Indiana University Bloomington

CSCI B 565

Data Mining

IU Bus Route Optimization - Group Name : Problem Solver

Author:
Pratish Merchant
Sagar Bhandare
Mrunal Pagnis

Supervisor:
Dr. Dalkilic

December 18, 2015

Abstract

IU bus route optimization goal is to plan an efficient schedule of a fleet of school buses that must pick up students from various bus stops and drop them to their desired locations in as minimum time as possible. In this project, we are considering the IU bus data at Indiana University Bloomington. Computational experiments are performed using real data. Results lead to increased bus utilization and reduction in transportation times with on-time delivery to various stops. The proposed decision-aid tool has shown its usefulness for actual decision-making: it outperforms current routing by reducing the total time taken by a bus on an average to travel from one stop to other.

Also, this project is to investigates the possibilities of using mathematical optimization techniques when planning school transports. It includes a mathematical formulation of the problem, proposes methods to solve the problem and by testing and evaluation it concludes on the pros and cons of using the developed techniques for a practical planning case.

Contents

1 Introduction

Indiana University (IU) is a large University that consists of many departments, which are distributed in a equally large campus that are spread across a geographically large region.Fortunately ,IU understands the difficulties with transportation faced by students and address this by providing a dedicated transportation system called the Indiana University Campus Bus.

Indiana University Bus System has a well defined time based schedule , which provides frequent pickups and drops from Monday to Friday , with relaxed services on weekends.This service tries to cover all the major destinations on campus as well as some far-off from campus like Stadium.

In recent times , IU has been a popular choice amongst students , both domestic as well as international , and hence the number of students visiting the campus each year has been on rise . This trend has been astutely observed by the IU Transit System and have decided to improve this system by observing trends and using data mining to effectively come up with a better solution which can help cope up with the increasing demand of students that are visiting in a optimal way.

For the purpose of this project we have used the following tools :

1. Raw Data : Micrsoft Access
2. Databases Used : Mongodb , MySql
3. Computation : Spring, Java , R , Weka

2 Problem Description

2.1 Presentation of case study

IU Campus Bus has currently total of 27 buses, of which ,approximately 18 buses are running at peak time during the day . These buses run 4 different routes categorized as routes : A, B , E and X. The minimum interval between service of any two buses is of 15 minute interval of service during the day. This interval can vary greatly upto about 30 minute intervals after 8.00 pm during weekdays.A bus can never pass another bus ahead of it , which adds to the delay of the buses. Due to the limitation of size of buses , not all stops can be considered as turnaround or relief points , but a certain predetermined stops are identified to act as one, which would help to reduce the overall delay.

The overall goal of IU Campus Bus is to use concepts of Data Mining to extract data that will help to accommodate the new demand of increasing students. Thus the above problem when broken down to smaller , simpler problems would amount to :

1. Calculating the variance between scheduled time of departure and the actual time that the bus departs and consequently , increase the travel time that a bus otherwise spends in delay.

2. Finding the factors that are directly or indirectly influencing the delay of the buses such as passenger count, weather etc.

3. Determining the average time between any two stops during any particular time of a day for all the routes.

Besides these primary objectives insight , we can additionally aim for the following sub-objectives :

1. Since IU Campus Bus is constrained due to lack of drivers , we have to efficiently optimize the count of passengers by maximizing it against a minimum frequency of service.

2. Aim to explore the possibility of dynamic scheduling by either route optimization or schedule optimization , or both if permitted by time constrains.

3 Data Description and Processing

3.1 Existing model

The first step in analysis is to procure as much data as possible . In this regards , the data the supplied in its very raw nature . The data was dispersed across various spreadsheets , database and text files . This meant cleaning the data and creating a unified database which houses data at one place .

The main files used for data extraction and analysis were :

1. Doublemap+Data+Succesful+importV2.accdb
2. SP15+Servicebook+Draft
3. Ridership+Spring+2015.mdb
4. RidershipSchedule2015.xlsx
5. GPS data

The next step was to analyze and extract the available data . Lets consider each file and its extraction process:

1. **Doublemap+Data+Succesful+importV2.accdb**
   (a) This Access Database consists of total of 6 tables , namely :
      i. Intervaldata2014-2015 : Table which provides information on time taken between two stops for a particular bus on a specific route at a particular time and data . This data is largely used to track the movement of buses during the semester.
      ii. Route ID : This table maps the Route ID for the corresponding routes.
      iii. Schedule Data : This table summarizes the expected timings which are required to be followed by drivers for a particular route subtype. It basically maps to the "SP15+ServicebookDraft" data.
      iv. Stop ID : This table contains mapping between stop names and their corresponding ID's.

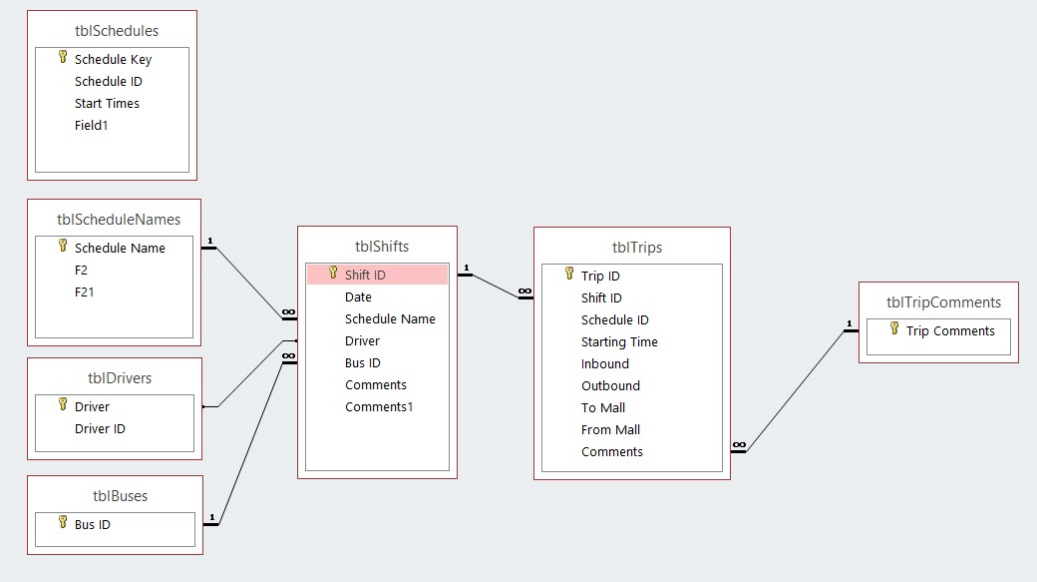

Figure 1: Relational View of Database

v. Weather Data : Weather data from taken from NOAA which gives temperature ranges and whether there was precipitation for a given day.

vi. Work Record : This table contains work information related to a driver's shift namely , the clock in/ clock out time , shift type and date time for a particular bus.

(b) The next step was to map data of "to" field of "Intervaldata2014-2015" to corresponding "stop" from "Stop ID" table.

(c) Also convert numeric value "354" to its corresponding route 'A' using data from "Route ID" table.

2. **SP15+Servicebook+Draft**
Contains scheduled data for all routes , listing all the major stops in the journey and the timings to depart from the stop for an entire day. Different sub routes are stored in different sheets . The scheduled data is for all days of the week.

3. **Ridership+Spring+2015.mdb**

(a) This contains 6 important tables , which are relationally dependent on each other :

i. tblSchedules
ii. tblScheduleNames
iii. tblDrivers
iv. tblBuses
v. tblShifts
vi. tblTrips

(b) Here tblShifts is the center table relating to trips , schedules and bus id.

(c) From this table , the two most important tables are tblTrips and tblShifts.

(d) Table tblShifts helps provide mappping between Bus Id and Schedule which is to be followed. Here have to clean the datetime field to make it usable.

(e) Table tblTrips is used to find the Inbound and Outbound flow of the passengers. The field lacks Bus ID column , which has to be added by matching it with Shift ID table using VLookup, along with Date field to store it in a relational database.

4. **RidershipSchedule2015.xlsx**
Contains passenger count information for specific days , during specific time intervals. Howvever , data related to many dates is missing and the accurate nature of data itself isnt reliable . Hence , this table was not used for future calculations.

5. **GPS data**
This contains vast amount of data related to the position of the bus . This has been used as reference guide for clearly visualizing the route buses take , but most data is not pertinent to Bloomington and remaining data is of redundant nature to be barely usable . Hence , most of the data has not been considered for calculations.

Next step after cleaning the data is transformation into csv format. We try to export all the tables in CSV format , which can be easily loaded into MySql database. We apply transforms like changing the format of date time suitable to be imported in MySql database.

Once we have the CSV files , we can load the extracted values into a relation database like MySql . While adding the Date Time values in MySql ,we apply certain cleaning again to keep data in a consistant and clean format :

1. We added another column which coverts the datetime to milliseconds . This way the handling of time values becomes relatively easy .
2. We also apply certain cleaning while loading values in MySql to handle some empty /null values encountered.
3. Remove some unwanted columns which are not useful for calculations.

Now ,the analysis and ETL phase is complete and we are ready for a creating a relational model base on the data that is created .The overall logic for mapping goes as follows :

1. Map Interval Data to Schedule data , based on the bus id , date and route of the bus.
2. However , there is no direct dependency between these two tables , as interval data only consists of the Route Id and Scheduled Data consists of Sub Route ID .
3. Therefore , we make use of another table called Shifts , which maps a Bus ID to a Sub Route ID for a particular date and together the combination of these tables , gives us exact interval data to scheduled data mapping.

3.2 Modeling the Data

The relation model described in previous section is realized through java classes in object oriented way. This makes it possible to process the data and generate new models for analysis. The modelling of current data involves following database tables :

1. clean_interval_data
2. clean_schedule_data
3. weather
4. passenger_counts

3.2.1 Defining the models

Following logical models of data are realized through Java in the project:

Schedule Model

Model: Schedule
Description: Represents schedule of any bus on a particular route.

Attributes:

1. Scdule_Id
2. route_id
3. Load time of bus
4. 1^{st} major stop reach time
5. 2^{nd} major stop reach time
6. nth major stop reach time

Trip Model

Model: Trip
Description: Represents data of single trip of bus. The trip can be defined as path of bus from loading point back to the same point.

Attributes:

1. route_id
2. Passenger Count
3. Load time of bus
4. 1^{st} major stop reach time
5. 2^{nd} major stop reach time
6. nth major stop reach time

Trip-Delay Model

Model: Trip-Delay
Description: Represents data of delay in single trip of bus.

Attributes:

1. trip_id
2. Load time delay of bus
3. Delay in 1^{st} major stop reach time
4. Delay in 2^{nd} major stop reach time
5. Delay in n^{th} major stop reach time

Daily Trip Summary Model

Model: Daily Trip Summary
Description: Represents summary of all the trips of a particular bus on a given route and given shift for given day. Also holds data about weather conditions on that day.

Attributes:

1. Date
2. Shift Id
3. Bus Id
4. List of Trip
5. List of Trip-Delays
6. Weather Condition

Weather Model

Model: Weather
Description: Represents data of weather condition for given day.

Quantifier : Good, Average, Bad, Extreme Attributes:

1. Date
2. Wind speed
3. Temperature
4. Visibility
5. Rain or Snow

3.2.2 Loading the data into models:

After defining the model the next step is to load them with the data for processing. Loading of data into models is order sensitive and should be done in fixed order. Models are loaded in the following order:

1. Schedule: This model is first created and populated by using table clean_schedule_data. The data is stored in the form of list of trips per route.
2. Trips: For each shift in the schedule clean_interval_data table is queried to find and load all actual trips for given bus_id. These trips are stored in memory for next model.
3. Daily Trip Summary : This model uses both of the above models.Trips are separated on the basis of date, route and bus_id. List of Trip-Delays is populated by comparing actual trip with the schedule model loaded already. Weather data is associated based on date.

3.3 Processing the models:

The models so generated are used to analyse the data and calculate the delay at all major stops for all routes and shifts. Following variances are calculated using the models:

1. Variance at each major stop: The variance of any bus in reaching a major stop on the route.
2. Variance of particular bus in any trip.
3. Variance observed on a particular day on particular route.

The results are presented in 'Analysis and Discussion' section in tabular format.

3.3.1 Generating training data for classifier

The raw data from the 'Daily trip summary' model is exported as a CSV file to be used as training data for delay prediction classifier based on Random Forest method in R . The output is a file conatining delays of majot stop for Route A which contains predicted data based on different values of weather , passenger count and time of day.The same data is used by R-scripts written to explore the factors responsible for delay in the route of bus.

4 Analysis and Discussion

The data generated by Java code is used by R-Scripts for further analysis. At the end of analysis we got very promising correlations for the delay of bus which I believe can help us to maximize the in transit time of the bus. We have proposed the suggestions based on these results. But, before jumping to the results lets look at how the data looks like and operations performed on data.

1. First look of generated data
   We have plotted delay at every major stop of few routes below to give an idea of how data is spread. Let's consider route A first. Route A has four major stops Stadium, Wells, Jordan and IMU, then it returns to Stadium again. We have plotted delay at each point for all trips below for comparison. Let's have a look:

   Each stop has delay up to 5-7 minutes for each delayed trip. The outliers are clearly visible. We checked manually, the outliers marks the days where either bus was broken down or due to technical errors in data entry. The points marked in green are the ones we need to concentrate on. To do that lets separate outliers from the data and refine the data. After applying cutoff here is new distribution of data.

   Now, in figure 2 we can see the delays uniformly distributed above the mean. The bus is early or late by 10 minutes at max in most of the delayed trips.

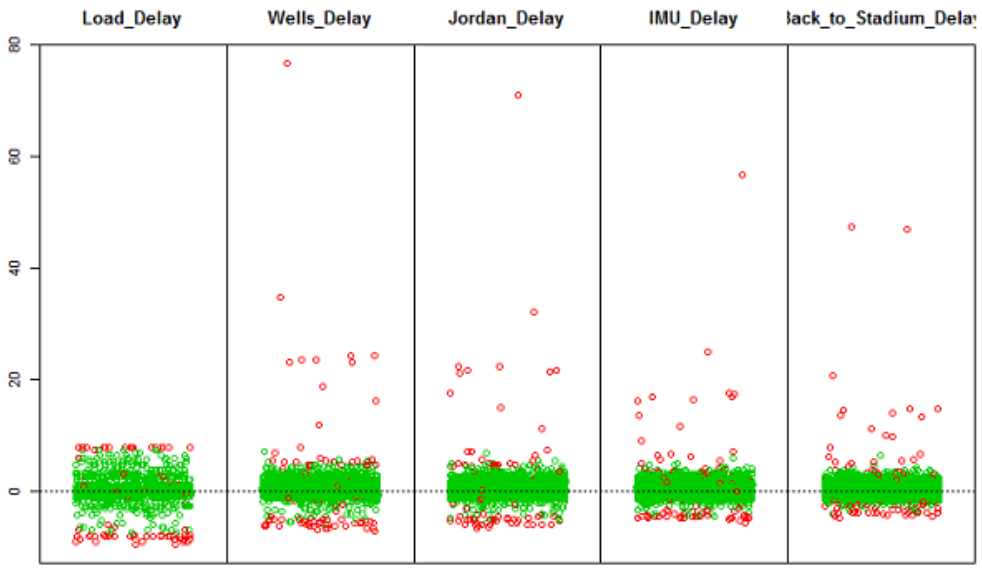

Figure 2: First look of data for Route A

Similar process is applied to route B as well. Figure(3) and Figure(4) represent

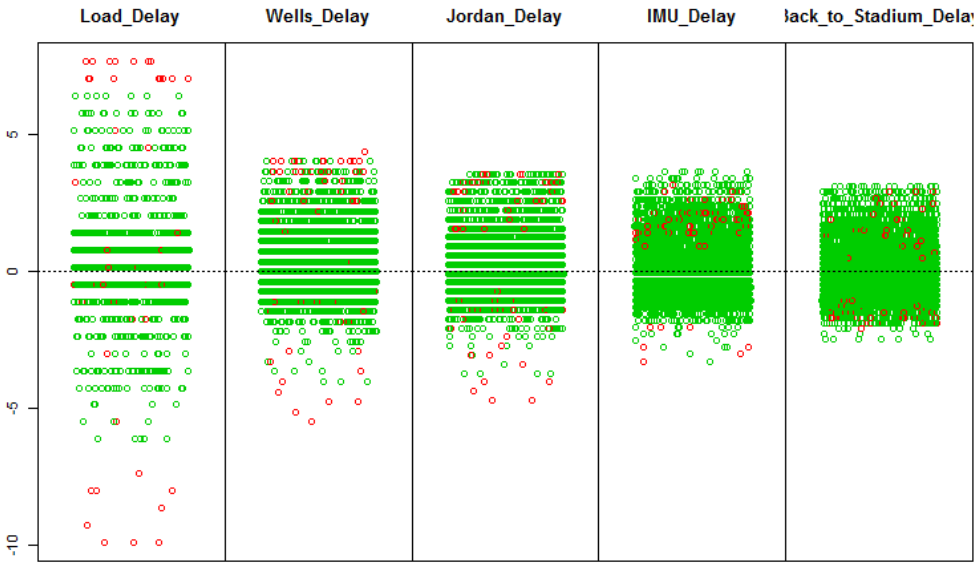

Figure 3: Route A delay at major stops

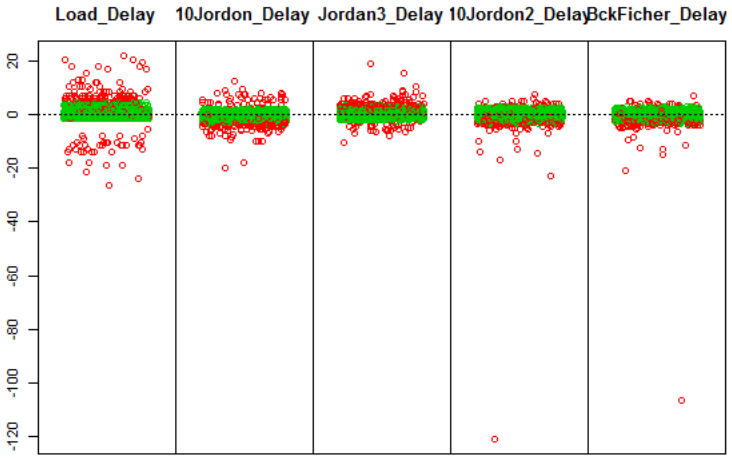

Figure 4: First look of data for Route B

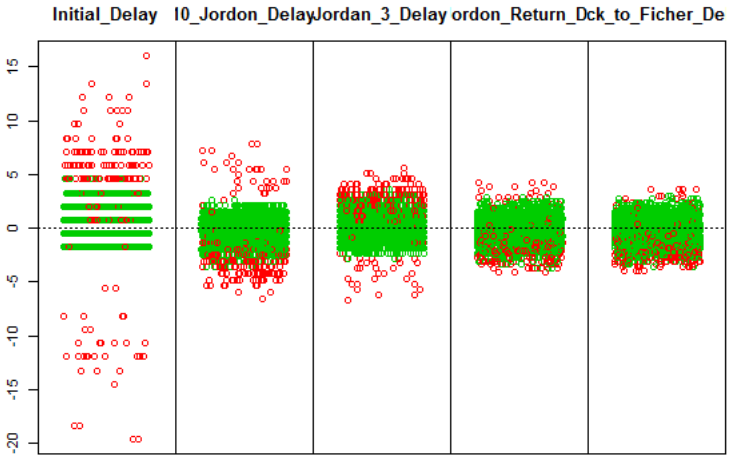

Figure 5: Route B delay at major stops

2. Statistical Analysis of delays In order to understand the reasons for delays first we need to know how much is the delay in various perspective. We calculated following variances to understand data better:

   (a) Mean and Variance at each major stop: The mean variance of any bus in reaching a major stop on the route

Table 1: Mean delay at each major stop on route A

Initial_Delay	Wells_Delay	Jordan_Delay	IMU_Delay	Back_to_Stadium_Delay
1.784	2.024	3.247	-0.610	-1.552

Table 2: Variance at each major stops on Route A

Initial_Delay	Wells_Delay	Jordan_Delay	IMU_Delay	Back_to_Stadium_Delay
3.248	8.254	9.861	17.302	26.283

Table 3: Mean Delay at each major stops on Route B

Initial_Delay	Jordon10_Delay	Jordan3_Delay	Jordon10_Return_Delay	Back_to_Fischer_Delay
1.454	-1.779	0.843	-1.952	0.340

Table 4: Variance at each major stops on Route B

Initial_Delay	Jordon10_Delay	Jordan3_Delay	Jordon10_Return_Delay	Back_to_Fischer_Delay
1.886295	4.699887	5.705647	11.1763	14.41642

   (b) Variance and mean of each bus in all trips for route A

Table 5: Mean of each bus at each major stop in all trips for route A

	Bus ID	Initial_Delay	Wells_Delay	Jordan_Delay	IMU_Delay	Back_to_Stadium_Delay
1	640	1.828	1.981	3.136	-0.704	-1.663
2	645	1.760	2.302	3.569	-0.405	-1.126
3	657	2.319	2.809	4.106	-0.117	-1.085
4	647	1.882	2.116	3.343	-0.451	-1.360
5	650	1.731	1.920	3.094	-0.819	-1.866
6	636	1.495	2.229	3.448	-0.571	-1.352
7	642	1.773	1.933	3.256	-0.595	-1.514
8	654	1.780	1.431	2.761	-0.908	-2.138
9	659	2.449	2.469	3.837	0.184	-0.633
10	638	1.722	1.950	3.145	-0.517	-1.419
11	661	1.706	1.471	2.714	-1.185	-2.983
12	648	1.730	2.051	3.258	-0.641	-1.506
13	651	1.782	1.909	3.327	-0.109	-1.455
14	649	1.396	1	2.396	-1.333	-2.396
15	639	1.250	1.550	3.250	0.050	-0.825
16	652	1.717	2.043	3.271	-0.598	-1.406
17	656	1.438	1.420	2.446	-1.491	-2.696
18	653	1.775	1.675	2.737	-1.250	-2.500
19	662	1.491	1.264	2.651	-1.208	-2.406
20	655	1.987	2.697	3.829	-0.171	-0.961
21	660	1.844	1.797	3.094	-0.703	-1.641
22	643	2.057	2.343	3.686	-0.114	-1.343
23	658	2.051	2.356	3.220	-0.966	-2.085
24	637	1.459	1.659	2.765	-1.024	-2.388
25	641	3.071	3.271	4.400	0.371	-0.557
26	644	2.625	3.938	5.375	1.125	0.625

Table 6: Variance of each bus at each major stop in all trips for route A

	BUS ID	Initial_Delay	Wells_Delay	Jordan_Delay	IMU_Delay	Back_to_Stadium_Delay
1	640	3.513	8.016	9.119	16.698	26.425
2	645	1.796	7.256	8.845	14.441	23.018
3	657	3.230	10.243	11.838	15.932	23.821
4	647	3.140	8.259	10.088	17.603	26.790
5	650	3.497	8.139	9.844	17.560	26.158
6	636	1.849	8.409	9.269	15.959	23.250
7	642	1.667	6.825	9.024	17.704	27.409
8	654	1.007	4.896	6.461	17.732	29.046
9	659	2.586	7.046	9.889	21.236	35.237
10	638	5.755	9.509	11.132	18.234	26.702
11	661	2.498	6.980	9.341	17.406	23.932
12	648	3.865	10.037	11.323	18.133	26.986
13	651	1.914	8.714	11.706	23.840	35.141
14	649	2.627	7.191	8.968	21.418	34.585
15	639	11.474	15.382	16.038	28.254	37.071
16	652	2.719	8.004	9.457	16.536	25.074
17	656	1.600	6.570	8.466	16.522	26.015
18	653	2.809	6.247	7.538	13.582	20.633
19	662	6.919	10.387	12.382	22.947	31.958
20	655	1.480	7.494	8.704	17.904	25.292
21	660	1.912	4.641	6.689	16.149	22.615
22	643	0.703	4.997	7.457	15.575	25.467
23	658	1.842	6.716	7.692	13.688	27.286
24	637	3.132	8.394	10.182	18.785	25.526
25	641	11.314	14.201	15.635	20.150	26.018
26	644	3.983	11.796	11.983	16.250	19.583

(c) Mean and variance at each major stop based on day of week

Table 7: Variance of delay at each major stop based on day of week

	Day	Initial_Delay	Wells_Delay	Jordan_Delay	IMU_Delay	Back_to_Stadium_Delay
1	Sunday	2.491	7.115	8.604	15.68	24.419
2	Monday	2.992	8.096	9.211	15.26	23.936
3	Wednesday	2.841	8.569	10.489	18.405	27.39
4	Saturday	2.963	6.788	8.503	16.13	26.285
5	Tuesday	3.86	9.815	11.527	19.506	27.885
6	Thursday	3.639	9.33	11.128	19.388	27.354
7	Friday	3.992	8.124	9.552	16.713	26.413

Table 8: Mean delay at each major stop based on day of week

	Day	Initial_Delay	Wells_Delay	Jordan_Delay	IMU_Delay	Back_to_Stadium_Delay
1	Sunday	1.691	1.816	3.023	-0.784	-1.787
2	Monday	1.847	1.964	3.156	-0.825	-1.798
3	Wednesday	1.767	2.087	3.355	-0.679	-1.509
4	Saturday	1.706	1.797	2.991	-0.66	-1.781
5	Tuesday	1.872	2.108	3.326	-0.63	-1.569
6	Thursday	1.847	2.289	3.517	-0.301	-1.238
7	Friday	1.779	2.138	3.399	-0.386	-1.157

3. Inferences from statistical data

   The mean and variance tables and graphs gives a good insight into the distribution of delay. We can derive following inferences from the results

   (a) Few of the bases higher delay than the others
   e.g. Bus IDs 639, 641 have highest variance at each stop for route A. This suggests that either the condition of bus is not up to the mark or there are some operational issues

   (b) The delay is higher on few days of the week
   Few days of week faces higher delay than other. E.g. The variance is very high on Tuesday and Thursday. The class schedules for Kelly school, SOIC department can be checked to verify if they have more classes scheduled on these days.Apart from that we

also know that many restaurants from IMU offer food discounts on Tuesday and Thursday e.g. 'Baha Fresh' in IMU has discount on Tuesday and Thursday. This may lead to some increase in passenger count for those days.

(c) The delay is higher on particular stops on the route than others. E.g. IMU and Jordan10 faces more delay compared to other stops.

(d) Weather don't have much impact on the delay of the bus. We plotted weather condition against the delay at each stop. We didn't get much convincing results for it.

5 Proposed Changes

Based on our analysis , we propose the following changes that can be implemented :

1. We suggest that some buses which have always shown a delay on average : e.g. Buses 639 and 641 on Route A , should undergo maintainance changes as they might be suffering from performance issues , which have not been discovered.

2. We find a general trend in delay of buses , such that some buses are particularly delayed on some specific days. For example , for Route A, we find the trend for delays of buses on Route A to be more frequent on Tuesdays and Thursdays.

3. We find some stops on routes ,for e.g. Returing Back to Stadium ,IMU on Route A and 10th Jordon on Route, have a longer delay time in general . This means that location of the stop or the service frequency of the stop needs to be discussed.

References

[1] Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2. http://CRAN.R-project.org/package=stargazer

[2] Baker, C.B., J.K. Eischeid, T.R. Karl, and H.F. Diaz, 1994: The quality control of long-term climatological data using objective data analysis. Preprints of AMS Ninth Conference on Applied Climatology, Dallas, TX., January 15-20, 1995.

[3] DoubleMap - Indiana University. (n.d.). Retrieved December 18, 2015, from http://iub.doublemap.com/map/

[4] "Campus Bus." Campus Bus. N.p., n.d. Web. 18 Dec. 2015.