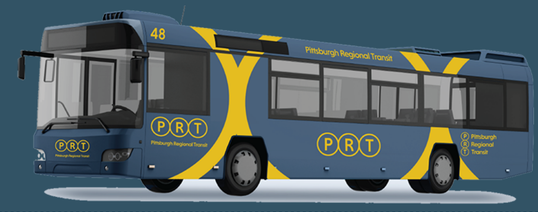
Does PRT Timeliness Affect Ridership?

Yewon Kim, Neha Dutt, Angelina Jia,
Mahitha Ramachandran

Mock Data Jam Project
Fall 2023



Question & Hypothesis

Question: Does the timeliness of a PRT bus route on a specific day affect the ridership on that bus? Is there any correlation between timeliness and average ridership?

Hypothesis: The more "on time" a Pittsburgh Regional Transit bus is for a given route on the first of each month in 2022, the greater the average ridership would be for that specific bus route that month.

Data

Data Source: Western Pennsylvania Regional Data Center & Pittsburgh Regional Transit

- Port Authority Monthly On-Time Performance by Route
- Monthly Average Ridership by Route & Weekday

Methods

Data Filtering: Excel

- Filtered the dataset to show data from weekdays in 2022 only
- Bus Route Selection : Randomly selected two sets of 5 bus routes

Note: Timeliness is measured in the dataset by a percentage so that the closer the values are to 1 the less the difference was between expected and actual arriving time

Scatter plots: R

- Created a scatter plot for each set grouped (color-coded) by bus route
- Added a linear regression line

Regression analysis: Excel

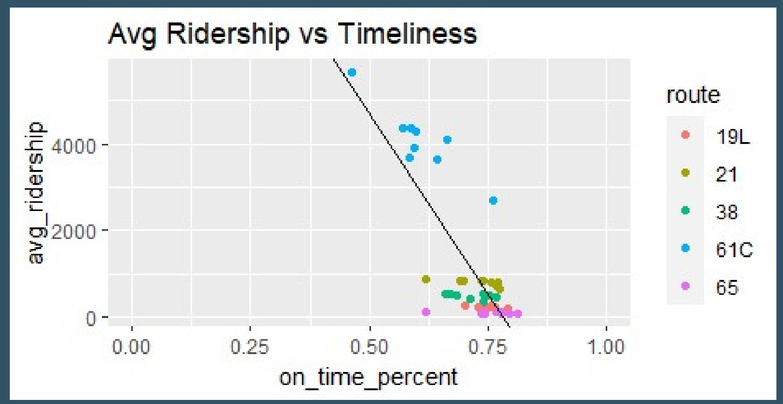
- Simple linear regression run for each set and later for altered sets

Challenges

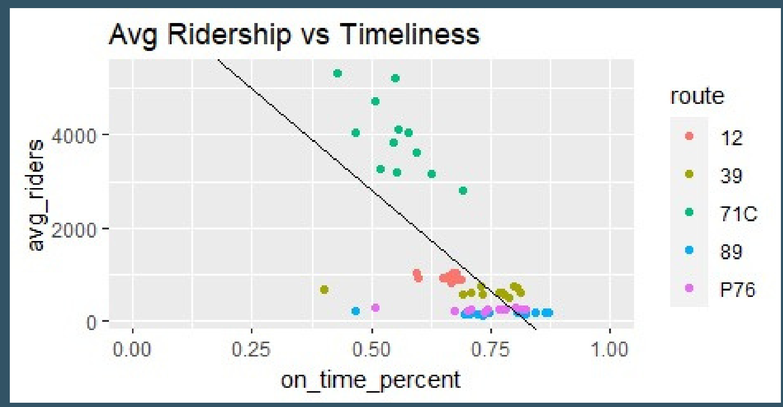
The first challenge we ran into was handling the enormous data sets, which included timeliness and ridership data for every single PRT route for every month since January of 2017. Early on, we decided that focusing only on one year of data (2022) and bus routes was best. We attempted to run data analysis using R on the entire data set of bus routes in 2022, but kept running into issues because of the sheer size of the data. Our solution was to take a random sample of the bus routes and analyze that instead, which we ended up doing twice to further interrogate our initial findings. The next challenge was understanding what we could and could not conclude from the analysis, which is discussed in the Conclusions section.

Results

	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	Regression Statistics									
4	Multiple R	0.799294								
5	R Square	0.638871								
6	Adjusted I	0.630472								
7	Standard I	938.9082								
8	Observatic	45								
9										
10	ANOVA									
11		df	SS	MS	F	gnificance F				
12	Regressio	1	67060218	67060218	76.07093	4.62E-11				
13	Residual	43	37906588	881548.6						
14	Total	44	1.05E+08							
15										
16		Coefficients	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%	
17	Intercept	13004.72	1367.878	9.507218	3.91E-12	10246.13	15763.31	10246.13	15763.31	
18	X Variable	-16692	1913.806	-8.72186	4.62E-11	-20551.5	-12832.4	-20551.5	-12832.4	
19										



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.666101							
5	R Square	0.443691							
6	Adjusted R	0.4341							
7	Standard E	1098.059							
8	Observatio	60							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*ignificance F*			
12	Regression	1	55775582	55775582	46.25862	6.31E-09			
13	Residual	58	69932556	1205734					
14	Total	59	1.26E+08						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	7130.894	884.6607	8.060598	4.85E-11	5360.053	8901.735	5360.053	8901.735
18	on_time_p	-8667.26	1274.34	-6.80137	6.31E-09	-11218.1	-6116.39	-11218.1	-6116.39
19									



Conclusions

The first random set of bus routes displayed a **negative relationship**, suggesting higher on-time percent correlated with lower average ridership. This contradicts our original hypothesis. Thus, we decided to take another random set of 5 bus routes to see if this relationship held. As seen in the graph, it did, and in both cases the **p-value for the slopes were very low**, meaning it would be very unlikely to get this outcome by chance. However, by grouping and color-coding the bus routes, it becomes clear that this relationship looks, in both sets, mostly set by one of the routes (61C in the first set and 71C in the second). If we remove these two routes from the sets, the regression slopes become -1937 and -1183, respectively, and their R square values are around 0.1, meaning that there are very weak relationships between the two variables. The 61C and 71C are among the busiest PRT bus routes, with much higher average ridership than the other routes in their sets, meaning the negative relationship that they display is noteworthy, but perhaps not representative of all routes. Thus, it would be remiss to exclude them. We can instead hypothesize that **perhaps higher ridership actually contributes to lower timeliness because of the time it takes for people to enter and exit the bus**. Further data and analysis would be required to make concrete conclusions on this, as we can make no conclusion about causation from our data. However, our analysis does indicate that there is some statistical significance in relating average ridership and timeliness of PRT bus routes. The implications of this are vast and call for further analysis, especially amidst upcoming changes to the transit routes and current issues surrounding public transit.