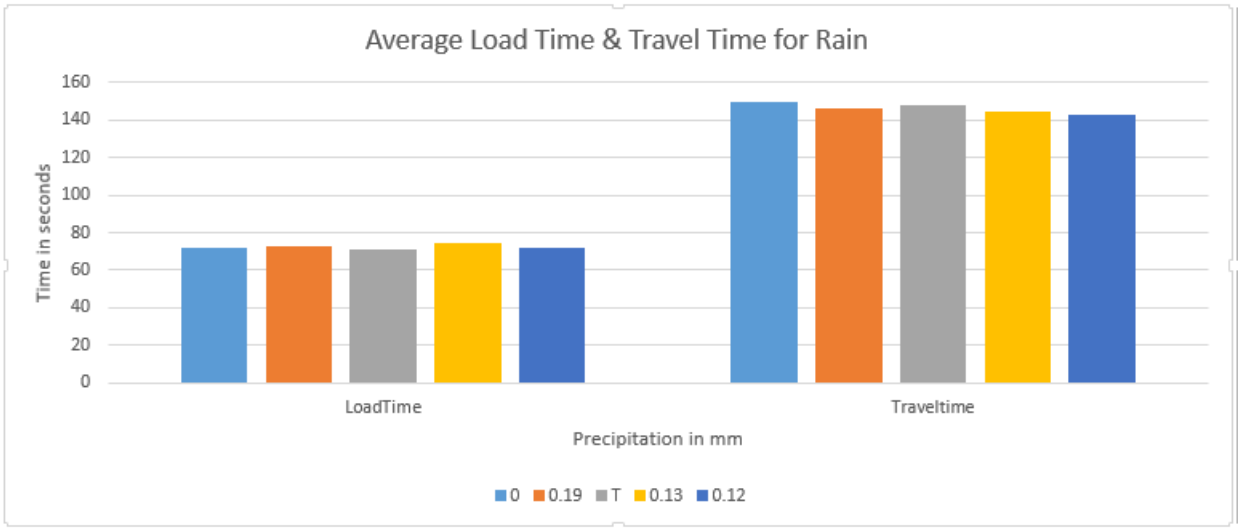


Graph 9: Effect of weather on route X.

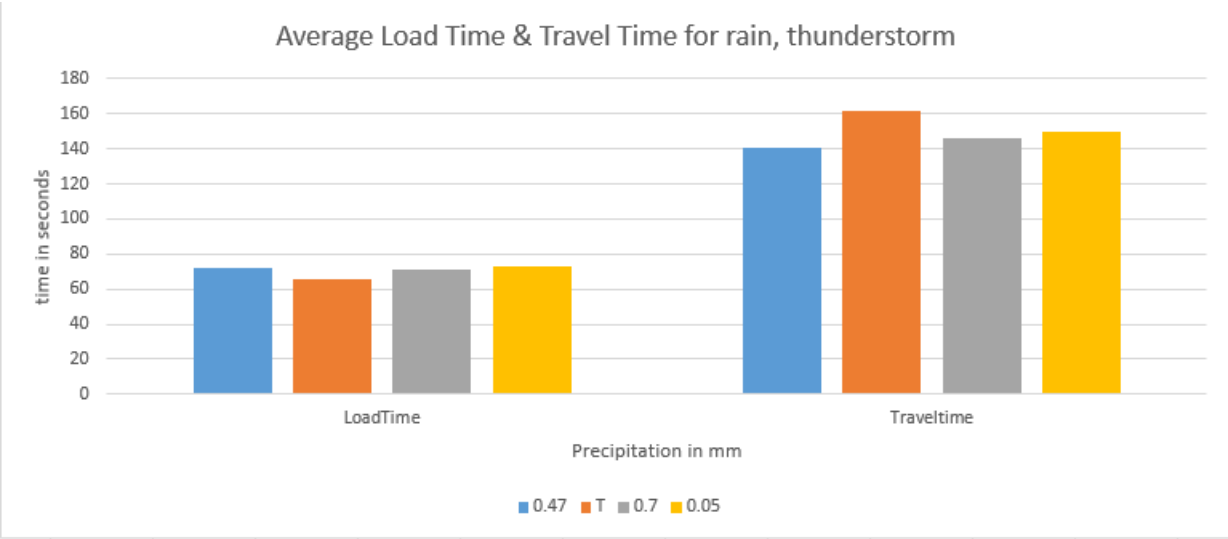
Travel time show in all the graphs is average travel time between two stops .

Effect of precipitation and weather on load time and travel time

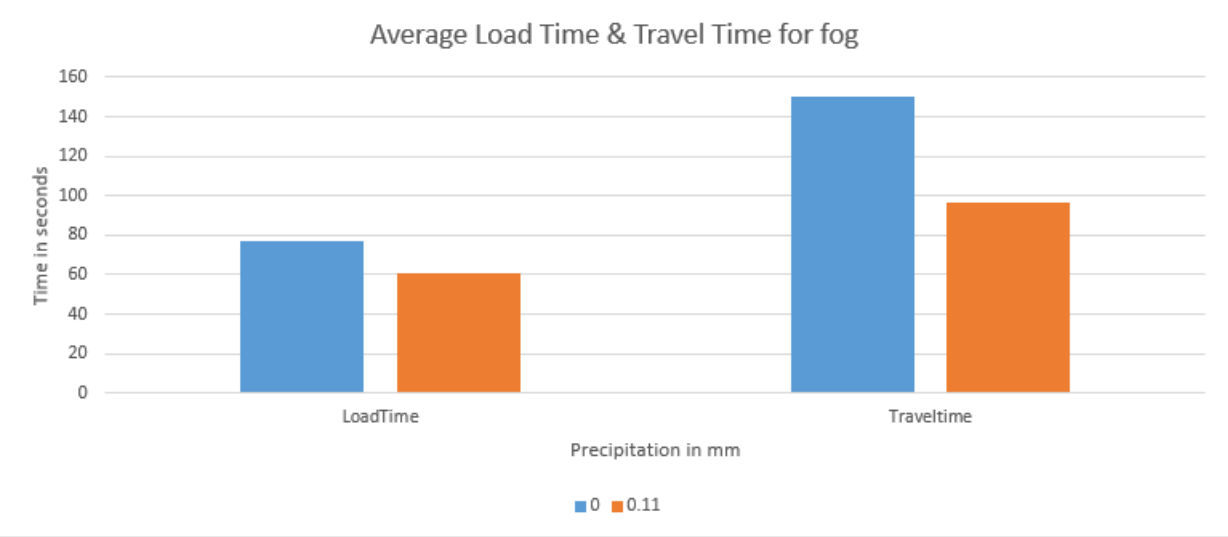
We have tried to find the role of precipitation in determining travel time an load time .



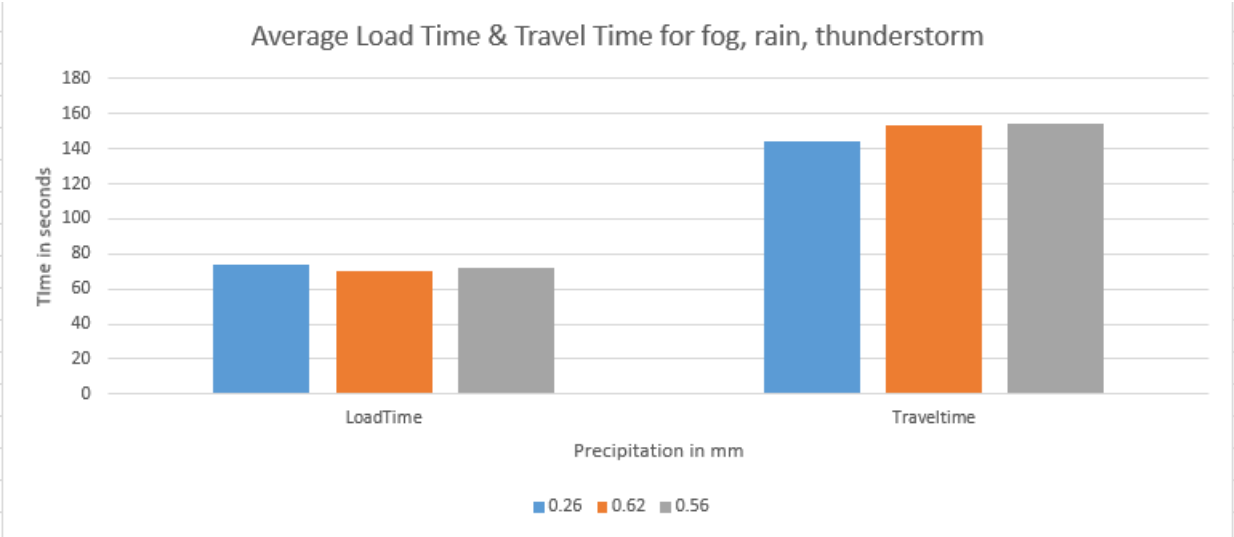
Graph 10: Effect of rain at different level of precipitation .



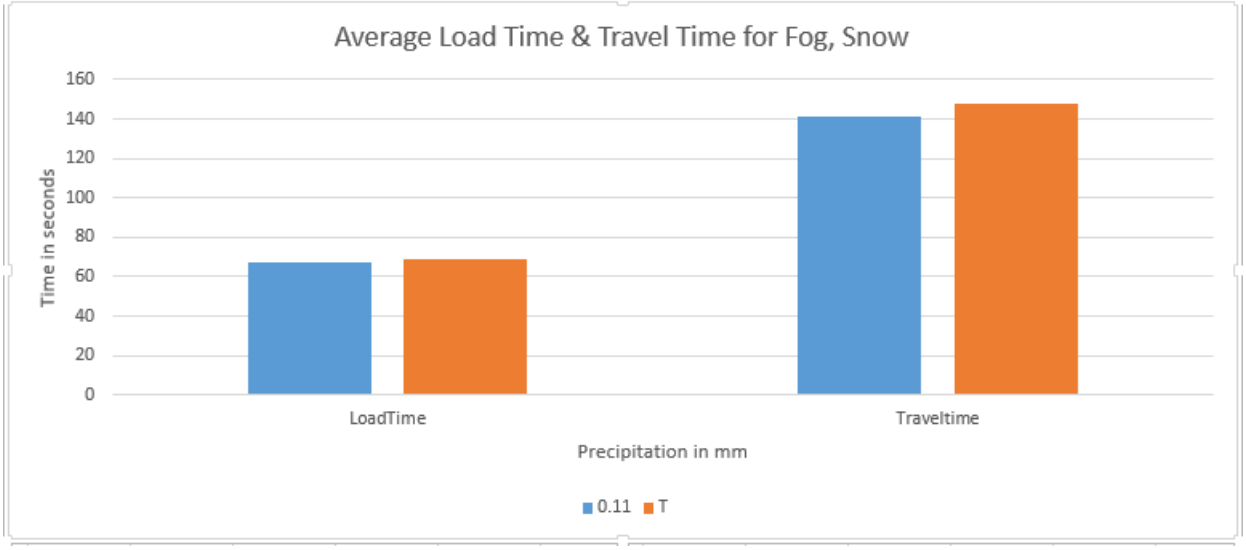
Graph 11: Effect of rain and thunder storm at different level of precipitation.



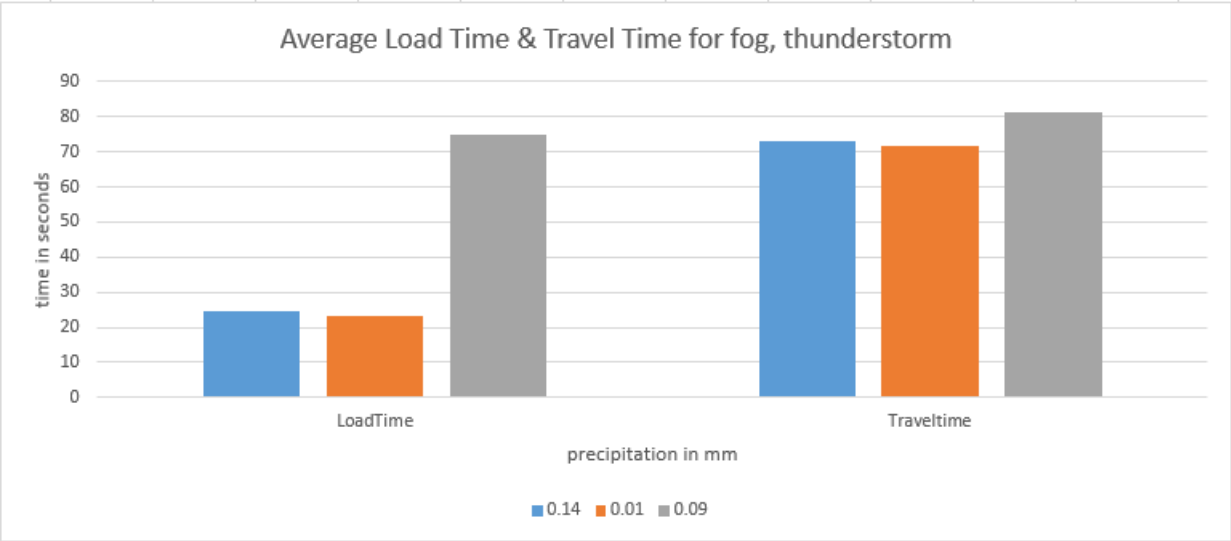
Graph 12: Effect of fog at different level of precipitation.



Graph 13: Effect of fog,rain and thunder storm at different level of precipitation.



Graph 14: Effect of snow and fog at different level of precipitation.



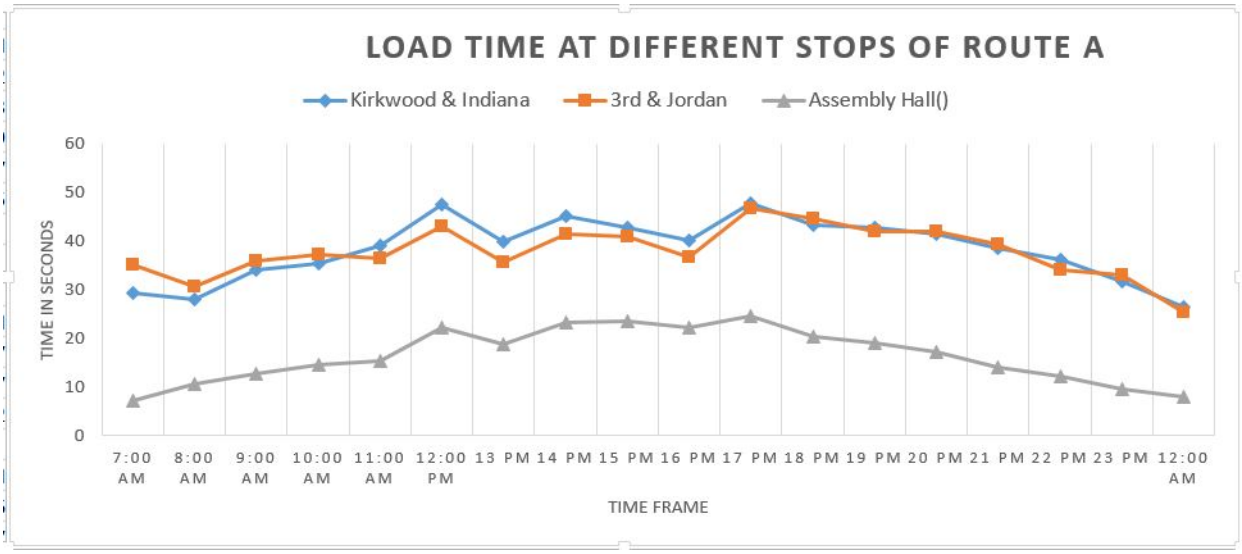
Graph 15: Effect of fog and thunderstorm at different level of precipitation.

Travel time show in all the graphs is average travel time between two stops .

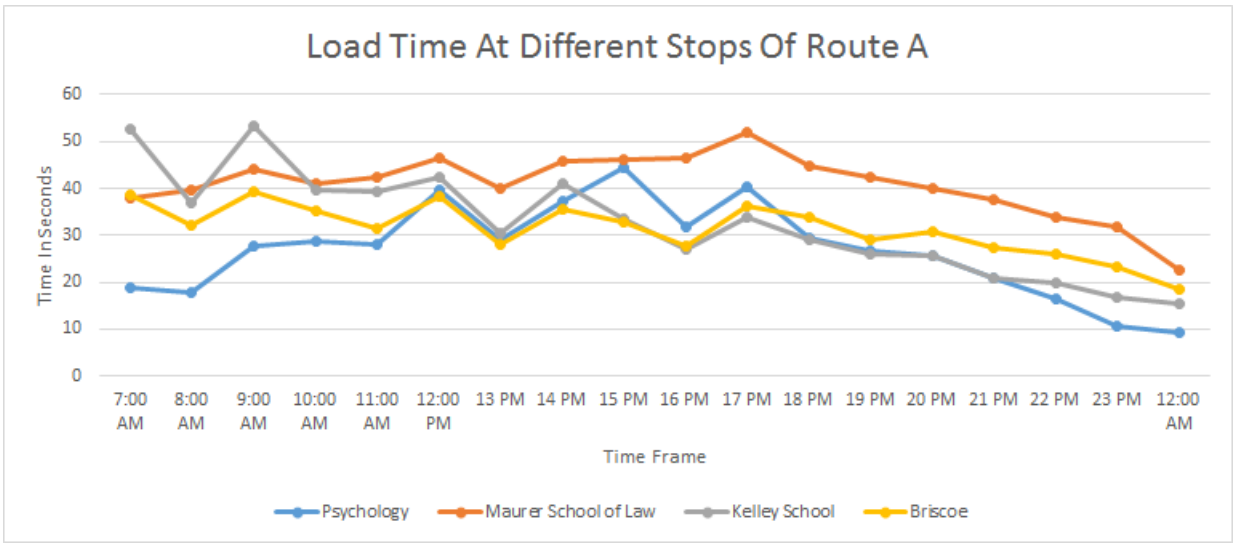
5.3 Load Time at various stops

We have tried to show how load time varies on different stops at the different time of the day.This will give us an idea about the time at which a stop has been used maximum or least used .

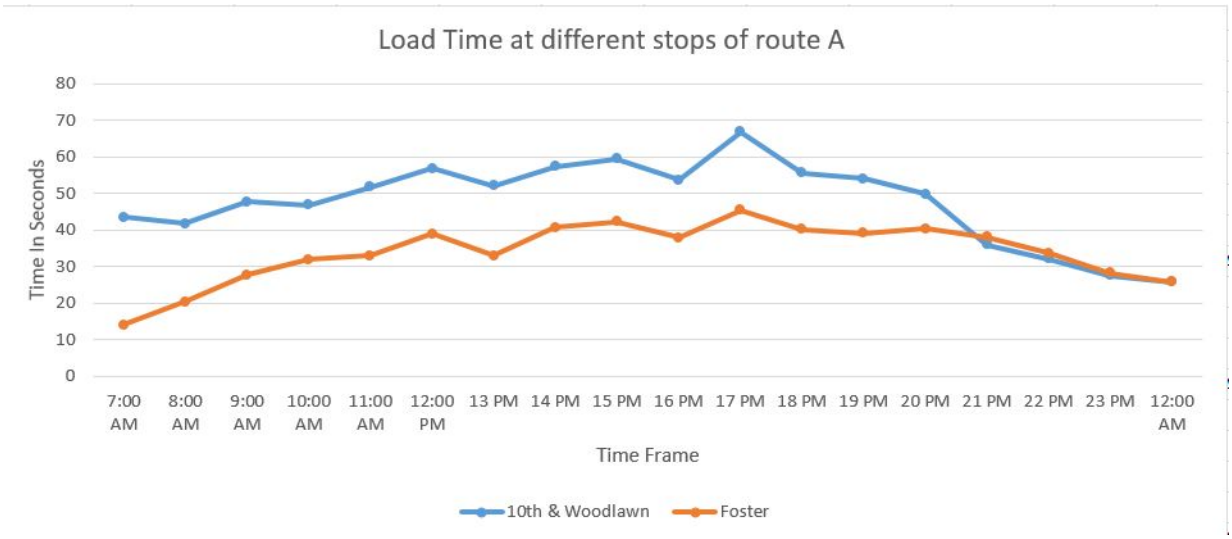
Load Time at various stops of route A



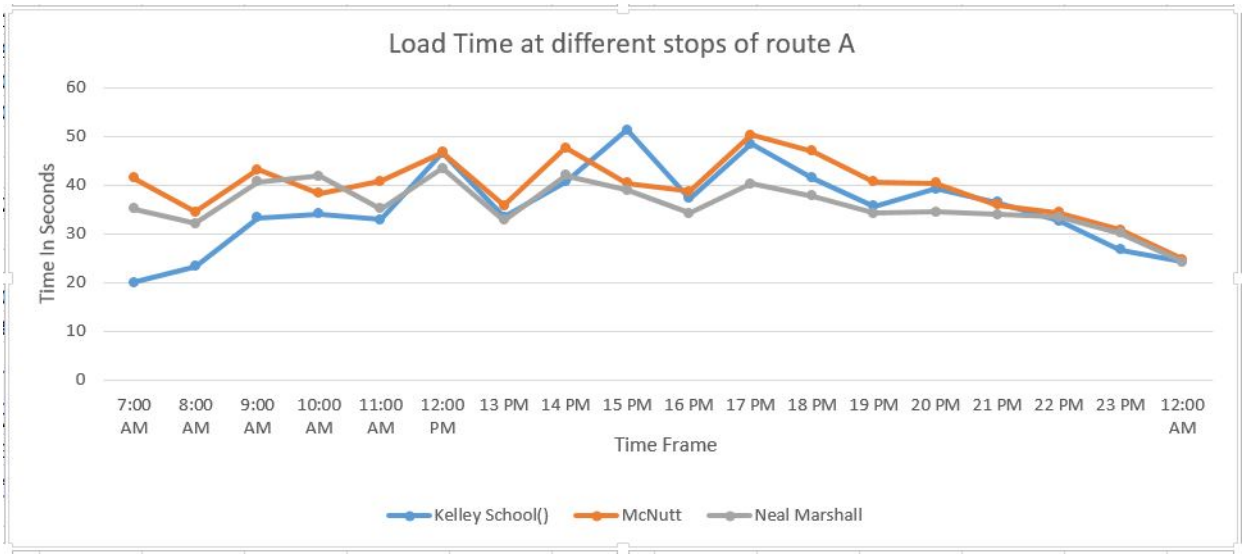
Graph 16: Load Time at various stops of root A.



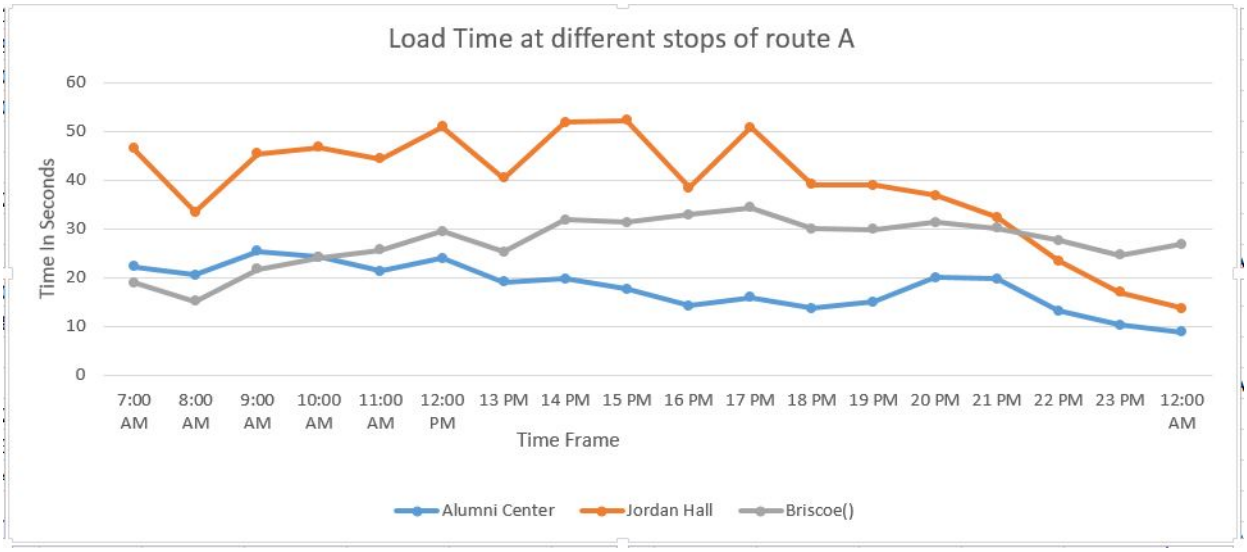
Graph 17: Load Time at various stops of root A.



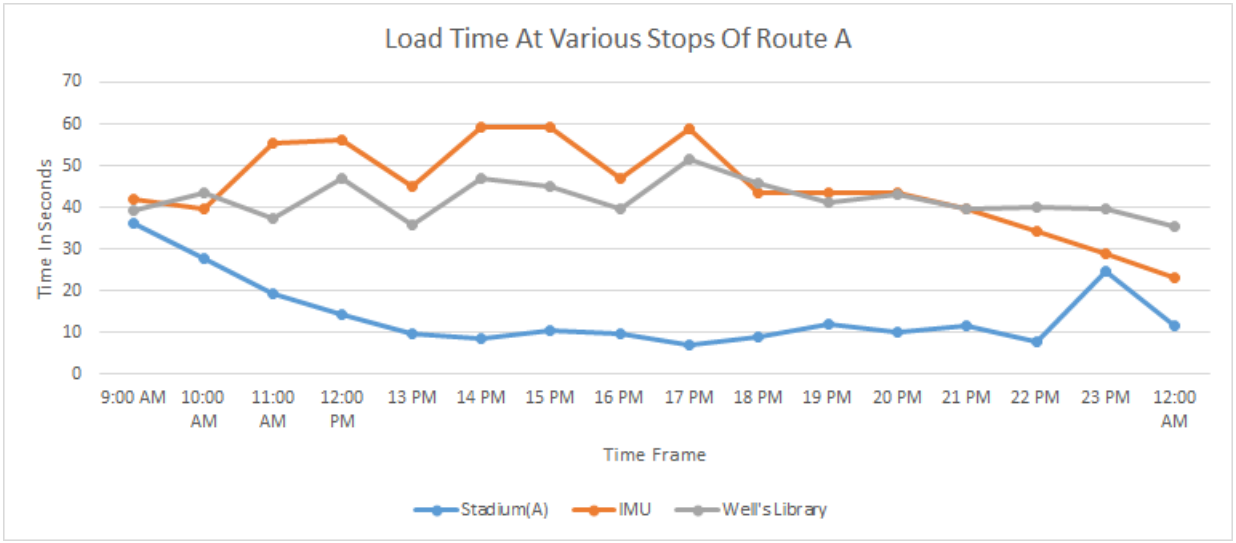
Graph 18: Load Time at various stops of root A.



Graph 19: Load Time at various stops of root A.



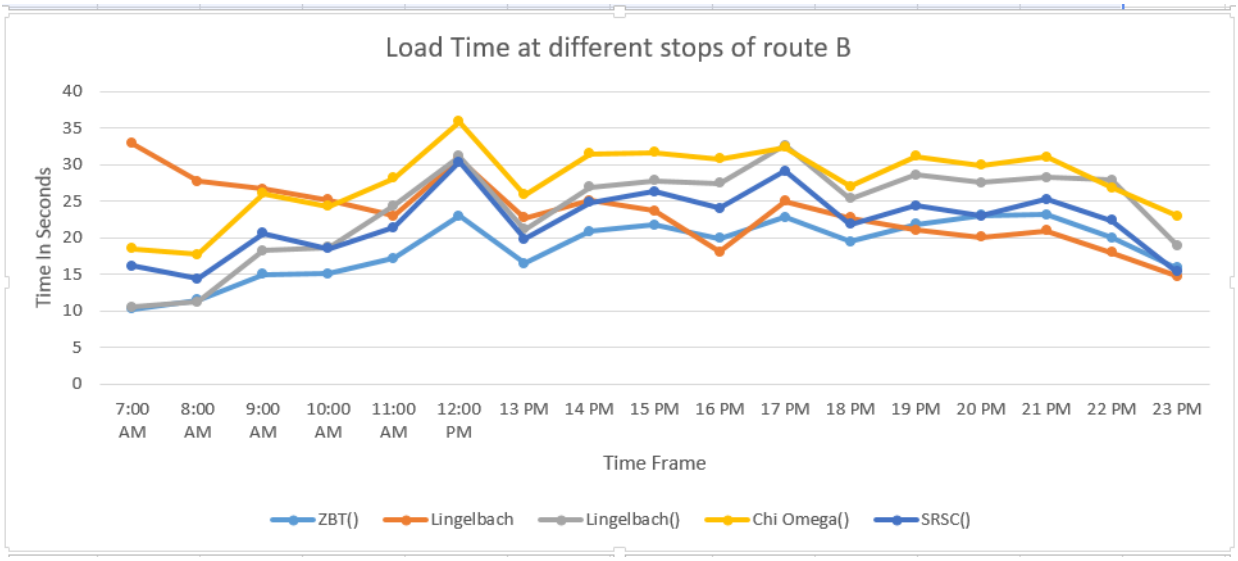
Graph 20: Load Time at various stops of root A.



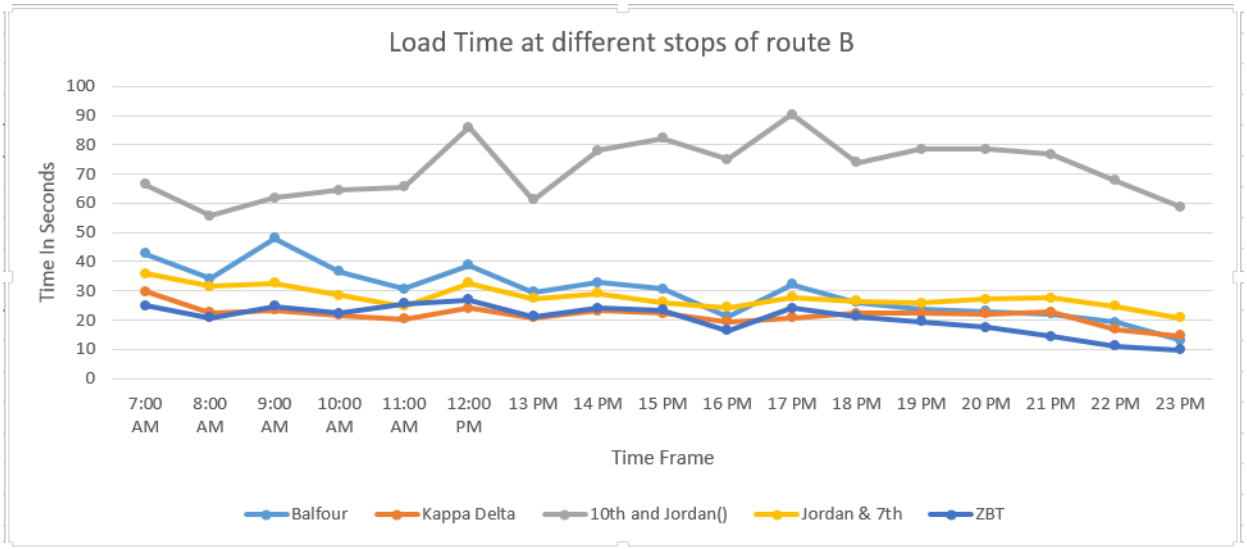
Graph 21: Load Time at various stops of root A.

It is evident from above graphs that for route A, busiest time for most of the stop is from 10:00 AM to 5:00 PM.

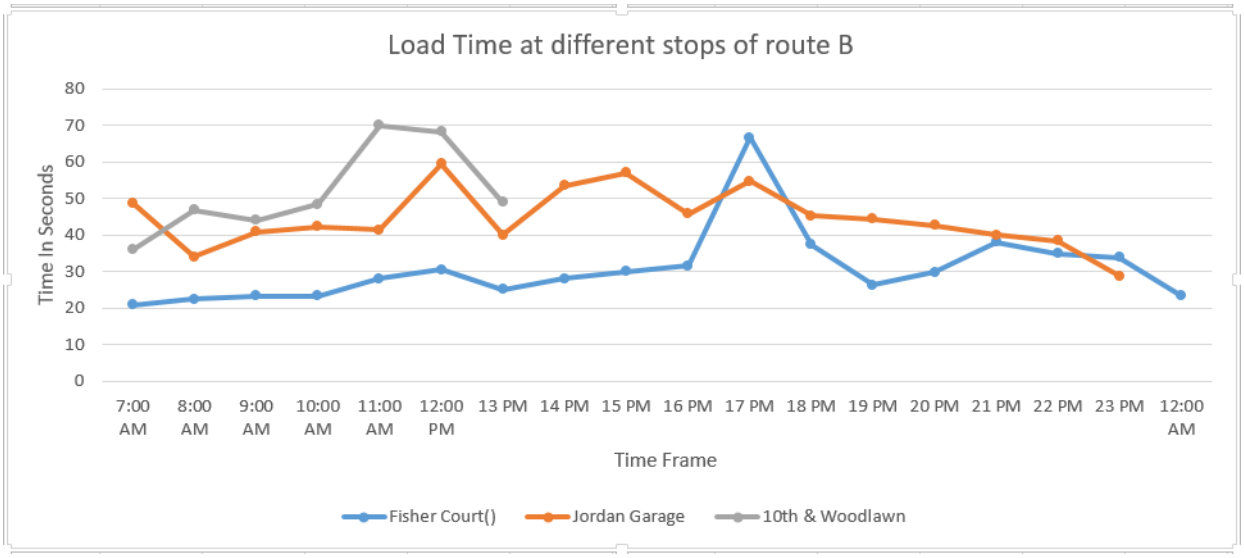
Load Time at various stops of route B



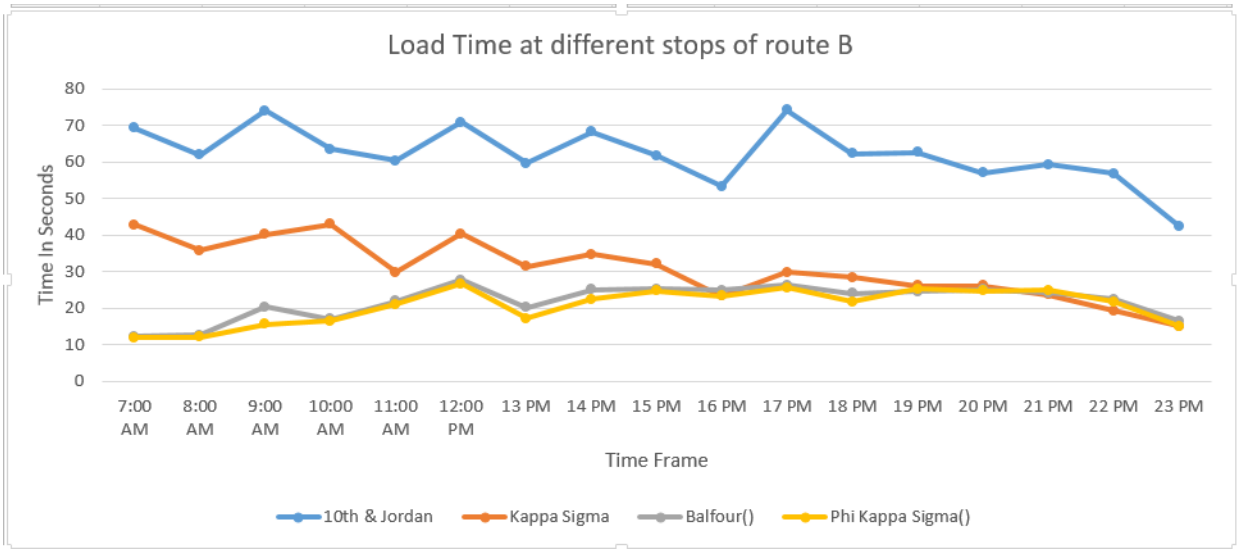
Graph 22: Load Time at various stops of root B.



Graph 23: Load Time at various stops of root B.



Graph 24: Load Time at various stops of root B.

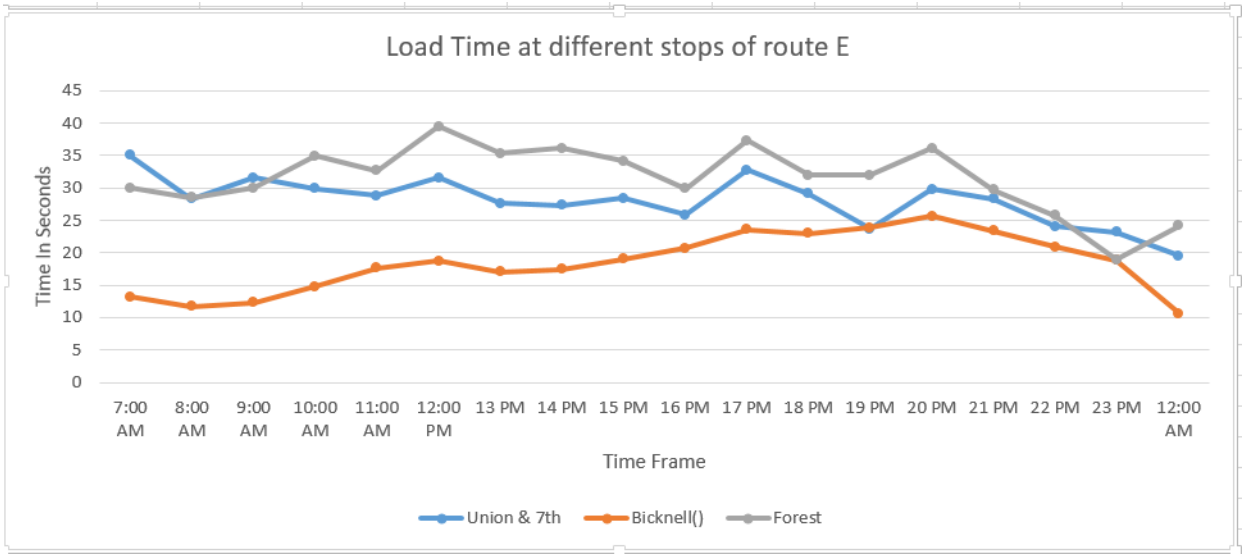


Graph 25: Load Time at various stops of root B.

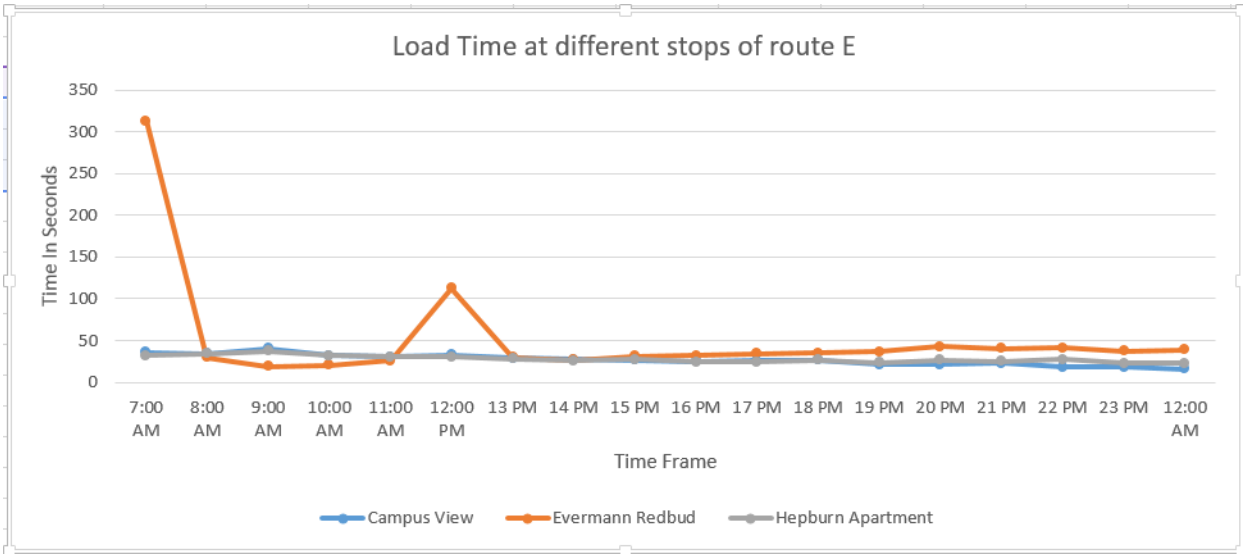
It is evident from above graphs that for route B, busiest stop is 10th and jordan. And there is a sudden surge in load time at the fisher court at around 5:00 PM .While Phi Kapa Sigma ,Lingebach and ZBT are least used stop

between 7:00 AM to 8:00 AM

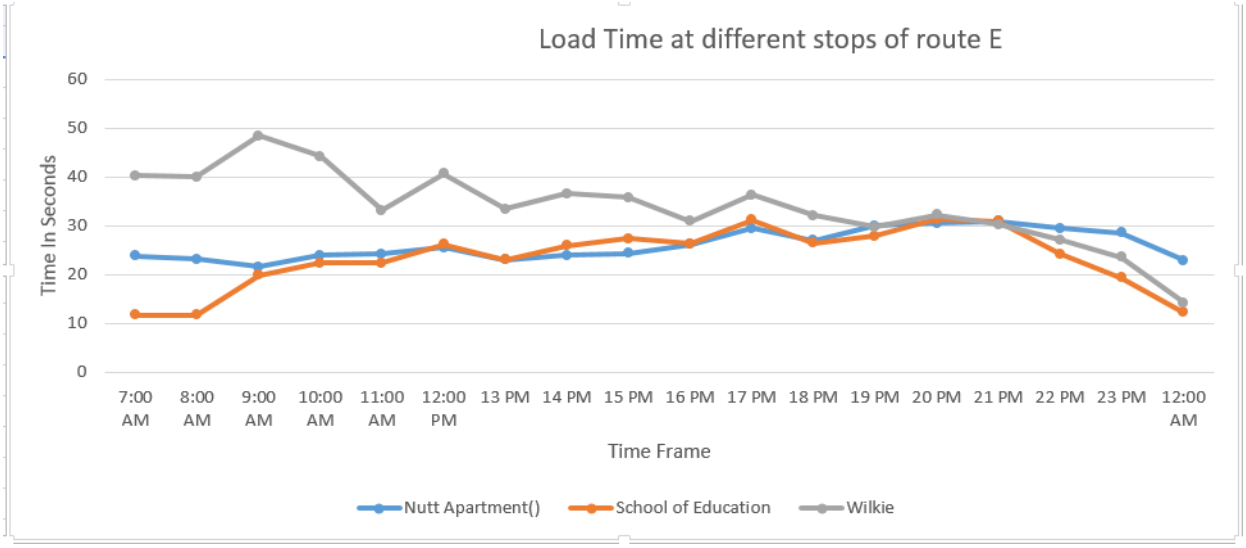
Load Time at various stops of route E



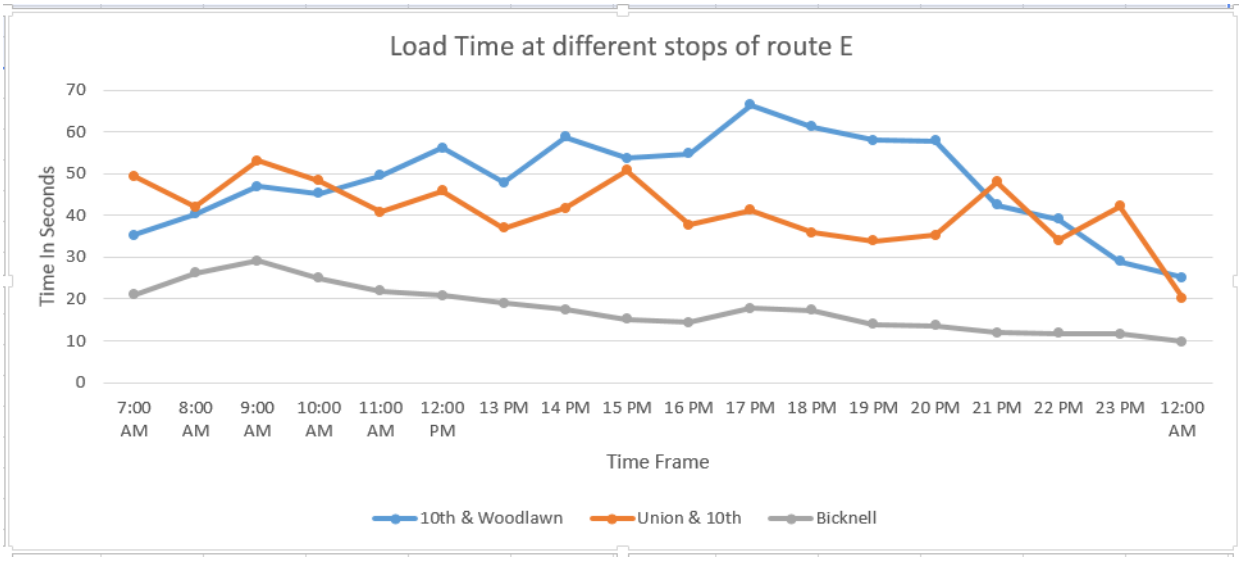
Graph 26: Load Time at various stops of root E.



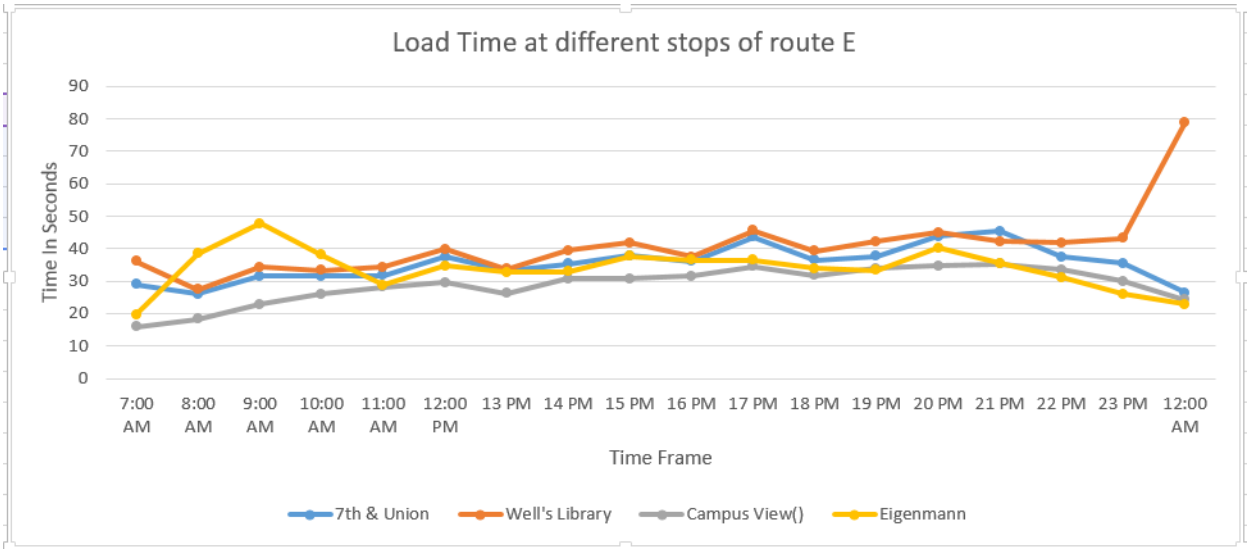
Graph 27: Load Time at various stops of root E.



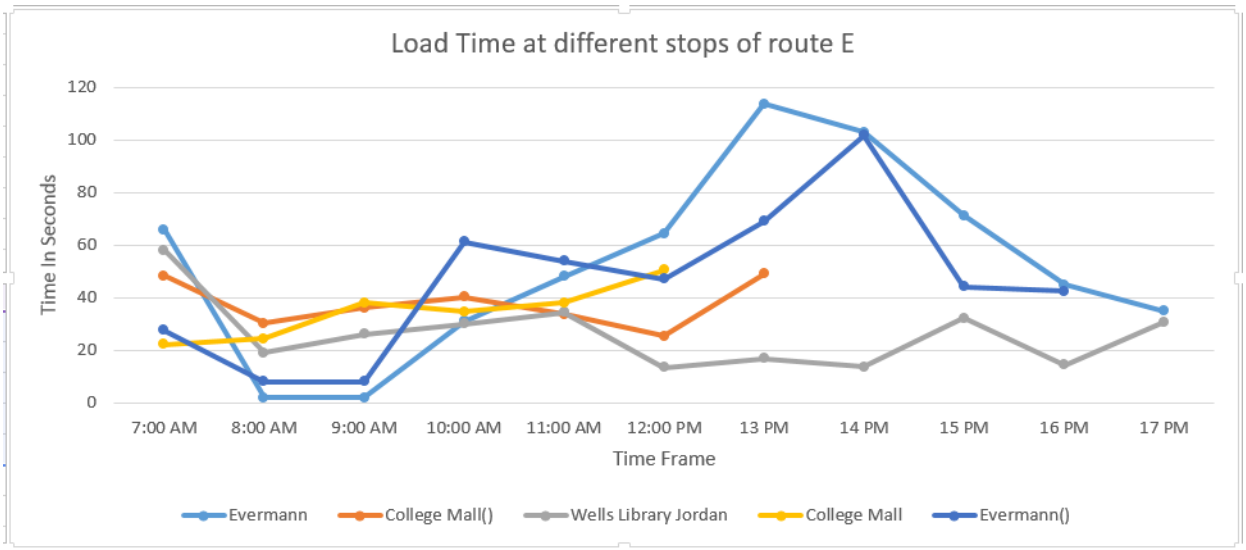
Graph 28: Load Time at various stops of root E.



Graph 29: Load Time at various stops of root E.



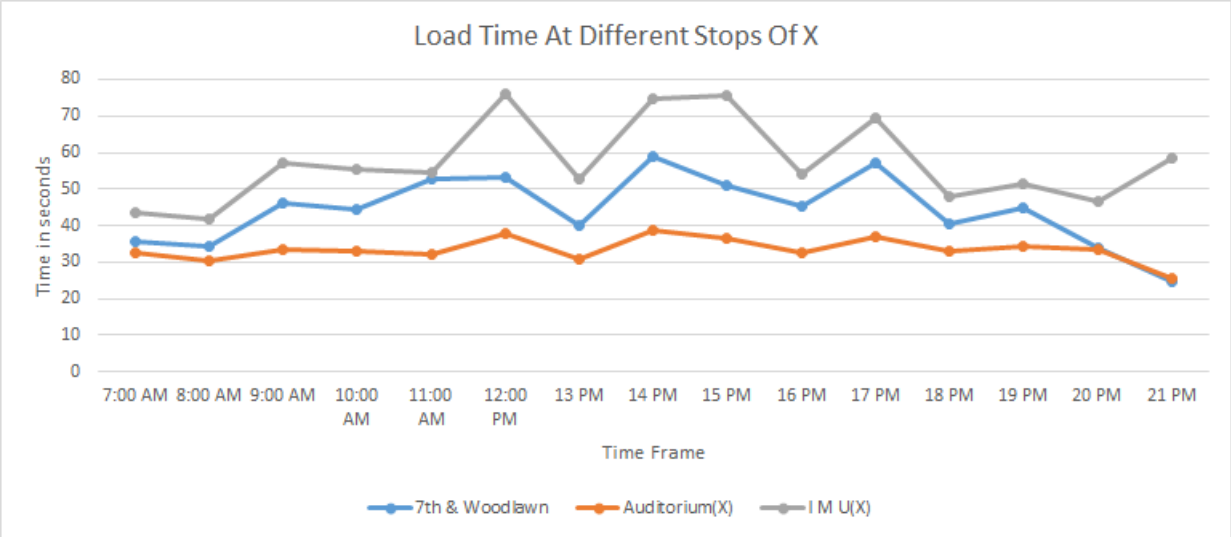
Graph 30: Load Time at various stops of root E.



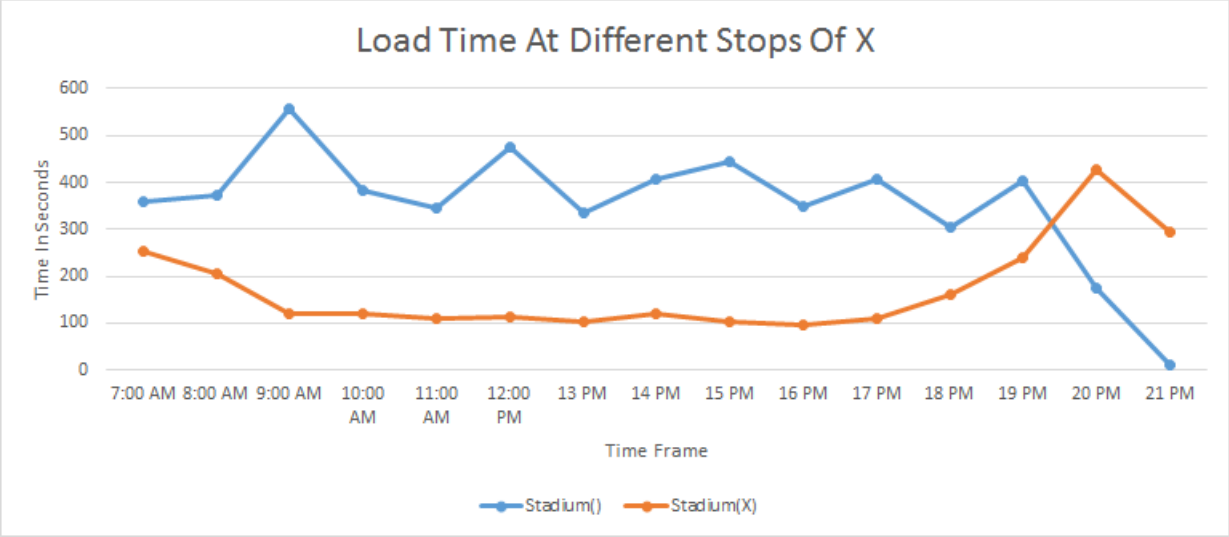
Graph 31: Load Time at various stops of root E.

Load Time at Evenmann is lowest between 8:00 AM to 9:00 AM.

Load Time at various stops of route X



Graph 32: Load Time at various stops of root X.



Graph 33: Load Time at various stops of root X.

6 Proposed Changes

Based on our results we have following proposal .

1 It is evident from the graph of variance(Graph 4) that highest deviation from schedule arrival time is for route X .This can be either due to travel time or load time . If we look at the load time graph(32 and 33) of the stops belonging to route X we find that, load time is very high at stop stadium() and stadium(X). On an average it is 7 to 8 minutes .Also other stops like IMU(X) and 7th and woodlan experience high load time during most of the time of the day. So we propose to either , increase the frequency of buses on route X or decrease load time at stadium(X).

2 We have also found that load time, at any stop, tends to increase in rain, snow and blowing snow weather events .As load time is directly proposal to number of people boarding a bus,this implies that when it rains or snows more people tends to travel by bus . So we propose to increase the frequency of buses when these weather events takes place .

3 For route A ,as evident from load time at different stops graph , Assembly hall is least utilize during morning hour 7:00 AM to 8:00 AM. So we propose that this stop can be dropped from the stop list for morning schedule .

4 As it is evident from Graph 21 that stop Stadium(A) is least utilized after 1:00 PM . We propose that after 1:00 PM route A should start from Alumni Center.This will save fuel and will reduce operational cost .

5 As it is evident from the graph of route E (Graph 31) that Evermann is least used during 8:00 AM to 9:00 AM . So we proposed to remove this stop ,temporarily, from route E.

7 Concluding remarks

We have found that certain weather conditions affect passenger count more than other weather conditions.We have observed, from our result (effect of weather on load time), that haze, rain, snow and blowing snow increases load time. While travel time is highly affected by fog, haze, fog and haze, haze and blowing snow.We have also tried to establish effect of precipitation on

load time and travel time, but there seems to be no clear connection between these two(Graph 10 To 15) .
We have also found that variance in arrival time is highest for route X.This can be attributed to high load time at Stadium(X). For route B, the busiest stop is 10th and jordan. And there is a sudden surge in load time at the fisher court at around 5:00 PM.While Phi Kapa Sigma ,Lingebach and ZBT are least used stop between 7:00 AM to 8:00 AM(graph 22,23 and 25).For route A, the busiest time for most of the stop is from 10:00 AM to 5:00 PM.

8 References

8.1 PEOPLE

1. Professor Mehmet Dalkilic constantly guided us through emails and by constantly checking our progress during the class.

2. Logan (IU Bus Service Employee) visited us during the class and guided us with the data by clarifying various attributes of the data set.

3. Rebekah Ruth Cuevas, a close friend who by coincidence is also an IU Bus Service Employee; helped us occasionally in understanding the abbreviated data values.

8.2 BOOKS

1. Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach, Vipin Kumar.

2. SCJP Sun certified Programming for Java 6.

3. https://docs.mongodb.org/manual/reference/operator/aggregation/#aggregation-pipeline-operator-reference .

4. Beginning Python From Novice to Professional Second Edition by Magnus Lie .