Teaching Case – Is Bikeshare Really Equally Accessible and Affordable to Everyone in Boston?

Seungyeon Kim

Overview.

Research by the Urban Institute has shown that bikeshare infrastructure tends to be concentrated in high-income neighborhoods in Washington, D.C.[1] However, in my own city exploration analysis of the relationship between bikeshare usage and income by census tract in D.C., I found that usage remains popular among low- to middle-income populations (those with a ratio of income to poverty between 0.5 and 1.99), despite disparities in bikeshare access across income levels. This experience deepened my understanding of the potential of bicycles as an affordable transportation option for low-income communities—and the consequences of limited access for those who may need it most.[2]

Both Boston's Bluebike 2015 data and Washington, D.C.'s Capital Bikeshare 2025 data reveal an inverted U-shaped correlation between income and bikeshare usage. That is, usage tends to peak among middle- and lower-middle-income populations. For very low-income individuals, even public bikeshare systems may be financially out of reach; for high-income individuals, other transportation options may be more appealing than bikes.

Intrigued by the possibility that bikeshare programs could become more accessible to low-income users, I explored Boston's Bluebike system further. I found that the City of Boston has already taken steps to improve access by offering discounted memberships to individuals eligible for public benefit programs. In January 2018, Boston's Bluebikes launched a low-income membership initiative called "SNAP Card to Ride," which provides discounted memberships to Supplemental Nutrition Assistance Program (SNAP) recipients living in Greater Boston.[3]

This teaching case explores the demographic and socioeconomic characteristics of Bluebike usage before and after the launch of this program, focusing on the years 2015 and 2019. These years were selected to align with the ACS release periods and the corresponding Housing + Transportation (H+T) Index datasets.[4]

---

[1] Yipeng Su, Robin Wang, "Three Ways Bikeshare Can Counteract—Not Reinforce—DC's Disparities," *Urban Institute*, February 11, 2019, https://www.urban.org/urban-wire/three-ways-bikeshare-can-counteract-not-reinforce-dcs-disparities.

[2] Rauf Ahmed and Ahmad El-Geneidy, "Changes in equity of bikeshare access and use following implementation of income-eligible membership program & system expansion in Greater Boston," *Journal of Transport & Health* Vol 21 (2021), https://www.sciencedirect.com/science/article/abs/pii/S2214140521000839.

[3] City of Cambridge, "Boston, Brookline, Cambridge and Somerville Launch "SNAP Card to Ride" Bike Share," *Community Development Department*, January 19, 2018, https://www.cambridgema.gov/CDD/News/2018/1/hubwaysnapbikeshare.aspx.

[4] Center for Neighborhood Technology, "H+T Index: Housing and Transportation Affordability Index," accessed April 18, 2025, https://htaindex.cnt.org/map/.

The analysis draws on:

- Bluebike trip data from 2015 and 2019,
- American Community Survey (ACS) data from 2011–2015 and 2015–2019,
- Housing + Transportation (H+T) Index for 2016 and 2022, which estimates housing and transportation costs by census tract based on ACS and other sectoral data.

The aim is to examine changes in bikeshare access before and after the implementation of the low-income membership program by analyzing:

1. The number of stations per census tract by income level, and
2. The volume of bikeshare usage per census tract by income level.
3. Regression models that examine the relationship between income level and bikeshare usage per census tract in 2015 and 2019.

While it is difficult to isolate the direct impact of the policy, this analysis seeks to indirectly assess whether any meaningful changes occurred following its implementation.

Data Preparation

This analysis used six key R packages for data cleaning, analysis, and visualization:

- *ggplot2*
- *dplyr*
- *readr*
- *sf*
- *tmap*
- *scales*

The data preparation process is organized into three sections based on the main datasets and was performed separately for the years 2015 and 2019. The overall objective was to clean the record-level data and merge them by GEOID or GISJOIN to create census tract-level datasets for income-based analysis.

1. Bluebike Data

- Loaded Bluebike trip data for 2015 and 2019. Each year's monthly data was combined using *bind_rows()* as part of a previous assignment.
- Summarized total bike usage (*total_users*) by *start_station_name* using *group_by()* and *summarise()* to prepare for later aggregation at the census tract level.
- Converted the dataset into an sf object using *st_transform()* to match the coordinate reference system of the tract shapefile.
- Joined the Bluebike sf object with the census tract shapefile using *st_join().*

2. ACS Data

- Loaded 5-year ACS data (2011–2015 and 2015–2019) and shapefiles including household income, income-to-poverty ratio, commute time to work, and race variables.
- Renamed variables with more intuitive names using the NHGIS codebook. (See Appendix 2. for a complete list of renamed variables.)

3. Housing + Transportation (H+T) Index Data

- Loaded the H+T Index data from 2016 and 2019 and renamed relevant columns for consistency in R.

4. Merging Datasets

- Merged Bluebike, ACS, and H+T Index datasets using either GEOID or GISJOIN with *merge() or join()* functions.
- Created new census tract-level variables using *mutate(), summarise(),* and *ntile():*
  - *income_quantile* (based on median income)
  - *total_stations and total_usage (using sum())*
  - *n_tracts (using n())*
  - *avg_transport_cost and avg_income (using mean())*
  - *avg_usage (calculated as total usage divided by number of stations)*

Analysis.

General Overview in 2015

First, Table 1 presents **the total number of Bluebike stations** (**156 stations**) and **total trip volume (1,107,442 users) in 2015**. This table was created using the *summarise()* function in dplyr, by aggregating bikeshare trip record level data to compute the total number of stations and total users.

<Table 1. Total Stations and Bike Usage Volume 2015 (record level)>

Total Number of Bluebike Stations	Total Number of Users
156	1,107,442

Second, we explore bikeshare usage by income levels. The following table is particularly interesting: while the average household income varies significantly across income quantiles, the average annual household transportation cost remains relatively similar. This indicates that transportation places a heavier financial burden on lower-income households, as it takes up a larger share of their income.

In terms of infrastructure, **wealthier neighborhoods tend to have more bikeshare stations**: the lowest income quantile has 35 stations, while the highest income quantile has 42. However, **when looking at total bike usage, the 3rd income quantile records the highest number of trips (351,859)**, followed by the 4th quantile (268,109).

If we shift our focus to **average usage per station (Total Usage ÷ Total Stations),** it becomes even clearer that middle-income neighborhoods are the most active users of Bluebikes. **The 3rd quantile has the highest usage per station (approx. 9,022)**, followed by the 2nd quantile (7,523). The 1st quantile which is the lowest income group shows the least number of average usage (5,788) per station.

This table was created by spatially joining the 2015 Bluebike trip data with census tract boundaries using the *st_join()* and *st_within()* functions. The Bluebike station-level data was then reorganized with *dplyr* functions *(group_by(), summarise())* to calculate the number of stations and total bike usage per tract.

This output was subsequently merged with ACS 2011–2015 data (including income, poverty ratio, race, and commute time) and the 2016 H+T Index using either GEOID or GISJOIN identifiers. Finally, the merged dataset was grouped by income quantile, and aggregate variables such as *total_stations, n_tracts, total_usage, avg_usage, avg_transport_cost,* and *avg_income* were computed using functions like *mutate(), filter(), summarise(), sum(),* and *mean()*. All variables presented in the table are newly aggregated.

<Table 2. Income Quantiles and Blubike Usage 2015 (census tract level)>

Income Quantile	Average Income*	Average Transport Cost**	Total Stations*** (A)	Number of Tracts	Total Usage (B)	Average Usage per Station (B/A)
1	$ 29,083	$ 7,411	35	23	202,604	5,788
2	$ 57,049	$ 8,481	34	23	255,801	7,523
3	$ 82,283	$ 7,869	39	23	351,859	9,022
4	$ 112,321	$ 8,436	42	22	268,109	6,383

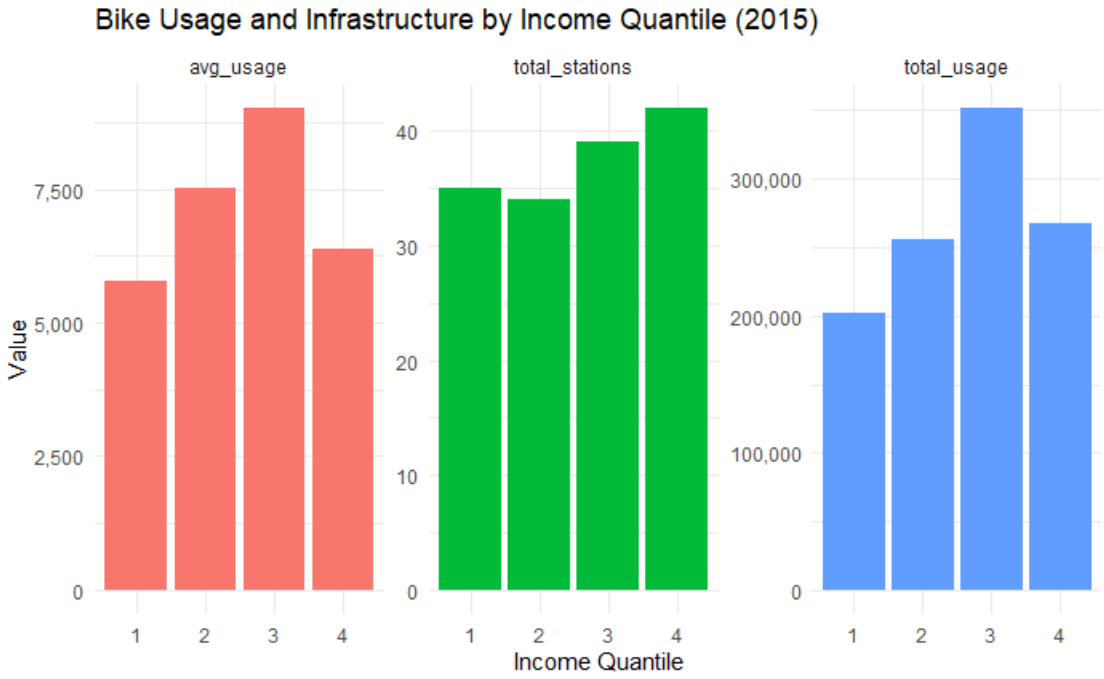
*Average income reflects the mean of median household income across tracts in each quantile.

**Average transportation cost reflects the mean annual household transport expenditure (for the average income level) in each quantile.

***Six stations with missing income values (NA) were excluded

To visualize the table results, bar charts were created *using ggplot, geom_col, facet_wrap* functions.

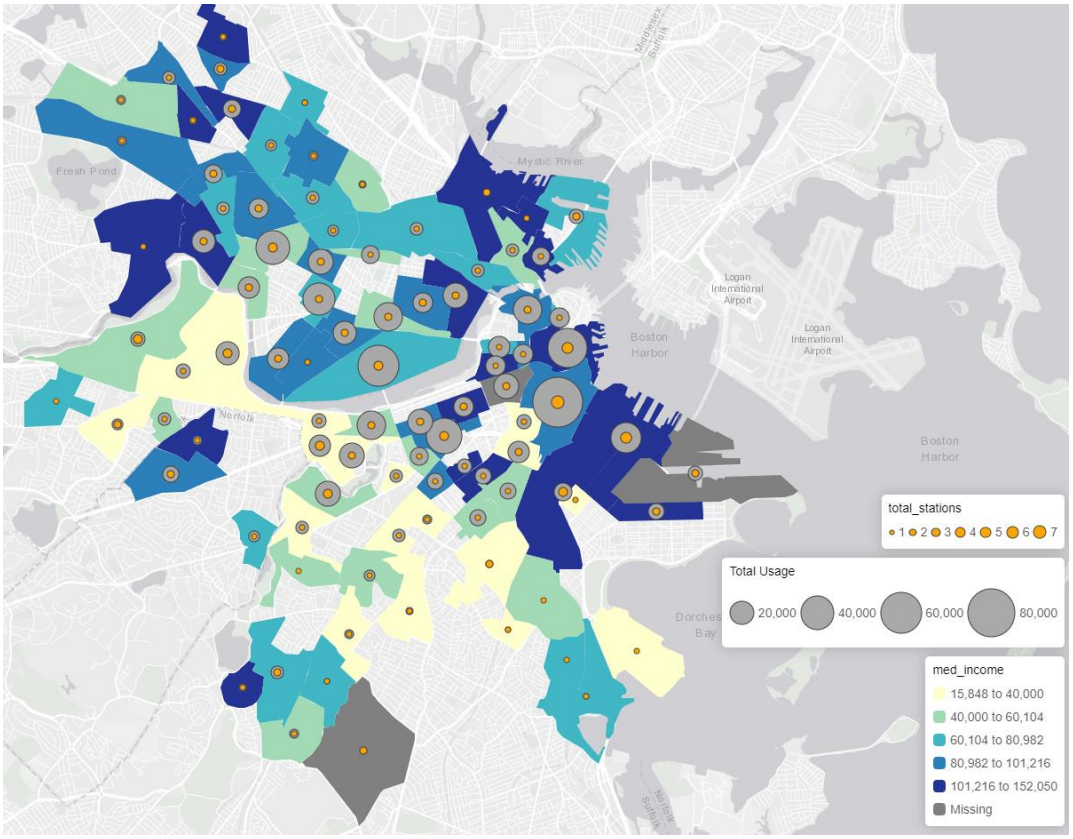
<Figure 1. Bluebike Usage and Infrstructure by Income Quantile 2015 (census tract level)>



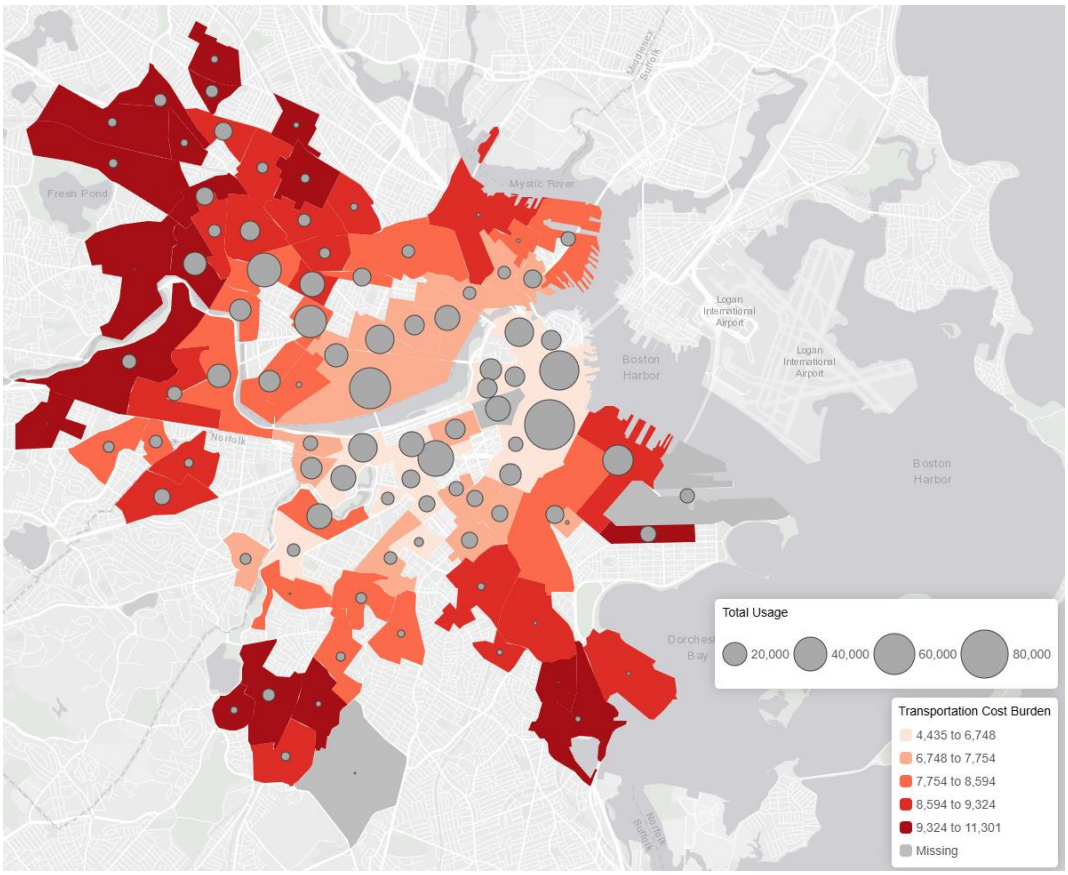
Two maps were created using *tmap* functions (*tm_shape, tm_fill, tm_bubbles, tm_layout*) to visualize this analysis. One key insight from the map is that, although absolute transportation costs are often similar or higher in wealthier areas, the relative cost burden is much heavier in lower-income neighborhoods.

This is why areas with the highest transport cost don't necessarily align with the lowest income tracts since we're comparing absolute values. Still, it's worth noting that **bike usage tends to be lower in suburban areas with higher transport cost burdens, and higher usage appears in central Boston**, likely due to better infrastructure and multimodal connectivity as shown in the map.

<Figure 2. Median Income, Station Count, and Bike Usage 2015 (census tract level)>



<Figure 3. Transportation Cost and Bike Usage 2015 (census tract level)>



General Overview in 2019

Let's now turn to the 2019 dataset. The record level Bluebike trip data covers the full calendar year, ending in December 2019, approximately two years after the launch of the discounted membership program. Tables and maps were produced using the same methodology described in the 2015 analysis.

Compared to 2015, the data clearly reflects the City of Boston's efforts to expand bikeshare access. Within just two years, 255 new stations were installed across the system, increasing total station count from 156 to 411**. This expansion, combined with the affordability initiative, resulted in a dramatic increase in total usage, more than doubling the number of trips by 2019**.

<Table 3. Total Stations and Bike Usage Volume 2019 (record level)>

Total Number of Bluebike Stations	Total Number of Users
411	2,522,771

The next table provides further insight into Boston's efforts to enhance bikeshare accessibility among lower-income populations. With persistent income gaps across quantiles and relatively similar average transportation costs, **the 1st (lowest) and 3rd (middle) income quantiles had the highest number (68) of stations installed** in their census tracts in 2019.

Notably, **in the 1st income quantile, total bike usage increased** from 202,604 trips in 2015 **to 303,821 trips in 2019**. Although average usage per station declined from 5,788 to 4,467, probably due to the increase in the number of stations, the improvement in total usage still reflects broader access.

However, despite the increased infrastructure, **the lowest-income areas still recorded the lowest total and average usage among all quantiles.**

An important shift compared to 2015 is that the **4th income quantile (highest-income tracts) rose to the second highest in both total and average usage**. Again, to visualize the table results, bar charts were created *using ggplot, geom_col, facet_wrap* functions.

<Table 4. Income Quantiles and Blubike Usage 2019 (census tract level)>

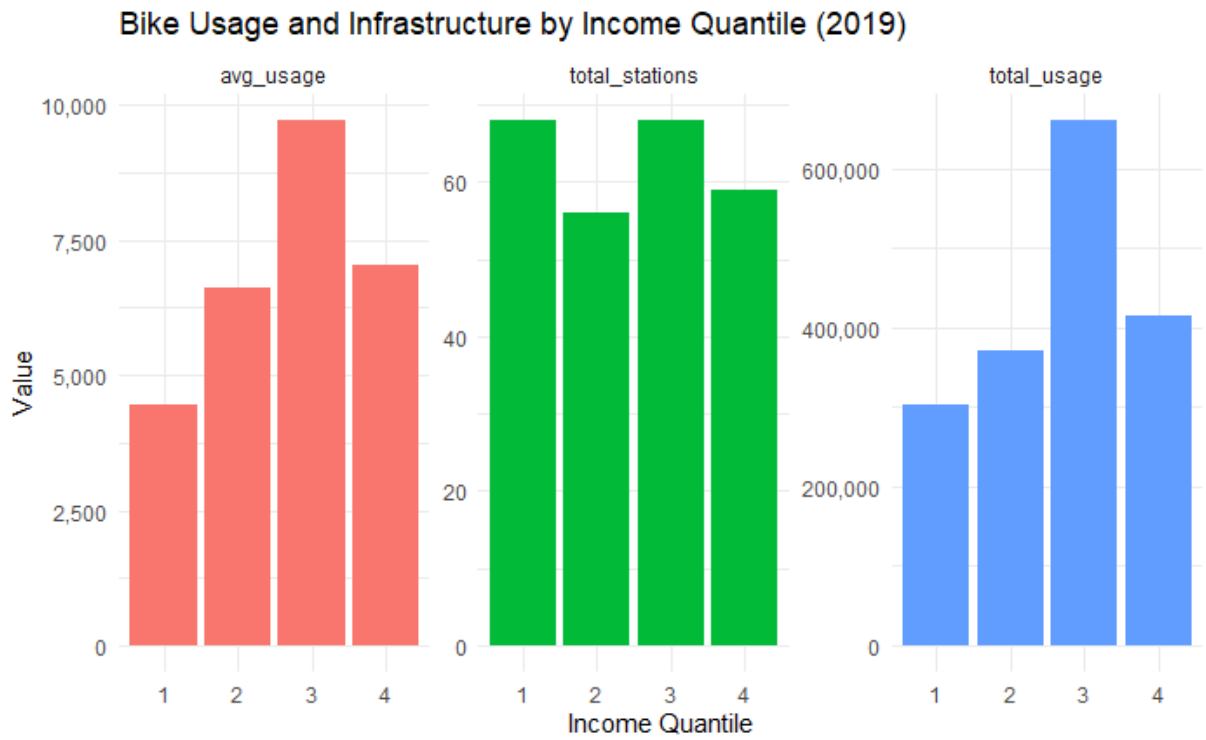
Income Quantile	Average Income*	Average Transport Cost**	Total Stations*** (A)	Number of Tracts	Total Usage (B)	Average Usage per Station (B/A)
1	$ 36,597	$ 9,766	68	34	303,821	4,467
2	$ 69,208	$ 10,234	56	33	369,991	6,606
3	$ 97,455	$ 10,207	68	33	661,299	9,724
4	$ 135,428	$ 10,526	59	33	414,832	7,031

*Average income reflects the mean of median household income across tracts in each quantile.

**Average transportation cost reflects the mean annual household transport expenditure (for the average income level) in each quantile.

***48 stations with missing HT index values (NA) were excluded

<Figure 4. Bluebike Usage and Infrstructure by Income Quantile 2019 (census tract level)>

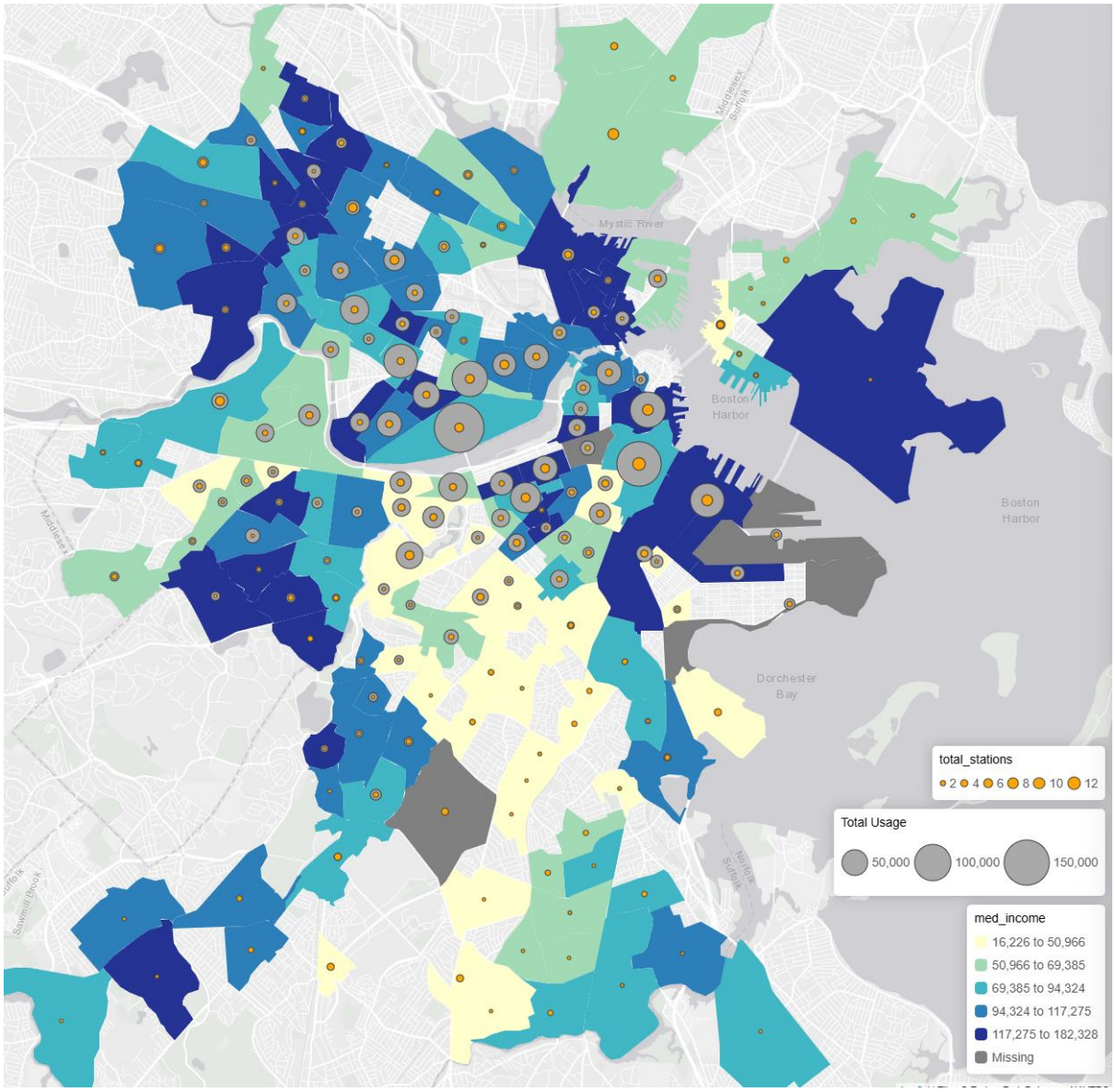


Two maps were created using the same methodology (*tmap*) as in the 2015 analysis. As shown in Figure X, there were more Bluebike stations installed across all income quantile tracts in 2019. However, **similar to 2015, suburban areas still showed lower levels of bike usage** compared to central Boston, regardless of income level.

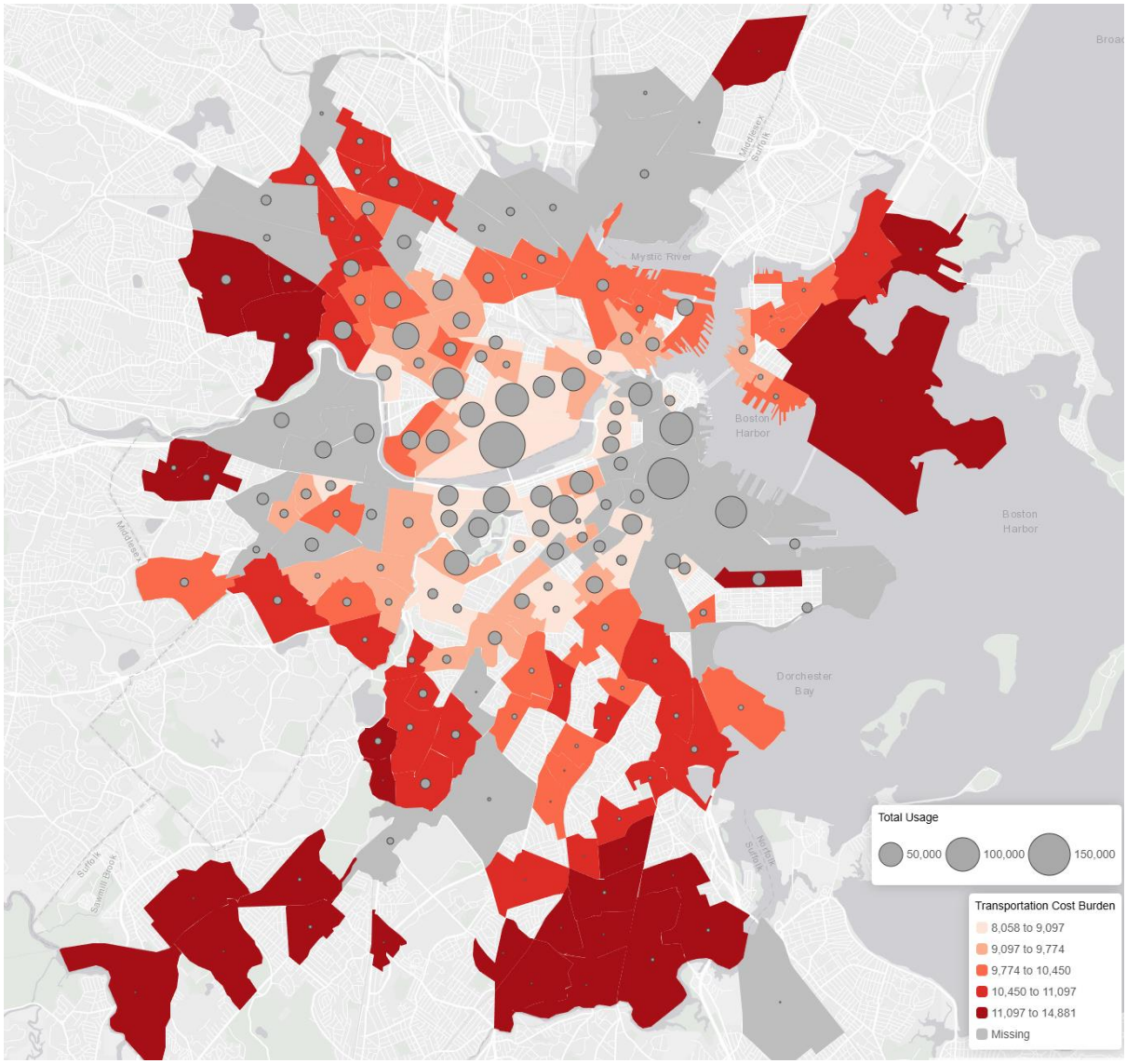
This suggests that **in addition to income, proximity and connectivity to the city center play a key role in encouraging bikeshare usage**.

It is also noteworthy that **the Bluebike network expanded into more remote, low-income tracts that bear a high transportation cost burden** and are located farther from the city center.

<Figure 5. Median Income, Station Count, and Bike Usage 2019 (census tract level)>



<Figure 6. Transportation Cost and Bike Usage 2015 (census tract level)>



Regression Models 2015 and 2019

Based on the general analysis, I constructed regression models using the *lm()* function in R. The dependent variable (Y) is total bikeshare usage by census tract, and the independent variables (X) include the following:

ACS 2011 – 2015 and 2015 - 2019 Data

- Household median income *(med_income)*
- Ratio of Income to Poverty Line *(poverty_under_0_5 ~ poverty_2_00_over)*
- Race *(white, black, native_american, pacific_islander)*
- Commute time to work *(less than five ~ ninety_or more)*

HT Index 2016 and 2022

- Transportation Cost For the Area Median Income *(t_cost_ami)*

After fitting the initial model with all variables, I applied a stepwise reduction approach by removing the variable one by one with the highest p-value, until only statistically significant predictors remained in the final model.

Key Findings from the 2015 Regression Model

The 2015 regression model identified several variables that were statistically significant in explaining total bikeshare usage (Adjusted $R^2 = 0.46$) at the census tract level:

- **Median Income**: Median income is positively associated with bikeshare usage, while the quadratic term is negative. This suggests an **inverted U-shaped relationship**, where usage increases with income up to a certain point, then declines in higher-income areas.
- **Commute Time**: Among commute time variables, the **5-to-9-minute** category shows a positive and statistically significant relationship with bike usage. This implies that **short-distance commuters are more likely to use bikeshare**.
- **Transportation Cost**: There is a **negative relationship** between transportation cost and bike usage. In other words, **census tracts with lower transportation cost burdens tend to have higher bike usage**. This aligns with spatial observations from the Figure x. These low-cost areas are typically located near the city center, where bikeshare infrastructure and multimodal connectivity are more robust.
- **Race**: None of the racial demographic variables show statistically significant relationships with bikeshare usage in 2015, indicating **no clear racial pattern** in bikeshare adoption at that time.

```
Call:
lm(formula = total_usage ~ med_income + I(med_income^2) + five_to_nine +
    sixty_to_eightynine + t_cost_ami, data = merged_df)

Residuals:
   Min     1Q Median     3Q    Max
-19376  -5783  -1915   4420  46340

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.713e+04  8.392e+03   2.041 0.044374 *
med_income          4.860e-01  1.524e-01   3.188 0.002004 **
I(med_income^2)    -2.489e-06  1.030e-06  -2.416 0.017825 *
five_to_nine        4.714e+01  8.593e+00   5.486 4.15e-07 ***
sixty_to_eightynine -2.272e+01  1.249e+01  -1.819 0.072440 .
t_cost_ami          -3.630e+00  9.102e-01  -3.988 0.000141 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10620 on 85 degrees of freedom
  (결측으로 인하여 3개의 관측치가 삭제되었습니다.)
Multiple R-squared:  0.499,     Adjusted R-squared:  0.4695
F-statistic: 16.93 on 5 and 85 DF,  p-value: 1.39e-11
```

Key Findings from the 2019 Regression Model

The 2019 regression model (Adjusted R^2 = 0.47) reveals several notable changes compared to 2015, particularly in the significance of **poverty ratio**, **race**, and **commute time** in explaining total bikeshare usage by census tract.

- **Median Income**: The relationship between median income and bike usage remains consistent with the 2015 model. This suggests that the inverted U-shaped relationship—where bikeshare usage increases with income up to a point—persisted over time.
- **Poverty Ratio**: Poverty ratio categories became statistically significant in 2019. This reinforces the earlier finding that bikeshare usage is **lowest among the very poor and very wealthy**. These variables may have become significant as new stations were installed in areas with high or low income levels, expanding the dataset to include tracts that previously lacked bikeshare infrastructure.
- **Commute Time**: In addition to the 5–9 minute commute category, the **10–14 minute commute category also became statistically significant**. This may indicate an increase in bikeshare usage among commuters traveling from slightly farther distances, possibly reflecting the **expansion of bikeshare infrastructure beyond the city center**.
- **Transportation Cost**: The **negative relationship between transportation cost burden (t_cost_ami) and bikeshare usage** became more pronounced in 2019. This suggests that residents in areas with higher transportation burdens are less likely to bike, even with improved access, possibly due to longer distances or less connected environments.
- **Race**: A notable shift from 2015 is that **race variables became statistically significant predictors of bikeshare usage**. Census tracts with higher shares of Black and Asian populations were associated with higher levels of bikeshare usage, compared to 2015 where no such relationship was found. **This shift may reflect increased infrastructure coverage or the affordable membership benefitting certain racial groups**; however, it is important to interpret this result cautiously, as multiple contextual factors could be contributing to the observed association.

```
Call:
lm(formula = total_usage ~ med_income + I(med_income^2) + poverty_under0_5 +
    poverty_2_00_over + black + asian + five_to_nine + ten_to_fourteen +
    t_cost_ami, data = merged_df_19)

Residuals:
   Min     1Q Median     3Q    Max
-38920  -9549   -588   6652  91195

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.241e+04  1.735e+04   2.445 0.015911 *
med_income         4.960e-01  2.191e-01   2.263 0.025364 *
I(med_income^2)   -2.025e-06  1.058e-06  -1.914 0.057919 .
poverty_under0_5  -2.049e+01  8.325e+00  -2.462 0.015218 *
poverty_2_00_over -6.429e+00  1.842e+00  -3.490 0.000672 ***
black              4.344e+00  1.801e+00   2.412 0.017352 *
asian              1.220e+01  5.577e+00   2.188 0.030585 *
five_to_nine       3.753e+01  1.744e+01   2.152 0.033371 *
ten_to_fourteen    4.698e+01  1.692e+01   2.776 0.006364 **
t_cost_ami        -5.160e+00  1.414e+00  -3.649 0.000387 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16590 on 123 degrees of freedom
  (결측으로 인하여 36개의 관측치가 삭제되었습니다.)
Multiple R-squared:  0.5078,    Adjusted R-squared:  0.4718
F-statistic:  14.1 on 9 and 123 DF,  p-value: 2.039e-15
```

Conclusion

While this analysis was initially motivated by the City of Boston's launch of a discounted bikeshare membership in 2018, the results offer only **indirect evidence** of its impact. Increases in bikeshare usage among low-income and racially diverse communities may suggest **positive trends in access and inclusion**, yet the relationship remains complex and not uniformly significant across all measures.

Through this process, it also became evident that **infrastructure expansion played a critical role** during the same period. The addition of over 250 new stations between 2015 and 2019 likely contributed to improved access across neighborhoods—particularly in areas that were previously underserved. This suggests that **physical proximity and network density** may be as influential as affordability in driving usage.

The absolute number of bikeshare users in low-income census tracts increased significantly during this period. However, the regression analysis indicates that **income-based disparities in access and usage persisted**, highlighting the continued need to address structural barriers.

Although stronger statistical associations between low-income status and bikeshare usage were expected, the findings reflect the **short-term and exploratory nature** of this analysis. As such, it may not fully capture the multifaceted impacts of Boston's policy interventions.

Nevertheless, the data clearly reflect a city-wide effort to **expand bikeshare availability** and **reduce financial barriers to entry**. This short-term analysis provides a valuable foundation for future evaluations aimed at understanding the **long-term effectiveness and equity outcomes** of bikeshare initiatives in Boston and beyond.

Appendix 1. Data Dictionary

This data dictionary pertains to the **key datasets published by Metro Boston's public bike share program, the U.S. Census Bureau, and the Center for Neighborhood Technology (CNT): Blue Bike trip history data, ACS 2011-2015 and 2015-2019, and H+T Index 2016 and 2022**. Newly created dataframes and variables are marked with an asterisk (*).

Name	Type	Level	Description
ACS2015_19	DataFrame	Census Tract	A dataset containing household median income, racial composition, commute time to work, and income-to-poverty ratio data at the census tract level from American Community Survey 2015 - 2019.
asian	Number	Census Tract	Number of Asian population
avg_income*	Number	Census Tract	Average household median income of census tracts in each income quantile
avg_transport_cost*	Number	Census Tract	Average transportation cost of household median income of census tracts in each income quantile
avg_usage*	Number	Census Tract	Number of bike users by each station (total_usage divided by total_stations)
black	Number	Census Tract	Number of Black or African American population
bluebike15*	DataFrame	Record	A dataset containing trip data for 2015 in the Boston metropolitan area.
bluebike15_by_tract*	DataFrame	Census Tract	A dataset containing the number of total usage and total station by census tract.
bluebike15_sf*	Sf, DataFrame	Record	bluebike15_summary datasets converted into an sf object.
bluebike15_summary*	DataFrame	Record	A dataset subsetting the number of total users by start station and its geometry.
bluebike15_with_tract*	Sf, DataFrame	Census Tract	bluebike15 combined with census tract shapefile.
bluebike15_with_tract_clean*	DataFrame	Census Tract	bluebike15 combined with census tract shapefile and cleaned leaving only one tract to each station.
bluebike19*	DataFrame	Record	A dataset containing trip data for 2019 in the Boston metropolitan area.
bluebike19_by_tract*	DataFrame	Census Tract	A dataset containing the number of total usage and total station by census tract.
bluebike19_sf*	Sf, DataFrame	Record	bluebike19_summary datasets converted into an sf object.
bluebike19_summary*	DataFrame	Record	A dataset subsetting the number of total users by start station and its geometry.
bluebike19_with_tract*	Sf, DataFrame	Census Tract	bluebike19 combined with census tract shapefile.
bluebike19_with_tract_clean*	DataFrame	Census Tract	bluebike19 combined with census tract shapefile and cleaned leaving only one tract to each station.
COUNTY	Character	Census Tract	County Name
fifteen_to_nineteen	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 15 to 19 minutes
five_to_nine	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 5 to 9 minutes

fourty_to_fourtyfour	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 40 to 44 minutes
fourtyfive_to_fiftynine	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 45 to 59 minutes
GEOID	Character	Census Tract	Census Geographic Area Identifier
GISJOIN	Character	Census Tract	GIS Join Match Code
ht_index15	DataFrame	Census Tract	The Housing and Transportation (H+T) Affordability Index 2016. A dataset providing both the cost of housing and the cost of transportation.
ht_index19	DataFrame	Census Tract	The Housing and Transportation (H+T) Affordability Index 2022. A dataset providing both the cost of housing and the cost of transportation.
income_poverty	DataFrame	Census Tract	A dataset containing household median income and income-to-poverty ratio data at the census tract level from American Community Survey 2011 - 2015.
income_quantile*	Number	Census Tract	Income quantile (1 - 4)
less_than_five	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: Less than 5 minutes
med_income	Number	Census Tract	Median household income in the past 12 months
merged_df*	DataFrame	Census Tract	A dataset containing Blue Bike, ACS, and HT Index data merged for 2015.
merged_df_19*	DataFrame	Census Tract	A dataset containing Blue Bike, ACS, and HT Index data merged for 2019.
n_tracts*	Number	Census Tract	Number of census tracts belonging to each income quantile.
native_american	Number	Census Tract	Number of American Indian and Alaska Native population
ninety_or_more	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 90 or more minutes
overview_15*	DataFrame	Census Tract	A dataset containing bike usage, average income, number of bike stations, average transportation cost by income quantile.
overview_19*	DataFrame	Census Tract	A dataset containing bike usage, average income, number of bike stations, average transportation cost by income quantile.
pacific_islander	Number	Census Tract	Number of Native Hawaiian and Other Pacific Islander population
poverty_0_5to0_99	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. Under .51
poverty_1_00to1_24	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. 1.00 to 1.24
poverty_1_25to1_49	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. 1.25 to 1.49
poverty_1_50to1_84	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. 1.50 to 1.84
poverty_1_85to1_99	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. 1.85 to 1.99
poverty_2_00_over	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. 2.00 and over
poverty_under0_5	Number	Census Tract	Ratio of Income to Poverty Level in the Past 12 Months. Under .50

race_transportation	DataFrame	Census Tract	A dataset containing racial composition and commute time to work data at the census tract level from American Community Survey 2011 - 2015.
sixty_to_eightynine	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 60 to 89 minutes
start_station_latitude	Number	Record	The latitude of the bike station where the rental started.
start_station_longitude	Number	Record	The longitude of the bike station where the rental started.
start_station_name	Character	Record	The name of the bike station where the rental started.
STATE	Character	Census Tract	State Name
STATEAB	Character	Census Tract	State Postal Abbreviation
t_cost_ami	Number	Census Tract	Average household transit cost
ten_to_fourteen	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 10 to 14 minutes
thirty_to_thrityfour	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 30 to 34 minutes
thirtyfour_to_thritynine	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 35 to 39 minutes
total_stations*	Number	Record / Census Tract	The total number of Blue Bike stations.
total_usage*	Number	Census Tract	The total number of Blue Bike users aggregated by census tract level.
total_users*	Number	Record	The total number of Blue Bike users.
TRACTA	Character	Census Tract	Census Tract Code
twenty_to_twentyfour	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 20 to 24 minutes
twentyfive_to_twentynine	Number	Census Tract	Number of workers 16 years and over who did not work at home, whose travel time is: 25 to 29 minutes
usage_model_15_original*	List	Census Tract	Regression model 2015 with y variable of total_usage and x variables of median income, commute time to work, transportation cost
usage_model_19*	List	Census Tract	Regression model 2019 with y variable of total_usage and x variables of median income, income to poverty ratio, commute time to work, transportation cost, and race.
white	Number	Census Tract	Number of white population

Appendix 2. Annotated R Syntax

Code

Hide

```
library(ggplot2)
library(dplyr)
library(readr)
library(tidyr)
library(sf)
library(tmap)
library(scales)
```

A. Bluebike Analysis 2015

Data Cleaning - Load Bluebike trip, ACS 2011 - 2015, and HT index 2016 data.

1. The 2015 Bluebike trip data is loaded to bluebike15.

Hide

```
bluebike15 <- read_csv("Week 10/bluebike2015_clean.csv")
```

2. The bluebike15_summary dataframe was created to summarize geometry, total users, and average age by start station, and was then converted into an sf object (bluebike15_sf).

Hide

```
bluebike15_summary <- bluebike15 %>%
  group_by(start_station_name, start_station_latitude, start_station_longitude) %>%
  summarise(total_users = n()) %>%
  arrange(desc(total_users))
```

```
`summarise()` has grouped output by 'start_station_name', 'start_station_latitude'. You can o
verride using the `.groups` argument.
```

Hide

```
bluebike15_sf <- st_as_sf(bluebike15_summary, coords = c("start_station_longitude", "start_st
ation_latitude"), crs = 4326)
```

3. Household median income and income-to-poverty ratio data at the census tract level (2011–2015) were extracted from IPUMS NHGIS. The dataset was cleaned by filtering for Massachusetts, removing columns with NA values, and renaming income and poverty variables based on the codebook for clarity.

Hide

```
income_poverty<-read_csv("Week 10/income_2011-15.csv") %>%
  filter(STATE=="Massachusetts")  %>%
  select(STUSAB, GISJOIN, COUNTY, TRACTA, GEOID, ADNEE002:ADNEE008,  ADNKE001) %>% # Remove N
As and margins of error
  rename(poverty_under0_5 = ADNEE002,   # Ratio of Income to Poverty Level (ADNEE001 - 008)
         poverty_0_5to0_99 = ADNEE003,
         poverty_1_00to1_24 = ADNEE004,
         poverty_1_25to1_49 = ADNEE005,
         poverty_1_50to1_84 = ADNEE006,
         poverty_1_85to1_99 = ADNEE007,
         poverty_2_00_over = ADNEE008,
         med_income = ADNKE001  # Median Household Income (in 2015 Inflation-Adjusted Dollar
s)
         )

income_poverty$med_income <- as.numeric(income_poverty$med_income) # Convert med_income as nu
meric

income_poverty <- income_poverty %>%
  filter(med_income>0) # Filter income outliers below 0
```

Hide

```
race_transportation <- read_csv ("race_transportation_ACS2015/nhgis0008_ds215_20155_tract.cs
v")

race_transportation <- race_transportation %>%
  filter (STATE == "Massachusetts") %>%
  select(STUSAB, GISJOIN, COUNTY, TRACTA, GEOID, ADKXE002:ADKXE010,  ADLOE002: ADLOE013) %>%
  rename (white = ADKXE002,
          black = ADKXE003,
          native_american = ADKXE004,
          asian = ADKXE005,
          pacific_islander = ADKXE006,
          less_than_five = ADLOE002,
          five_to_nine = ADLOE003,
          ten_to_fourteen = ADLOE004,
          fifteen_to_nineteen = ADLOE005,
          twenty_to_twentyfour = ADLOE006,
          twentyfive_to_twentynine = ADLOE007,
          thirty_to_thrityfour = ADLOE008,
          thirtyfour_to_thritynine = ADLOE009,
          fourty_to_fourtyfour = ADLOE010,
          fourtyfive_to_fiftynine = ADLOE011,
          sixty_to_eightynine = ADLOE012,
          ninety_or_more = ADLOE013)
```

4. From the same NHGIS data source, a tract shapefile was downloaded and loaded as tract_sh to extract spatial data.

Hide

```
tract15_sh<-st_read("Week 10/nhgis0004_shape/nhgis0004_shapefile_tl2015_us_tract_2015")
```

5. HT Index 2016 was loaded as ht_index15, as it will be used for the 2015 analysis.

Hide

```
ht_index15 <- read_csv("htaindex2015_data_tracts_25.csv")
```

6. Remove “” with gsub function from tract variables of ht_index15.

Hide

```
ht_index15$tract <- gsub('"', '', ht_index15$tract)
```

Spatial joining and merging the datasets

7. Match the coordinate reference systems of the tract shapefile and bluebike15_sf, then perform a spatial join between them.

Hide

```
bluebike15_sf <- st_transform(bluebike15_sf, st_crs(tract15_sh))

bluebike15_with_tract <- st_join(bluebike15_sf, tract15_sh, join = st_within)

bluebike15_with_tract_clean <- bluebike15_with_tract %>%
  st_drop_geometry() %>%
  distinct(start_station_name, .keep_all = TRUE)  # Only leave one tract per one station
```

8. Summarize total usage and the number of stations per census tract and save it as bluebike15_by_tract. Finally, merge ht_index15 and ACS datasets (income_poverty, race_transportation) with bluebike15_by_tract using GEOID or GISJOIN.

Hide

```
bluebike15_by_tract <- bluebike15_with_tract_clean %>%
  group_by(GEOID) %>%
  summarise(total_usage = sum(total_users, na.rm = TRUE), total_stations = n_distinct(start_s
tation_name, na.rm = TRUE), GISJOIN = first(GISJOIN))


merged_df <- bluebike15_by_tract %>%
  left_join(ht_index15, by = c("GEOID" = "tract")) %>%
  left_join(income_poverty, by = "GISJOIN") %>%
  left_join(race_transportation, by = "GISJOIN")
```

Creation of Bluebike Trip Tables and Bar Charts

9. Create a summary table showing the total number of Bluebike stations and total bike users in 2015.

Hide

```
bluebike15_summary %>%
  ungroup() %>%
  summarise(total_stations = n(),                  #Total number of stations
            total = sum(total_users, na.rm = TRUE))  #Total number of bike users
```

10. Create a summary table showing the total number of stations, number of tracts, total usage, average usage per station, average transportation cost, and average income by income quantile.

Hide

```
overview_15<-merged_df %>%
  filter(!is.na(med_income)) %>%
  mutate(income_quantile = ntile(med_income, 4)) %>%
  group_by(income_quantile) %>%
  summarise(total_stations = sum(total_stations, na.rm = TRUE), n_tracts= n(), total_usage= s
um(total_usage),
            avg_usage = total_usage / total_stations, avg_transport_cost= mean(t_cost_ami), a
vg_income= mean(med_income))
```

11. Convert the overview_15 dataframe to long format to create bar charts with facet_wrap. Use ggplot2 and geom_col to generate the visualizations.

Hide

```
overview_15 <- overview_15 %>%
select(income_quantile, total_usage, avg_usage, total_stations) %>% # Select income_quantile,
total_usage, avg_usage, total_stations
pivot_longer(cols= c(total_usage, avg_usage, total_stations), names_to= "variable",values_to=
"value")


ggplot(overview_15, aes(x = factor(income_quantile), y = value, fill = variable)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ variable, scales= "free_y") +
  labs(x= "Income Quantile", y= "Value", title = "Bike Usage and Infrastructure by Income Qua
ntile (2015)") +
  scale_y_continuous(labels = label_comma())
```

Creation of Tmap

12. Join merged_df with the tract15 shapefile by GISJOIN, and ensure that tract15_sf is converted to an sf object for mapping.

Hide

```
tract15_sf <- left_join(merged_df, tract15_sh, by = "GISJOIN")
tract15_sf <- st_as_sf(tract15_sf)
```

13. Use tmap to create an interactive map of median income and station count in 2015. Display different income quantiles in distinct colors, and represent total usage and total stations using bubble sizes.

Hide

```
tmap_mode("view")
tm_shape(tract15_sf) +
  tm_fill("med_income", palette = "YlGnBu", style = "quantile") +
  tm_bubbles("total_usage",col = "darkgray", scale = 4,title.size = "Total Usage")  +
  tm_bubbles(size = "total_stations", col = "orange", scale = 1) +
  tm_layout(title = "Median Income and Station Count 2015")
```

14. Create a map using tmap to visualize areas by levels of transportation cost and bikeshare usage. Use color to represent transportation cost and bubble size to indicate total bike usage.

Hide

```
tm_shape(tract15_sf) +
  tm_fill("t_cost_ami", palette = "Reds", style = "quantile", title = "Transportation Cost Bu
rden") +
  tm_bubbles("total_usage", col = "darkgray", scale = 4, title.size = "Total Usage") +
  tm_layout(title = "Transportation Cost and Bike Usage 2015")
```

Regression Model

15. Construct a regression model with total_usage as the dependent variable and relevant independent variables. Iteratively remove statistically insignificant variables until only significant predictors remain in the model.

Hide

```
usage_model_15_original <- lm(total_usage ~ med_income +
                      I(med_income^2) +
                      five_to_nine +
                      sixty_to_eightynine +
                      t_cost_ami, data = merged_df)
summary(usage_model_15_original)
```

B. Bluebike Analysis 2019

Data Cleaning - Load Bluebike trip, ACS 2015 - 2019, and HT index 2022 data.

1. The 2019 Bluebike trip data is loaded to bluebike19. Column names are renamed appropriately for the R analysis.

Hide

```
bluebike19 <- read_csv("bluebike2019_combined.csv")
```

Hide

```
bluebike19 <- bluebike19 %>%
  rename (start_station_id = 'start station id',
          start_station_name = 'start station name',
          start_station_latitude = 'start station latitude',
          start_station_longitude = 'start station longitude')
```

2. The bluebike19_summary dataframe was created to summarize geometry, total users, and average age by start station, and was then converted into an sf object (bluebike19_sf).

Hide

```
bluebike19_summary <- bluebike19 %>%
  group_by(start_station_name, start_station_latitude, start_station_longitude) %>%
  summarise(total_users = n()) %>%
  arrange(desc(total_users))

bluebike19_sf <- st_as_sf(bluebike19_summary, coords = c("start_station_longitude", "start_st
ation_latitude"), crs = 4326)
```

3. Household median income, income-to-poverty ratio, race, and commute time to work data at the census tract level (2015–2019) were extracted from IPUMS NHGIS. The dataset was cleaned by filtering for Massachusetts, removing columns with NA values, and renaming the variables based on the codebook for clarity.

Hide

```
ACS2015_19 <-read_csv("2019Bluebike/nhgis0009_csv/nhgis0009_ds244_20195_tract.csv") %>%
  filter(STATE=="Massachusetts")  %>%
  select(STUSAB, GISJOIN, COUNTY, TRACTA, GEOID, ALUCE002:ALUCE006, ALU3E002:ALU3E013, ALWVE0
02:ALWVE008, ALW1E001) %>% # Remove NAs and margins of error
  rename(poverty_under0_5 = ALWVE002,   # Ratio of Income to Poverty Level (ADNEE001 - 008)
         poverty_0_5to0_99 = ALWVE003,
         poverty_1_00to1_24 = ALWVE004,
         poverty_1_25to1_49 = ALWVE005,
         poverty_1_50to1_84 = ALWVE006,
         poverty_1_85to1_99 = ALWVE007,
         poverty_2_00_over = ALWVE008,
         med_income = ALW1E001,  # Median Household Income (in 2015 Inflation-Adjusted Dollar
s)
         white = ALUCE002,
          black = ALUCE003,
          native_american = ALUCE004,
          asian = ALUCE005,
          pacific_islander = ALUCE006,
          less_than_five = ALU3E002,
          five_to_nine = ALU3E003,
          ten_to_fourteen = ALU3E004,
          fifteen_to_nineteen = ALU3E005,
          twenty_to_twentyfour = ALU3E006,
          twentyfive_to_twentynine = ALU3E007,
          thirty_to_thrityfour = ALU3E008,
          thirtyfour_to_thritynine = ALU3E009,
          fourty_to_fourtyfour = ALU3E010,
          fourtyfive_to_fiftynine = ALU3E011,
          sixty_to_eightynine = ALU3E012,
          ninety_or_more = ALU3E013)

ACS2015_19$med_income <- as.numeric(ACS2015_19$med_income) # Convert med_income as numeric

ACS2015_19 <- ACS2015_19 %>%
  filter(med_income > 0)
```

4. From the same NHGIS data source, a tract shapefile was downloaded and loaded as tract_sh to extract spatial data. The bluebike_with_tract dataset was then created by spatially joining bluebike_sf with tract_sh.

Hide

```
tract19_sh<-st_read("2019Bluebike/nhgis0009_shapefile_tl2019_us_tract_2019/US_tract_2019.sh
p")
```

5. HT Index 2022 was loaded as ht_index19, as it will be used for the 2019 analysis. (Released in 2022, but based on ACS 2015-2019)

Hide

```
ht_index29 <- read_csv("htaindex2022_data_tracts_25.csv")
```

6. Remove “” with gsub function from tract variables of ht_index19.

Hide

```
ht_index19$tract <- gsub('"', '', ht_index19$tract)
```

Spatial joining and merging the datasets

7. Match the coordinate reference systems of the tract shapefile and bluebike19_sf, then perform a spatial join between them.

Hide

```
bluebike19_sf <- st_transform(bluebike19_sf, st_crs(tract19_sh))

bluebike19_with_tract <- st_join(bluebike19_sf, tract19_sh, join = st_within)

bluebike19_with_tract_clean<- bluebike19_with_tract %>%
  st_drop_geometry()%>%
  distinct(start_station_name, .keep_all= TRUE)  # Only leave one tract per one station
```

8. Summarize total usage and the number of stations per census tract and save it as bluebike19_by_tract. Finally, merge ht_index19 and ACS datasets with bluebike19_by_tract using GEOID or GISJOIN.

Hide

```
bluebike19_by_tract <- bluebike19_with_tract_clean %>%
  group_by(GEOID) %>%
  summarise(total_usage = sum(total_users, na.rm = TRUE), total_stations = n_distinct(start_s
tation_name, na.rm = TRUE), GISJOIN = first(GISJOIN))


merged_df_19 <- bluebike19_by_tract %>%
  left_join(ht_index19, by = c("GEOID" = "tract")) %>%
  left_join(ACS2015_19, by = "GISJOIN")
```

Creation of Bluebike Trip Tables and Bar Charts

9. Create a summary table showing the total number of Bluebike stations and total bike users in 2019.

Hide

```
bluebike19_summary %>%
  ungroup() %>%
  summarise(
    total_stations = n(),                  # Total count of stations
    total = sum(total_users, na.rm = TRUE))  #Total number of bike users
```

10. Create a summary table showing the total number of stations, number of tracts, total usage, average usage per station, average transportation cost, and average income by income quantile.

Hide

```
overview_19<-merged_df_19 %>%
  filter(!is.na(med_income)) %>%
  mutate(income_quantile = ntile(med_income, 4)) %>%
  group_by(income_quantile) %>%
  summarise(total_stations = sum(total_stations, na.rm = TRUE), n_tracts= n(), total_usage= s
um(total_usage),
            avg_usage = total_usage / total_stations, avg_transport_cost= mean(t_cost_ami), a
vg_income= mean(med_income))
```

11. Convert the overview_19 dataframe to long format to create bar charts with facet_wrap. Use ggplot2 and geom_col to generate the visualizations.

Hide

```
overview_19 <- overview_19 %>%
select(income_quantile, total_usage, avg_usage, total_stations) %>% # Select income_quantile,
total_usage, avg_usage, total_stations
pivot_longer(cols= c(total_usage, avg_usage, total_stations), names_to= "variable",values_to=
"value")


ggplot(overview_19, aes(x = factor(income_quantile), y = value, fill = variable)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ variable, scales= "free_y") +
  labs(x= "Income Quantile", y= "Value", title = "Bike Usage and Infrastructure by Income Qua
ntile (2019)") +
  scale_y_continuous(labels = label_comma())
```

Creation of Tmap

12. Join merged_df with the tract19 shapefile by GISJOIN, and ensure that tract19_sf is converted to an sf object for mapping.

Hide

```
tract19_sf <- left_join(merged_df_19, tract19_sh, by = "GISJOIN")
tract19_sf <- st_as_sf(tract19_sf)
```

13. Use tmap to create an interactive map of median income and station count in 2015. Display different income quantiles in distinct colors, and represent total usage and total stations using bubble sizes.

Hide

```
tmap_mode("view")
tm_shape(tract19_sf) +
  tm_fill("med_income", palette = "YlGnBu", style = "quantile") +
  tm_bubbles("total_usage", col = "darkgray",scale = 4,title.size = "Total Usage")  +
  tm_bubbles(size = "total_stations", col = "orange", scale = 1) +
  tm_layout(title = "Median Income and Station Count 2019")
```

14. Create a map using tmap to visualize areas by levels of transportation cost and bikeshare usage. Use color to represent transportation cost and bubble size to indicate total bike usage.

Hide

```
tm_shape(tract19_sf) +
  tm_fill("t_cost_ami", palette = "Reds", style = "quantile", title = "Transportation Cost Bu
rden") +
  tm_bubbles("total_usage", col = "darkgray", scale = 4, title.size = "Total Usage") +
  tm_layout(title = "Transportation Cost and Bike Usage 2019")
```

Regression Model

15. Construct a regression model with total_usage as the dependent variable and relevant independent variables. Iteratively remove statistically insignificant variables until only significant predictors remain in the model.

Hide

```
usage_model_19 <- lm(total_usage ~ med_income +
                     I(med_income^2) +
                       poverty_under0_5 +
                       poverty_2_00_over +
                       black +
                       asian +
                       five_to_nine +
                       ten_to_fourteen +
                       t_cost_ami, data = merged_df_19)
summary(usage_model_19)
```