The Accuracy of PAT Bus Schedules for Carnegie Mellon University Commuter Stops

Stats 36-303: Team E
4/29/2011
Matt Belenky, Tim Higgins, John Sperger, Chao Wang, Tian Wu

Section 1: Introduction

1.1 Research Question and Motivation

Many students and professors at Carnegie Mellon are extremely reliant on the public transit system to get to work and school, however students frequently complain about the PAT bus system. The most common complaints are late buses, inaccurate schedules, and the frustration that occurs after waiting for a bus only to have multiple buses of the same route arrive at the same time. Waiting wastes time, causes frustration, and in the long run could lead commuters to find a way to travel that doesn't involve public transportation. The aim of this study is to first measure the degree to which these complaints are accurate, and if buses are systematically late to develop a model for predicting expected arrival time.

This study will be built on a strong general literature base in the area of public transportation and it will investigate the accuracy of bus time tables for the Forbes and Morewood intersection which is the most commonly used bus stop for commuters at Carnegie Mellon University. Bus departure times will be observed and compared to posted bus schedules. A number of potential factors that influence bus punctuality will also be measured including the weather, the time of day, and the level of light. Using these factors and the information collected on bus arrival times, a model will be created to predict when a bus will arrive given the scheduled arrival time.

1.2 Citations to Relevant Literature – An Overview

We compiled several relevant pieces of literature in order to investigate Pittsburgh public transit. First we have the *Users' Perceptive Evaluation of Bus Arrival Time Deviations in Stochastic Networks*. The paper has a different goal than we do, but it still provides us with some perspectives to approach the project and define the scope of it, such as how to define "bus being on time", how long are we going to wait and observe the bus etc. The *Bus Punctuality Statistics GB* gave us some brief idea that what factors may affect the accuracy of buses. From *MTA gets low marks for bus service in survey*, we come up with the idea to cluster our observation time into 4 clusters or even three in the later stage. These similarities in observation gave us a good starting point to begin the project.

Besides, we also realized that our own research has its unique characters with which we were able to predict our potential problems or limitation by comparing our research design. The most significant difference is that we only observe 6 bus routes at one stop (both inbound and outbound), which is relative;y small compared to other large-scale researches.

1.3 Quick Summary of Main Results

Over the course of our analysis we uncovered some useful information pertinent to Carnegie Mellon Students. We surveyed almost 480 buses which provided us information to make inference on the factors that affect the punctuality of the buses and potentially make some conclusions about the Pittsburgh public transit.

Section 2: Methodology

2.1 Target Population and Frame

The population we targeted is all of the buses which stop on campus that CMU students use to get to school. Our sampling frame is the list of port authority buses that stop at the Forbes and Morewood intersection which is the most commonly used bus stop by CMU students. The sample population is 61A, 61B, 61C, 61D, 67 and 69 that stop at the Forbes and Morewood bus stops.

These two stops are adequate for our purposes for several reasons. First, the other Forbes stops (such as Hamburg and Beeler) are very close to the Forbes and Morewood stops. If the bus is late at Hamburg, then it will likely be just as late 100 feet down the road at Forbes and Morewood. Second, the other stops close to campus (such as Fifth and Morewood, and the Wilkins stops) are used by a substantially smaller percentage of the student population and mostly for the purposes of shuttling to non residential locations. The areas in Shadyside in which students live are well serviced by the faster and more reliable loop buses.

Buses moving up and down Fifth will also be independent from buses moving up and down Forbes because many travel through different neighborhoods and thus face different traffic patterns. This would lead to trouble in data analysis because the independent routes could lead to a bimodal distribution, hurt the accuracy of any inference we would like to make about how late buses usually are, or increase the number of man hours beyond what is feasible for our group.

Due to the nature of public transit, the decision was made to use cluster sampling because attempting to observe specific buses would be extremely difficult. Instead of sampling individual buses, the bus schedule was broken down into one hour increments of potential observation.

These hours of observation were then stratified into groups of similar hours: the morning rush hours, lunch hours, afternoon, and the evening rush hours. The decision was made to exclude extremely early buses and night buses because they are outside the scope of normal commuting times and measuring them would either require more manpower than the team had available or reducing the number of observation hours for the more interesting strata.

2.2 Sample Size

We assume that there will be around 12 buses per hour at the bus stop.
We set ME= 0.5 min
SD=5 min
(From the selective research, 5min seems to be a good starting point for the standard deviation of the bus lateness.)
Z95%=1.96
N=2wks*7days/wk*12hrs/day*12buses/hr = 2016 buses
We are doing SRS without replacement.
$n_0=Z95\%^2*SD^2/ME^2 = 1.96^2*(5^2)/(0.5^2) = 384.16$
$n=N*n_0/(N+n_0) = 2016*384.16/(2016+384.16)= 322.6729$

However, as we started our observation, we noticed that there are more than 12 buses per hour. Thus, we modified our sample size to 20.

The revises sample size:
N=2wks*7days/wk*12hrs/day*20buses/hr = 3360 buses
We are doing SRS without replacement.
$n_0=Z95\%^2*SD^2/ME^2 = 1.96^2*(5^2)/(0.5^2) = 384.16$
$n=N*n_0/(N+n_0) = 3360*384.16/(3360+384.16)= 344.744$

Inflate the size by 10% for cluster effects
Estimated Sample size = 380 buses

2.3 Sample Design and Methods

Our sampling method is Stratified One-Stage Cluster Sampling. Our design was relatively simple. We developed four strata: 7am~10am “Morning Commute”, 10am~1pm “Lunch hours”, 1pm ~4pm “Afternoon hours” and “4pm to 7pm” is the “Evening Commute”. We then determined the necessary number of hours to sample, which is 380 buses according to our sample size calculation.

A sample was drawn by taking all of the potential hours of observation within a stratum over the two week period and assigning them a number in chronological order. A random number generator was then used to generate a set of numbers to determine which hours would be included in our sample.

The bus observations weren't performed at random times, we had four key time intervals which were split up among the five group numbers. This was done to gain a good understanding of what happens in certain parts of the day.

2.4 Response

Some key statistics or variables that we measured included the bus route number (the buses listed above were the bus routes of interest), whether the bus was going inbound or outbound, actual departure time for the bus from the bus stop and scheduled time to depart from the bus stop, the time difference between actual and scheduled, whether the bus was on-time or not, the light condition (dawn, light, dark, or overcast), road condition (dry, wet, or snow/ice), weather (normal or not raining/snowing, rain, or snow).

Since our design is based on observation, we do not have a true non-response rate. We experienced some difficulty with our sampling choices because of scheduling concerns, but this does not represent true non-response rates. However, there are circumstances where buses did not stop for some reason (such as two 61A buses are going to the same direction and one has picked up all the customers and another one did not have any passengers getting off), therefore, we record the bus arrival time to predict its punctuality, but it does not have an "actual departure time." The observation sheet is included in the Appendix 1.

2.5 Post-survey Processing

A small number of buses that stop at the Forbes/Morewood intersection were excluded from our study. These buses are 28X, 58, East Liberty Garage and West Miflin Garage. The primary objective of our study was to study buses used to commute to and from Carnegie Mellon. As such, there are a number of buses that stop at the Forbes/Morewood intersection that are not a part of our target population. A number of buses that stop at the Forbes/Morewood intersection, such as the 28x, are buses that are used to get to non residential areas. Other buses appear so infrequently that they could not be relied upon as a commuter and we would not be able to gather enough data to make any inferences. The departure times of these buses were recorded during observation periods, however they were removed from the sample as part of the post survey processing.

After recording in the observation sheet, we coded our results into another speadsheet, which is included in the Appendix 4. We quantified some of the categorical variables such as "day of the week," weather and stratum in order to make it easier for us to do statistical analysis. However, this also brings some disadvantages as we will discuss later in the weakness section. Some key variables were created using the raw data. Time difference eqauls the actual departure time minus the scheduled departure time. We also defined a bus to be on time if it arrives within 5 min of its scheduled time, either earlier or later. The code book is included in Appendix 3.

Section 3: Results

3.1 Introduction to Results

Our total sample size is 480 buses. Margin of error is 5 minutes after consulting Professor Brian Junker. As was discussed earlier, the sample size we wanted to get was 380 buses. Our actual sample size was a good outcome. The descriptive statistics section will talk about the demographic information of our bus sample. Our surveyed buses appear to be representative of the population of buses that appear at the bus stop and this leans towards potential validity for the study.

The goal of our study is to figure out how useful the bus schedules are and what factors affect the punctuality of the buses. In the following sections, there will be interesting results revealed about which buses are more/ less on time and what conditions cause buses to be more/ less on time.

3.2 Statistical Analyses

3.2.1 Descriptive Statistics

Minitab outputs for the descriptive statistics are included in the Appendix 5.
Our observed bus count excluding the buses not of interest was 480 and the next step was to produce a simple histogram where bus route would be labeled as the x axis and count (or number of buses) would be labeled as the y axis. Figure 1 shows the sample size for each route. Observing the histogram it shows that 61A, B, C, and D are all have a count near 100 with 61C being the highest at 105 and 61A being the lowest at 98, not big differences. Then we had our last two bus routes 67 and 69 with significantly lower number of bus stops with 42 and 34 respectively they are scheduled to come less frequently and less often used.

Another histogram was created using our four stratum or the four different time observation intervals (7-10am, 10am-1pm, 1-4pm, and 4-7pm). Observing Figure 2, we can see that the 1-4pm time interval had the highest volume of buses arriving there with a sample of 149. The 10a-1pm time slot had the lowest numbers of buses with a sample of just 74. And stratums 1 and 4

(7-10am and 4-7pm) had very close samples of 133 and 124 respectively. So, pre-experiment logic will say that our data does indeed confirm our assumption that during hours when people are heading home the number of buses will increase. And during hours that are later in the morning or closer to noon people are typically at work so there isn't much demand for buses.

From Output 1 we can see that our sample mean of time difference, which is actual departure time minus scheduled departure time, is 2.438 min with a standard deviation of 5.524 min. Time differences by routes are shown in Output 2. The means are all positive, which means that on average, buses all tend to come later than they are scheduled. There is no particular route that is much more on time than the others. While the mean time differences range from 1.876 to 3.439, the standard deviations range from 4.593 to 6.263, which is really big.

The mean of the absolute time differences is 4.325. It is larger than the time difference sample mean because the effect of earlier buses and late buses does not cancel out. It shows how much the buses on average deviate from the scheduled time, no matter early or late.

The proportion of buses being on time in our sample is 0.46. Basically, buses are on time for half of the time.

Output 5 shows the proportion of buses on time for different routes. 69 has the highest proportion which is 0.618, while 61D has the lowest proportion which is 0.337.

Output 6 shows the proportion of buses being on time by different road conditions. When the road is dry, the mean proportion is 0.544. When the road is wet, the mean proportion is a lot lower, which is 0.133. This is not surprising because we expect the driver to drive more slowly when the road is wet.

Figure 3 are the histograms of time difference and the absolute value of time difference. While the first histogram of time difference seems to be unimodel, the second one shows a dip from 1 to 2 min. Inspired by the graduate student Zachary Kurtz, we attribute this dip to the effect of traffic light at Forbes/Morewood.

3.2.2 Linear Regression

One of our study's goal was to build a model that can predict the lateness of a bus, provided the predictor variables that we are looking into, including day of the week, route, being inbound or outbound, road condition and weather, and time of the day (morning commute, lunch hours, afternoon hours, evening commute). The response variable is the time difference between the bus actual departure time and the scheduled departure time.

We did a full model with all the variables, models within each stratum and models for each route. The regression output summaries are included in Appendix 7. Here the models are briefly explained:

Full Model:
For Appendix 6 Linear Model 1, using 95% significance level, we only found two significant variables, which are Route 61D and Strata. The P-values for them respectively are 0.040 and 0.047. The coefficient of 61D is 1.62, which mean that we expect 61D to be 1.62 min later than 61A while 61A is the baseline for the categorical variable Route here. The coefficient for Strata is 0.565, which means that as the day goes on, moving from the previous stratum to the next stratum is expected to increase the lateness by 0.56 min.

Model for Strata 1:
For Appendix 6 Linear Model 2, using 95% significance level, we found Day and Road to be the two significant variables. The P-values for them respectively are 0.003 and 0.001. The coefficient for Day is 1.412, which means that as the week goes on, the next day is expected to have 1.412 more minutes of lateness than the previous day within the time frame from 7am to 10am. The coefficient for Road is 5.821, which means that we expect the wet road to cause 5.821 more minutes of lateness than dry road within the time frame from 7am to 10am.

Model for Strata 2:
For Appendix 6 Linear Model 3, using 95% sigificance level, we found Day and Light to be the two significant variables. The P-values for them respectively are 0.001 and 0. The coefficient for Day is -5.350, which means that as the week goes on, the buses are expected to come 5.350 minutes earlier than the previous day within the time frame from 10am to 1pm. The coefficient for Light is 16.638, which means that during dark hours, there tends to be 16.64 more minutes of delay when the light condition gets worse.

Model for Strata 3:
For Appendix 6 Linear Model 4, using 95% significance level, we did not find any variable significant.

Model for Strata 4:
For Appendix 6 Linear Model 5, using 95% significance level, we did not find any variable significant.

Model for 61 A:
For Appendix 6 Linear Model 6, using 95% significance level, we did not find any variable significant.

Model for 61 B:
For Appendix 6 Linear Model 7, using 95% significance level, we did not find any variable significant.

Model for 61 C:
For Appendix 6 Linear Model 8, using 95% significance level, we did not find any variable significant.

Model for 61 D:
For Appendix 6 Linear Model 9, using 95% significance level, we did not find any variable significant.

Model for 67:
For Appendix 6 Linear Model 10, using 95% significance level, we found Inbound/Outbound and Road to be the two variables that are significant. The P-values for them respectively are 0.006 and 0.031. The coefficient for Inbound/Outbound is 5.125, which means that the outbound 67's are expected to be 5.125 min later than the inbound 67's. The coefficient for Road is -9.98, which means that buses are expected to come 9.98 earlier with the wet road condition than the dry road condition. This is a surprising finding, which will be discussed later.

3.2.3 Logistic Regression

In addition to trying to predict whether a bus will come early or late quantitatively by predicting how many minutes it will be late, we also want to predict whether a bus will on time or not qualitatively. Thus, a binomial logistic regression is helpful here to find the probability of a bus being on time or not on time, provided the values of the predictor variables.

Full Logistic Model:
For Appendix 8 Model 1, using 95% significance level, we found Road and Weather to be the two significant variables. The P-values for them are both effectively 0. The coefficient for Road is -2.080, which means that wet road is going to decrease the probability of a bus being on time. The coefficient for Weather is 1.577, which means that rainy days are expected to increase the probability of a bus being on time. This is counter-intuitive. First, the correlation between Road and Weather should be high because the road is wet when it rains. We would expect the signs of their effects to be the same. Buses come earlier when it is raining also is very surprising. So we did logistic models with predictor variable only being Road and Weather respectively.

Here's is what we found:
For Appendix 8 Model 2, the P-value for Road is effectively 0. For Appendix 8 Model 3, the P-value for Weather is 0.754. This shows that the reason that Weather was significant in the Full

Logistic Model before was because of its high colinearity with the Road. The sign of their effects of being different might just be a coincidence.

Section 4: Discussion

4.1 Unexpected Results

One of the unexpected results is that more buses are likely to arrive on time during raining days. This might be because we misinterperated and the bus arrived at the time spot as the one that should appear while indeed, the bus ought to have arrived in the previous time spot. In other words, it was possible the bus arrived much later than usual. However, as was discussed in the logistic regression part, the effect of the Weather variable is not significant itself. It might have just been an accident.

4.2 Brief Answers to Research Questions

•How accurate are the bus schedules for the Forbes/Morewood Intersection?
According to our sample, the proportion of buses being on time is 46%, which indicates that half of the time, buses do not come as it is scheduled to.

•Can we identify what factors influence the punctuality of buses?
Most of the factors that we looked into are not significant. Road condition seems to be the only factor that is significantly affecting the on time accuracy of the buses.

4.3 Weaknesses

We were not able to record some factors that might be critical, such as the load of the bus, the years of experience of the drivers, etc.. For example, it will be hard for us to define "heavy load". We cannot count the number of passengers on board either. Yet, while the bus is heavily loaded, it takes longer for people to get on and get off, which really affects the punctuality of the buses.

When we did regressions, the way we coded some variables such as Day of the week and Stratum as ordered variables made it impossible for us to identify the different effects of different level. When using ordered variables, we assume that the effects of moving from one level to the next level are the same. For Day of week, we assume that moving from Monday to Tuesday will have a same effect on the punctuality of the buses as moving from Tuesday to Wednesday, which is likely to not be reasonable.

4.4 Take Home Message

From this study we validated our pre-experiment hypothesis that more buses are likely to come by the bus stops during rush hour (morning and afternoon) and less so in the middle of the day. But, the point wasn't to see how many buses come to a certain stop but how accurate those that did come were. By flipping a coin you may find out if the bus was on time or not. For future groups also interested in measuring bus accuracy we would suggest measuring more detailed and particular variables as mentioned in the "weaknesses" section above. For instance it would be wise to look into why buses are more on time when its raining. This is not only counter-intuitive but it could be because the bus already passed its stop and we're seeing the next one who's not on time. Or bus drivers are rushing to the bus stop because they don't want to be late. The bus already passing by could be a more legitimate reason but again it would be important to look into why rain and bus accuracy had such a high correlation. Human clustering is another important element future groups should look into measuring. Meaning the number of people going into each bus which would conceivably hold up other buses behind it leading to a clustering effect. Moreover, further analysis could be done just focusing on the number of people getting on each bus and comparing it to neighboring bus stops. Is the delay of the bus a result of human clustering at some bus stops and less so at others?

Another question worth asking is finding out how long are students waiting on average and how long do they wait for a bus before they start walking or some alternative transportation route. This would involve more work and a finely executed sample where there would be some interaction between the researchers and these people.

The take home message is to not accept our results for what they are but to further investigate. If people change the value of what they define as "on-time" then maybe the probability of the bus coming on time increases. There are several things worth researching and it would be interesting to delve into the details a bit more.

Works Cited (Sources)

Daskalakis, Nikolaos G. and Anthony Stathopoulos. "Users' Perceptive Evaluation of Bus Arrival Time Deviations in Stochastic Networks." *Journal of Public Transportation, Vol. 11, No. 4,* 2008. http://www.nctr.usf.edu/jpt/pdf/JPT11-4Daskalakis.pdf.

Farid, A. (2010). *Trip productivity evaluation of bus service: Medan Kidd Bus Station.* http://www.uniten.edu.my/newhome/uploaded/coe_civil/mutrfc2010/001%20PG%20presentation.pdf. February 25, 2011.

Macaffe, K. (2008). *Bus Punctuality Statistics GB: 2007*. Great Britain: Department for Transport.

http://www.dft.gov.uk/pgr/statistics/datatablespublications/public/buspunctuality/buspunctuality07. February 25, 2011

John Wirtz, Edwards and Kelcey, McCord, Mark and Mishalani, Rabi. *Passenger Wait Time Perceptions at Bus Stops*, from http://www.nctr.usf.edu/jpt/pdf/JPT%209-2%20Mishalani.pdf. January 23, 2011.

Brian Caulfield and Margaret O'Mahony, *A Stated Preference Analysis of Real-Time Public Transit Stop Information"* from http://www.nctr.usf.edu/jpt/pdf/JPT12-3Caulfield.pdf. February 5, 2011.

The Department of Transportation, UK, *Bus Punctuality Punctuatity.* http://www.dft.gov.uk/pgr/regional/buses/buspunctuality. UK Dept. of Treasury. February 8, 2011.

Weikel, Dan. *MTA gets low marks for bus service in survey*. http://articles.latimes.com/2009/aug/29/local/me-bus-riders29. February 5, 2011.

Appendices

Appendix 1: Observation Template

Date: Light Condition: Weather Condition:

Time: Road Condition:

	Inbound/outbound	Bus number/route	Departure Time	Notes
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				

Appendix 2: Codebook

Date – Month/Day/Year

Was the bus at the Inbound (Towards Downtown) or Outbound (Towards Squirrel Hill) Stop?
1 – Inbound
2 – Outbound

Day of the Week –
1 – Monday
2 – Tuesday
3 – Wednesday
4 – Thursday
5 – Friday
6 – Saturday
7 – Sunday

Bus Route - 28X, 67, 61A, 61B, 61C, 61D, 58, 69, East Liberty Garage, 56, West Mifflin Garage

Scheduled Departure Time – Pittsburgh bus schedules do not give arrival times for the Forbes/Morewood intersection. The closest scheduled stop is the Forbes/Craig intersection. 3 minutes was added to the Forbes/Craig scheduled time to determine the "scheduled" time for Forbes/Morewood if the bus was going in the direction of Squirrel Hill. 3 minutes was subtracted fromk the Forbes Craig/scheduled time if the bus was going in the direction of downtown.

Departure Time – time that the bus departed from the Forbes/Morewood intersection

Time Difference – Departure Time – Scheduled Time. A positive value indicates that a bus was late, a negative value indicates that the bus was early, and a 0 value indicates that the bus was exactly on time.

Absolute Time Difference - absolute value of the variable Time difference

On Time
0 – Not on Time – Bus was more than 5 minutes late or early – Absolute Time Difference > 5
1 – On Time – Bus was less than 5 minutes late or early – Absolute Time difference ≤ 5

Light Condition
1 – Dawn
2 – Light
3 – Dark
4 – Overcast

Road Condition
1 – Dry

2 – Wet
3 – Snow/Ice

Weather
1 – "Normal" not raining/snowing
2 – Rain
3 – Snow

Notes – any observations about the bus that might be important or potential grounds for excluding the bus
DNS – Did Not Stop

Stratum
1 – Buses scheduled to arrive between 7AM and 10AM
2 – Buses scheduled to arrive between 10AM and 1PM
3 – Buses scheduled to arrive between 1PM and 4PM
4 – Buses scheduled to arrive between 4PM and 7PM

Appendix 3: Coded data spreadsheet

date	day	ID	route	dep	sch	diff	ontime	light	road	weather	strat
3/14/2011	1	2	61A	1413	1405	8	0	2	1	1	3
3/14/2011	1	2	61D	1416	1411	5	1	2	1	1	3
3/14/2011	1	2	61B	1422	1420	2	1	2	1	1	3
3/14/2011	1	2	69	1422	1417	5	1	2	1	1	3
3/14/2011	1	2	61C	1425	1427	-2	1	2	1	1	3
3/14/2011	1	2	61A	1437	1435	2	1	2	1	1	3
3/14/2011	1	2	61D	1444	1447	-3	1	2	1	1	3
3/14/2011	1	2	67	1452	1448	4	1	2	1	1	3
3/14/2011	1	2	61C	1453	1457	-4	1	2	1	1	3
3/14/2011	1	2	61B	1454	1450	4	1	2	1	1	3
3/14/2011	1	2	61A	1512	1505	7	0	2	1	1	3
3/14/2011	1	2	69	1515	1517	-2	1	2	1	1	3
3/14/2011	1	2	61D	1519	1512	7	0	2	1	1	3
3/14/2011	1	2	61C	1522	1523	-1	1	2	1	1	3
3/14/2011	1	2	61A	1523	1527	-4	1	2	1	1	3
3/23/2011	3	1	61D	1507	1459	8	0	2	1	1	3
3/23/2011	3	1	61C	1508	1505	3	1	2	1	1	3
3/23/2011	3	1	69	1510	1513	-3	1	2	1	1	3
3/23/2011	3	1	61A	1512	1502	10	0	2	1	1	3
3/23/2011	3	1	61B	1515	1510	5	1	2	1	1	3
3/23/2011	3	2	69	1518	1517	1	1	2	1	1	3
3/23/2011	3	2	61A	1519	1505	14	0	2	1	1	3

Appendix 4 : Note on excluded bus routes

There were two bus stops we had to exclude from our data: 58 and 28x. Both the 58 and 28x were infrequent and would hinder our data instead of help make our conclusions valid. The 28x is primarily known as an airport shuttle and students are not going to the airport and returning

from it the next day. The 58 also had few stops it would make during the week so we would have two significant outliers that would only deter us from properly analyzing bus efficiency of buses that are frequently coming through Morewood and Forbes or Forbes and Craig. Because the number of 58s and 28x's making stops was so limited, they were usually on time with no clustering or lateness problems when they did come. Which is great and points to the fact that buses can be on time but this is statistically insignificant in terms calculating the overall bus accuracy. So looking at the big picture, the 58 and 28x are inconsequential to proving if buses are accurate or not, it takes away from our conclusion rather than help promote a certain argument.

Appendix 5: Descriptive Statistics

Output 1:
Descriptive Statistics: Time Difference

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Time Difference	480	0	2.438	0.252	5.524	-20.000	0.000	2.000	5.000

Output 2:
Descriptive Statistics: Time Difference by Bus Route

Variable	Bus Route	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	
Time Difference	61A	98	0	1.898	0.579	5.730	-20.000	0.000	1.000	
	61B	103	0	2.563	0.468	4.748	-19.000	0.000	3.000	
	61C	105	0	1.876	0.532	5.456	-20.000	-1.000	2.000	
	61D	98	0	3.439	0.612	6.063	-15.000	0.000	3.000	
	67	42	0	2.333	0.967	6.265	-7.000	-2.250	2.000	
	69	34	0	2.588	0.788	4.593	-10.000	0.000	2.000	

Output 3:
Descriptive Statistics: AbsTime

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
AbsTime	480	0	4.325	0.192	4.209	0.000	1.000	3.000	6.000	29.000

Output 4:
Descriptive Statistics: Proportion of buses On Time

Variable	X	N	Sample p
On Time	224	480	0.466667

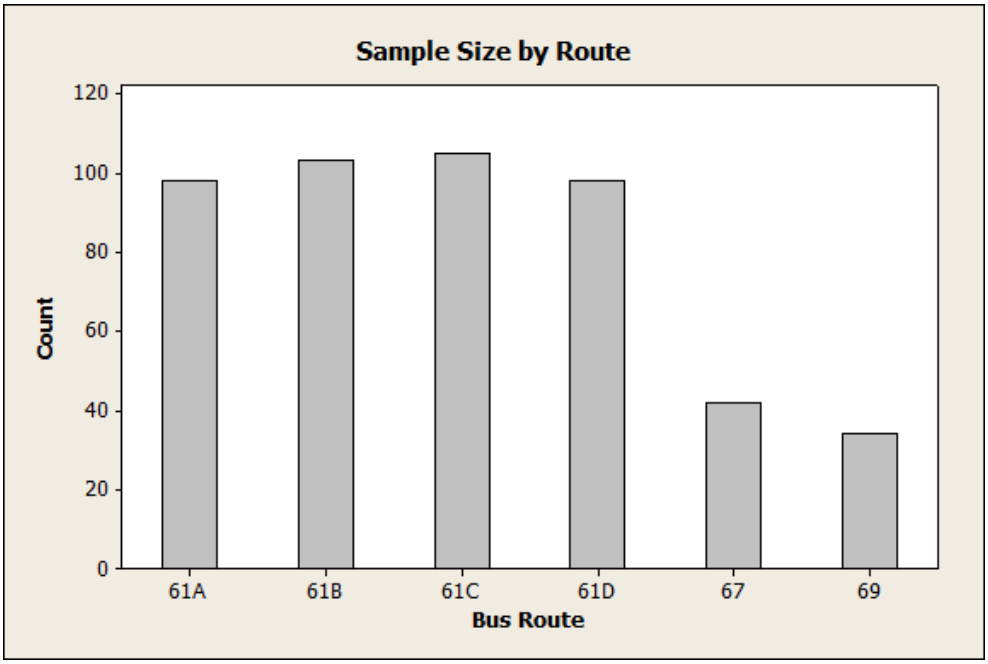
Output 5:
Descriptive Statistics: Proportion of buses On Time by Bus Route

Variable	Bus Route	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
On Time	61A	98	0	0.4592	0.0506	0.5009	0.0000	0.0000	0.0000
	61B	103	0	0.5049	0.0495	0.5024	0.0000	0.0000	1.0000
	61C	105	0	0.4762	0.0490	0.5018	0.0000	0.0000	0.0000
	61D	98	0	0.3367	0.0480	0.4750	0.0000	0.0000	0.0000
	67	42	0	0.5476	0.0777	0.5038	0.0000	0.0000	1.0000
	69	34	0	0.6176	0.0846	0.4933	0.0000	0.0000	1.0000

Output 6:
Descriptive Statistics: On Time by Road Condition

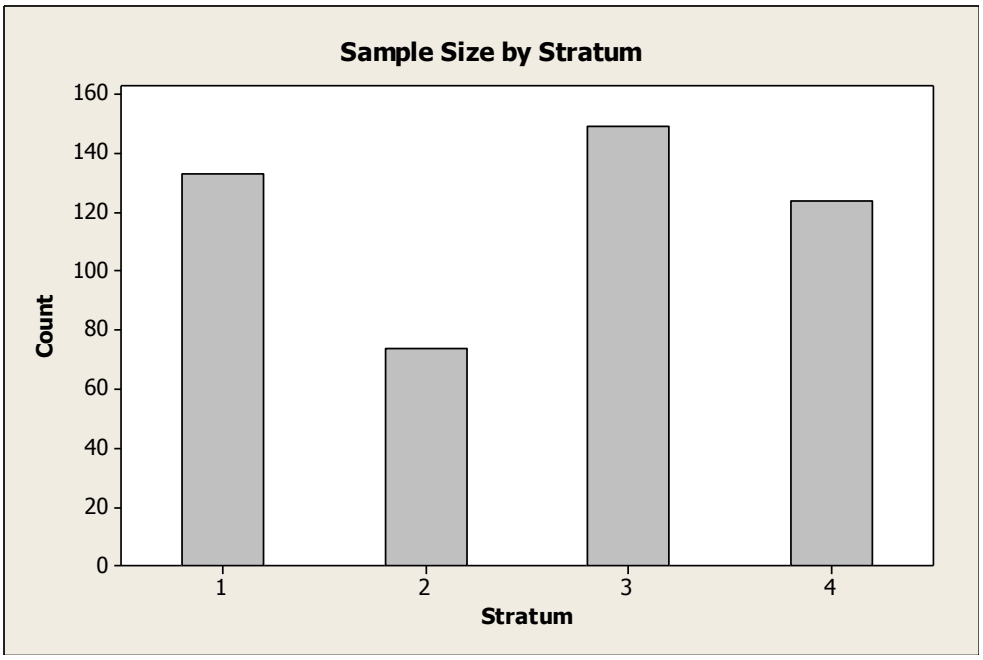
Variable	Road Condition	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
On Time	1	390	0	0.5436	0.0253	0.4987	0.0000	0.0000	1.0000
	2	90	0	0.1333	0.0360	0.3418	0.0000	0.0000	0.0000

Figure 1:



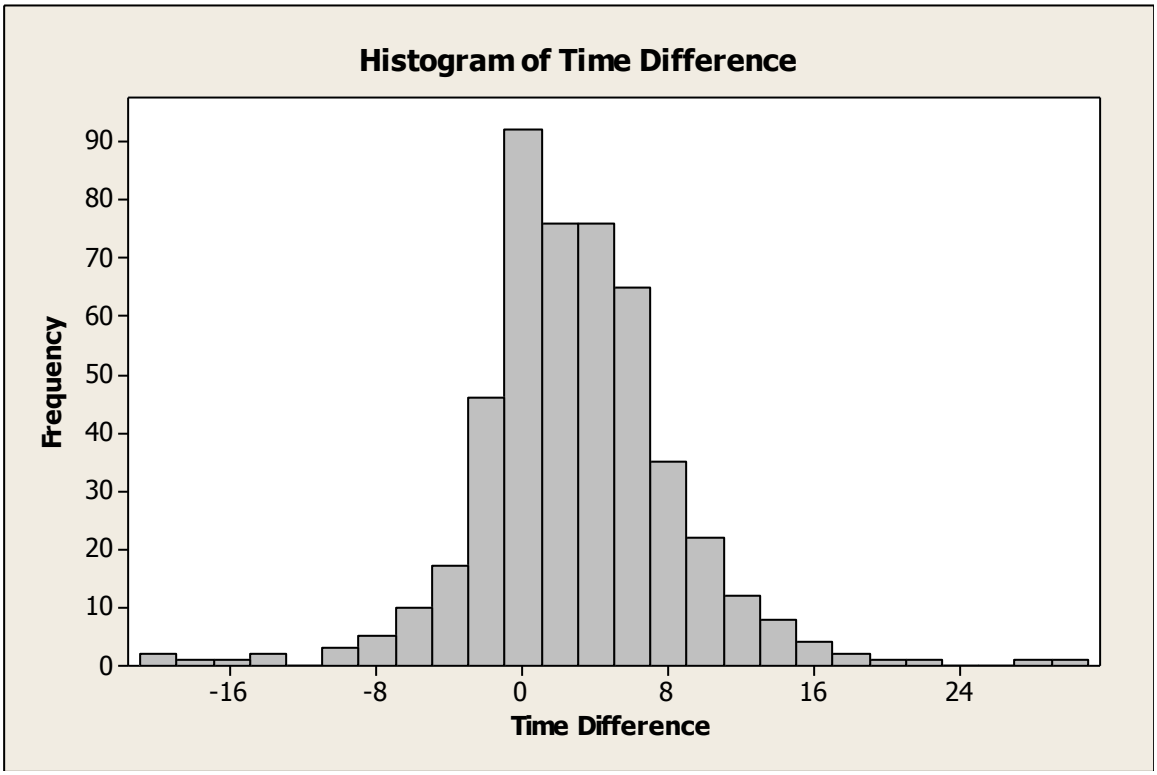
Route	61A	61B	61C	61D	67	69
n	98	103	105	98	42	34

Figure 2:



Stratum	1	2	3	4
n	133	74	149	124

Figure 3:



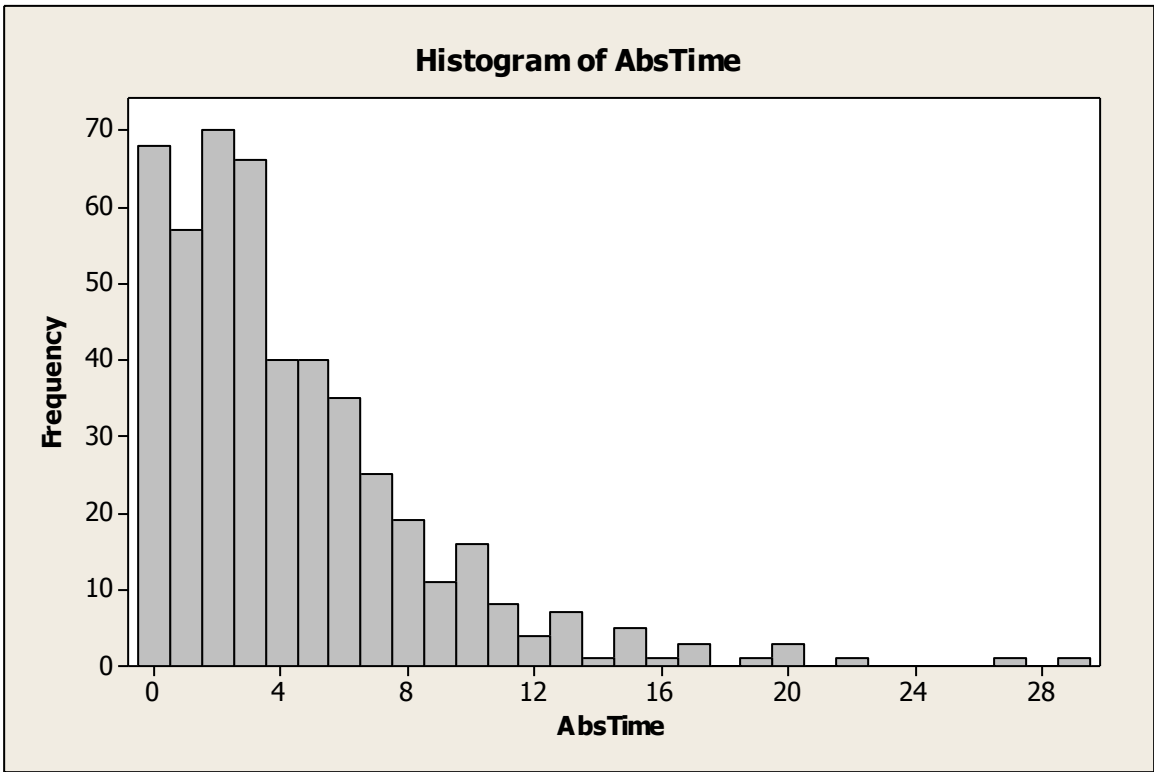
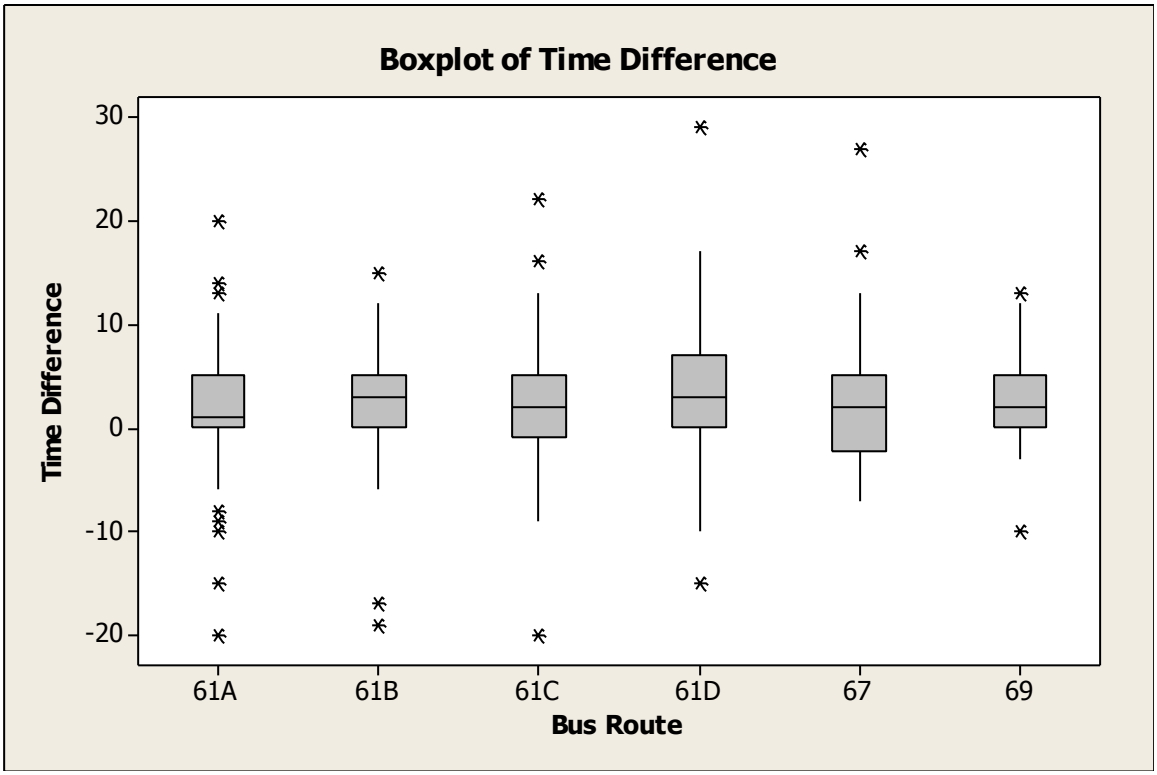
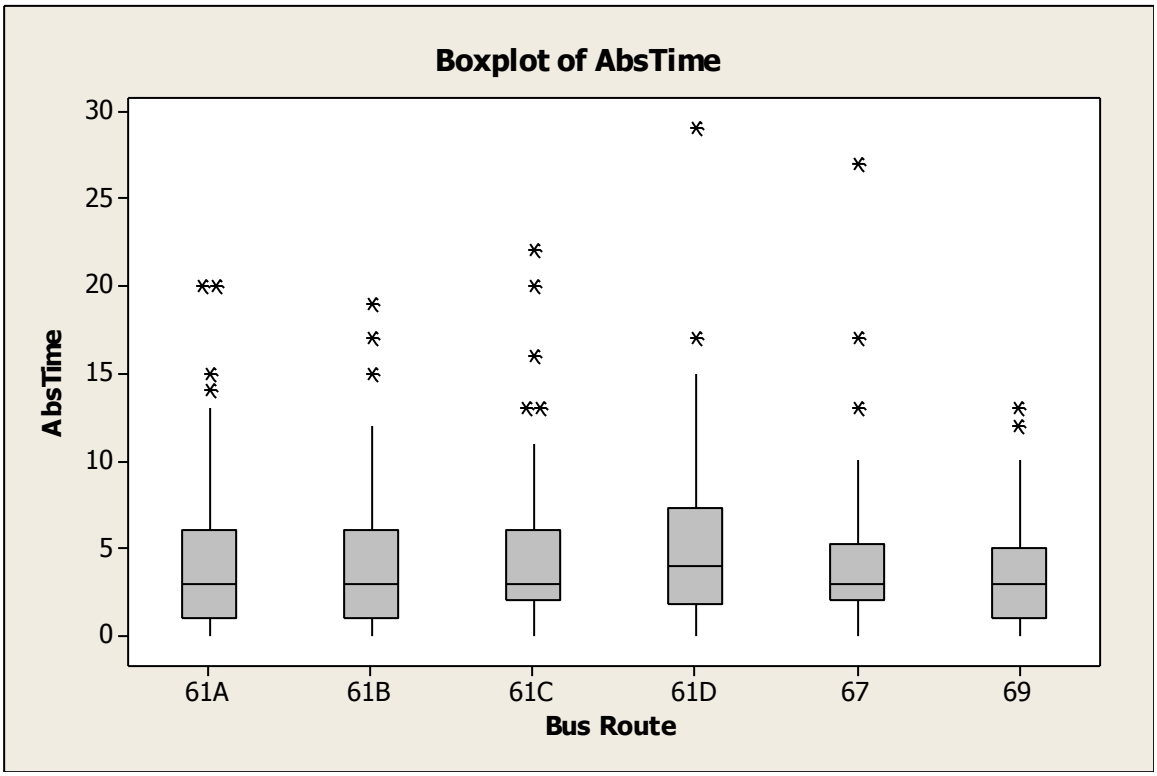
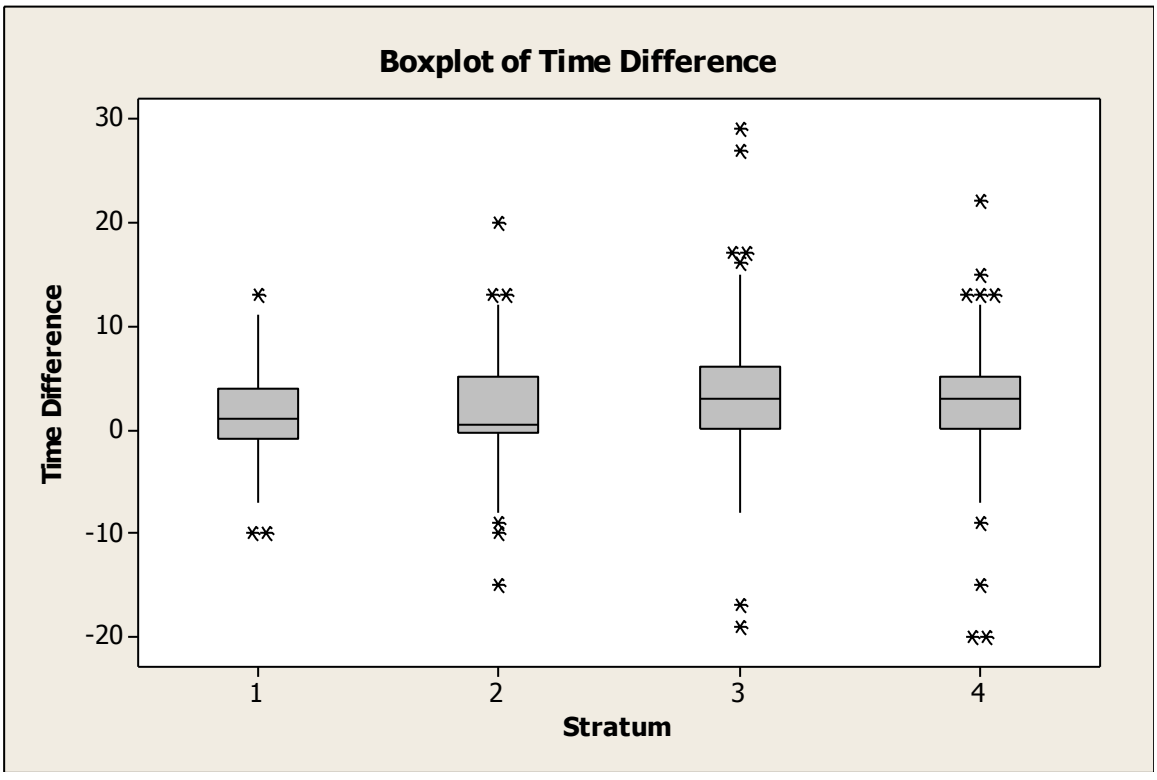
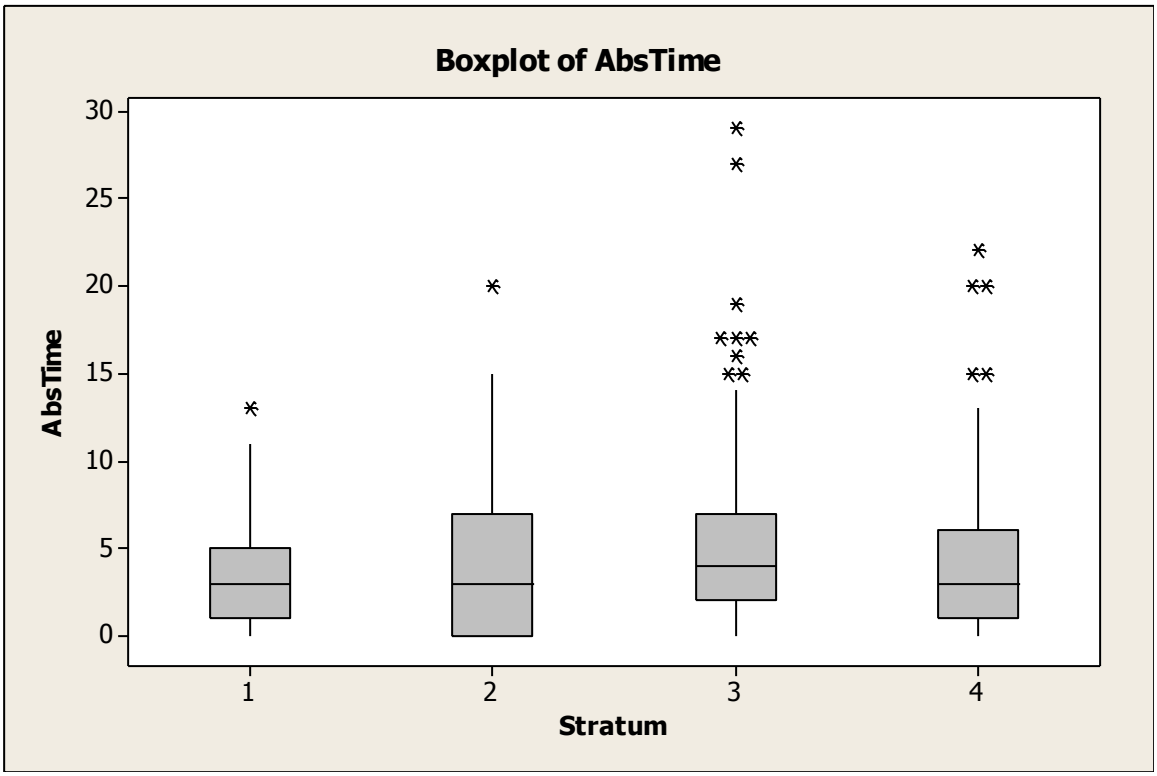

Histogram of AbsTime
Frequency
AbsTime
0
10
20
30
40
50
60
70
0
4
8
12
16
20
24
28

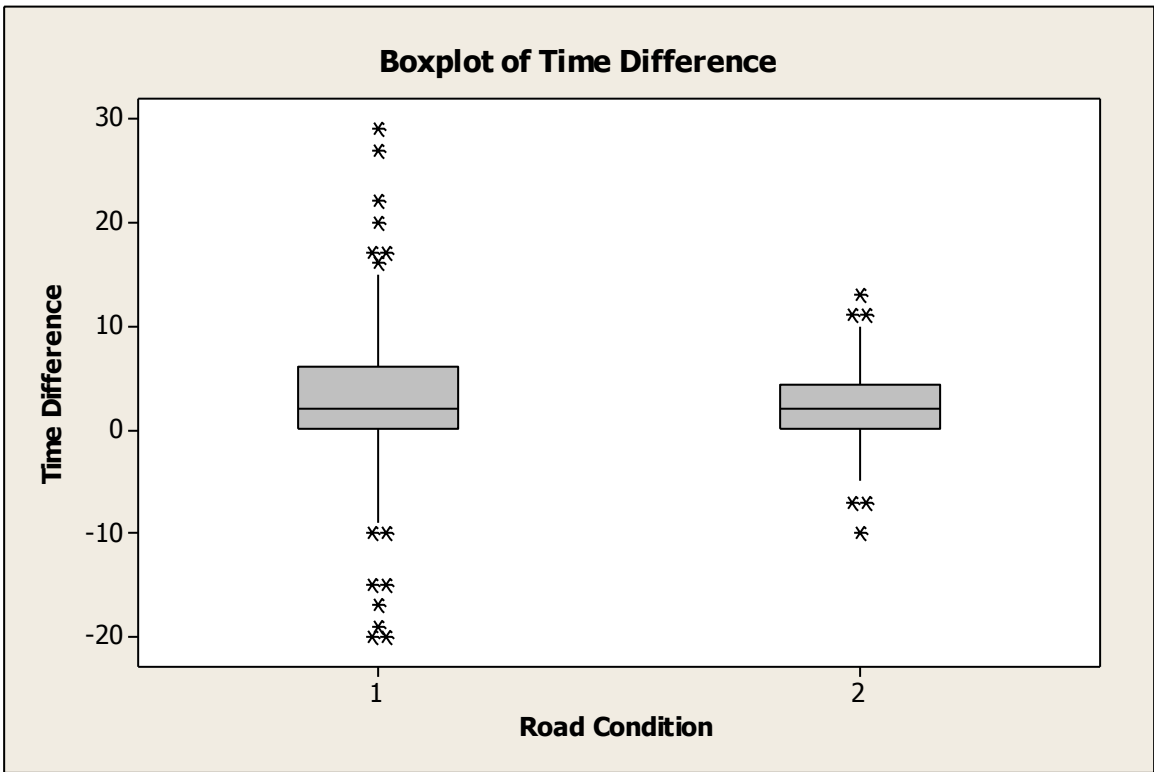
Figure 4:

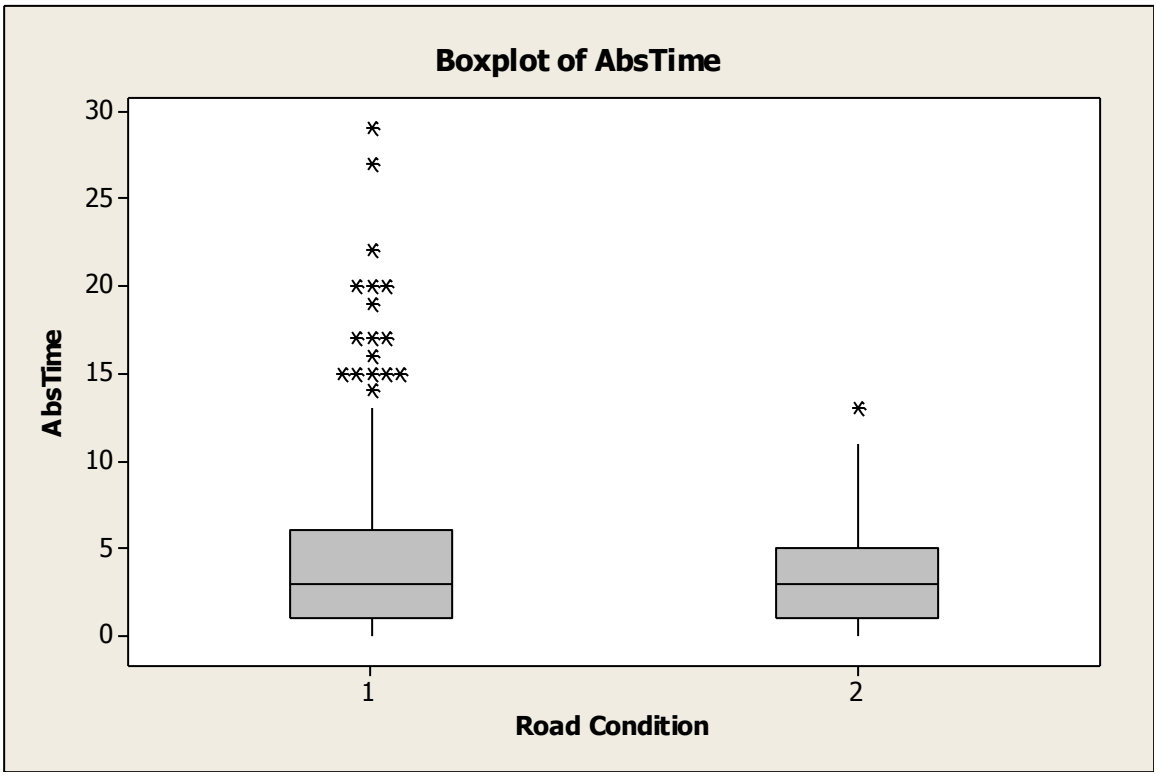

Boxplot of Time Difference
Time Difference
Bus Route
30
20
10
0
-10
-20
61A
61B
61C
61D
67
69


Boxplot of AbsTime
AbsTime
30
25
20
15
10
5
0
61A
61B
61C
61D
67
69
Bus Route


Boxplot of Time Difference
Time Difference
30
20
10
0
-10
-20
1
2
3
4
Stratum


Boxplot of AbsTime
AbsTime
30
25
20
15
10
5
0
1
2
3
4
Stratum


Boxplot of Time Difference
Time Difference
30
20
10
0
-10
-20
1
2
Road Condition



Appendix 6: Linear Regression

Linear Model 1 (Full linear model)

lm(formula = diff ~ day + IO + route + light + road + weather + strat)

Residuals:

Min	1Q	Median	3Q	Max
-23.0084	-2.8374	-0.2025	2.7634	25.4879

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.70383	1.99824	0.352	0.7248	
day	0.02472	0.18210	0.136	0.8921	
IO	0.60815	0.50982	1.193	0.2335	
route61B	0.79226	0.78033	1.015	0.3105	
route61C	-0.07899	0.77478	-0.102	0.9188	
route61D	1.62208	0.78820	2.058	0.0401	*
route67	0.30310	1.02035	0.297	0.7666	
route69	0.89782	1.10953	0.809	0.4188	
light	0.09975	0.39679	0.251	0.8016	

road -0.02919 1.04335 -0.028 0.9777
weather -1.38480 1.05122 -1.317 0.1884
strat 0.56454 0.28294 1.995 0.0466 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.507 on 468 degrees of freedom
Multiple R-squared: 0.02799, Adjusted R-squared: 0.005143
F-statistic: 1.225 on 11 and 468 DF, p-value: 0.2672

Linear Model 2 (Linear Model for Strata 1)

lm(formula = diff ~ day + IO + light + road + weather, data = datastrat1)

Residuals:
Min 1Q Median 3Q Max
-12.0538 -2.6287 0.5748 2.3193 10.9462

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.26088 3.53377 -2.904 0.00435 **
day 1.41150 0.45913 3.074 0.00258 **
IO -0.03102 0.71626 -0.043 0.96553
light -0.87500 0.51100 -1.712 0.08926 .
road 5.82112 1.74871 3.329 0.00114 **
weather NA NA NA NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.059 on 128 degrees of freedom
Multiple R-squared: 0.09381, Adjusted R-squared: 0.06549
F-statistic: 3.313 on 4 and 128 DF, p-value: 0.01280

Linear Model 3 (Linear Model for Strata 2)
lm(formula = diff ~ day + IO + light + road + weather, data = datastrat2)

Residuals:
Min 1Q Median 3Q Max
-14.1190 -2.1190 0.2313 2.6673 14.8810

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.280	4.068	0.315	0.75394	
day	-5.350	1.612	-3.318	0.00144	**
IO	1.332	1.176	1.132	0.26131	
light	16.638	3.852	4.319	5.07e-05	***
road	NA	NA	NA	NA	
weather	NA	NA	NA	NA	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.031 on 70 degrees of freedom
Multiple R-squared: 0.2575, Adjusted R-squared: 0.2257
F-statistic: 8.094 on 3 and 70 DF, p-value: 0.0001058

Linear Model 4 (Linear Model for Strata 3)

lm(formula = diff ~ day + IO + light + road + weather, data = datastrat3)

Residuals:

Min	1Q	Median	3Q	Max
-22.92861	-3.41773	0.08205	2.96863	26.24698

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.66113	3.76038	0.708	0.480
day	-0.27836	0.31998	-0.870	0.386
IO	1.17559	1.05377	1.116	0.266
light	0.02292	0.93190	0.025	0.980
road	-0.01612	2.31992	-0.007	0.994
weather	NA	NA	NA	NA

Residual standard error: 6.287 on 144 degrees of freedom
Multiple R-squared: 0.01838, Adjusted R-squared: -0.008888
F-statistic: 0.674 on 4 and 144 DF, p-value: 0.611

Linear Model 5 (Linear Model for Strata 4)

lm(formula = diff ~ day + IO + light + road + weather, data = datastrat4)

Residuals:

```
     Min       1Q   Median       3Q      Max
-23.20126 -2.23988 -0.01411  2.59089 18.79874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5122     4.7578   0.318    0.751
day           0.5772     0.4840   1.193    0.235
IO           -0.2897     1.0153  -0.285    0.776
light        -0.2539     1.2954  -0.196    0.845
road         -0.9000     2.0751  -0.434    0.665
weather       0.5005     1.8308   0.273    0.785

Residual standard error: 5.62 on 118 degrees of freedom
Multiple R-squared: 0.03855,    Adjusted R-squared: -0.002187
F-statistic: 0.9463 on 5 and 118 DF,  p-value: 0.4539
```

Linear Model 6 (Linear Model for 61A)

```
lm(formula = diff ~ day + IO + light + road + weather + strat,
    data = data61A)

Residuals:
     Min       1Q   Median       3Q      Max
-20.72429 -2.70167  0.05873  2.99786 16.97907

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.24877    4.63975   0.269    0.788
day         -0.02354    0.40545  -0.058    0.954
IO           1.83438    1.19214   1.539    0.127
light        0.49056    0.91191   0.538    0.592
road        -1.94390    2.50672  -0.775    0.440
weather     -0.30675    2.58599  -0.119    0.906
strat       -0.24290    0.69941  -0.347    0.729

Residual standard error: 5.822 on 91 degrees of freedom
Multiple R-squared: 0.03171,    Adjusted R-squared: -0.03213
F-statistic: 0.4967 on 6 and 91 DF,  p-value: 0.8093
```

Linear Model 7 (Linear Model for 61B)

```
lm(formula = diff ~ day + IO + light + road + weather + strat,
```

data = data61B)

Residuals:

Min	1Q	Median	3Q	Max
-21.05086	-2.26439	0.07799	2.36540	13.06220

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.36547	3.75679	-0.363	0.717
day	0.30739	0.36719	0.837	0.405
IO	0.11306	0.96164	0.118	0.907
light	-0.12367	0.78327	-0.158	0.875
road	2.43847	1.99887	1.220	0.225
weather	-0.44844	1.97271	-0.227	0.821
strat	0.07266	0.53695	0.135	0.893

Residual standard error: 4.809 on 96 degrees of freedom
Multiple R-squared: 0.02924, Adjusted R-squared: -0.03143
F-statistic: 0.4819 on 6 and 96 DF, p-value: 0.8203

Linear Model 8 (Linear Model for 61C)

lm(formula = diff ~ day + IO + light + road + weather + strat,
data = data61C)

Residuals:

Min	1Q	Median	3Q	Max
-22.22352	-2.80438	-0.01424	2.62625	18.77938

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3826	4.4635	0.310	0.757
day	0.1754	0.3895	0.450	0.653
IO	-0.9971	1.1023	-0.905	0.368
light	0.7074	0.9484	0.746	0.458
road	-1.4485	2.4601	-0.589	0.557
weather	-0.0528	2.2788	-0.023	0.982
strat	0.5111	0.5698	0.897	0.372

Residual standard error: 5.495 on 98 degrees of freedom
Multiple R-squared: 0.044, Adjusted R-squared: -0.01453

F-statistic: 0.7518 on 6 and 98 DF, p-value: 0.6095

Linear Model 9 (Linear Model for 61D)

```
lm(formula = diff ~ day + IO + light + road + weather + strat,
    data = data61D)

Residuals:
    Min      1Q  Median      3Q     Max
-19.6808  -3.4788  -0.4294   3.4405  23.6948

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.9424     4.9377   1.811   0.0734 .
day          -0.5979     0.4347  -1.376   0.1723
IO           -0.2050     1.2342  -0.166   0.8685
light        -0.9717     0.8693  -1.118   0.2666
road          0.5799     2.4195   0.240   0.8111
weather      -4.4655     2.6554  -1.682   0.0961 .
strat         1.5961     0.6986   2.285   0.0247 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.995 on 91 degrees of freedom
Multiple R-squared: 0.08295,    Adjusted R-squared: 0.02248
F-statistic: 1.372 on 6 and 91 DF,  p-value: 0.2344
```

Linear Model 10 (Linear Model for 67)

```
lm(formula = diff ~ day + IO + light + road + weather + strat,
    data = data67)

Residuals:
    Min      1Q  Median      3Q     Max
-8.99391 -2.88061 -0.02298  2.23228 20.00570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0800     5.6615  -0.191  0.84982
day          -0.4241     0.6131  -0.692  0.49361
IO            5.1251     1.7463   2.935  0.00586 **
light         2.4689     1.5363   1.607  0.11703
```

```
road          -9.9804     4.4273  -2.254  0.03054 *
weather       -1.3598     4.3360  -0.314  0.75569
strat          1.5755     1.1325   1.391  0.17294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.464 on 35 degrees of freedom
Multiple R-squared: 0.3506,     Adjusted R-squared: 0.2393
F-statistic: 3.149 on 6 and 35 DF,  p-value: 0.01415
```

Linear Model 11 (Linear Model for 69)

```
lm(formula = diff ~ day + IO + light + road + weather + strat,
    data = data69)

Residuals:
    Min      1Q  Median      3Q     Max
-12.9043  -2.2701  -0.5598   2.2878   7.9005

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.71317    6.13493  -1.257    0.219
day           1.06947    0.83260   1.284    0.210
IO            1.59108    1.67296   0.951    0.350
light        -0.04054    1.33965  -0.030    0.976
road          2.68069    3.03570   0.883    0.385
weather      -0.64902    2.92846  -0.222    0.826
strat         0.58314    0.86969   0.671    0.508

Residual standard error: 4.746 on 27 degrees of freedom
Multiple R-squared: 0.1263,     Adjusted R-squared: -0.06782
F-statistic: 0.6507 on 6 and 27 DF,  p-value: 0.6892
```

Appendix 7: Logistic Regression

Logistic Model 1 (Full Logistic Model)

```
glm(formula = ontime ~ day + IO + route + light + road + weather +
    strat, family = binomial, data = data)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
```

-1.8141 -1.1266 -0.3053   1.0515   2.3587

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.91254	0.78296	1.166	0.243813	
day	0.12820	0.07129	1.798	0.072129	.
IO	-0.26907	0.20161	-1.335	0.182015	
route61B	0.17634	0.30807	0.572	0.567055	
route61C	0.06046	0.30525	0.198	0.843005	
route61D	-0.63461	0.31520	-2.013	0.044081	*
route67	0.22353	0.38924	0.574	0.565779	
route69	0.82365	0.45887	1.795	0.072664	.
light	-0.31587	0.17436	-1.812	0.070056	.
road	-2.08018	0.48897	-4.254	2.10e-05	***
weather	1.57716	0.47365	3.330	0.000869	***
strat	0.04533	0.12127	0.374	0.708559	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 663.29 on 479 degrees of freedom
Residual deviance: 576.62 on 468 degrees of freedom
AIC: 600.62

Number of Fisher Scoring iterations: 5

Logistic Model 2 (Logistic Model with Road being the only predictor variable)

glm(formula = ontime ~ road, family = binomial, data = data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.252	-1.252	-0.535	1.104	2.007

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.2214	0.3708	5.991	2.09e-09	***
road	-2.0466	0.3263	-6.272	3.57e-10	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 663.29 on 479 degrees of freedom
Residual deviance: 608.37 on 478 degrees of freedom
AIC: 612.37

Number of Fisher Scoring iterations: 4

Logistic Model 3 (Logistic Model with Weather being the only predictor variable)

glm(formula = ontime ~ weather, family = binomial, data = data)

Deviance Residuals:
   Min   1Q Median   3Q   Max
-1.159 -1.117 -1.117   1.239   1.239

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.24109   0.35483 -0.679   0.497
weather   0.09832   0.31333   0.314   0.754

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 663.29 on 479 degrees of freedom
Residual deviance: 663.19 on 478 degrees of freedom
AIC: 667.19

Number of Fisher Scoring iterations: 3

Appendix 8: R Code for Regression Anlysis

```
data=read.csv("bus master.csv",header=T)
date=data[,1]
day=data[,2]
IO=data[,3]
route=data[,4]
dep=data[,5]
sch=data[,6]
diff=data[,7]
```

```
ontime=data[,8]
light=data[,9]
road=data[,10]
weather=data[,11]
strat=data[,12]

### Linear Regressions
lm1full=lm(diff~day+IO+route+light+road+weather+strat)
summary(lm1full)

### New data frame and linear regressions for each strata and bus route
datastrat1=data.frame(data[which(strat==1),])
lmstrat1=lm(diff~day+IO+route+light+road+weather,data=datastrat1)
summary(lmstrat1)

datastrat2=data.frame(data[which(strat==2),])
lmstrat2=lm(diff~day+IO+route+light+road+weather,data=datastrat2)
summary(lmstrat2)

datastrat3=data.frame(data[which(strat==3),])
lmstrat3=lm(diff~day+IO+route+light+road+weather,data=datastrat3)
summary(lmstrat3)

datastrat4=data.frame(data[which(strat==4),])
lmstrat4=lm(diff~day+IO+route+light+road+weather,data=datastrat4)
summary(lmstrat4)

lmstrat1nor=lm(diff~day+IO+light+road+weather,data=datastrat1)
summary(lmstrat1nor)
lmstrat2nor=lm(diff~day+IO+light+road+weather,data=datastrat2)
summary(lmstrat2nor)
lmstrat3nor=lm(diff~day+IO+light+road+weather,data=datastrat3)
summary(lmstrat3nor)
lmstrat4nor=lm(diff~day+IO+light+road+weather,data=datastrat4)
summary(lmstrat4nor)

data61A=data.frame(data[which(route=="61A"),])
lm61A=lm(diff~day+IO+light+road+weather+strat,data=data61A)
summary(lm61A)
```

```
data61B=data.frame(data[which(route=="61B"),])
lm61B=lm(diff~day+IO+light+road+weather+strat,data=data61B)
summary(lm61B)

data61C=data.frame(data[which(route=="61C"),])
lm61C=lm(diff~day+IO+light+road+weather+strat,data=data61C)
summary(lm61C)

data61D=data.frame(data[which(route=="61D"),])
lm61D=lm(diff~day+IO+light+road+weather+strat,data=data61D)
summary(lm61D)

data67=data.frame(data[which(route=="67"),])
lm67=lm(diff~day+IO+light+road+weather+strat,data=data67)
summary(lm67)

data69=data.frame(data[which(route=="69"),])
lm69=lm(diff~day+IO+light+road+weather+strat,data=data69)
summary(lm69)

### Logistic Regressions
logrfull=glm(formula = ontime ~ day + IO+route+light+road+weather+strat, family =
binomial,data = data)
summary(logrfull)

logrroad = glm(formula=ontime~road,family=binomial,data=data)
summary(logrroad)
logrweather= glm(formula=ontime~weather,family=binomial,data=data)
summary(logrweather)
```