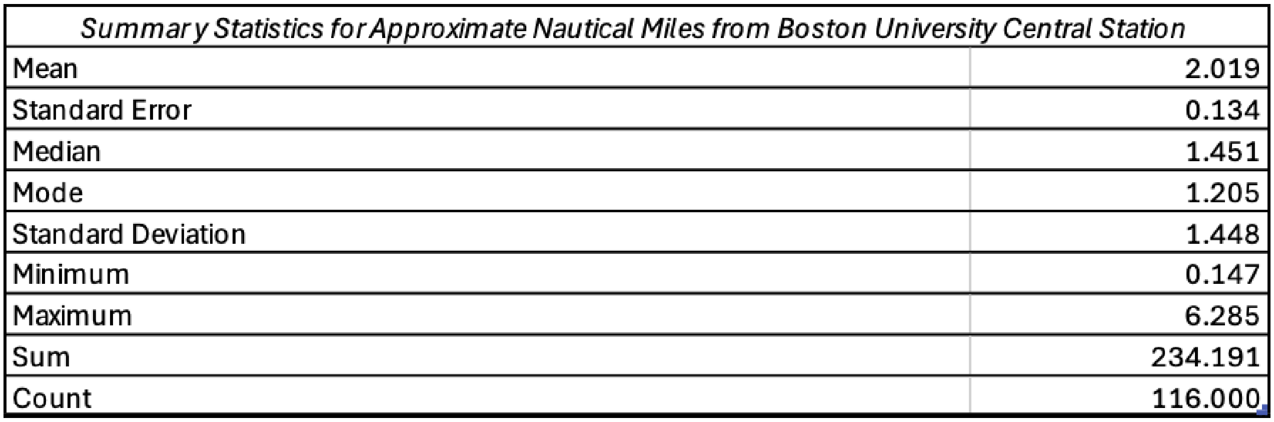
BA223 Project Team 11 - Rebekah Lewi, Oliver Coleman, Tanisha Agrawal, Vivaan Mehta

PART A

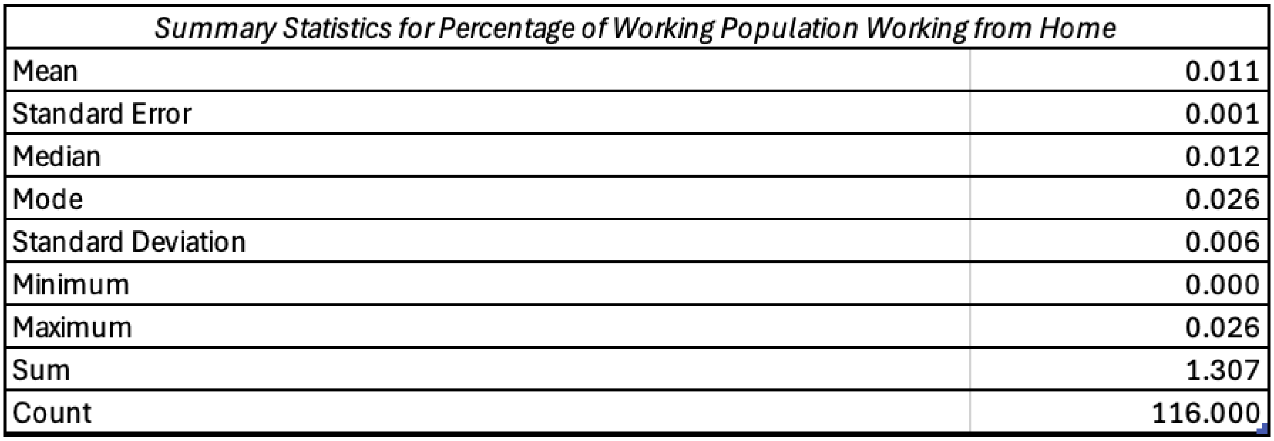
The two variables we have decided to create for this data set are “Approximate Nautical Miles from Boston University Central Station” and “Percentage of Working Population Working from Home”. The variable “Approximate Nautical Miles from Boston University Central Station”, abbreviated as “Miles BUC”, was calculated using the distance formula provided to us by Professor Macgarvie, =ACOS((SIN(C2*PI()/180)*SIN(42.34989*PI()/180)+COS(C2*PI()/180)*COS(42.34989*PI()/180)*COS((-71.1068)*PI()/180-D2*PI()/180))) *3443.8985. The values 42.34989 and -71.1068 are the longitude and latitude coordinates respectively for the Green Line Station Boston University Central which we used as a midpoint for approximating the coordinates for Boston University as a whole. We predicted that as the distance between the start station and Boston University increased the number of total Blue Bike trips would likely decrease. This is because the stations that are close to BU are in a densely populated area and are likely to have lots of short local trips such as students and staff traveling to and from class. Additionally, the correlation value for this variable and trips was -.67, which not only supports our prediction of a negative predictor but was also the fourth strongest correlation in the entire dataset. The variable “Percentage of Working Population Working from Home” was calculated using two of the variables already provided in the dataset. We effectively created a better variable for at home workers by dividing the number of people who worked from home by the total working population over the age of 16 to create a percentage. We predicted that as the percentage of at home workers increased, the number of total trips from that start station would decrease. This is because areas with a higher concentration of remote workers are losing out on commuting trips which we expected to make up many of the total trips. Again, the correlation value for this variable confirmed our priors with a negative relationship.

Table 1. Summary Statistics for Approximate Nautical Miles from BU Central Station



Summary Statistics for Approximate Nautical Miles from Boston University Central Station	
Mean	2.019
Standard Error	0.134
Median	1.451
Mode	1.205
Standard Deviation	1.448
Minimum	0.147
Maximum	6.285
Sum	234.191
Count	116.000

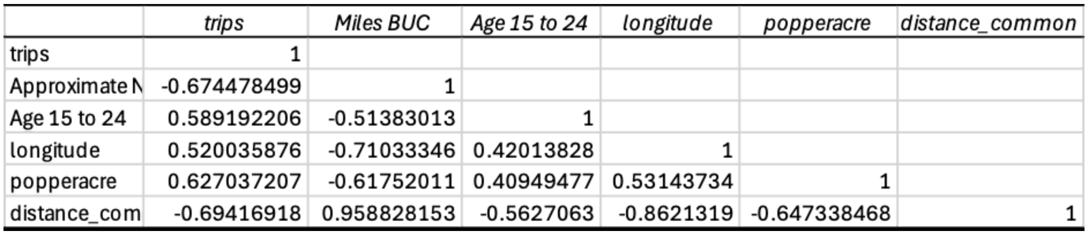
Table 2. Summary Statistics for Percentage of Working Population Working from Home



Summary Statistics for Percentage of Working Population Working from Home	
Mean	0.011
Standard Error	0.001
Median	0.012
Mode	0.026
Standard Deviation	0.006
Minimum	0.000
Maximum	0.026
Sum	1.307
Count	116.000

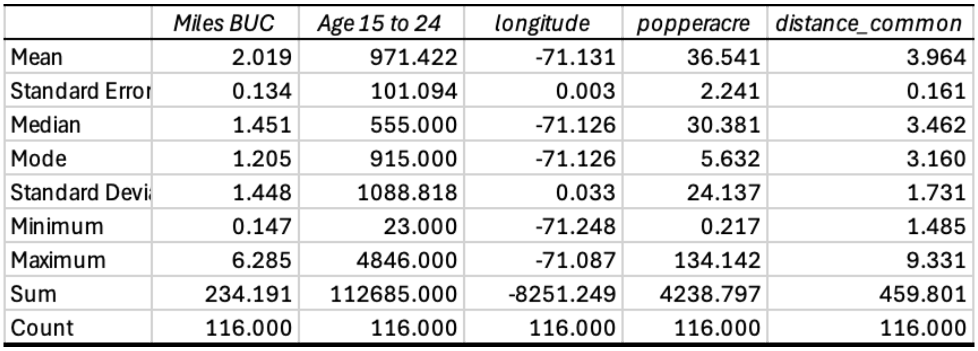
Based on this information, as well as the correlation coefficients for the rest of the dataset, we initially predicted that the 5 most important variables for predicting the number of trips would be: Approximate Nautical Miles from Boston University Central Station, Percentage of Working Population Working from Home, Ages 15 to 24, Population Per Acre, and Distance from Boston Common. We believed that the variables we created would both have negative relationships for the reasons we already explained. For the variable “Ages 15 to 24” we combined two variables that were already in the dataset to create a new variable with an extremely high and positive correlation. For this reason, we expected trips to increase for those areas with a higher number of people ages 15 to 24 years old. As for Population Per Acre, we expected that more densely populated areas would have a higher number of short trips and therefore higher total trips. Lastly, we expected that as the distance from Boston Common increased, the total number of trips would decrease, because areas closer to Boston Common have a higher amount of commercial activity and therefore more reason for pedestrians to bike around. When we ran an initial regression we came out with an Adjusted R Squared value of .581. The variables Approximate Nautical Miles from Boston University Central Station, Percentage of Working Population Working from Home, and Distance from Boston Common all came out with insignificant p-values. We hypothesized that having two distance-based factors was making our regression less significant so for the next test we removed the two variables we had created ourselves and replaced them with “longitude” and “Workers16yearsandover”. This time we came up with a slightly lower Adjusted R Squared of .580 and found that Workers16yearsandolder was highly insignificant For the next regression we tried replacing it by creating our on dummy variable to separate Green Line vs Orange Line using the function: =IF(LEFT(K2,1)="G",1,0). Once again we found a lower Adjusted R Squared. We tried 2 more regressions, each time replacing this variable with another and leaving the rest the same. Ultimately, we found that on our 6th Regression we were able to raise the Adjusted R Squared to .583 by substituting Approximate Nautical Miles from Boston University Central Station back into the regression. The final equation of the regression is Predicted Trips= - 5872702 + longitude(-82802.602) + Age15to24(1.452) + pooperacre(79.303) +Approximate Nautical Miles from Boston University Central Station(1812.506) + distance_common(- 4457.324). The standard errors for each in order are: 3386786.160, 47694.100, .528, 23.550, 1947.89, and 2400.338. From these coefficients we learn that the variables longitude and distance_common have negative relationships with trips while the rest are positive. It is surprising that distance_common and distance from BU Central would have different relationships because they are both dense areas which we expected to decrease trips.

Table 3: Correlation Matrix



	trips	Miles BUC	Age 15 to 24	longitude	popperacre	distance_common
trips	1					
Approximate N	-0.674478499	1				
Age 15 to 24	0.589192206	-0.51383013	1			
longitude	0.520035876	-0.71033346	0.42013828	1		
popperacre	0.627037207	-0.61752011	0.40949477	0.53143734	1	
distance_com	-0.69416918	0.958828153	-0.5627063	-0.8621319	-0.647338468	1

Table 4: Summary Statistics



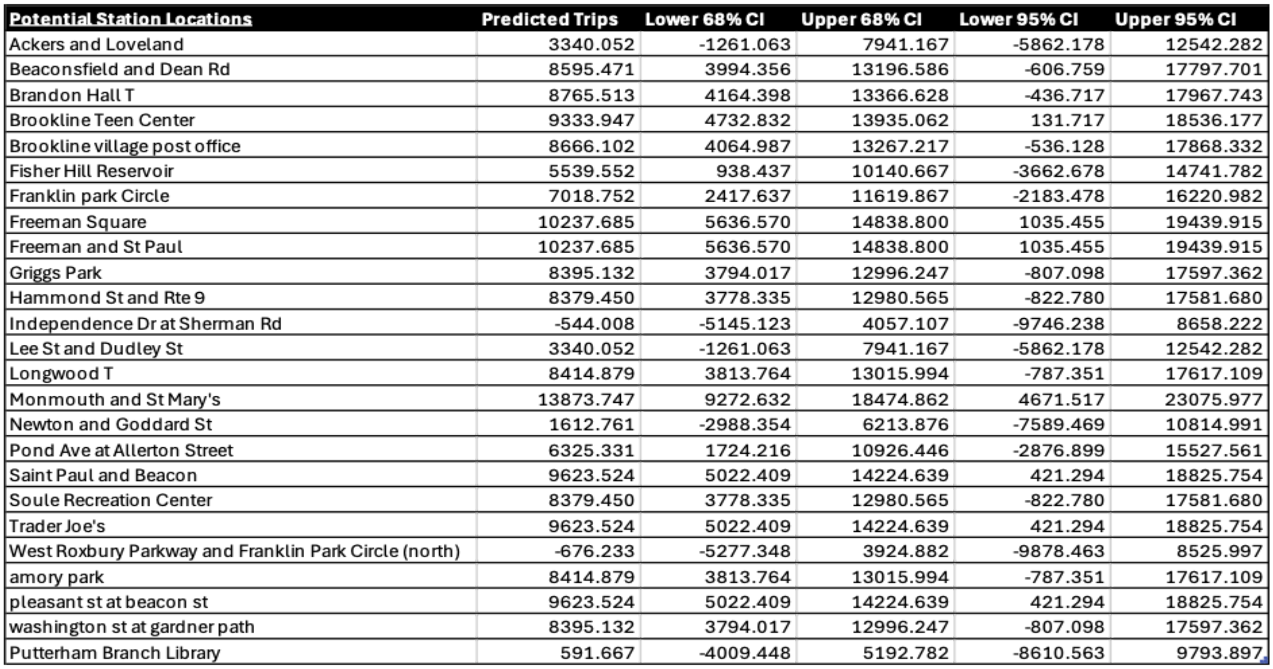
	Miles BUC	Age 15 to 24	longitude	popperacre	distance_common
Mean	2.019	971.422	-71.131	36.541	3.964
Standard Error	0.134	101.094	0.003	2.241	0.161
Median	1.451	555.000	-71.126	30.381	3.462
Mode	1.205	915.000	-71.126	5.632	3.160
Standard Devi	1.448	1088.818	0.033	24.137	1.731
Minimum	0.147	23.000	-71.248	0.217	1.485
Maximum	6.285	4846.000	-71.087	134.142	9.331
Sum	234.191	112685.000	-8251.249	4238.797	459.801
Count	116.000	116.000	116.000	116.000	116.000

As shown in Table 5, we next applied this regression model to potential new stations to try and predict their total Blue Bike trips. Two of the values our regression predicted for the stations Independence Dr at Sherman Rd and West Roxbury Parkway and Franklin Park Circle (north) were negative. While logically the idea of a station having negative trips is impossible, this doesn't necessarily mean the regression is flawed. It reflects a common limitation of linear models, which can often generate predictions outside the feasible range when the intercept is a large negative number. Because the predicted number of trips cannot be negative, we've interpreted these results to indicate an approximately zero expected demand for these stations. Additionally, for the functionality of this regression we must also assume that adding a new Blue Bike station does not reduce the trips of other neighboring stations. The addition of new stations would redistribute customers, pulling some from existing stations; however, in practice this would distort our model to overpredict trips. This is explained by the strangely high numbers for some of these new stations. Table 5 also provides the Confidence Intervals for the potential new stations. Some CI are wide and include negative values, therefore all negative values should be effectively treated as 0. This does not mean the model is invalid, but rather that there is a significant unexplained variation in trips making up about .40 of the Adjusted R Squared value.

These unexplained variations may occur for several reasons, one possibility being a cluster of new stations. Our regression assumes that adding multiple new stations to a similar area will increase total trips by the same amount for each station. If we add 4 new stations, the number of predicted total trips will increase 4 times. This prediction is unreliable as a new close network of stations could cause trips to increase by even more or by less than what the model predicts. In one scenario, the added concentration of stations increases total riders by even more than predicted. This is because an increase in stations in the same area allows for shorter trips to become more accessible. As the number of Blue Bike docks increase, so too does the riders trust that a dock will be open. In another reality, this cluster of stations shifts the customer demand between all of them. This would mean that each station receives fewer total trips as the number of trips does not increase but rather distributes itself among more stations.

Finally, to maximize our model's effectiveness we must assume the omission of other potential variables does not cause significant change in the prediction of trips. For example, there was no data in the dataset revolving around the presence of bike lanes or the change in elevation levels. One would assume that an increase in bike lanes around a station would increase the total trips of that station. This would conflict with variables already in our regression and cause their coefficients to have an upward bias. Distance from Boston Common and population density are both variables that are also likely to have a larger presence of bike lanes and are therefore susceptible to this bias. If the elevation of terrain in Boston came back to show that areas around BU Central had greater change in elevation, we would likely over-predict trips, as the number of trips around BU Central would be less due to the surrounding hilly area.

Table 5. Station Trip Predictions with Confidence Intervals



Potential Station Locations	Predicted Trips	Lower 68% CI	Upper 68% CI	Lower 95% CI	Upper 95% CI
Ackers and Loveland	3340.052	-1261.063	7941.167	-5862.178	12542.282
Beaconsfield and Dean Rd	8595.471	3994.356	13196.586	-606.759	17797.701
Brandon Hall T	8765.513	4164.398	13366.628	-436.717	17967.743
Brookline Teen Center	9333.947	4732.832	13935.062	131.717	18536.177
Brookline village post office	8666.102	4064.987	13267.217	-536.128	17868.332
Fisher Hill Reservoir	5539.552	938.437	10140.667	-3662.678	14741.782
Franklin park Circle	7018.752	2417.637	11619.867	-2183.478	16220.982
Freeman Square	10237.685	5636.570	14838.800	1035.455	19439.915
Freeman and St Paul	10237.685	5636.570	14838.800	1035.455	19439.915
Griggs Park	8395.132	3794.017	12996.247	-807.098	17597.362
Hammond St and Rte 9	8379.450	3778.335	12980.565	-822.780	17581.680
Independence Dr at Sherman Rd	-544.008	-5145.123	4057.107	-9746.238	8658.222
Lee St and Dudley St	3340.052	-1261.063	7941.167	-5862.178	12542.282
Longwood T	8414.879	3813.764	13015.994	-787.351	17617.109
Monmouth and St Mary's	13873.747	9272.632	18474.862	4671.517	23075.977
Newton and Goddard St	1612.761	-2988.354	6213.876	-7589.469	10814.991
Pond Ave at Allerton Street	6325.331	1724.216	10926.446	-2876.899	15527.561
Saint Paul and Beacon	9623.524	5022.409	14224.639	421.294	18825.754
Soule Recreation Center	8379.450	3778.335	12980.565	-822.780	17581.680
Trader Joe's	9623.524	5022.409	14224.639	421.294	18825.754
West Roxbury Parkway and Franklin Park Circle (north)	-676.233	-5277.348	3924.882	-9878.463	8525.997
amory park	8414.879	3813.764	13015.994	-787.351	17617.109
pleasant st at beacon st	9623.524	5022.409	14224.639	421.294	18825.754
washington st at gardner path	8395.132	3794.017	12996.247	-807.098	17597.362
Putterham Branch Library	591.667	-4009.448	5192.782	-8610.563	9793.897

PART B

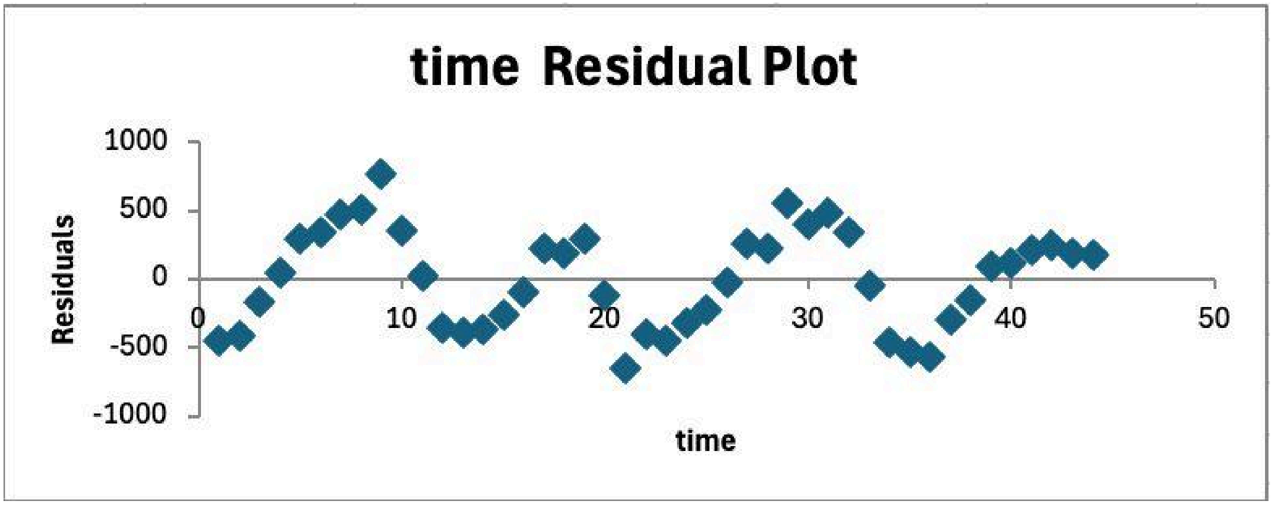
4a.

Regression Equation
Trips = 3.9977(Time) + 654.10

A simple linear regression was estimated using Trips as the dependent variable and Time as the only explanatory variable. The resulting model yields an intercept of 654.10 and a time coefficient of 3.9977, indicating a very modest average monthly increase in ridership. Model performance, however, is weak: the R^2 is 0.0203, the adjusted R^2 is effectively zero, and the p-value on Time (0.356) indicates that the trend is not statistically significant. These results show that time alone does not meaningfully account for variation in monthly trips at the 1200 Beacon St. station.

The residual diagnostics further highlight the model's limitations. Rather than displaying random variation, the residuals exhibit a smooth, recurring seasonal pattern, negative residuals during winter months and positive residuals during summer months. This consistent cyclical structure underscores the presence of strong seasonality in bike usage that the simple linear trend fails to capture. As a result, the analysis demonstrates that a time-only specification is insufficient, and that the inclusion of monthly indicator variables is necessary to appropriately model trip behavior over time.

Figure 1: Residual Pattern Over Time



4b. Monthly Dummy Variable Model

When a model is estimated using only monthly dummy variables (with January as the omitted baseline), the goodness of fit improves substantially relative to the time-only regression. In this specification, the R^2 increases from 0.02 to 0.871, and the adjusted R^2 rises to 0.827, indicating that the model now explains the vast majority of the variation in monthly trip counts at the 1200 Beacon St. station. The highly significant overall F-statistic ($p < 0.0000000004$) further confirms that the monthly indicators collectively possess strong explanatory power.

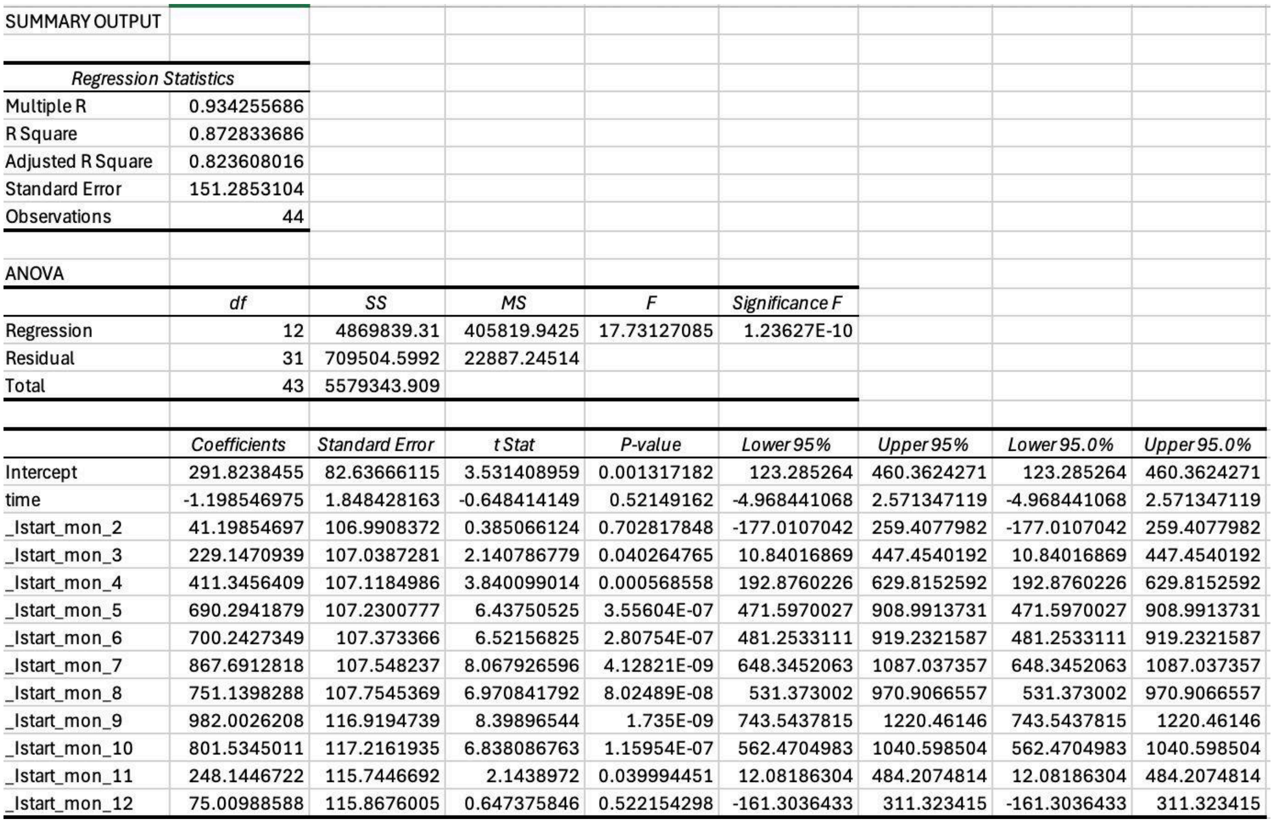
This sharp improvement underscores a key point: ridership at this station is driven primarily by seasonality rather than by a steady linear trend. Several summer and early-fall months—including July (≈ 860.5), August (≈ 742.75), and September (≈ 970.42), exhibit large positive coefficients relative to January, while winter months remain comparatively low. These patterns capture the predictable rise and decline in bike usage across the year. In contrast to the nearly flat and statistically insignificant time-trend model, the monthly dummy regression aligns closely with the actual seasonal behavior of riders, producing a far more accurate and interpretable representation of trip patterns.

4c. Time Trend with Monthly Dummy Variables

When monthly dummy variables are added to the regression alongside the time trend, the estimated effect of Time changes substantially relative to the earlier specifications. In the time-only model, Time had a positive coefficient of approximately 4.0, suggesting a gradual increase in ridership; however, that estimate was statistically insignificant and accounted for almost none of the variation in trips. After introducing monthly indicators, the Time coefficient becomes much smaller and turns slightly negative (–1.20), while remaining statistically insignificant ($p \approx 0.52$).

This shift illustrates an important point: the apparent upward trend observed in the simple regression was largely the result of recurring seasonal peaks during the summer months. Once the model explicitly controls for these predictable seasonal patterns, there is no evidence of a meaningful long-run increase in ridership at the 1200 Beacon St. station over the 2022–2025 period. In effect, the seasonally adjusted trend is essentially flat, ridership exhibits strong seasonal fluctuations, but no underlying upward or downward trajectory over time.

Table 6: Regression Statistics of Monthly Dummy Variables



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.934255686
R Square	0.872833686
Adjusted R Square	0.823608016
Standard Error	151.2853104
Observations	44

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	12	4869839.31	405819.9425	17.73127085	1.23627E-10
Residual	31	709504.5992	22887.24514		
Total	43	5579343.909			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	291.8238455	82.63666115	3.531408959	0.001317182	123.285264	460.3624271	123.285264	460.3624271
time	-1.198546975	1.848428163	-0.648414149	0.52149162	-4.968441068	2.571347119	-4.968441068	2.571347119
_Istart_mon_2	41.19854697	106.9908372	0.385066124	0.702817848	-177.0107042	259.4077982	-177.0107042	259.4077982
_Istart_mon_3	229.1470939	107.0387281	2.140786779	0.040264765	10.84016869	447.4540192	10.84016869	447.4540192
_Istart_mon_4	411.3456409	107.1184986	3.840099014	0.000568558	192.8760226	629.8152592	192.8760226	629.8152592
_Istart_mon_5	690.2941879	107.2300777	6.43750525	3.55604E-07	471.5970027	908.9913731	471.5970027	908.9913731
_Istart_mon_6	700.2427349	107.373366	6.52156825	2.80754E-07	481.2533111	919.2321587	481.2533111	919.2321587
_Istart_mon_7	867.6912818	107.548237	8.067926596	4.12821E-09	648.3452063	1087.037357	648.3452063	1087.037357
_Istart_mon_8	751.1398288	107.7545369	6.970841792	8.02489E-08	531.373002	970.9066557	531.373002	970.9066557
_Istart_mon_9	982.0026208	116.9194739	8.39896544	1.735E-09	743.5437815	1220.46146	743.5437815	1220.46146
_Istart_mon_10	801.5345011	117.2161935	6.838086763	1.15954E-07	562.4704983	1040.598504	562.4704983	1040.598504
_Istart_mon_11	248.1446722	115.7446692	2.1438972	0.039994451	12.08186304	484.2074814	12.08186304	484.2074814
_Istart_mon_12	75.00988588	115.8676005	0.647375846	0.522154298	-161.3036433	311.323415	-161.3036433	311.323415

4d .

For November 2025, only the November dummy (_Istart_mon_11) equals 1, so the prediction is:
Predicted Trips = 291.823846 + (–1.198547 × 45) + 248.144672
= 291.823846 – 53.934615 + 248.144672
= ≈ 486 trips.

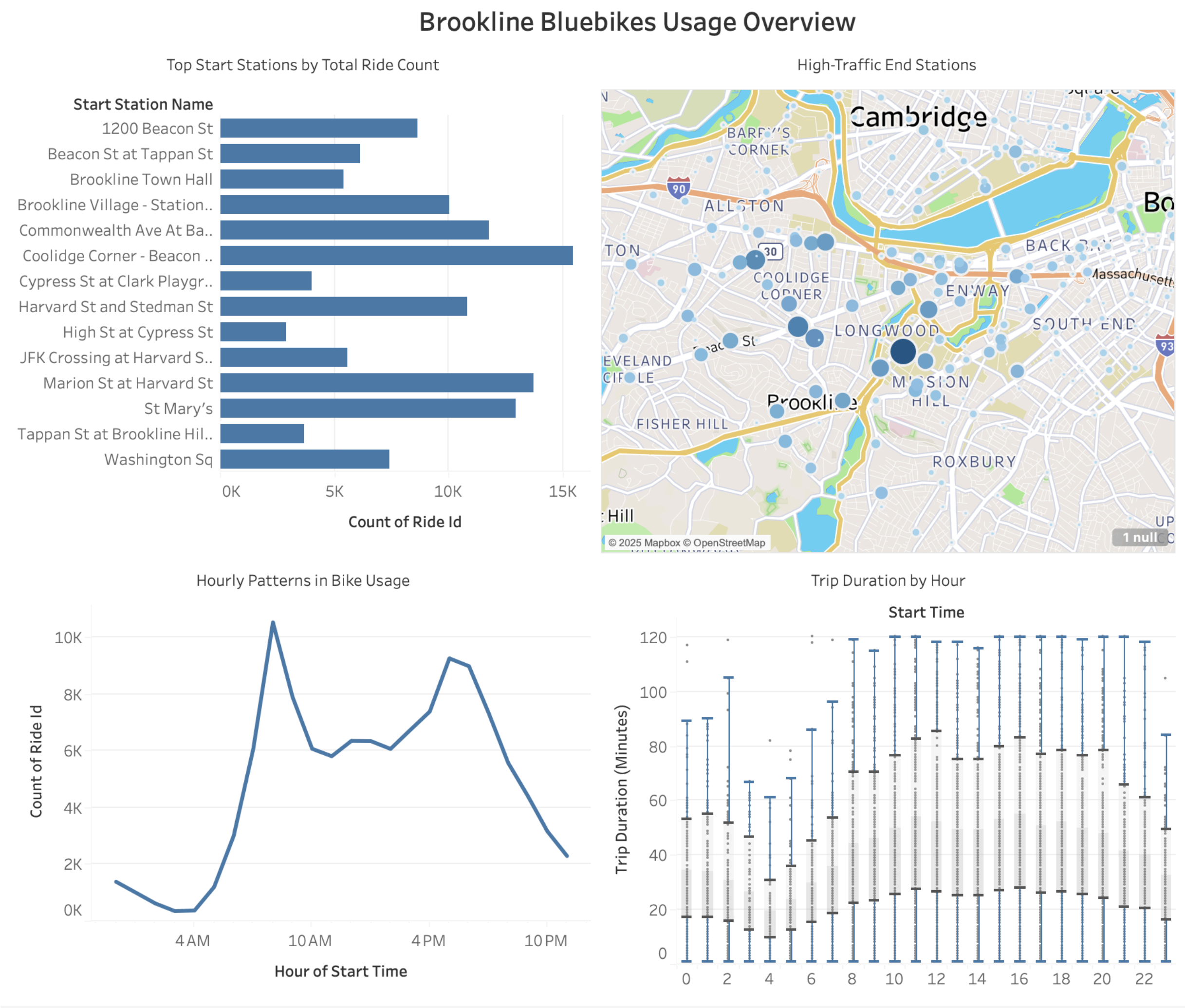
For June 2026, only the June dummy (_Istart_mon_6) is active, producing:
Predicted Trips = 291.823846 + (–1.198547 × 52) + 700.242735
= 291.823846 – 62.324444 + 700.242735
= ≈ 930 trips.

4e.

In the full model for my station, the seasonally adjusted growth rate is the coefficient on Time, which is –1.20. This indicates that after accounting for predictable seasonal swings, the underlying trend at 1200 Beacon St is essentially flat and slightly negative over time. In contrast, Brookline as a whole shows a positive seasonally adjusted growth rate of +3.49 trips per month. This means the station is underperforming relative to the broader system: while Brookline's overall demand is slowly rising, 1200 Beacon St is not experiencing that same long-run growth once seasonality is removed, suggesting that its ridership depends heavily on seasonal patterns rather than sustained underlying increases in demand.

PART C

Figure 2: Screenshot of Tableau Dashboard of Brookline Bluebikes Usage Overview



The Tableau workbook (.twbx) file is also uploaded