Transit Analysis Final Project

Maxine Taubman[1] and Kimi Holsapple[1]

[1]University of California, Santa Cruz
[2]University of California, Santa Cruz
{*mtaubman, kholsapp*}*@ucsc.edu*

May 2022

1 Introduction

The goal of this project is to calculate and analyze trends in average minutes late of the loop and metro buses. Through this project, we will review the data collection process and how data was cleaned, interpret the data, explore possible models, analyze results and validation, review our original goals and goals met, reflect on our original timeline and end result, and explore possible improvements, limitations, and future work.

Understanding traffic patterns are critical to city planning, fare adjustments, and road capacity. Traffic affects overall quality of life and city pollution. Increased traffic congestion is linked to increased vehicle emissions and decreases air quality index making health conditions like asthma aggravated in cities with high pollution. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243514/: :text=' Additionally waiting in traffic for long periods and the increased amount of time that it takes to travel decreased the overall quality of life with regions with heavy traffic. These also feed into economic considerations as using you vehicle more, by sitting in congested roadways increases gas use and decreases property values for homes near roadways with heavy traffic [https://www.irtba.org/QualityofLife]. From the perspective of our campus and transportation offices, understanding these patterns are important to event planning and class scheduling. In this paper we will try to analyze traffic patterns on campus.

2 Problem Description

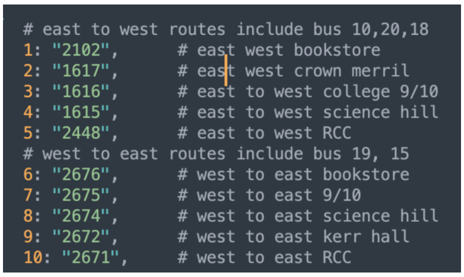
How long, on average, will the bus be late? We interpret this to be our wait time. Minutes late is calculated for both the loop and metro buses.

3 Data Description

We have collected 2 data sets, their methods of collection and are described in detail below.

3.1 Metro

The metro data sets includes data web scraped from cruzmetro.com/arrivals and contains a weeks worth of data starting Tuesday April 19th, 2022. It follows 10 bus stops, 5 on eastbound routes and another 5 on westbound rounds. Their stop ids were recorded and they are defined in the python dictionary below:



```
# east to west routes include bus 10,20,18
1: "2102",      # east west bookstore
2: "1617",      # east west crown merril
3: "1616",      # east to west college 9/10
4: "1615",      # east to west science hill
5: "2448",      # east to west RCC
# west to east routes include bus 19, 15
6: "2676",      # west to east bookstore
7: "2675",      # west to east 9/10
8: "2674",      # west to east science hill
9: "2672",      # west to east kerr hall
10: "2671",     # west to east RCC
```

Furthermore, the bus type was also recorded. This keeps track of what bus arrived at that specific spot for that time and allows the ability to include the fact that different route types could affect lateness. The buses collected were the eastbound buses 10,20,18 and the westbound buses 15 and 19.

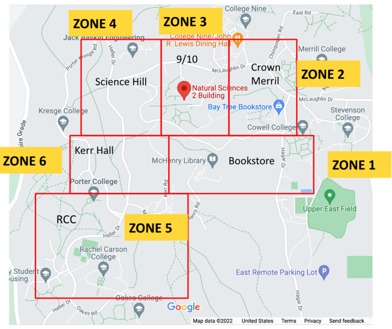
Santa_Cruz_Metro_complete.csv

This data set has 13 features in total and was collected by a combination of Google transit API and web scraped values from the Santa Cruz metro station. Null values were eliminated and scraped text values were cleaned to remove irrelevant information. Wait times for each bus arrival record were created by finding the difference for that bus and route's scheduled arrival time.

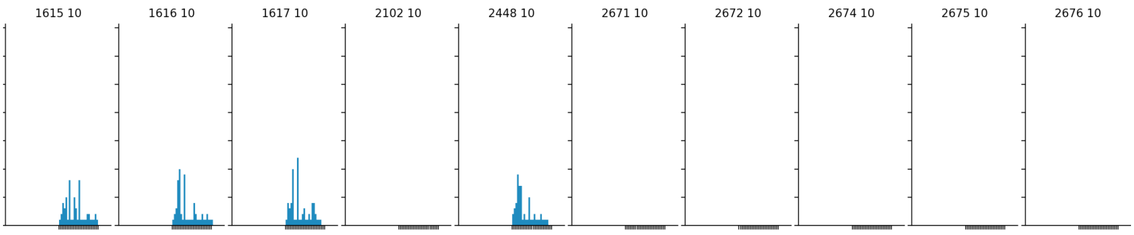
metro_bus_with_schedule_poisson.csv

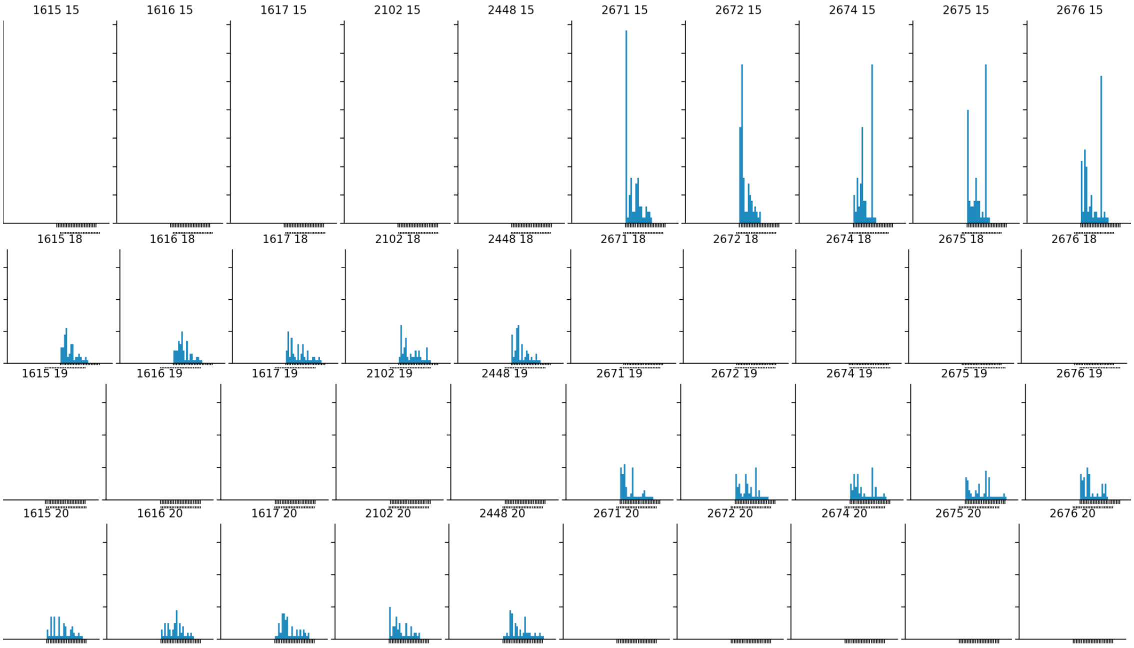
This data set was created from the above data set with a resulting 6 features. These in order are: time of arrival, zone_id, dayofweek, bus_type, and minutes_late. The meaning of each feature is explained below.

We determined it pertinent that location and type of bus would become important indicators of lateness alongside their arrival time. In order to find the zones, campus bus stops ids were mapped to 1 of 6 zones. Each zone corresponds to the nearest classroom locations for that location and tries to encode where students would get off on the bus to go to a specific class. For instance, students with classes in Baskin Engineering would most likely, if taking the bus to get to class, get off at a Science Hill bus stop. This meant that stop ids 1615 a 2674 would map to zone id 4. The zone divisions are shown below.



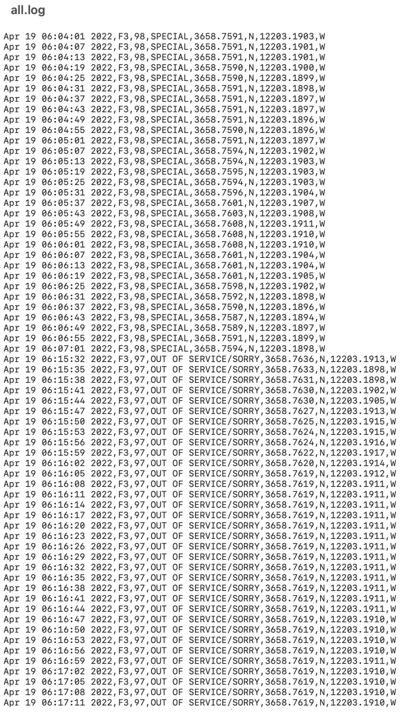
The data set for metro bus's was found to also be triangularly or exponentially shaped in distribution. Using the Python Seaborn library, bus frequency was plotted against bus lateness, or how many minutes the bus was late. This created a histogram with at most 13 buses belong to a single bin and 0 being the minimum amount. The distribution ranged from 0 to 59 minutes late with, 30 minutes being the halfway point. A histogram was created for every stop and for every bus type. It can clearly be seen that the eastbound routes (on the left half) tend to range in lateness from 0 to 30 minutes while the westbound routes on the right seem to have two peaks of lateness: one group that is a little late (under 30 minutes) and another group that is very late (over 30 minutes).





3.2 Loop

Kerry Veenstra provided a log file that contained the actual arrival times of the different loop buses on campus. The data was collected from three stations: Taps/Base, Opers, and College 8(Rachel Carson). Veenstra collected data for a week. The log file contained the date and time, bus ID, bus route (loop, special, out of service, etc), and latitude and longitude, as shown in Figure 1.



all.log

```
Apr 19 06:04:01 2022,F3,98,SPECIAL,3658.7591,N,12203.1903,W
Apr 19 06:04:07 2022,F3,98,SPECIAL,3658.7591,N,12203.1901,W
Apr 19 06:04:13 2022,F3,98,SPECIAL,3658.7591,N,12203.1901,W
Apr 19 06:04:19 2022,F3,98,SPECIAL,3658.7590,N,12203.1900,W
Apr 19 06:04:25 2022,F3,98,SPECIAL,3658.7590,N,12203.1899,W
Apr 19 06:04:31 2022,F3,98,SPECIAL,3658.7591,N,12203.1898,W
Apr 19 06:04:37 2022,F3,98,SPECIAL,3658.7591,N,12203.1897,W
Apr 19 06:04:43 2022,F3,98,SPECIAL,3658.7591,N,12203.1897,W
Apr 19 06:04:49 2022,F3,98,SPECIAL,3658.7591,N,12203.1896,W
Apr 19 06:04:55 2022,F3,98,SPECIAL,3658.7590,N,12203.1896,W
Apr 19 06:05:01 2022,F3,98,SPECIAL,3658.7591,N,12203.1897,W
Apr 19 06:05:07 2022,F3,98,SPECIAL,3658.7594,N,12203.1902,W
Apr 19 06:05:13 2022,F3,98,SPECIAL,3658.7594,N,12203.1903,W
Apr 19 06:05:19 2022,F3,98,SPECIAL,3658.7595,N,12203.1903,W
Apr 19 06:05:25 2022,F3,98,SPECIAL,3658.7594,N,12203.1903,W
Apr 19 06:05:31 2022,F3,98,SPECIAL,3658.7596,N,12203.1904,W
Apr 19 06:05:37 2022,F3,98,SPECIAL,3658.7601,N,12203.1907,W
Apr 19 06:05:43 2022,F3,98,SPECIAL,3658.7603,N,12203.1908,W
Apr 19 06:05:49 2022,F3,98,SPECIAL,3658.7608,N,12203.1911,W
Apr 19 06:05:55 2022,F3,98,SPECIAL,3658.7608,N,12203.1910,W
Apr 19 06:06:01 2022,F3,98,SPECIAL,3658.7608,N,12203.1910,W
Apr 19 06:06:07 2022,F3,98,SPECIAL,3658.7601,N,12203.1904,W
Apr 19 06:06:13 2022,F3,98,SPECIAL,3658.7601,N,12203.1904,W
Apr 19 06:06:19 2022,F3,98,SPECIAL,3658.7601,N,12203.1905,W
Apr 19 06:06:25 2022,F3,98,SPECIAL,3658.7598,N,12203.1902,W
Apr 19 06:06:31 2022,F3,98,SPECIAL,3658.7592,N,12203.1898,W
Apr 19 06:06:37 2022,F3,98,SPECIAL,3658.7590,N,12203.1896,W
Apr 19 06:06:43 2022,F3,98,SPECIAL,3658.7587,N,12203.1894,W
Apr 19 06:06:49 2022,F3,98,SPECIAL,3658.7589,N,12203.1897,W
Apr 19 06:06:55 2022,F3,98,SPECIAL,3658.7591,N,12203.1899,W
Apr 19 06:07:01 2022,F3,98,SPECIAL,3658.7594,N,12203.1898,W
Apr 19 06:15:32 2022,F3,97,OUT OF SERVICE/SORRY,3658.7636,N,12203.1913,W
Apr 19 06:15:35 2022,F3,97,OUT OF SERVICE/SORRY,3658.7633,N,12203.1898,W
Apr 19 06:15:38 2022,F3,97,OUT OF SERVICE/SORRY,3658.7631,N,12203.1898,W
Apr 19 06:15:41 2022,F3,97,OUT OF SERVICE/SORRY,3658.7630,N,12203.1902,W
Apr 19 06:15:44 2022,F3,97,OUT OF SERVICE/SORRY,3658.7630,N,12203.1905,W
Apr 19 06:15:47 2022,F3,97,OUT OF SERVICE/SORRY,3658.7627,N,12203.1913,W
Apr 19 06:15:50 2022,F3,97,OUT OF SERVICE/SORRY,3658.7625,N,12203.1915,W
Apr 19 06:15:53 2022,F3,97,OUT OF SERVICE/SORRY,3658.7624,N,12203.1915,W
Apr 19 06:15:56 2022,F3,97,OUT OF SERVICE/SORRY,3658.7624,N,12203.1916,W
Apr 19 06:15:59 2022,F3,97,OUT OF SERVICE/SORRY,3658.7622,N,12203.1917,W
Apr 19 06:16:02 2022,F3,97,OUT OF SERVICE/SORRY,3658.7620,N,12203.1914,W
Apr 19 06:16:05 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1912,W
Apr 19 06:16:08 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:11 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:14 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:17 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:20 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:23 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:26 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:29 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:32 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:35 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:38 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:41 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:44 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:16:47 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:16:50 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:16:53 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:16:56 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:16:59 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1911,W
Apr 19 06:17:02 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:17:05 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:17:08 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
Apr 19 06:17:11 2022,F3,97,OUT OF SERVICE/SORRY,3658.7619,N,12203.1910,W
```

Figure 1: log file provided by Kerry Veenstra

I then put the log file into Excel, where columns and rows organized the data. This original, uncleaned data has 8 columns and over $98,000$ rows.

All entries containing "OUT OF SERVICE" were removed, as well as all other routes besides entries labeled as "LOOP". The sheet was then ordered based on "Bus ID". We simplified the latitude and longitude and assigned them to the three zones: Opers, Taps/Base, and College 8. A "Time Only" column was created, as well as "Day Only", and scheduled departure was added. The Loop Website from the University states that buses depart from the base every $10 - 20$ minutes starting at 7:25am. We assume, for consistency and simplicity, that buses depart every 15 minutes. This scheduled departure time was assigned to each record that displayed a bus departing from Taps/Base. Then, a new column was created to show the difference between actual and scheduled departure. Positive values indicate that the bus was late, and negative values indicate that the bus was early. The sheet was then filtered to have only entries that contained a "Minutes Late" value; entries that had a blank entry for the difference of actual and scheduled departure were removed. The difference of actual departure time and scheduled time is calculated, and from this we find the average minutes late of the loop bus: 2.81, or about 3 minutes. Here is the cleaned data set.

We set the Taps/Base zone as shown in the Figure 2.

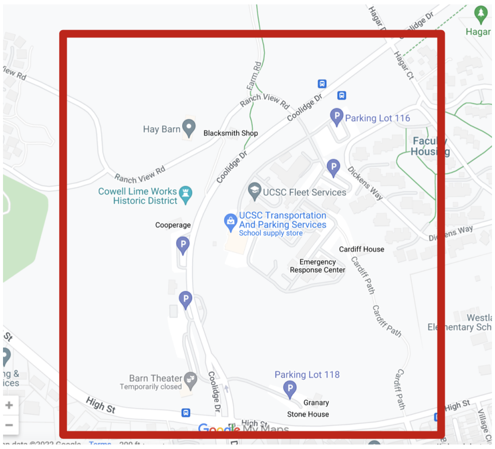


Figure 2: Taps Zone

To visualize any patterns, we created a scatter plot of the loop buses. We can see two patterns. First, there seems to be three groups of loops based on minutes late. There is a group of buses that are early, a group that is late, and a group that is even later. The second pattern shown is that there seems to be three peaks during the day.

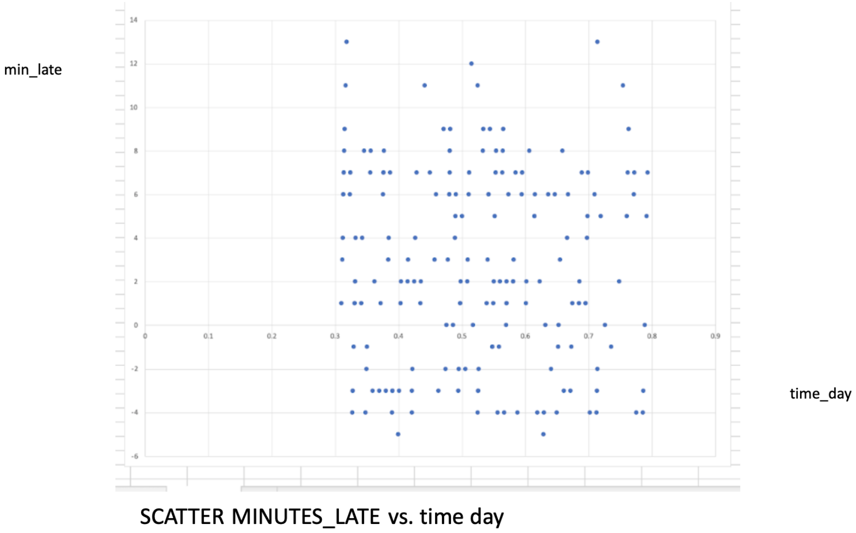


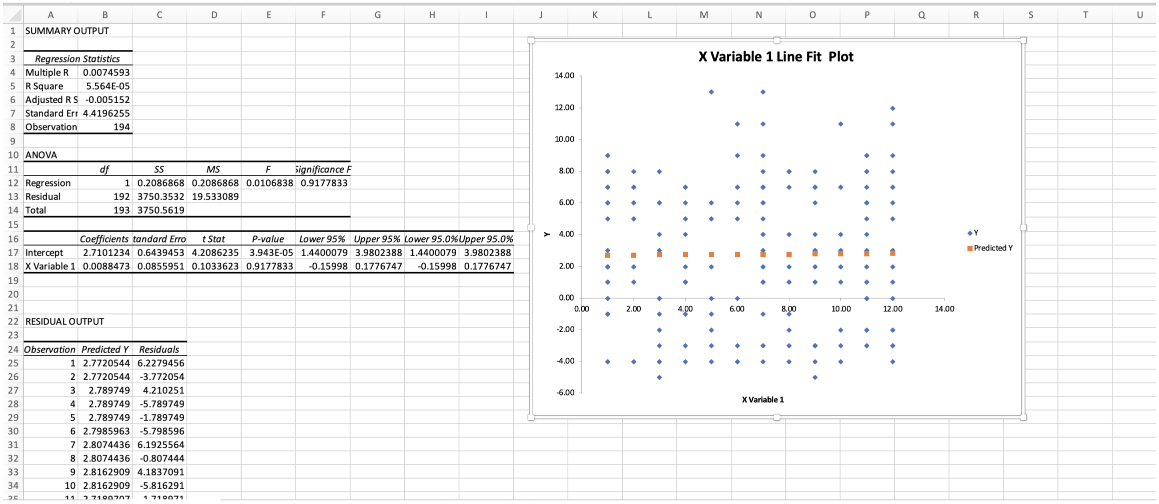
Figure 3: Scatter Plot

4 Solution Description

Three models were explored in this project. Each model after the first is meant to be an improvement on the last as we tried to employ an iterative process in our methodology. Each is described below.

4.1 Linear Regression

The first model that was used was Linear Regression. This was computed in Excel.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.0074593
R Square	5.564E-05
Adjusted R S	-0.005152
Standard Err	4.4196255
Observation	194

ANOVA

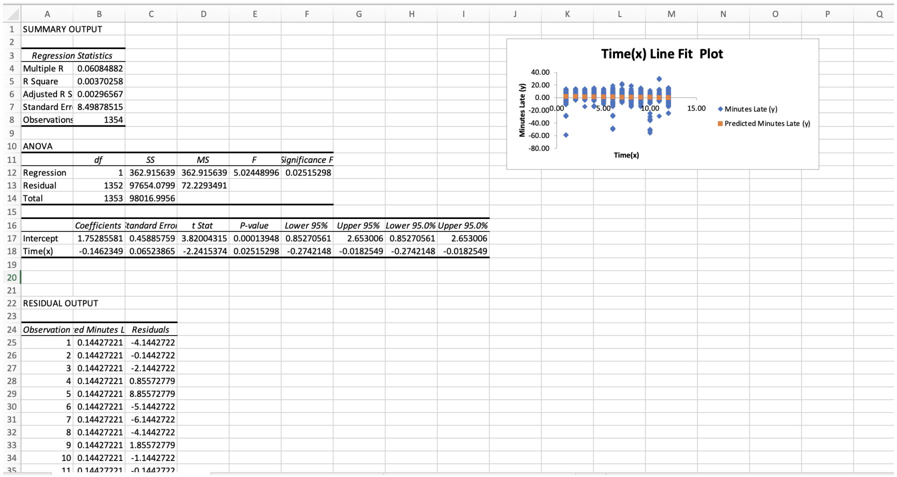
	df	SS	MS	F	Significance F
Regression	1	0.2086868	0.2086868	0.0106838	0.9177833
Residual	192	3750.3532	19.533089		
Total	193	3750.5619			

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2.7101234	0.6439453	4.2086235	3.943E-05	1.4400079	3.9802388	1.4400079	3.9802388
X Variable 1	0.0088473	0.0855951	0.1033623	0.9177833	-0.15998	0.1776747	-0.15998	0.1776747

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals
1	2.7720544	6.2279456
2	2.7720544	-3.772054
3	2.789749	4.210251
4	2.789749	-5.789749
5	2.789749	-1.789749
6	2.7985963	-5.798596
7	2.8074436	6.1925564
8	2.8074436	-0.807444
9	2.8162909	4.1837091
10	2.8162909	-5.816291
11	2.7189707	1.718971

Notice that our R^2 is very small. This small R^2 implies that time has a small affect on the variation in minutes late, which we know is not true.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.06084882
R Square	0.00370258
Adjusted R S	0.00296567
Standard Err	8.49878515
Observations	1354

ANOVA

	df	SS	MS	F	Significance F
Regression	1	362.915639	362.915639	5.02448996	0.02515298
Residual	1352	97654.0799	72.2293491		
Total	1353	98016.9956			

	Coefficients	Standard Erroi	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1.75285581	0.45885759	3.82004315	0.00013948	0.85270561	2.653006	0.85270561	2.653006
Time(x)	-0.1462349	0.06523865	-2.2415374	0.02515298	-0.2742148	-0.0182549	-0.2742148	-0.0182549

RESIDUAL OUTPUT

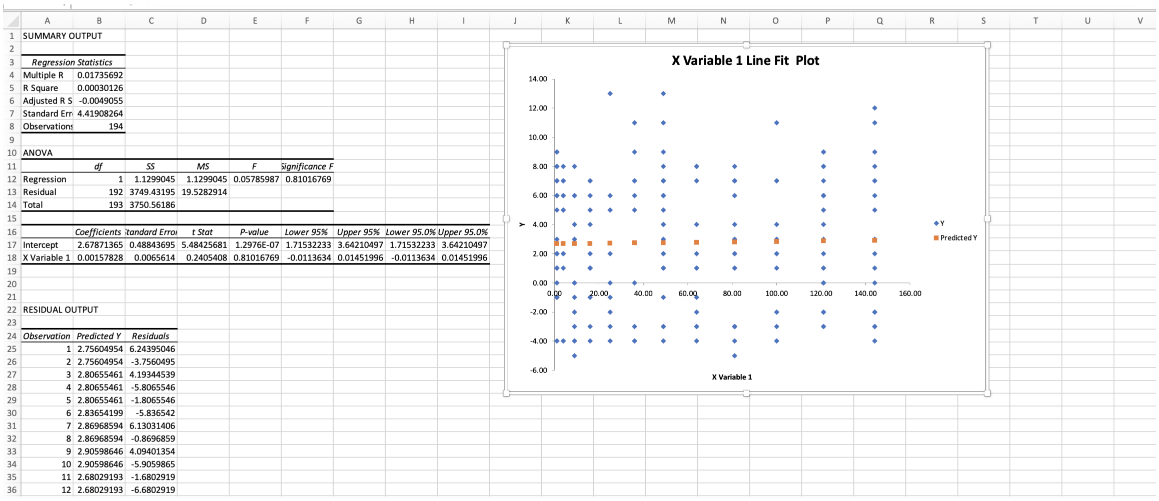
Observation	ed Minutes L	Residuals
1	0.14427221	-4.1442722
2	0.14427221	-0.1442722
3	0.14427221	-2.1442722
4	0.14427221	0.85572779
5	0.14427221	8.85572779
6	0.14427221	-5.1442722
7	0.14427221	-6.1442722
8	0.14427221	-4.1442722
9	0.14427221	1.85572779
10	0.14427221	-1.1442722
11	0.14427221	-0.1442722

Again, notice that our R^2 is very low.

Thus we need a model that produces a larger R^2.

4.2 Quadratic Regression

The next model that was tested was Quadratic Regression. This was also computed in Excel.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.01735692
R Square	0.00030126
Adjusted R S	-0.0049055
Standard Err	4.41908264
Observations	194

ANOVA

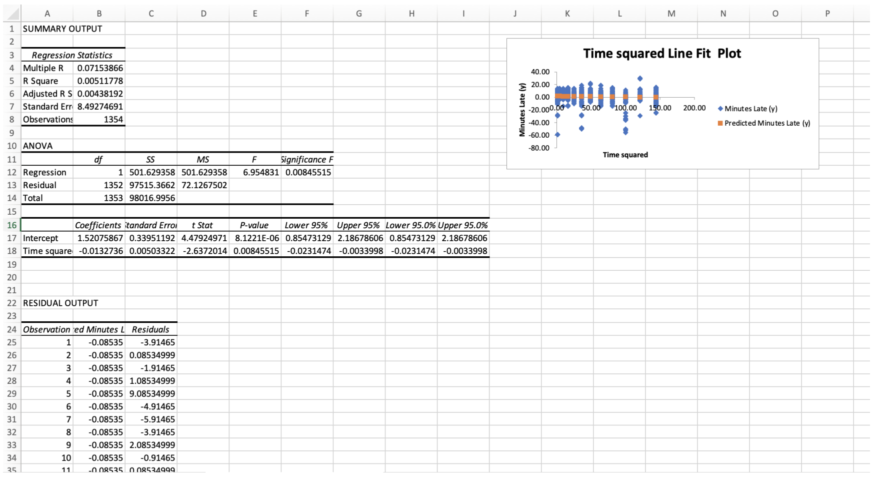
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.1299045	1.1299045	0.05785987	0.81016769
Residual	192	3749.43195	19.5282914		
Total	193	3750.56186			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.67871365	0.48843695	5.48425681	1.2976E-07	1.71532233	3.64210497	1.71532233	3.64210497
X Variable 1	0.00157828	0.0065614	0.2405408	0.81016769	-0.0113634	0.01451996	-0.0113634	0.01451996

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	2.75604954	6.24395046
2	2.75604954	-3.7560495
3	2.80655461	4.19344539
4	2.80655461	-5.8065546
5	2.80655461	-1.8065546
6	2.83654199	-5.836542
7	2.86968594	6.13031406
8	2.86968594	-0.8696859
9	2.90598646	4.09401354
10	2.90598646	-5.9059865
11	2.68029193	-1.6802919
12	2.68029193	-6.6802919

Note that our R^2 again is very small.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.07153866
R Square	0.00511778
Adjusted R S	0.00438192
Standard Err	8.49274691
Observations	1354

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	501.629358	501.629358	6.954831	0.00845515
Residual	1352	97515.3662	72.1267502		
Total	1353	98016.9956			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.52075867	0.33951192	4.47924971	8.1221E-06	0.85473129	2.18678606	0.85473129	2.18678606
Time square	-0.0132736	0.00503322	-2.6372014	0.00845515	-0.0231474	-0.0033998	-0.0231474	-0.0033998

RESIDUAL OUTPUT

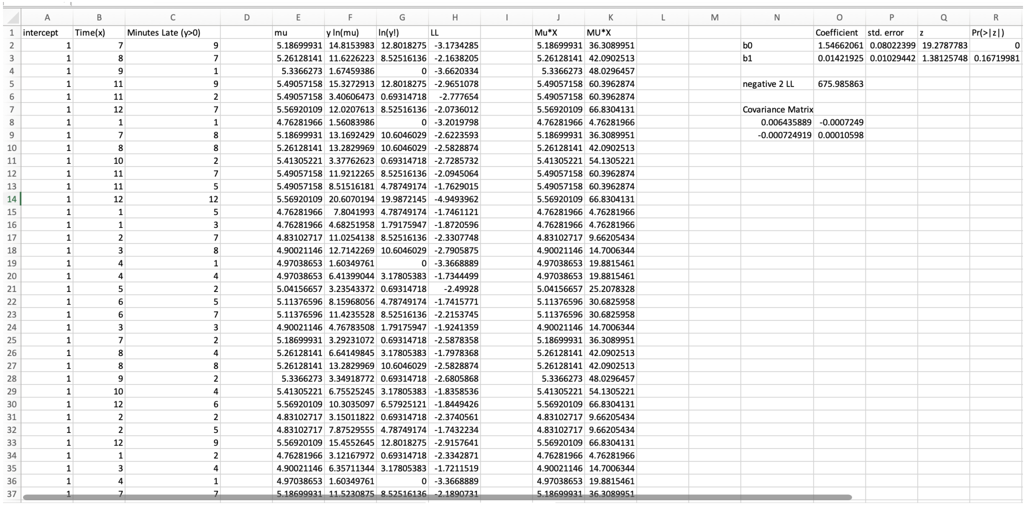
Observation	*ed Minutes L*	*Residuals*
1	-0.08535	-3.91465
2	-0.08535	0.08534999
3	-0.08535	-1.91465
4	-0.08535	1.08534999
5	-0.08535	9.08534999
6	-0.08535	-4.91465
7	-0.08535	-5.91465
8	-0.08535	-3.91465
9	-0.08535	2.08534999
10	-0.08535	-0.91465
11	-0.08535	0.08534999

The metro quadratic regression also gives a very low R^2 value, and so we explore a different model.

4.3 Poisson Regression

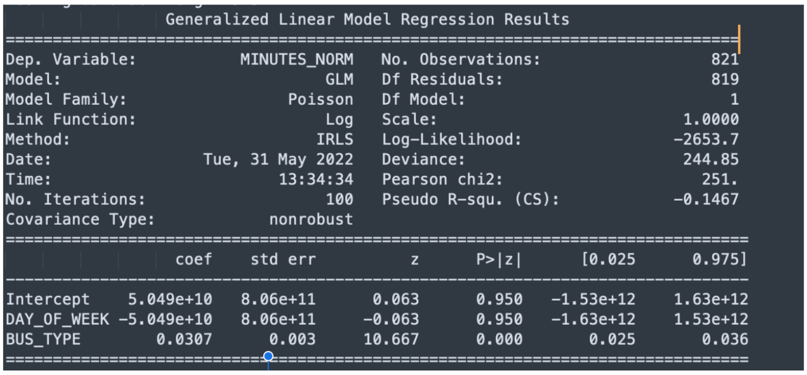
We assume for the Poisson Regression model, that the buses departing are independent events (in reality, this is not the case). We also assume that our variables must be counts, which means that we only consider positive minutes late values. In order to also display the early values, we would need to separate models.

For the loop buses, Poisson regression was computed in Excel.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	intercept	Time(x)	Minutes Late (y>0)		mu	y ln(mu)	ln(y!)	LL		Mu*X	MU*X				Coefficient	std. error	z	Pr(>|z|)
2	1	7	9		5.18699931	14.8153983	12.8018275	-3.1734285		5.18699931	36.3089951			b0	1.54662061	0.08022399	19.2787783	0
3	1	8	7		5.26128141	11.6226223	8.52516136	-2.1638205		5.26128141	42.0902513			b1	0.01421925	0.01029442	1.38125748	0.16719981
4	1	9	1		5.3366273	1.67459386	0	-3.6620334		5.3366273	48.0296457							
5	1	11	9		5.49057158	15.3272913	12.8018275	-2.9651078		5.49057158	60.3962874			negative 2 LL	675.985863			
6	1	11	2		5.49057158	3.40606473	0.69314718	-2.777654		5.49057158	60.3962874							
7	1	12	7		5.56920109	12.0207613	8.52516136	-2.0736012		5.56920109	66.8304131			Covariance Matrix				
8	1	1	1		4.76281966	1.56083986	0	-3.2019798		4.76281966	4.76281966			0.006435889	-0.0007249			
9	1	7	8		5.18699931	13.1692429	10.6046029	-2.6223593		5.18699931	36.3089951			-0.000724919	0.00010598			
10	1	8	8		5.26128141	13.2829969	10.6046029	-2.5828874		5.26128141	42.0902513							
11	1	10	2		5.41305221	3.37762623	0.69314718	-2.7285732		5.41305221	54.1305221							
12	1	11	7		5.49057158	11.9212265	8.52516136	-2.0945064		5.49057158	60.3962874							
13	1	11	5		5.49057158	8.51516181	4.78749174	-1.7629015		5.49057158	60.3962874							
14	1	12	12		5.56920109	20.6070194	19.9872145	-4.9493962		5.56920109	66.8304131							
15	1	1	5		4.76281966	7.8041993	4.78749174	-1.7461121		4.76281966	4.76281966							
16	1	1	3		4.76281966	4.68251958	1.79175947	-1.8720596		4.76281966	4.76281966							
17	1	2	7		4.83102717	11.0254138	8.52516136	-2.3307748		4.83102717	9.66205434							
18	1	3	8		4.90021146	12.7142269	10.6046029	-2.7905875		4.90021146	14.7006344							
19	1	4	1		4.97038653	1.60349761	0	-3.3668889		4.97038653	19.8815461							
20	1	4	4		4.97038653	6.41399044	3.17805383	-1.7344499		4.97038653	19.8815461							
21	1	5	2		5.04156657	3.23543372	0.69314718	-2.49928		5.04156657	25.2078328							
22	1	6	5		5.11376596	8.15968056	4.78749174	-1.7415771		5.11376596	30.6825958							
23	1	6	7		5.11376596	11.4235528	8.52516136	-2.2153745		5.11376596	30.6825958							
24	1	3	3		4.90021146	4.76783508	1.79175947	-1.9241359		4.90021146	14.7006344							
25	1	7	2		5.18699931	3.29231072	0.69314718	-2.5878358		5.18699931	36.3089951							
26	1	8	4		5.26128141	6.64149845	3.17805383	-1.7978368		5.26128141	42.0902513							
27	1	8	8		5.26128141	13.2829969	10.6046029	-2.5828874		5.26128141	42.0902513							
28	1	9	2		5.3366273	3.34918772	0.69314718	-2.6805868		5.3366273	48.0296457							
29	1	10	4		5.41305221	6.75525245	3.17805383	-1.8358536		5.41305221	54.1305221							
30	1	12	6		5.56920109	10.3035097	6.57925121	-1.8449426		5.56920109	66.8304131							
31	1	2	2		4.83102717	3.15011822	0.69314718	-2.3740561		4.83102717	9.66205434							
32	1	2	5		4.83102717	7.87529555	4.78749174	-1.7432234		4.83102717	9.66205434							
33	1	12	9		5.56920109	15.4552645	12.8018275	-2.9157641		5.56920109	66.8304131							
34	1	1	2		4.76281966	3.12167972	0.69314718	-2.3342871		4.76281966	4.76281966							
35	1	3	4		4.90021146	6.35711344	3.17805383	-1.7211519		4.90021146	14.7006344							
36	1	4	1		4.97038653	1.60349761	0	-3.3668889		4.97038653	19.8815461							
37	1	7	7		5.18699931	11.5230875	8.52516136	-2.1890731		5.18699931	36.3089951							

This was additionally implemented in Python, where a Poisson regression was created using the sci-kit learn library.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            MINUTES_NORM   No. Observations:                  821
Model:                             GLM   Df Residuals:                      819
Model Family:                  Poisson   Df Model:                            1
Link Function:                     Log   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -2653.7
Date:                 Tue, 31 May 2022   Deviance:                       244.85
Time:                         13:34:34   Pearson chi2:                     251.
No. Iterations:                    100   Pseudo R-squ. (CS):            -0.1467
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    5.049e+10   8.06e+11      0.063      0.950   -1.53e+12    1.63e+12
DAY_OF_WEEK -5.049e+10   8.06e+11     -0.063      0.950   -1.63e+12    1.53e+12
BUS_TYPE       0.0307      0.003     10.667      0.000       0.025       0.036
==============================================================================
```

5 Result Analysis

The loop bus is, on average, 2.81 or about 3 minutes late. On average, the metro bus will be 16.71 or about 17 minutes late.

6 Validation Analysis

The metrics we used for our validation includes R squared, Chi squared, and log likelihood. R squared is a technique that measures the variance of a dependent variable that's explained by an independent variable. An R squared above 0.9 is a good standard to show high correlation. This is an ratio of the residuals over the total dispersion of points from their mean value. Chi squared is another statistical test and tries to show if the relationships are significant. The larger the Chi squared value the greater the differences are not by chance between what is expected and what is observed. Finally log likelihood is another goodness of fit test with higher values showing better fit.

z-score is also used to understand the distance from the mean and give our coefficients some

6.1 Linear Regression

R squared for the loop was 0.00005 and for the metro was 0.003. These values are very small, which implied that time of day did not heavily influence lateness.

6.2 Quadratic Regression

Again, R squared for the loop was 0.0003 and for the metro were 0.005. These values were again incredibly small, and there was no improvement between the linear and quadratic models.

6.3 Poisson Regression

For the final model, we analyze chi squared and log likelihood for the python implementation and the z score for coefficient values in the loop.

For the Python Poisson we see that Chi squared is 251, a very large number, with the log likelihood being -2653. Due to the very low log likelihood it seems unlikely that this implementation was done correctly. More work needs to be done to refine this implementation to improve these scores and well fit the data.

The loop bus on the other hand seems to have a log likelihood score of 657. This could imply that the time for this simple case and the regression is in fact a good fit. However

7 Original Goals and Goals Met

Our original goals were:

- Identify interesting factors that could contribute to bus lateness: class sizes, class times, number of classes, weather, bus direction, time, bus density (arrival rate in a set time interval), number of lights and stops between routes.
- Determine what factors cause the longest delays and how to classify wait times for buses with certain factors
- Calculate the average wait times per stop – Maybe there's a strategy to consistently get the bus at a reasonable time

The goals that were met included:

- Calculate and analyze trends in average minutes late of the loop and metro buses
- Identify interesting factors that could contribute lateness
- Clean data
- Find and finish model

8 Original Timeline and End Result

Our original timeline included:

- By the end of the week (Sunday, the 24th) get all the data (class schedule and metro bus(all campus metro)-week's worth, and possibly historical data for loop-if professor responds in time), all campus stops and metro station stop.
- (4/25 - 5/1) Next week, goals are to process data, combine data, and possibly normalize it and work on report. Necessary to get through this to know what model we are going to use and have all members learn it. (start data section of report in meantime.)
- (5/2 - 5/16) work on model and visualization
- (5/17 - 5/23) Week of 23rd, combine all findings into analysis
- Report and presentation done by finals week

Collecting and cleaning data took more time than was originally planned.

9 Resources Used

- Loop website
- Metro website
- Excel
- Python
- Kerry Veenstra

10 Conclusion

We conclude that the mean minutes late of the metro bus is 16.7 minutes, and the mean for the loop bus is 2.81 minutes. Average lateness could be explained by variables such as time of day, day of the week, and bus type. More work is needed to create more robust models. Poisson Regression seems to be the most accurate model so far, but more exploration in this type of model is needed to make more accurate predictions.

In the future, we hope to improve our implementation. This would include possibly making an application, as well as exploring more models. It would also be beneficial to include more bus arrival data (not just spring week in April). Lastly, we would like to consider other input factors such as how many students drive, the average traffic including index (cars), and how many students live on/off campus.

11 References

Kerry Veenstra
`https://taps.ucsc.edu/buses-shuttles/campus-shuttles.html`
`https://online.stat.psu.edu/stat501/lesson/15/15.4`
`https://www.youtube.com/watch?v=zsRyEJUDPvc&t=339s`
`https://timeseriesreasoning.com/contents/poisson-regression-model/`
`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243514/#:~:text=Traffic`

12 Git Repo