Examining Equity Across BlueBikes

Bryce Russell-Benoit

Overview

For the last few months, I have been working to examine the drivers of BlueBikes ridership in Boston. This question came to me during my discussion with the high school students that visited class early in the semester. They only just became eligible to ride BlueBikes due to their age so it's not too surprising that none of them had used BlueBikes before. However, they all had a generally negative view of the system. One student asked specifically why she would pay so much to use a BlueBike when she could drive wherever she needed to go. Many of the areas in the southern part of Boston are not very well-connected by transit and amenities are generally more spread out which makes owning a car pretty necessary for a family. If you already have a car payment and insurance, it doesn't make much financial sense to pay for a BlueBikes membership on top of other transportation costs. I hypothesized that this effect would be greater in low-income areas that are trying to minimize costs but cannot live without a car.

To test this hypothesis, I used the publicly available data from BlueBikes which contains information about all rides taken in a given month. The most relevant data about the rides includes where rides start and end along with the date, time, and duration. This makes it fairly easy to determine which stations have the most ridership and I can theoretically compare that ridership with census data such as income. However, census data is tied to where people live, and I don't have a simple way to determine what is a person's "home" station and what is their destination station. In this analysis I will work to disentangle BlueBikes rides to count rides where people live and then run a linear regression to determine if demographic or economic factors like income have an impact on whether people choose to use BlueBikes.

Data Preparation

Import and Cleaning

First, I need to import the BlueBikes dataset. In this study I will be looking at data from September 2019 since it is a fairly "normal" month without major transportation interruptions. Other months/years may have slightly different formatting or available data, but this instruction should apply with minor changes. Here I use read.csv() to import the file from where it's located on my computer and then save it to a dataframe object that I call "BlueBike".

```r
# Read the CSV file for the Sept 2019 BlueBikes data
BlueBike <- read.csv(file=BlueBikefilepath)
```

Next, I want to verify that my data is imported correctly by using the head() function to read the first 6 rows.

```r
# Output the first 6 rows
head(BlueBike)
```

tripduration	starttime	stoptime	start.station.id
916	2019-09-01 00:00:21.2560	2019-09-01 00:15:38.0670	9
394	2019-09-01 00:00:33.0140	2019-09-01 00:07:07.1410	189
480	2019-09-01 00:00:52.4870	2019-09-01 00:08:52.9530	9
800	2019-09-01 00:01:05.5390	2019-09-01 00:14:25.8490	46
758	2019-09-01 00:01:13.4250	2019-09-01 00:13:52.2810	56
261	2019-09-01 00:01:33.7780	2019-09-01 00:05:55.5560	124

Everything in the data looks good, but there is some important information here I need to extract. The starttime and stoptime columns are long strings of characters that can't easily be interpreted. I need to convert them to a more interpretable format using lubridate. This will allow me to easily extract information like the hour of the day or day of the week which I will do later. Based on the format of the datetime, I will use ymd_hms() to convert the starttime and stoptime columns.

```r
# Import necessary libraries
# You may need to use install.packages("*Package_Name*") before running this
library(tidyverse)
library(lubridate)

# Convert starttime and stoptime
BlueBike$starttime <- ymd_hms(BlueBike$starttime)
BlueBike$stoptime <- ymd_hms(BlueBike$stoptime)
```

With this updated datetime column, I can easily extract the day of the week and hour that the ride started and ended for further analysis. I will do this with the wday() function and the hour() function respectively.

```r
# Create columns for day of the week
BlueBike$startday <- wday(BlueBike$starttime, label=TRUE)
BlueBike$stopday <- wday(BlueBike$stoptime, label=TRUE)

# Create columns for hour of the day
BlueBike$starthour <- hour(BlueBike$starttime)
BlueBike$stophour <- hour(BlueBike$stoptime)
```

Now with these additional columns, I can start to filter to get to the critical data I need.

Filtering

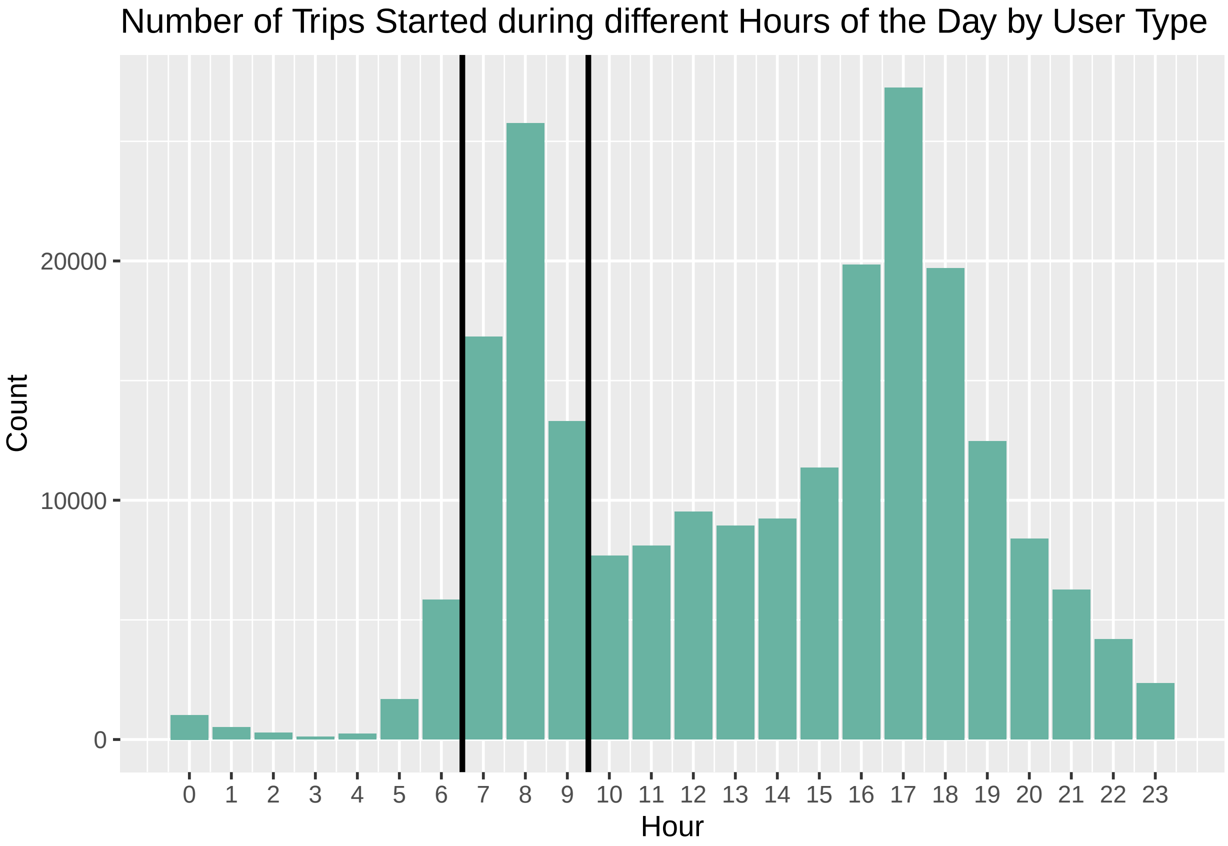
In order to compare BlueBikes ridership to Census data I will need to attribute ride counts to users "home" stations. One way to do this is to assume a commuter pattern so that I count 1 ride where a trip starts in the morning and 1 ride where a trip ends in the evening. This way, rides will be counted at the station closest to where people live rather than where they work.

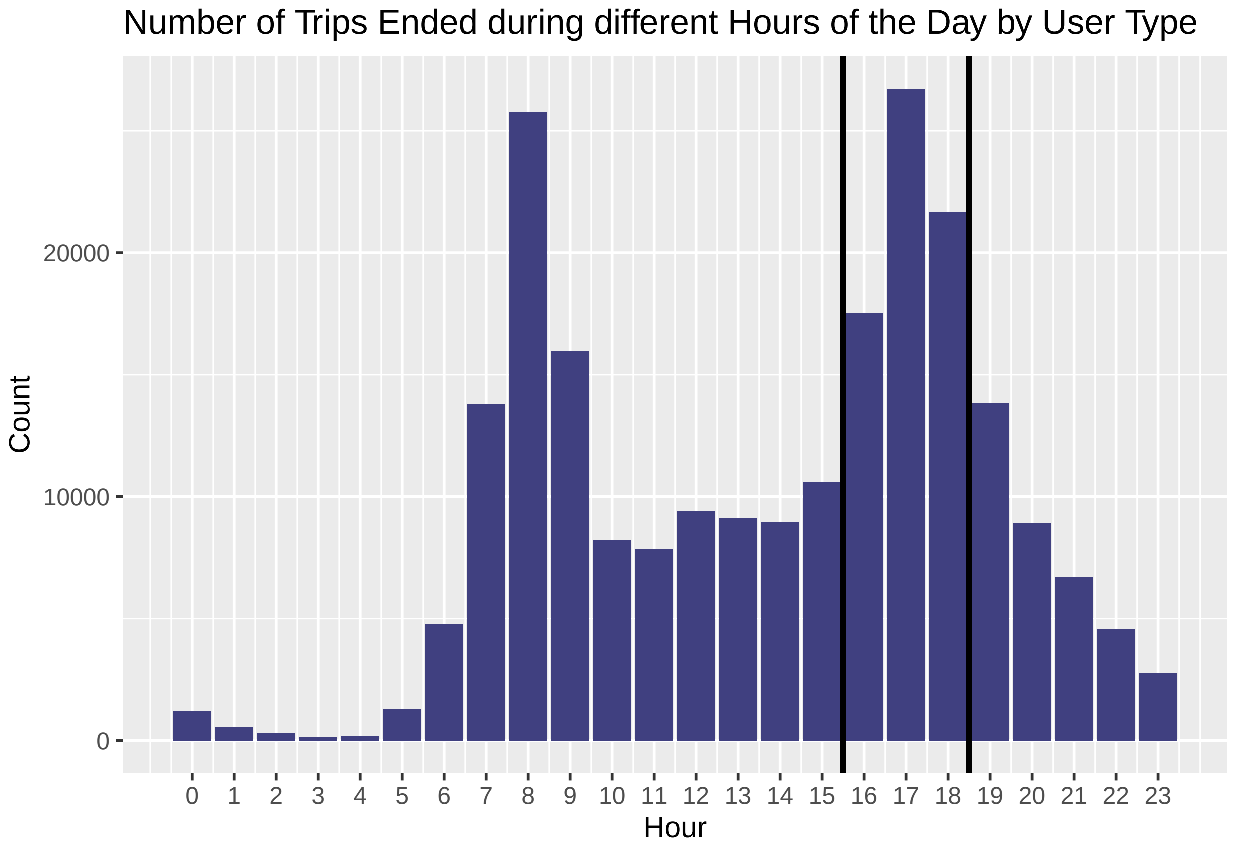
In order to focus in on commuter behavior I can filter the data a little more. Trips that last over an hour are less likely to be routine trips to and from home so I can filter those out. I can also use the usertype column which indicates whether the user is a monthly subscriber which is much more economical for commuters. Finally, I can remove weekends where 9-5 work is less common. I can add all of these conditions with filter().

```r
# Import needed library
library(dplyr)
```

```
# Filter to short trips from Subscribers on weekends
BlueBike <- filter(BlueBike, tripduration <= 3600,
                   usertype == "Subscriber",
                   startday != "Sat", startday != "Sun")
```

In order to determine home stations, I'll need to examine what are the main commuting hours using ggplot() and geom_bar().





There seems to be two 3 hour peaks in the morning between 7:00 and 9:59 and in the evening between 16:00 and 18:59. Now I will use these hours to try to only count the home and away stations. I'll start by creating 2 different dataframes for these morning and evening rides with only the name and ID columns for counting.

```
# Set the column names needed for this aggregation
key_columns <- c("start.station.id", "start.station.name",
                 "end.station.id", "end.station.name")

# Filter to morning rides starting between 7:00 and 9:59
MorningRides <- BlueBike[BlueBike$starthour <= 9 & BlueBike$starthour >= 7, key_columns]
# Filter to evening rides ending between 16:00 and 18:59
EveningRides <- BlueBike[BlueBike$stophour <= 18 & BlueBike$stophour >= 16, key_columns]
```

Now I'll use the aggregate() function to count the amount of time each station was the start station in the morning or the end station in the evening. Then I will combine these two aggregations together using merge() and add up the counts.

```
# Calculate frequency of morning start station names and ID using aggregate()
start_station_counts <- aggregate(MorningRides$start.station.id,
                                  by = list(MorningRides$start.station.id,
                                            MorningRides$start.station.name),
                                  FUN = length)

# Calculate frequency of evening end station IDs using aggregate()
end_station_counts <- aggregate(EveningRides$end.station.id,
                                by = list(EveningRides$end.station.id),
                                FUN = length)

# Rename the columns
colnames(start_station_counts) <- c("Station_ID", "Name", "Start_Count")
colnames(end_station_counts) <- c("Station_ID", "End_Count")

# Merge into a single dataframe
commuter_flow <- merge(start_station_counts, end_station_counts, by="Station_ID")

# Add columns together to get a total count
commuter_flow$Total_Count <- commuter_flow$Start_Count + commuter_flow$End_Count
```

Lets check in with how this looks.

```
# Output the first 6 rows
head(commuter_flow)
```

Station_ID	Name	Start_Count	End_Count	Total_Count
3	Colleges of the Fenway - Fenway at Avenue Louis Pasteur	216	217	433
4	Tremont St at E Berkeley St	273	361	634
5	Northeastern University - North Parking Lot	247	338	585
6	Cambridge St at Joy St	389	501	890
7	Fan Pier	39	70	109
8	Union Square - Brighton Ave at Cambridge St	268	262	530

Station Data

If I want to join the BlueBikes station data with Census data, I will need to know which Census tract each station resides in so I'll have to pull in some outside data just like I did in the beginning. I will also filter this station list down to stations only within Boston that existed in 2019.

```
# Read the CSV file for BlueBikes station data
Stations_GIS <- read.csv(file= Stationsfilepath)

# Filter the dataset to only stations that were added in or before 2019
Stations_GIS <- filter(Stations_GIS,
                       Deployment.Year <= 2019,
                       District == "Boston")
```

Now I can add the flow data onto this so it can be joined with the Census data.

```
# Merge in flow data
Stations <- merge(Stations_GIS, commuter_flow, by="Name")
```

Now I can add up ridership between all stations within each Census tract using aggregate().

```
# Aggregate ride counts at the tract level
Tract_flow <- aggregate(Stations$Total_Count, by = list(Stations$CT_ID_10), FUN = sum)
colnames(Tract_flow) <- c("CT_ID_10", "Tract_Count")
```

```r
# Output the first 6 rows
head(Tract_flow)
```

CT_ID_10	Tract_Count
25025000100	928
25025000201	290
25025000502	311
25025000504	303
25025000602	840
25025000701	409

Census Data

I will add more data that includes information about population, demographics, income, and more from BARI. Then I will add the ride counts at the tract level.

```r
# Read the CSV file for Census income data and skip first row data key
Census <- read.csv(file= Censusfilepath)

# Add Tract_flow data to Census data by the tract ID. Include all tracts for mapping
Census <- merge(Census, Tract_flow, by="CT_ID_10", all = TRUE)
```

```r
# Output the first 6 rows
head(Census)
```

CT_ID_10	TotalPop	PopDen	Age1834	White	Black	MedHouseIncome	GINI	MedGrossRent
25001010100	2973	307.434	0.106	0.895	0.019	59063	0.528	1248
25001010206	3617	182.823	0.155	0.892	0.019	74639	0.423	1196
25001010208	1122	53.548	0.053	0.950	0.034	68367	0.405	1131
25001010304	2394	337.943	0.114	0.914	0.011	85263	0.489	966
25001010306	2507	364.198	0.091	0.917	0.022	70071	0.381	1360
25001010400	3066	509.457	0.069	0.892	0.068	66364	0.446	1269

Now I have 3 main data sets I can use for my analysis:

- BlueBikes
  - All BlueBikes trips filtered to Subscriber trips on Weekdays
- Stations
  - Station names and locations with ridership counts
- Census
  - Census tracts with indicators and ridership counts

Analysis

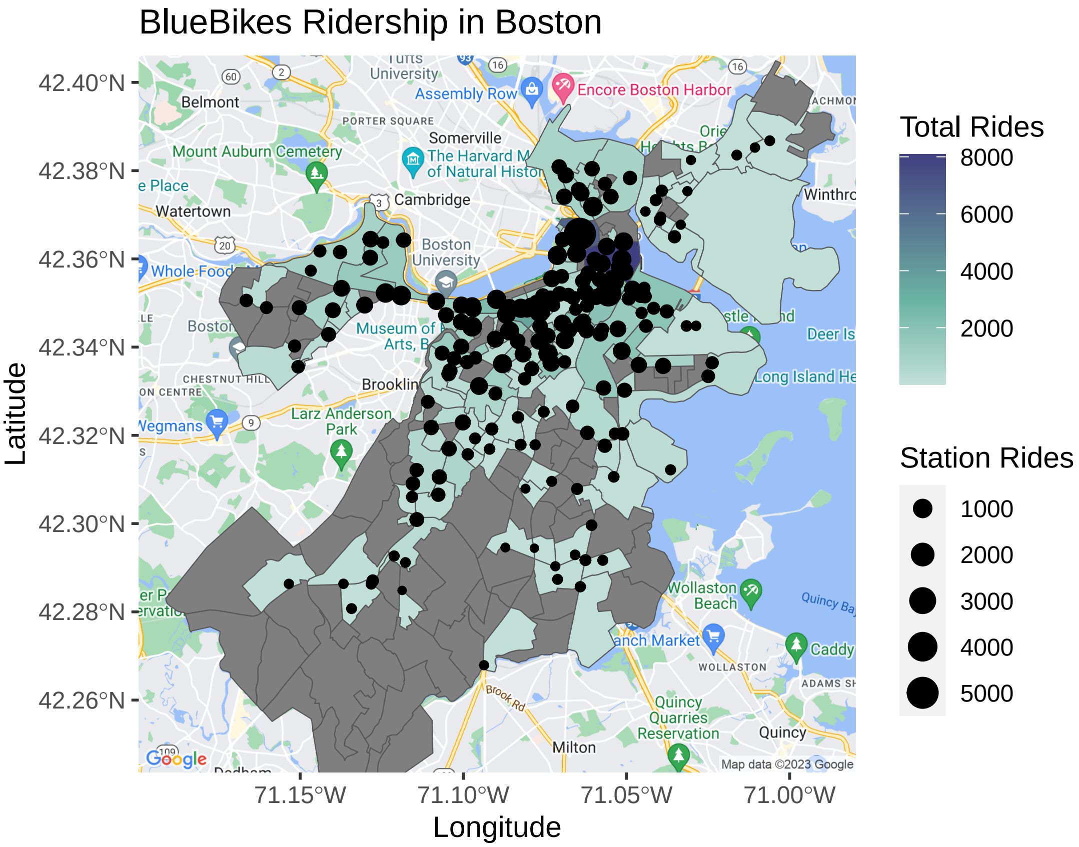
Descriptive Statistics

First I'll want to map the distribution of ridership to see where the most rides are taken in Boston. After merging my dataset with a shapefile for mapping, I'll need to set the dimensions of my map using get_map() from ggmap.

```r
# Get basemap for Boston
Boston <- get_map(location=c(left = -71.193799,
                           bottom = 42.22,
                           right = -70.985746,
                           top = 42.43), source="google", maptype = "roadmap")

Bostonmap <- ggmap(Boston)
```

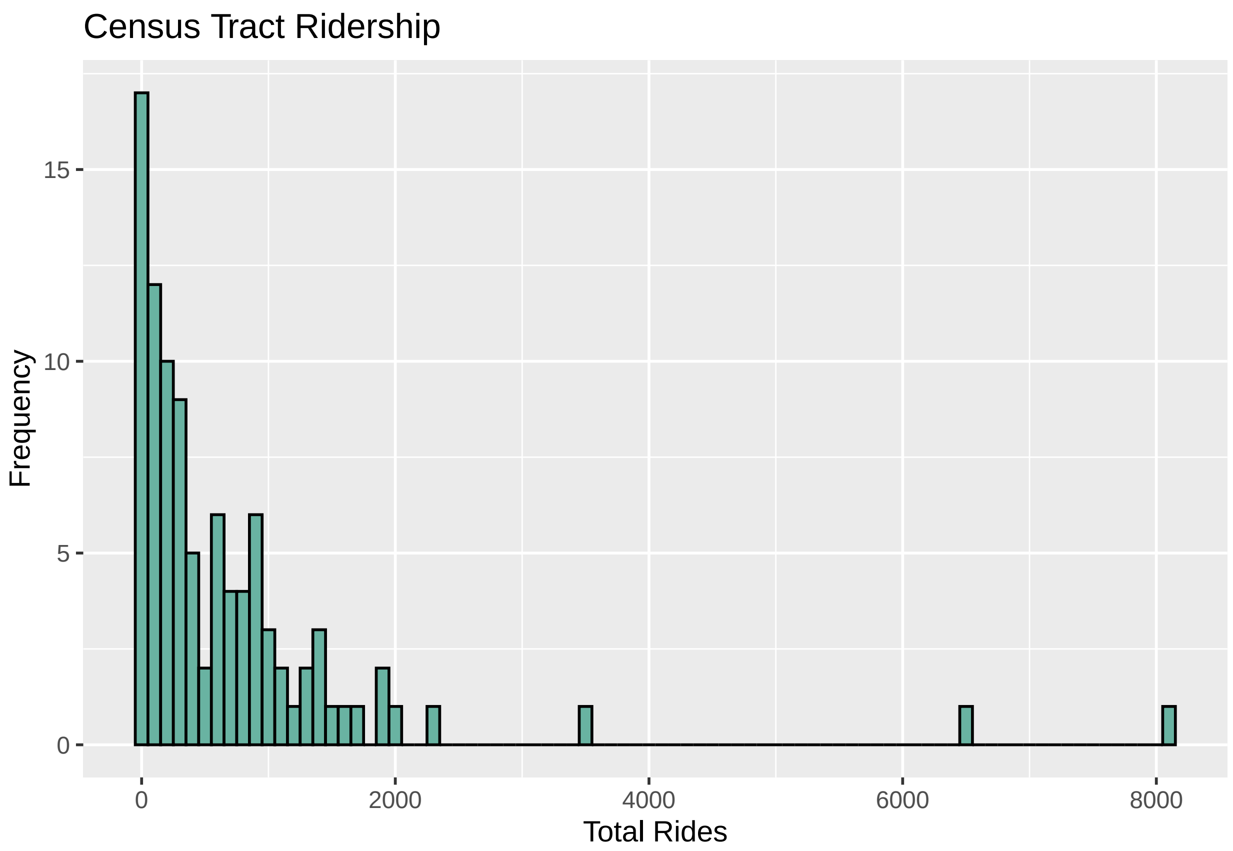
Now I can use ggplot and ggmap's geom_sf() to display the Census tracts and geom_point() to display the individual BlueBikes stations.

```r
# Start with basemap
Bostonmap +

    # Map Census tracts and color using Tract_Count
    geom_sf(data=tracts_full_geo, aes(fill=Tract_Count),inherit.aes = FALSE) +
    scale_fill_gradient2(low="white", mid="#69b3a2", high="#404080", midpoint=3000) +

    # Add stations using latitude and longitude
    geom_point(data=stations_full_geo,
               aes(x=Longitude, y=Latitude, size=Total_Count), inherit.aes = FALSE) +
    # Limit dot size for readability
    scale_size(range = c(1,5)) +

    # Set title and labels
    labs(title="BlueBikes Ridership in Boston",
       x= "Longitude", y="Latitude",
       size="Station Rides", fill="Total Rides")
```



Overall, BlueBikes seems to get less popular as distance from the downtown increases, but it's not an even gradient. There seem to be a few major outliers for ridership in downtown, this will be easier to see in an histogram using ggplot's geom_histogram().

```
# Ridership Histogram at the tract level
ggplot(Census, aes(x = Tract_Count)) +
    geom_histogram(binwidth = 100, fill = "#69b3a2", color = "black") +

    # Set title and labels
    labs(title = "Census Tract Ridership",
       x = "Total Rides",
       y = "Frequency")
```



The vast majority of Census tracts had less than 1,000 rides in September, but there are 3 extreme outliers on the higher end that were visible on the map as Government Center, North Station/TD Garden, and the Prudential Center. All of these areas contain major transit hubs where riders will start trips in the morning after taking a train into this part of the city. These rides appear to come from residents due to my filtering, but they are artificially inflated by multimodal commuters.

Regression

Before I do a regression analysis, the extreme outliers identified in the map and histogram should be removed. The stations in these tracts operate differently due to their extreme transit connectivity that makes it unlikely that the bulk of riders are residents.

```r
# Remove extreme outliers due to last mile commuters downtown
Census <- Census[Census$Tract_Count < 3000, ]
```

Now I'll run an initial regression with all factors that I hypothesize could be relevant or that could illuminate an important disparity using lm(). I'm excluding factors about how people commute as that would have too direct of a correlation and not tell me anything about why people choose to use BlueBikes.

```r
# Create a linear model to predict ridership
lm <- lm(Tract_Count ~ TotalPop + PopDen + # Potential users
             MedHouseIncome + MedGrossRent + GINI + # Economic indicators
             Age1834 + White + Black, # Demographic data
         data=Census)
```

```r
# Output model summary
summary(lm)
```

Table 5:

	Dependent variable:
	Tract_Count
TotalPop	0.017
	(0.034)
PopDen	−0.002
	(0.003)
MedHouseIncome	0.004
	(0.003)
MedGrossRent	0.186
	(0.150)
GINI	3,425.938***
	(628.233)
Age1834	895.324**
	(423.410)
White	39.671
	(513.879)
Black	−262.249
	(424.481)
Constant	−2,111.146***
	(486.315)
Observations	88
R^2	0.496
Adjusted R^2	0.445
Residual Std. Error	413.133 (df = 79)
F Statistic	9.708*** (df = 8; 79)
Note:	*p<0.1; **p<0.05; ***p<0.01

Not all factors are significant and there's likely some colinearity between them. I will use backwards elimination to remove factors with the highest p-value or *Pr(>|t|)* one at a time until all remaining factors are significant with p-values less than 0.05.

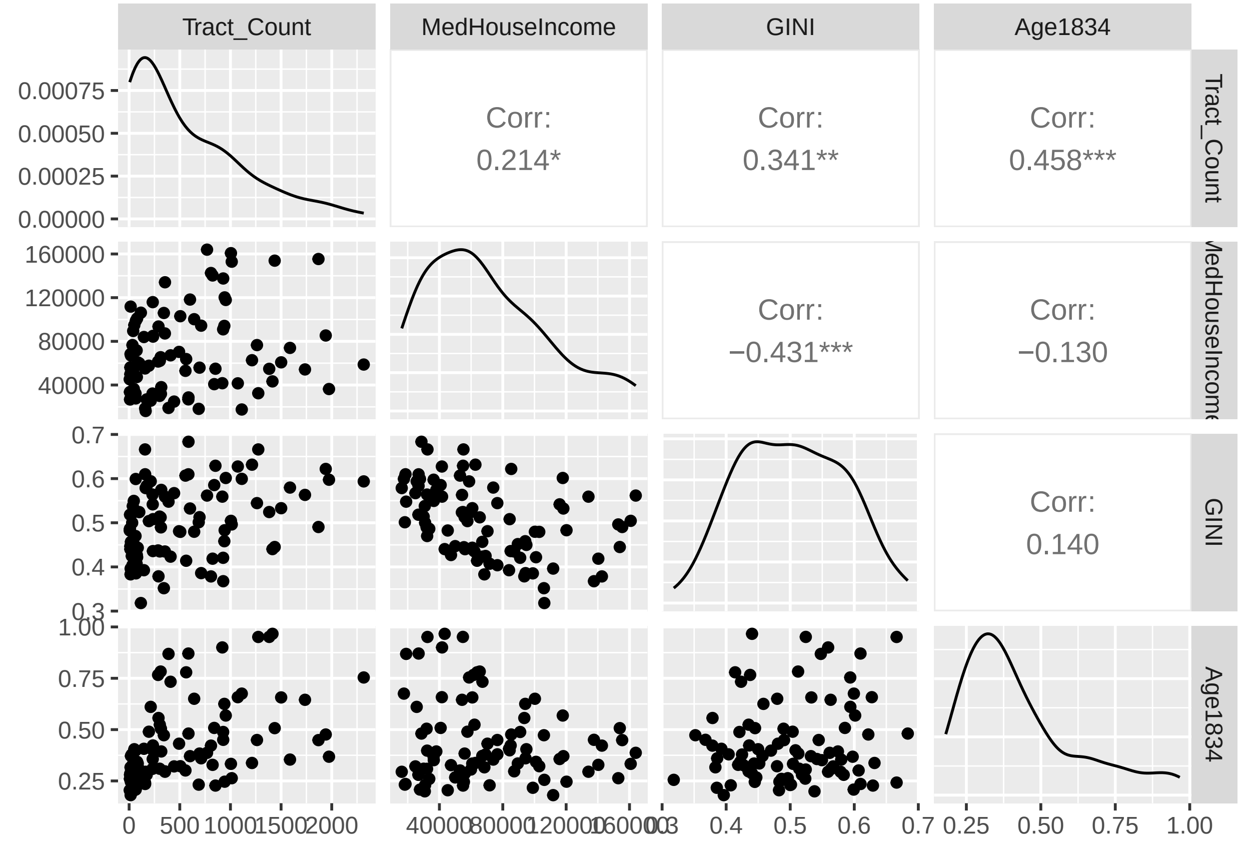
```r
# Create a linear model to predict ridership
lm <- lm(Tract_Count ~
             MedHouseIncome + GINI + # Economic indicators
             Age1834, # Demographic data
         data=Census)

# Output model summary
summary(lm)
```

After backwards elimination is complete, all that remains is Median Household Income, GINI Index, and percent of the population between the ages of 18 and 34. To understand this modela bit better, I need to examine relationships between these factors using ggpairs() which will show correlation coefficients on the top right half and scatter plots in the bottom left half with density plots in between.

Table 6:

	Dependent variable:
	Tract_Count
MedHouseIncome	0.007***
	(0.001)
GINI	3,286.853***
	(590.927)
Age1834	1,270.126***
	(222.662)
Constant	$-$2,139.024***
	(353.812)
Observations	89
R^2	0.477
Adjusted R^2	0.458
Residual Std. Error	407.288 (df = 85)
F Statistic	25.829*** (df = 3; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# Import library for colinearity plot
library (GGally)

# View colinearities
ggpairs(data=Census,
        columns=c("Tract_Count", # Target
                  "MedHouseIncome", "GINI", # Economic indicators
                  "Age1834"))  # Demographic data
```



All factors have an individually strong correlation with the target although median income is the smallest.

It appears that a higher income does have a significant impact on ridership. When all other factors are held constant, a $1,000 increase in median income should result in approximately 70 more rides in September. Income is also a part of the GINI coefficient which is a measure of income inequality or how abnormal the income distribution is. In this model, a 1% increase in income inequality would result in about 33 more rides in September with all other factors held constant. Finally we have the proportion of the population between ages 18 and 34 which is on the same scale as GINI. That means that a 1% increase in that proportion results in about 10 more rides in September with other factors held constant.

All factors are extremely significant and reasonably impactful, resulting in a model that explains about 46% of the variation in ridership with just 3 variables.

Conclusion

Unfortunately, BlueBikes do seem to have some major differences in use due to economic factors. Lower income neighborhoods are significantly less likely to use BlueBikes. It seems quite easy to justify this since incomes tend to go up as one gets closer to the urban core of Boston. It seems like BlueBikes might just be less useful to people living farther outside of the city in less dense areas. However, population and population density were quickly removed from my model due to insignificance which should've been able to reveal that relationship if that was the key. It seems more likely that there is a genuine difference that makes BlueBikes less attractive in low income areas due to its cost or other forms of disinvestment such as low-quality bike infrastructure or none at all.

The GINI coefficient is slightly more abstract and harder to integrate into the conceptual model. GINI and median income and negatively correlated, but both of them have a positive impact. It looks like the correlation becomes weaker at higher income levels though as the GINI coefficients vary a lot more. It seems like these tracts with high income and high GINI coefficients tend to perform better on ridership than high income tracts with low GINI coefficients. It's possible that this is representing new-build areas with subsidized income-restricted housing that are more bikeable, but this would require a much closer investigation to prove.

I'm not surprised to find that the people ages 18-34 are more likely to use BlueBikes since I've personally noticed that college students or college-age people tend to be some of the most common groups I see riding BlueBikes. Not only do many universities offer discounts for BlueBikes, but most university students don't have cars and don't want to transport a bike between semesters which makes it an especially attractive option for commuting to and from class.

This process of joining and analyzing data has provided a glimpse into why people choose to use BlueBikes. BlueBikes is one small part of a larger transportation system in Boston, but it provides a very useful connection that allows multi-modal transportation without needing to worry about bike storage and maintenance. However, there are some aspects of the system that are holding it back from working for more people in the city. Residents who are uninterested in BlueBikes aren't avoiding it because they just don't like it. People don't use systems that don't work for them. That may be because it's confusing, inconvenient, unsafe, expensive, unreliable or any number of reasons. By identifying income and even age as potential factors that limit or encourage use, we can start thinking more creatively about how to make this system work for more people and make Boston a more equitable and accessible city.

Appendix: Data Dictionary

New Columns in Bold

Record Level (Rides)

- start.station.id – Station ID for where the BlueBikes trip started
- start.station.name – Station name for where the BlueBikes trip started
- end.station.id – Station ID for where the BlueBikes trip ended
- end.station.name – Station name for where the BlueBikes trip ended

Station Level

- Station_ID – Station ID from record level
- Name – Station name from record level
- **Start_Count** – Count of times the station was the origin of a BlueBikes trip
- **End_Count** – Count of times the station was the destination of a BlueBikes trip
- **Total_Count** – Sum of Start_Count and End_Count or the total number of times the station was used during a BlueBikes ride
- CT_ID_10 – Census Tract ID number
- Deployment.Year – The year the station was deployed
- District – The town/city where the station is located

Census Tract Level

- CT_ID_10 – Census Tract ID number
- **Tract_Count** - Count of all commuter trips at stations within a Census tract
- TotalPop - Total population
- PopDen - Population per square mile
- MedHouseIncome - Median household income
- MedGrossRent - Median gross rent in USD
- GINI - GINI index, a commonly-used measure of income inequality
  - GINI has a range from 0 = perfect equality to 1 = perfect inequality
- Age1834 - Percentage of residents with ages between 18 and 34
- White - Proportion of White Non-Hispanic residents
- Black - Proportion of Black Non-Hispanic residents