No Easy Solution: An Observational Study on the Relationship Between Various Factors, Traffic Congestion, and Highway Fatalities

Harper Kates

2023-03-01

1. Abstract

Traffic congestion is an important subject to study, as the negative effects and complexities of this issue are well-documented. Prior research hypothesizes that public transportation is negatively correlated to traffic congestion. This study started with the same hypothesis but took many other factors into consideration when modeling traffic congestion. This study uses three different city-level traffic congestion-related response variables across the US, as well as a highway fatality-related variable at the state level. The results show that variables including region of the US, binge drinking, and number of people per household all play noticeable roles in explaining traffic congestion in cities, while variables such as median income, gas usage, and miles driven play noticeable roles in explaining highway fatality rate in states. The results also found that public transport ridership is positively correlated with traffic congestion, which can be explained by the fact that cities that naturally have more traffic congestion will invest in public transportation as a potential solution. Overall, the study does not find any significant evidence to confirm or reject the hypothesis that public transportation is good for traffic, but it introduces many new variables that appear to have relationships with traffic congestion and/or highway fatality rate.

2. Background and significance

Traffic congestion is something that almost everyone has experienced at least sometime in their lives, some people more than others. There are many negative effects of traffic congestion, but there does not exist an easy "solution" to this problem. Some of these negative effects can be practical, such as losing hours due to congestion and potentially being late to important events. It can also have psychological effects, as seen in "road rage," where even the most mild-mannered driver can become irritable and unforgiving towards other drivers. In fact,

there is scientific evidence that traffic congestion is heavily correlated with increased levels of stress and anxiety, which leads to a decline in mental health ("'I Am Sick and Tired of This Congestion'_ Perceptions of Sanandaj Inhabitants on the Family Mental Health Impacts of Urban Traffic Jam | Elsevier Enhanced Reader" n.d.).

The closest we have to a "solution" to traffic congestion is investing more in public transportation, but this may not work as well in less densely populated areas. This discrepancy in public transportation effectiveness is because public transportation works over a specific area, so it will be more effective for densely populated areas than for less densely populated areas. This effect can sometimes make up for the natural development of traffic congestion that more often appears in densely populated areas. Some public transportation experts state that the threshold for implementing reliable public transportation systems is around 3,000 people per square mile ("Excerpt: Many Cities Have Transit. How Many Have Good Transit? | Kinder Institute for Urban Research" n.d.). However, this number can potentially change as many newer cities are starting to fall short of this mark due to rising income and lower transport costs (Malpezzi, n.d.). After studying the effects of public transportation, one can ask why cities are designed to be less population-dense in the first place if this lack of density lowers the effectiveness of public transportation.

Traffic congestion is a valuable subject to study, but why did I specifically choose traffic congestion as my topic of choice? I have always been intrigued (mainly frustrated) by traffic congestion in the area where I grew up. For reference, I grew up in the Nashville metropolitan area, an area of the US with high traffic congestion despite being less population-dense than other big cities. This level of congestion is not only present on highways; it is arguably worse within major cities such as Nashville, Murfreesboro, or Franklin. Furthermore, the way cities in the area have been constructed (more highways and suburbs, less walkable cities) yields traffic congestion in pretty much every part of the city, not just the city center. There have been countless examples of my car going at an average speed lower than my walking speed for minutes at a time, which prompts me to ask this question time and time again: How is this an acceptable way to live?

A less emotionally charged reason as to why I chose this subject is because there are plenty of statistics out there pertaining to not only traffic metrics, but other factors that may explain/relate to said traffic metrics. These other factors can be demographics, health factors, education, etc. Studying traffic patterns will inevitably involve statistics in some form, so this subject would be perfect for a statistics-based project. It is important to make the distinction between correlation and causation. Many things are said to "cause" traffic congestion, but some of these factors may just be simple correlations. I decided to do an observational study on traffic factors, so no conclusions on causation can be made, but I will make note of factors that can be reasonably argued to have a causal relationship with traffic metrics.

I have personally heard many good things about public transportation and how it can be a remedy for traffic congestion, but, being from an area with poor public transportation systems, I had never experienced this positive effect. This means that I have preconceived positive notions about public transportation without any real experience to back it up. This

project will be a perfect opportunity to prove or disprove this notion, but I have to be careful not to implement any confirmation bias in my findings. Like traffic metrics, there are plenty of metrics out there to measure public transit ridership, whether it be over time, per capita, etc. In this paper I will explore whether or not public transportation truly is related to good traffic flow, as well as determine other potential factors for explaining traffic congestion.

3. Methods

Since there is no one source that includes every variable I need, the data came from many different sources. One source, INRIX, contained many traffic metrics by city, which will all be response variables. The ones I chose were hours lost due to traffic congestion per person per year (or just hours lost due to traffic for short), change in traffic due to COVID (change in congestion from 2019 to 2020), and last-mile speed (average speed in mph within 1 mile from the city center). INRIX is a transportation data analysis firm that gives reliable information on countless traffic metrics not only throughout the US, but around the world. Unfortunately, the page has been updated since I retrieved the data (also, the change due to COVID column was removed), so if this study were to be replicated, it would require a different data set. Therefore, it is important to note that the data was retrieved on 9/5/22. Also, I intended to look at traffic metrics by state, but since INRIX only includes relatively large cities in their data set, some states have no cities. This means that I will actually be using 2 separate data sets, one for cities and the other for states. The main state-level data will come from the Bureau of Transportation Statistics (BTS), an official government-licensed database that includes metrics on highway fatalities, public transport ridership, total gas usage and miles driven, number of registered licenses, and number of total vehicles. The response variable in question is highway fatalities per capita, or highway fatality rate. The rest of the data sets include potential explanatory variables.

Most of the city and state-level demographic data comes from the US Census Bureau's 2021 American Community Survey. This data includes the basic demographic factors (age, race, gender, etc.), but for the state-level data, I also created a variable that measures the proportion of people between the ages of 25 and 64, which is what I would consider the prime driving age range. Drivers younger than 25 generally have worse judgment; younger drivers are generally less experienced and "are more likely than older drivers to perform risky driving behaviors such as speeding, close following, and smaller gap acceptance," (Williams 2006). In contrast, elderly drivers, defined as older than 65, tend to have deteriorated motor skills, health problems, and are more likely to experience side effects of medications (Lyon et al. 2020). Therefore, it would make sense that states with higher proportions of people in their "prime driving age" would generally be less prone to car accidents and/or fatalities, assuming all other factors are kept constant.

The city data set also includes health-related variables, which come from the CDC and are measured not by city but by census tract. I decided to take the average of each census tract for each city for each variable, and after some quick cleaning, I was able to merge this data set

into the full data set. The CDC data set includes 28 different health-related crude prevalence variables, including crude prevalence of binge drinking, getting less than 8 hours of sleep per day, high blood pressure, chronic obstructive pulmonary disease, and older people being up to date on their core medical procedures, as well as a few others. I thought this set of variables would be important because health-related factors would almost certainly have something to do with change in traffic due to COVID; I hypothesized that those with pre-existing conditions during the COVID outbreak would be less likely to travel.

As discussed earlier, population density (people per square mile) will be an integral part of this project, so finding a reputable source is paramount. I decided to use the data from simplemaps.com, which cites government-backed sources such as the previously mentioned US Census Bureau and American Community Survey.

Another one of my preconceived ideas heading into this project was the general importance of setting a reasonable speed limit when planning traffic flow. Hence, state speed limits for urban and rural areas also had to be considered when creating the full state data set. The source I used was the Insurance Institute for Highway Safety (IIHS), a non-profit organization renowned for its scientific use of crash testing and general knowledge about highway safety. This is also perfect for the data set used in my project since the main metric for the state data set looks at highway fatalities.

Median income could also have some effect on highway fatality rate; it would be reasonable to hypothesize that poorer areas generally have less resources to focus on road construction and thus can have higher frequencies of car accidents. The same is true for income growth, but it is less clear what direction the correlation would be. Growth in income can either help with or worsen traffic congestion depending on how willing the state transportation authorities are to invest in beneficial traffic modifications. Both of these income-related variables come from the Federal Reserve Economic Data, a database, owned by the Federal Reserve Bank of St. Louis, that has been measuring various economic factors over time since 1991. Since the data set for my project is a snapshot in time, I only want to look at median income in 2021 and income growth from the preceding period (2020).

Like the city data set, the state data set also includes a variable relating to alcohol. However, instead of looking at prevalence of binge drinkers, this variable measures the per capita alcohol consumption for each state. As publisher of the data set John Elflein explains, "although New Hampshire consumes the highest amount of alcohol per capita, it reports lower rates of binge drinking than other states," ("Total Alcohol Consumption Per Capita by U.S. State 2020" n.d.). This highlights the difference between alcohol consumption and binge drinking, which can potentially tell us that alcohol consumption per capita will most likely explain less about highway fatality rate than binge drinking for traffic congestion. Even so, I decided to include this variable just to determine if prevalence of binge drinking and alcohol consumption per capita truly have different effects, as John Elflein implies.

Finally, I decided to add group quarters statistics to the data set, since more people living in group quarters tends to indicate a higher population density, in which the average person has to

travel fewer miles and uses less high-speed roadways, reducing the chance of highway fatalities. According to the US Census Bureau, this variable can be broken down into two categories: institutionalized, which includes correctional facilities, mental hospitals, and nursing facilities; and non-institutionalized, which includes college dorms and military quarters. Due to the variety of institutionalized group quarters, simply using institutionalized group quarters (or proportion) as a variable does very little to explain anything of value. Instead, we will look at the following variables (measured on a per-100000-people basis): people living in adult correctional facilities, people living in nursing facilities, people living in college dorms, and people living in military quarters. Each of these variables has its own explanation as to why it could potentially be useful.

Theoretically, having a higher proportion of people in prison can indicate one of two things: either criminal behavior is more prevalent, or law enforcement is stricter. If more crimes are being committed, then this could indicate that the state in question has a higher prevalence of reckless driving. However, if law enforcement is stricter, this could indicate that police officers are more likely to give speeding tickets, which can deter people from driving unreasonably fast and potentially reducing car crashes. Next, people who live in college dorms are generally less likely to travel since most of their required resources are in walking distance. However, it is important to note that college students are in an age range (about 18-22 for 4-year colleges) prone to poor judgment decisions in driving, so there are arguments for both positive and negative correlations between the proportion of people living in college dorms and highway fatality rate.

Finally, the proportion of people who live in military quarters can also reasonably go either way in terms of correlation to highway fatality rate. On one hand, people who live in military quarters are generally less free to travel, as, like college students, most of their resources do not require private transportation. Furthermore, states with more people in military quarters may also have more retired veterans, who may be disabled and may require public transportation, which is hypothesized to be good for reducing highway fatalities. On the other hand, a study conducted by University of Wyoming psychology professor Joshua D. Clapp shows that male veterans with PTSD are more likely to exhibit erratic driving behavior: "Veterans with a diagnosis of PTSD reported performance deficits, exaggerated safety/caution, and hostile/aggressive driving behavior at a greater frequency than student motorists sampled for the development of the [Driving Behavior Survey]," (Clapp et al. 2019). Even so, there are drawbacks to this study, as, for example, it only looks at male veterans with PTSD in contrast to the veteran population as a whole. Therefore, it is fair to hypothesize that states with a higher proportion of people in military quarters are likely to have fewer fatalities per capita, if all other variables are kept constant.

4. Results

In any professional statistics project, exploratory data analysis (EDA) is one of the most important steps. EDA is used to determine what variables may be important to use in a potential

statistical model, but it is important to note that not every variable deemed important by EDA will eventually end up in the final model. This is because EDA mainly looks at correlations between variables, which can be misleading since an explanatory variable with a strong correlation to the response variable may not add much to the final model, usually since it explains variation that has already been explained by other variables. So, while EDA is a vital step in developing a statistical model, it does have some drawbacks that may cause contradictions between EDA and the final result. The plan for this project is to generate 4 total models, one for each of these response variables: city-level hours lost due to traffic per person per year, city-level change in traffic due to COVID, city-level last-mile speed, and state-level highway fatality rate.

First, I transformed some of the variables into proportions to account for population, since looking at raw numbers would most likely show the same result as simply looking at population. One of the variables in particular measures the crude prevalence of not having access to health insurance, which is more confusing to interpret than simply having access to health insurance, so I converted this rate statistic into its proportional inverse to create a more understandable variable. Next, I decided to construct a correlation matrix, which shows the correlations between every numerical variable in the data set. Finding the variables with the strongest correlations (highest absolute values in the table) with each of the traffic metrics can give us a good idea of what variables may be important to use in each of the final models. It is also important to have some sort of explanation behind each of these strong correlations rather than simply chalking it up to pure variation in data. Here is a small section of the correlation matrix:

Variable	Hours Lost Due to Traffic Congestion	Population	Public Transport Ridership per Capita	Population Change from 2020 to 2021
Hours Lost	1	.634	.716	-.555
Population	.634	1	.452	-.330
Public Transport Ridership/Capita	.716	.452	1	-.614
Population Change	-.555	-.330	-.614	1
Population Density	.746	.641	.858	-.623

The explanatory variables with the strongest correlations with hours lost due to traffic (starting with the strongest by correlation coefficient) are population density (.746), public transport ridership per capita (.716), population (.634), and population change (positive number means the city grew in population and vice versa) from 2020 to 2021 (-.555). Density and population

make the most sense as to why they are correlated with traffic congestion, since denser/more populated areas naturally tend to have more traffic congestion. However, the negative correlation between population change and hours lost seems to suggest that either rapidly-growing cities have less traffic congestion (which makes no intuitive sense), or, the more likely conclusion, people are moving away from cities with high traffic congestion. The most important correlation to discuss, however, is public transport ridership's positive correlation with hours lost. This is a prime example of correlation not meaning causation, as it would be foolish to conclude that public transportation actually hurts traffic simply by looking at a correlation matrix. The reason this correlation exists is because denser cities, which naturally have high traffic congestion, are more likely to seek investment in public transportation as a means to fix the problem. Therefore, public transportation is not a cause, but a result of high traffic congestion.

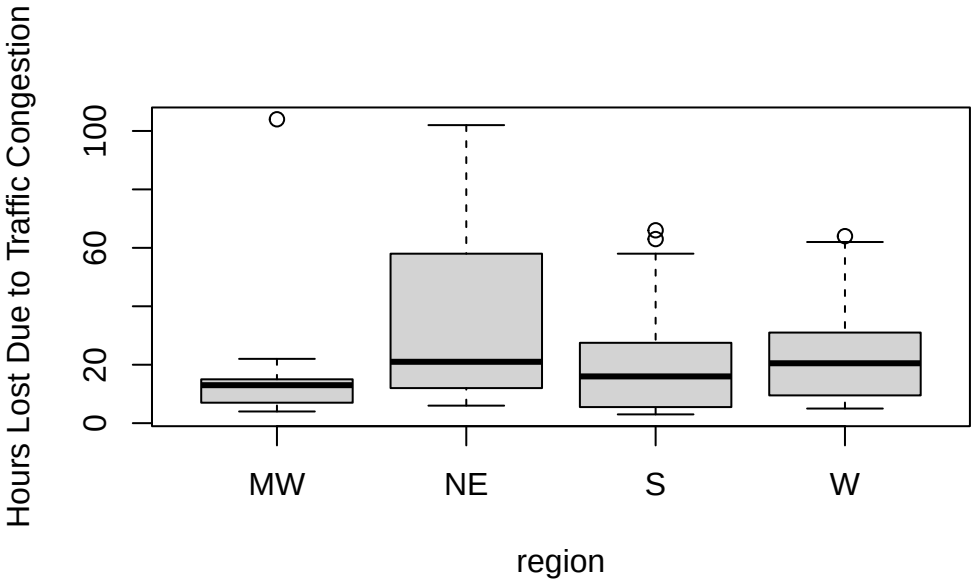
The explanatory variables with the strongest correlations (in terms of correlation coefficient) with change in traffic due to COVID are crude prevalence of various health factors (all negative), crude prevalence of having health insurance (-.420), and people per household (.451), but none of these correlations are particularly strong, as none of them go past 0.5 in absolute value. The health factors in question are cholesterol screening (-.468), pap smear use (-.422), and getting colonoscopy-related tests (-.409, specific to ages 50-75). The negative correlations between the health factors and change in traffic due to COVID seems to suggest that cities with stricter lockdown policies (and thus a stronger decrease in traffic during COVID) also have a higher proportion of people with preexisting conditions, thus increasing the motive for enforcing the lockdown policies in the first place. As for average household size, it would make sense to assume that cities with larger households tend to be more family-oriented. Observational studies have shown that "one-person households…are more prevalent in central cities than suburbia and…married-couple households…tend to reside in suburbia instead of central cities," (Jung and Yang 2016). The result of more traditional families living in the suburbs is a greater reliance on transportation, which cannot completely be halted by the impact of COVID due to the fact that using transportation is a necessity for these families. This is why larger families' rate of transportation has not fallen off due to COVID, explaining (to the best of my ability) why cities with more people per household have not seen as severe of a decrease in traffic due to COVID. Since none of the correlations are particularly strong, it is more reasonable to suggest that this may be a result of random variation in data, as explaining some of the correlations yields theories that require significant thought to process. This might suggest that, of the four total models, the one that uses change in traffic due to COVID as the response variable might be the weakest in proving any sort of point.

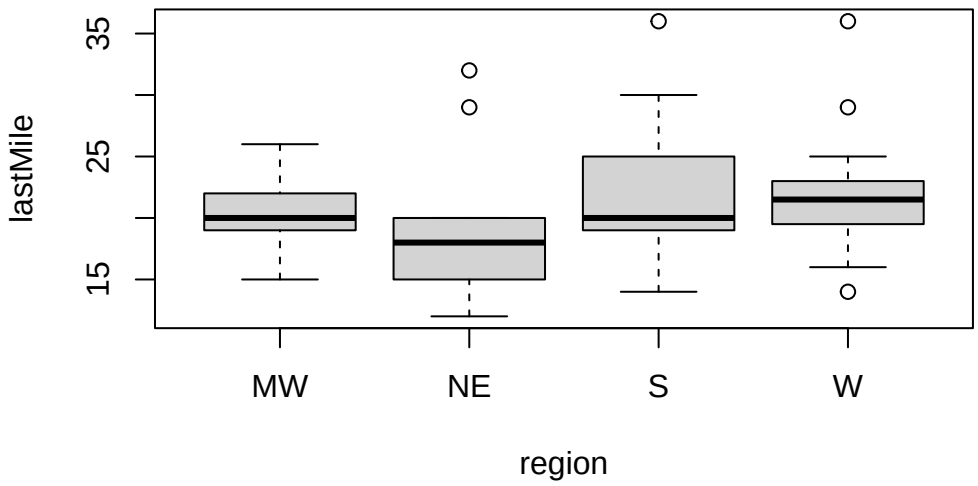
The explanatory variables with the strongest correlations with last-mile speed are population density and public transport ridership per capita (both negative); no other variable comes close. Intuitively, it makes sense that densely-populated cities have slower last-mile speeds, so the population density correlation makes the most sense. As for public transport ridership per capita, the same reasoning from the hours lost correlation will be used in the last-mile correlation: public transportation is not a cause, but a result of high traffic congestion. While this response variable has only two reasonably important explanatory variables, both of these

variables are important for what this project will uncover; they also appear in the hours lost model, suggesting that they have some sort of serious importance in this project.

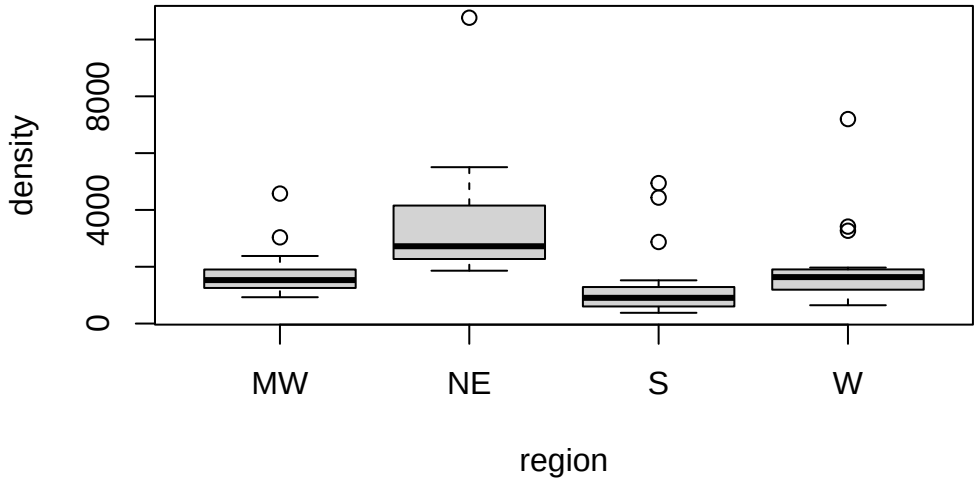
The next step of EDA is to compare explanatory and response variables using tables, boxplots, scatterplots, or other means. This is different from simply looking at a correlation matrix for two reasons: first, looking at graphs can help one detect potentially non-linear relationships (suggesting that the variable needs to be transformed), and second, graphs can be split into groups of points determined by a categorical variable, potentially showing different correlations for different groups. The "groups" in this case will be regions of the US, split into Northeast, Midwest, South, and West. If there is a significant difference in slope between two numerical variables in a graph between two (or more) regions, this indicates that an interaction term in the final model may be useful, if the non-interaction component is also significant enough to appear in the model.

First, I looked at the boxplots between region and each response variable, just to get an idea of which regions have good or bad traffic congestion metrics. According to hours lost and last-mile speed, Northeastern cities have the worst traffic, and there is no clear difference in change in traffic due to COVID between any two regions.

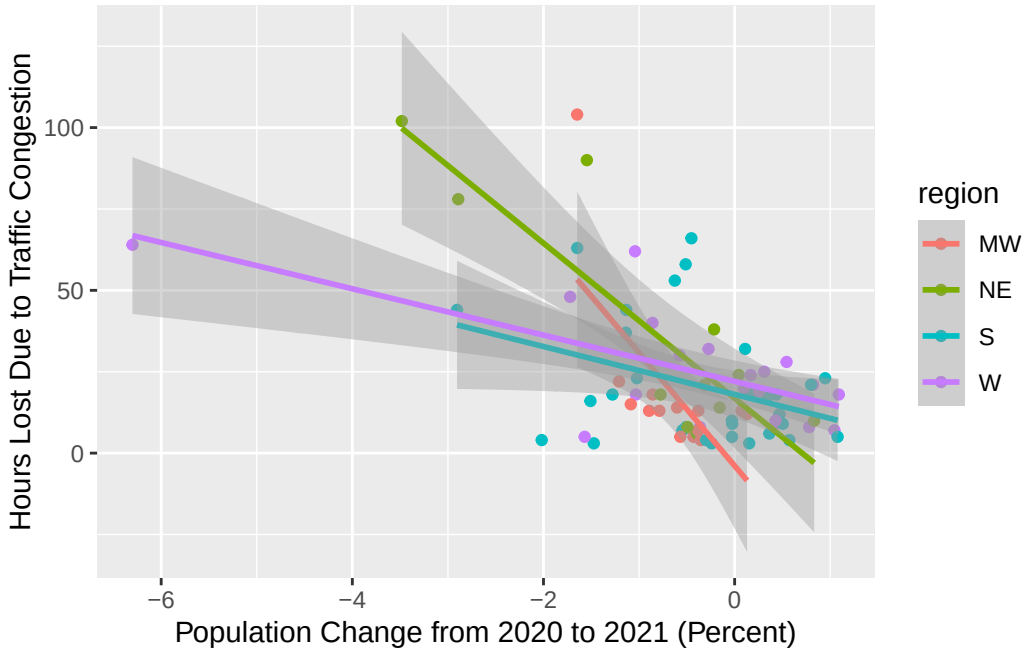


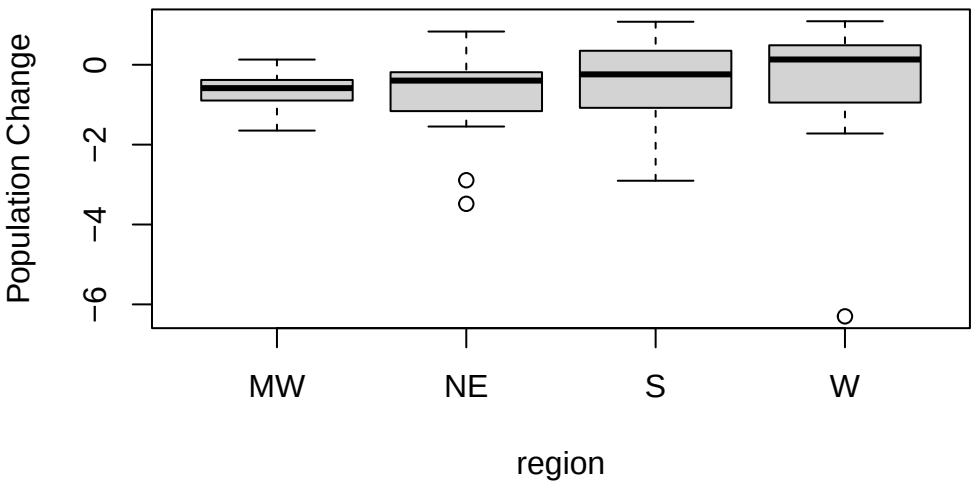


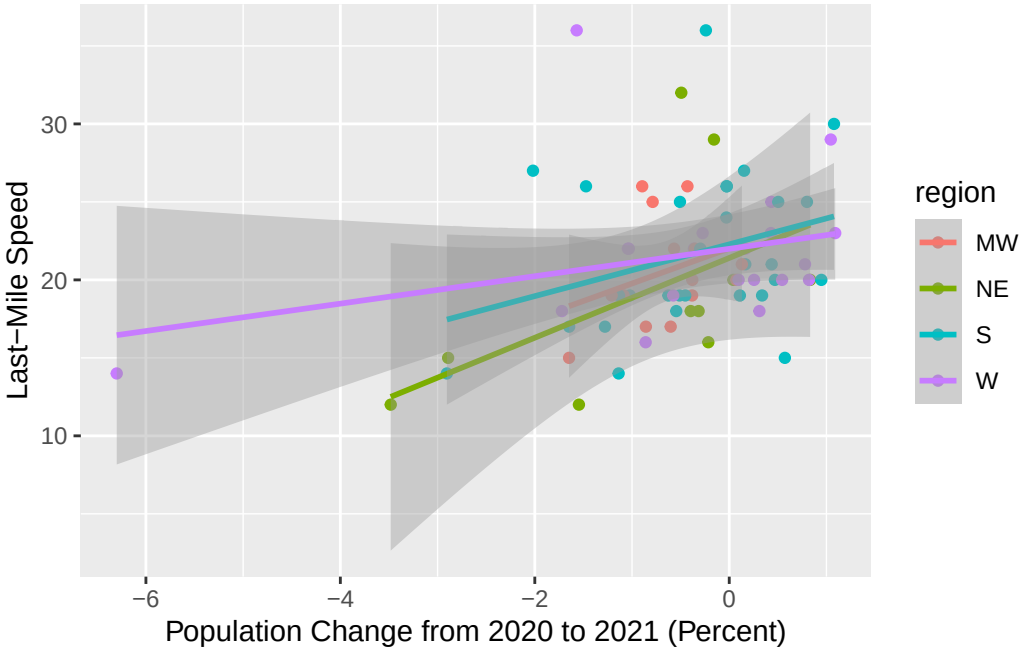
Based on what the correlation matrix tells us, Northeastern cities having poor traffic congestion metrics may indicate that cities in this region are densely-populated (since density appears to come up regularly).



As expected, Northeastern cities do indeed have the highest population density, but the density graph brings up another interesting issue: while Southern cities clearly have lower population densities than other regions, their traffic congestion metrics are still not as good as would be expected from low-city-density regions. Clearly, this is something that needs to be investigated further, but for now, I will continue to look at other variables.

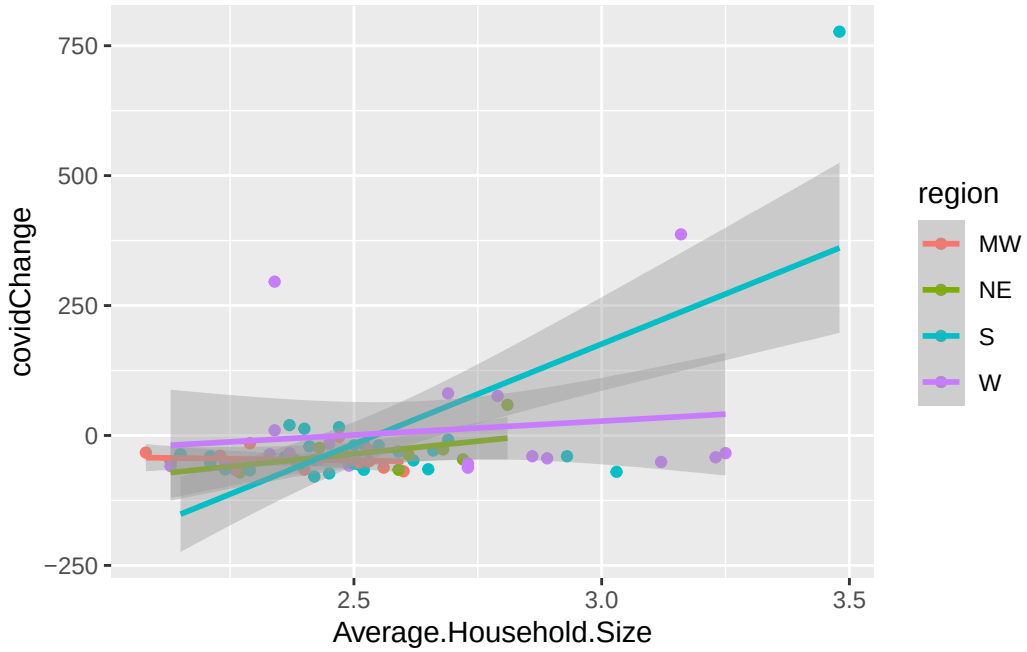


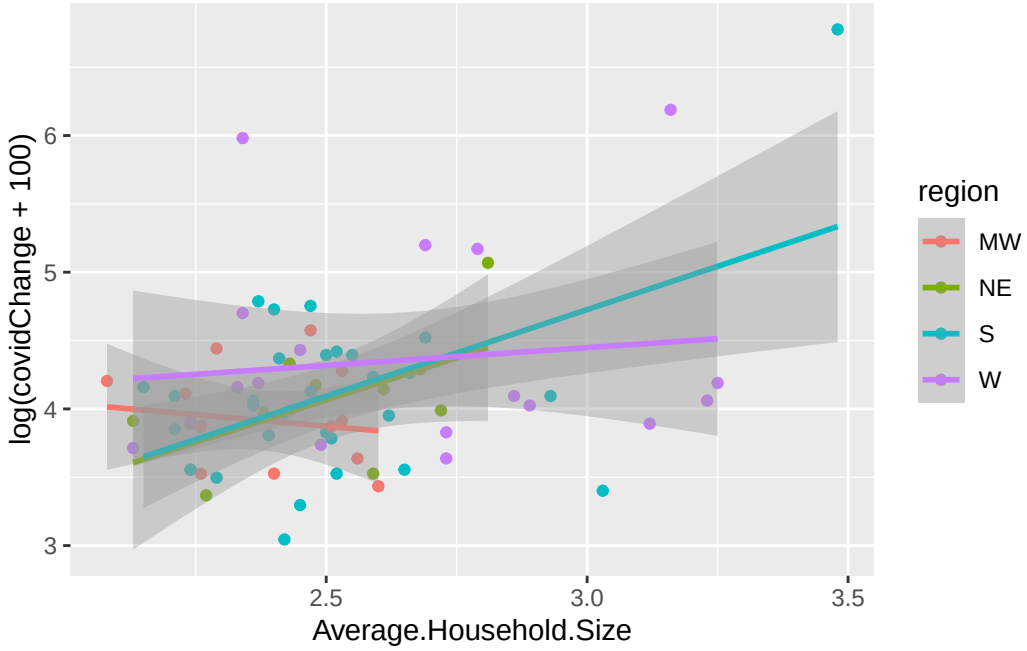




Continuing with the variables with high raw correlation coefficients, the graphs between population change and hours lost due to traffic confirm that, even when dividing by region, population change is universally negatively correlated with hours lost. However, no region in particular has a noticeable difference in population change, which makes things more interest-

ing when considering that Midwestern and Northeastern cities have noticeably steeper slopes in the aforementioned graph.





Average household size has one of the strongest correlations with change in traffic due to

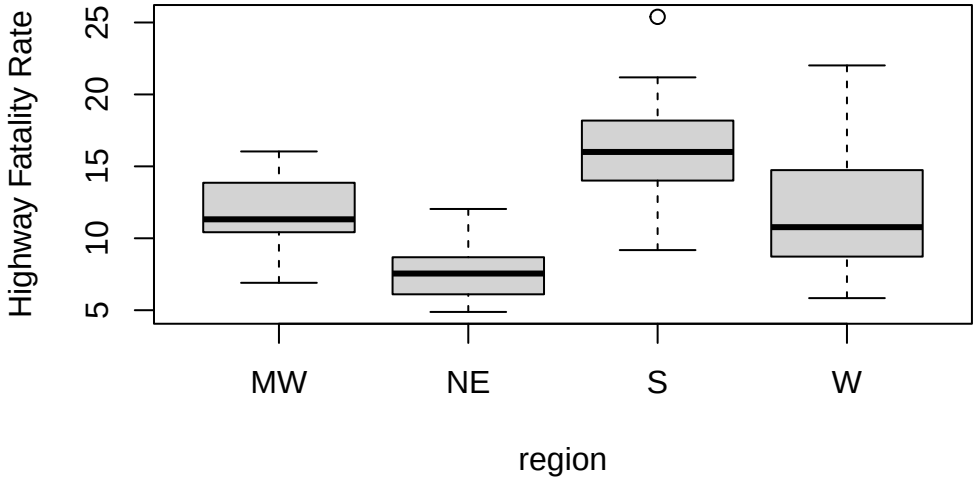
COVID, so looking at a region-split graph between the two variables might be important. However, the graph shows that the correlation is very likely skewed by a few outliers, so a log-transformation is required to make the graph more interpretable. Since change in traffic due to COVID is in percent, it can be negative, with a theoretical minimum of -100, so I added 100 to the variable before using a log-transformation since it is impossible to take the log of a negative number (or 0, but no relatively large city can realistically have a change of -100%). The resulting graph with the log transformation is more legible; the correlations are easier to see and therefore explain. While Northeastern and Southern cities have positive correlations between household size and change in traffic due to COVID, this correlation is negative for Midwestern cities; this might suggest that an interaction term may be present in the model. Northeastern and Southern cities, in terms of density and public transportation, are generally designed in polar opposite fashions, so it is definitely interesting to see an area where both of these regions experience the same effect. However, it is entirely possible that this could simply be more random variation, since none of the correlations by region are particularly strong. Furthermore, the positive correlation for Southern states only exists due to one point that skews the data (Brownsville, TX), so making any kind of conclusion would be a relatively futile effort. This further reinforces the weaknesses of the potential change in traffic congestion due to COVID model, but it is still a good idea to try the model anyway to determine if it gives any valuable information.

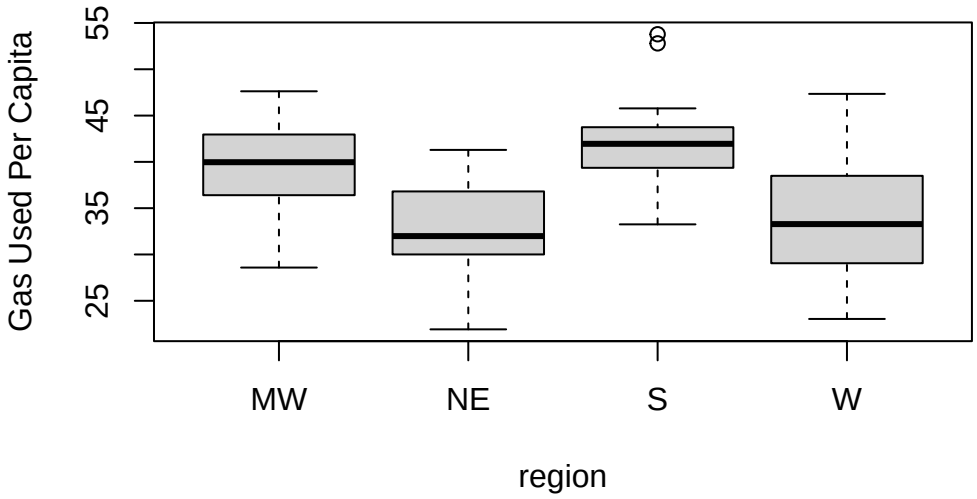
Now that the EDA for city data is complete, I will explain my EDA for state data. The metric I chose to use for states is highway fatality rate (highway fatalities per 100000 people). This is more fair than city-based metrics like hours lost and last-mile speed since it puts every state, regardless of how many or few major cities they have, on an even playing field. After converting most of my raw variables into per capita statistics, as well as adding income growth (change in median income from 2020 to 2021) and population growth (change in population from 2020 to 2021), I created another correlation matrix, this one using the variables from the state data set. Here is a small section of the correlation matrix for the state data:

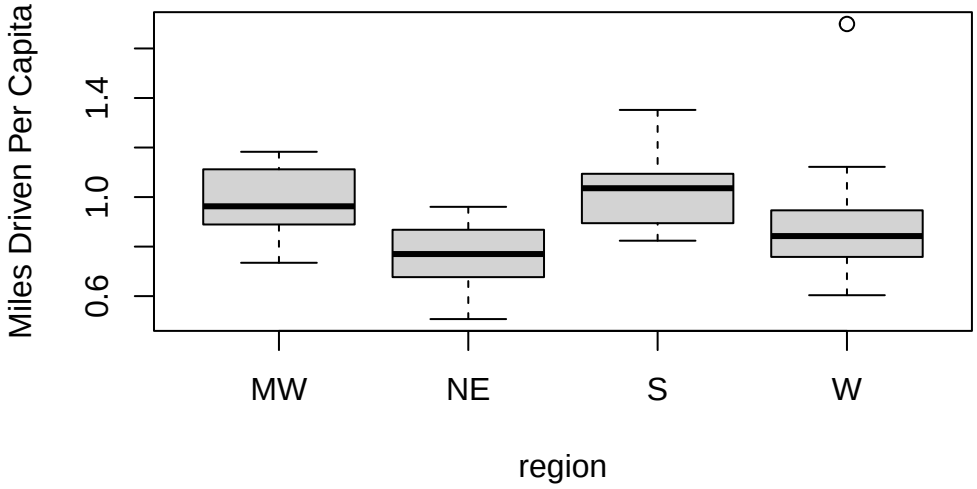
Variable	Fatality Rate	Highway Miles Driven Per Capita	Gas Usage Per Capita	Median Income
Fatality Rate	1	.838	.789	-.760
Miles Per Capita	.838	1	.867	-.534
Gas Usage Per Capita	.789	.867	1	-.555
Median Income	-.760	-.534	-.555	1
Prison Population Proportion	.711	.446	.476	-.720

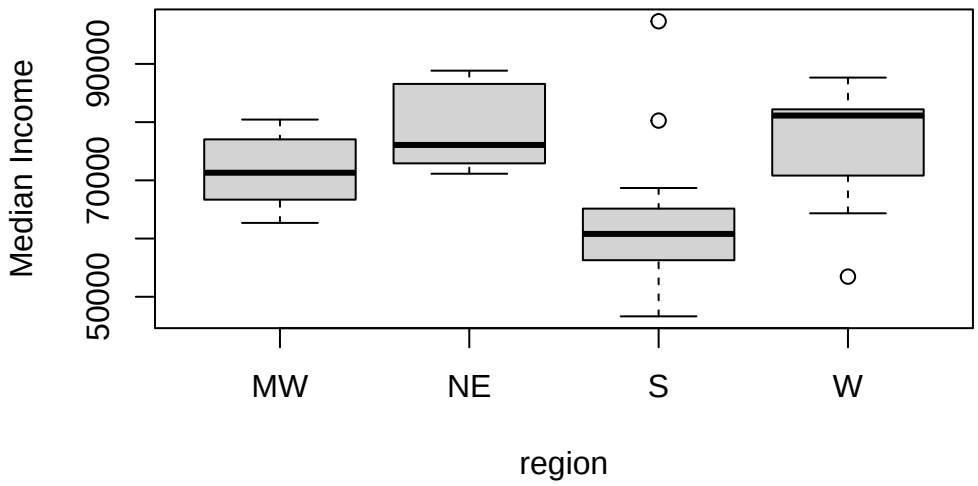
Variable	Fatality Rate	Highway Miles Driven Per Capita	Gas Usage Per Capita	Median Income
Public Transit Units Per Capita	-.546	-.612	-.693	.341

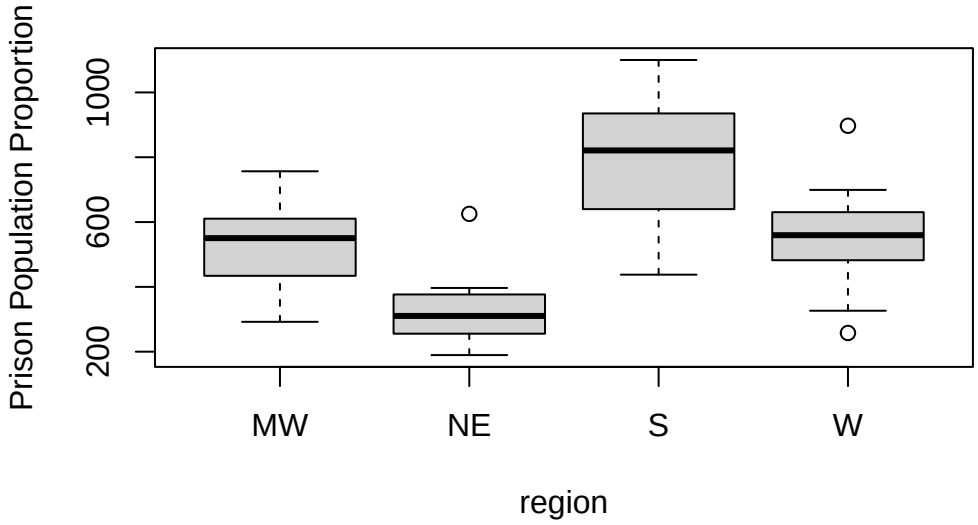
The variables with the strongest correlations with highway fatality rate are highway vehicle miles per 100 people (.838), gas usage (in gallons) per 100 people (.789), median income (-.760), proportion of people in adult correctional facilities (.711), and public transit units per person (-.546).




Gas Used Per Capita
25
35
45
55
MW
NE
S
W
region


Miles Driven Per Capita
0.6
1.0
1.4
MW
NE
S
W
region

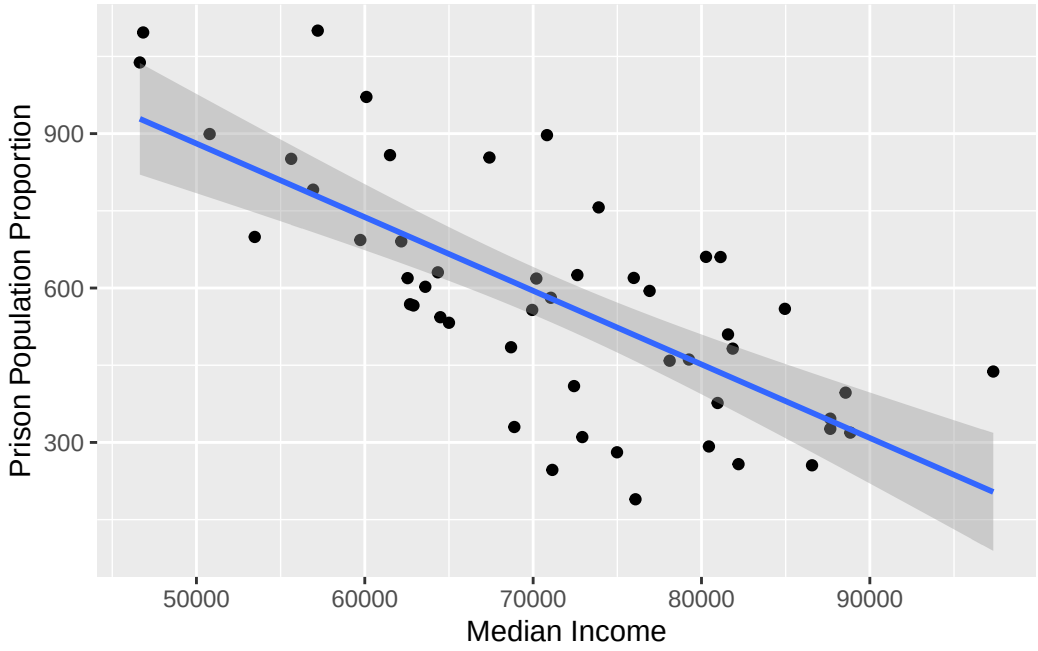




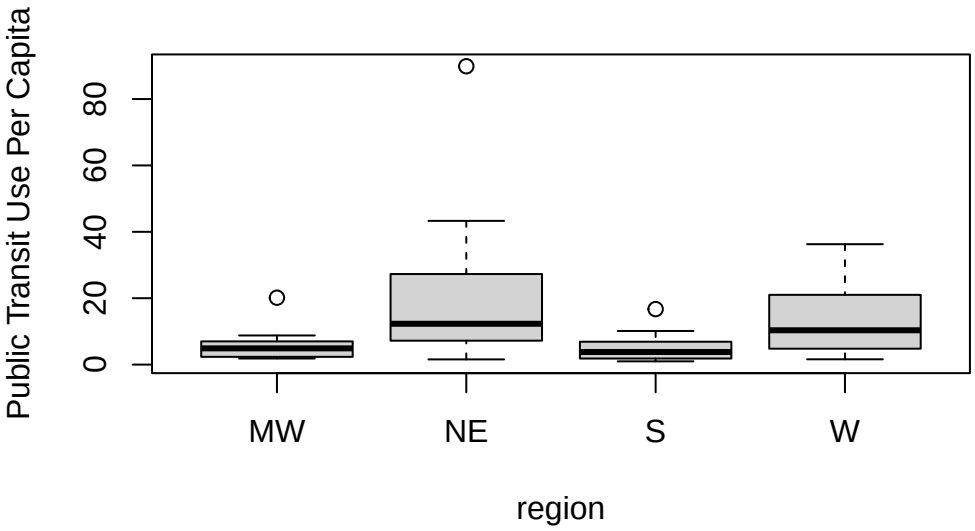
First, looking at highway fatality rate by region paints a concerning picture: Southern states clearly have the highest highway fatality rates of all regions. This alone should raise some concerns as to how Southern roads or cities are constructed, especially since Southern drivers generally drive more miles and use more gas than drivers in every other region except in

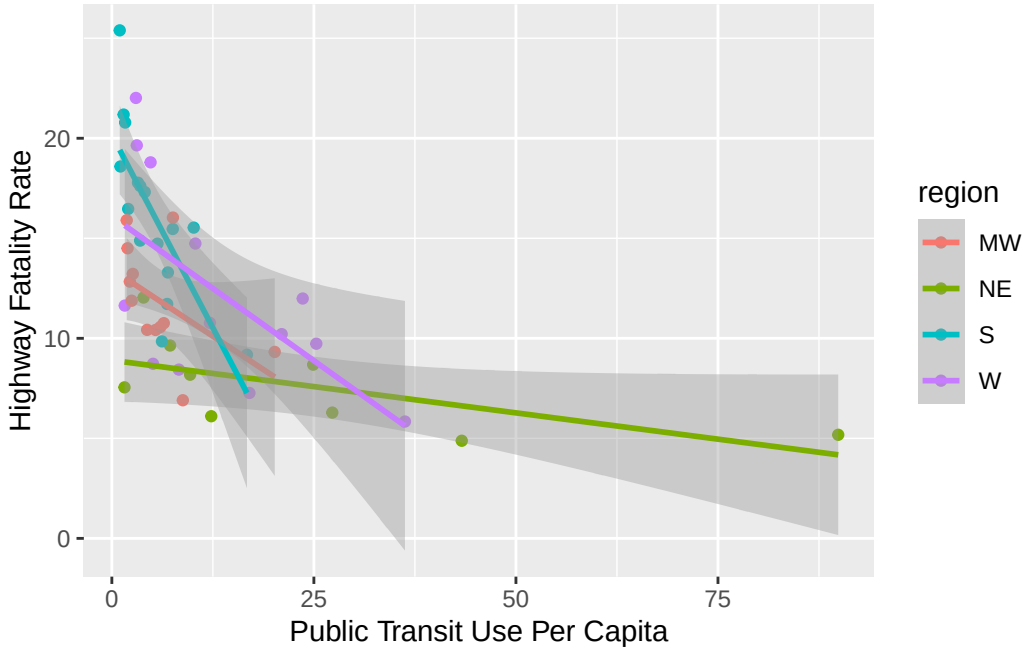
the Midwest. This indicates that Southerners are incentivized, if not required, to drive longer distances to get to their destination, which, conveniently enough, corresponds to the graph from the city data set that shows that Southern cities are the least densely populated. Furthermore, Southern states generally have a lower median income than in other regions, which is important because median income is strongly negatively correlated with highway fatality rate. Adding still to the list of problems with Southern states, they also have the highest proportion of people in adult correctional facilities, and, as stated earlier, this is a variable that is strongly positively correlated with highway fatality rate. This raises another important question: Are these problems all related to each other? If so, what does it mean for the US, or, more specifically, the South?

Further analysis shows that median income is negatively correlated with proportion of people in prison (-.720), which shows that two of the aforementioned problems may be related to each other. This correlation is backed up by intensive research on the US incarceration crisis; "poverty and excessive legal punishments contribute significantly to the United States' high rate of imprisonment, which has disproportionately affected low-income and minority populations," ("Incarceration and Poverty in the United States" n.d.). Since this research also notes that minority populations are being incarcerated at a disproportionately high rate, it might make sense to add race-based variables to the analysis. However, the correlation matrix shows that the proportion of white people (by state) has little to no correlation with median income (-.0383), prison population proportion (-.126), or highway fatality rate (.0853). Obviously, this does not mean that race has absolutely nothing to do with any of these problems; measuring these variables on an individual county, city, or even suburb level, rather than state-level, may yield more informative results. Even so, we have surmised that median income is heavily related to proportion of people in prison, not just via our data, but also through prior advanced research.



This scatterplot reinforces the idea that median income is heavily negatively correlated with prison population proportion (correlation coeff. -.720), so it is entirely possible that only one of these variables will be required. If I had to choose between either variable, I would use median income instead of prison population proportion, because a) using incarceration statistics that have little correlation to race-based data contradicts prior research that proves otherwise, and b) studies show that median income has a reasonable relationship with population density, which in turn is related to highway fatality rate. A study from 1999 from the Journal of Energy and Development found that there is "a significant positive correlation between median household income and population density for 1980…and 1990," (Vandegrift and DiCaro 1999), and that "for U.S. states, high incomes are associated with high population densities and people migrate toward states with high incomes," (Vandegrift and DiCaro 1999). Another more recent study from Slippery Rock University shows that this can be due to more employment opportunities in densely populated areas: the relationship between density and employment is positive indicating that an increase in population or housing density increases employment in a metropolitan area," (Hummel 2020). Therefore, prison population proportion, median income, population density, and fatality rate all have significant and reasonable connections to each other, according to our state-based data and the studies of prior researchers.





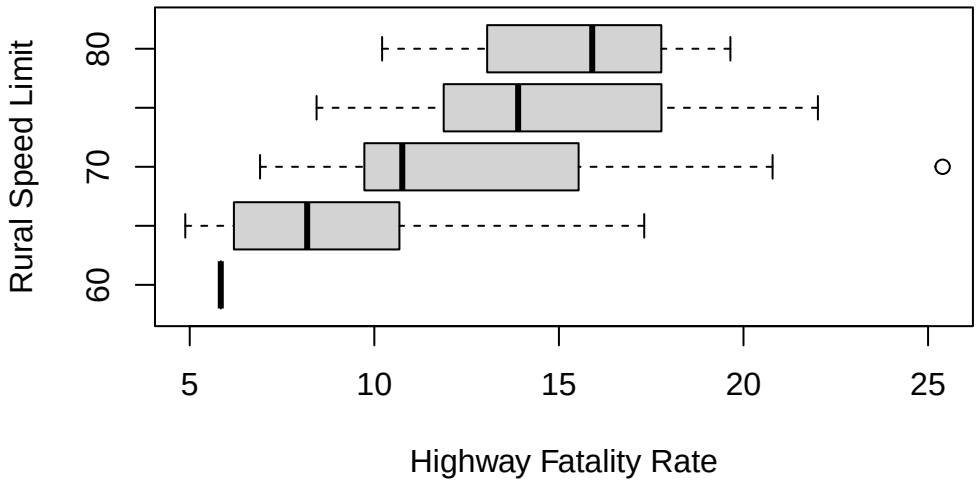
The state-level data set also has a variable that measures frequency of public transport ridership per capita, which has been shown (in the city data set) to be valuable in explaining variance in traffic congestion. The basic boxplots show that Southern and Midwestern states have lower frequencies of public transport ridership per capita, while Northeastern and West-

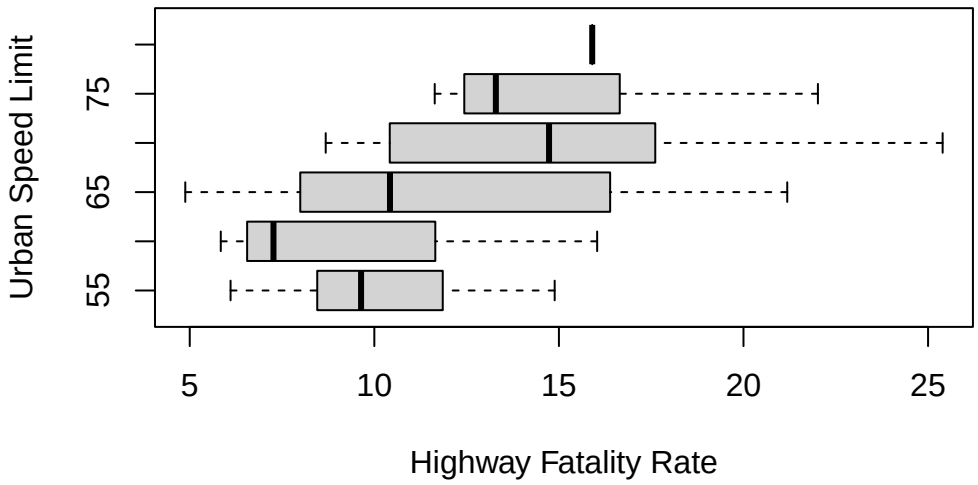
ern states have higher frequencies; this is something that has already been shown. As expected, there is a negative correlation between public transport ridership per capita and highway fatality rate, which intuitively makes sense because states (or metropolitan areas) that use more public transit are less likely to use highways and more likely to focus their transportation efforts on cities, as cities are more densely populated than highways and can benefit more from public transit. (The slope is less steep for Northeastern states because there is a relatively small difference in highway fatality rate between states in the Northeast, limiting how steep the slope can be.)

One of my biggest ideas going into this project was that speed limits are one of the most important factors determining highway fatality rates. This stems from conventional wisdom that "speed kills;" driving faster reduces one's margin of error, meaning that one slip-up when driving at a high speed can have more devastating results than at lower speeds. I decided to test this theory by introducing urban (and to a lesser extent rural) highway speed limit to my EDA. Since people are going longer distances on rural highways, rural highways end up having more total miles driven, which can indicate that rural highway speed limit would be more important than urban speed limit. However, the reason I focused more on urban speed limits is because urban highways are more densely-packed, providing a potential "breeding ground" for highway fatalities in comparison to the less dense rural highways. Since the speed limits (courtesy of IIHS) are in increments of 5, they can essentially be treated as categorical variables for now, so I used a table instead of a boxplot to compare speed limit to region.

	ruralLimit				
	60	65	70	75	80
region					
MW			8	3	1
NE		7	1	1	
S		2	10	4	
W	1	2	2	6	2
#Total cases	1	11	21	14	3

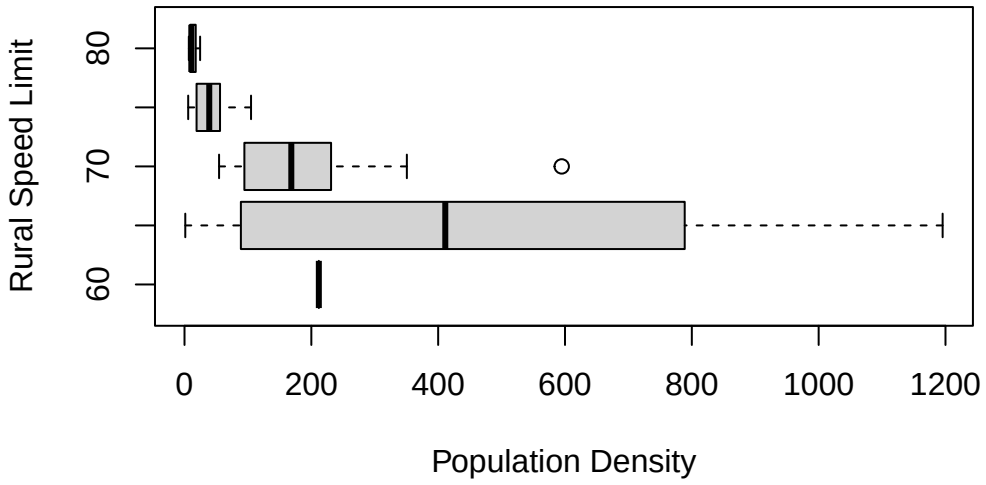
	urbanLimit					
	55	60	65	70	75	80
region						
MW	3	1	2	3	2	1
NE	4		3	1	1	
S	2		4	9	1	
W	2	2	6		3	
#Total cases	11	3	15	13	7	1

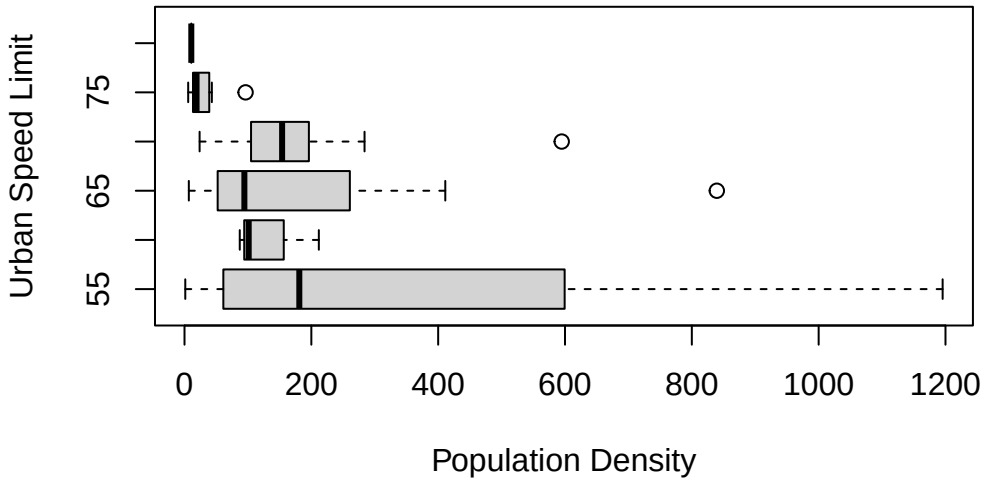




The tables show that Northeastern states have the lowest urban speed limits of each region, but they also show something interesting: over half of all Southern states (9 of 16) have an urban highway speed limit of 70 mph. Next, I looked at the boxplots for each speed limit compared to highway fatality rate, and I found that fatality rate appears to increase as speed

limit increases. This immediately reinforced my suspicions that speed limit had something to do with highway fatalities, and I started to come up with what I believed to be an easy solution: simply reduce the speed limit in areas with high fatality rates (more specifically, go from 70 to 65 mph in certain Southern states).





First, the fact that Northeastern states have the lowest speed limits makes it reasonable to suggest that high-population density areas generally have lower speed limits; this makes intuitive sense because people in less densely-populated areas can drive much faster due to the relative lack of cars on the highways. The boxplots back up this notion, as states with an urban highway speed limit of 55 mph (lowest in the data set) are by far the most densely populated. This is further backed up by the correlation matrix, which shows that the non-speed limit variable most strongly correlated with each speed limit variable is population density (-.530 for rural speed limit, -.375 for urban speed limit).

The next step in analysis is to create advanced statistical models that intend to show more informative results than in the EDA section. As discussed prior, there will be 4 models, 3 for the city-level data, and 1 for the state-level data. The first model uses hours lost due to traffic per person per year as the response variable, so only variables that can reasonably have some sort of relationship to hours lost will be used. Since the response variable is not Normally distributed, a Poisson model, which uses whole numbers for the response variable, will be the starting model. Also, since there are many possible variables that can end up becoming variables in the resulting model, a backwards regression strategy will be done. This strategy starts by creating a model with every variable, then gradually taking away the least significant variable until every variable is significant. This process is accelerated by using the stepAIC function, which removes variables based on how they affect the model's AIC (lower is better). AIC is calculated using log-likelihood, and rewards explaining more variation in the data while also penalizing excessive use of variables. Since stepAIC is an automated function, it will be used very often throughout the rest of this project. Using stepAIC on the full Poisson model yields a model with 10 different variables, but, according to the goodness-of-fit test, this model fails. The goodness-of-fit test uses chi-squared distributions to compare the values in the model to the values in the data; if the resulting p-value is less than 0.05, the test fails, meaning the model does not fit the data.

Most of the time, finding the optimal model takes many attempts, and the "best" model may end up being something far from expected. The first optimal model ended up being a negative binomial model, which generally serves the same function as the quasi-poisson model, a variation of the Poisson model that includes a dispersion parameter that accounts for high variances in the explanatory variables relative to the means. The only difference between the negative binomial and quasi-poisson models is that the variance of the negative binomial model is a quadratic function of the mean, while the variance of the quasi-poisson model is a linear function of the mean; this difference can cause variables to be weighted differently, which can affect how well each model fits the data. Backwards regression was used for negative binomial, and, surprisingly, the full model passed the goodness-of-fit test, which is a good sign that the correct model type had finally been used. After running stepAIC, we get a model with 8 different variables that also fits the data. This indicates that the optimal model has been found, so the next step in analysis is to interpret each coefficient.

```
Call:
glm.nb(formula = hrsLost2020 ~ Est2021 + Average.Household.Size +
    BINGE_CrudePrev + popChange + Stdev + region + forpct + riderrate,
    data = model1keys, init.theta = 3.838934112, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6137  -0.7923  -0.1778   0.4807   2.1086

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             2.542e+00  1.043e+00   2.438  0.01476 *
Est2021                 2.774e-07  6.817e-08   4.069 4.72e-05 ***
Average.Household.Size -6.451e-01  3.378e-01  -1.910  0.05615 .
BINGE_CrudePrev         1.338e-01  3.546e-02   3.775  0.00016 ***
popChange              -1.167e-01  7.410e-02  -1.575  0.11532
Stdev                  -3.199e-05  1.498e-05  -2.136  0.03267 *
regionNE                5.542e-01  2.693e-01   2.058  0.03960 *
regionS                 6.928e-01  2.350e-01   2.948  0.00320 **
regionW                 6.784e-01  2.586e-01   2.624  0.00870 **
forpct                  1.966e-02  9.552e-03   2.058  0.03961 *
riderrate               4.239e-03  2.135e-03   1.985  0.04714 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.8389) family taken to be 1)

    Null deviance: 193.381  on 75  degrees of freedom
Residual deviance:  76.422  on 65  degrees of freedom
AIC: 582.89

Number of Fisher Scoring iterations: 1

              Theta:  3.839
          Std. Err.:  0.733

 2 x log-likelihood:  -558.892
```

Negative binomial automatically log-transforms the response variable. This means that the coefficient actually measures the change in the log of the response as the explanatory variable in question increases by 1 unit. The 8 different variables in the model are population, people per household, crude prevalence of binge drinking, population change, wealth disparity, region

(as a factor), percentage of foreign-born people, and public transport ridership per capita. I expected the model to contain an interaction term, but it did not significantly strengthen the model, so I decided to leave it out. Population change may be insignificant in terms of p-value, but removing this variable from the model causes multiple other variables to become insignificant, which makes the model weaker.

Population has a coefficient of .0000002774, which is still positive, but very small; this makes sense when considering the relative lack of impact one person has on a city-wide phenomenon. It also makes sense that the coefficient is positive because, as shown in the EDA, larger cities in terms of population tend to have more city-wide traffic congestion. This number means that, with all other variables kept constant, a city with exactly one more person will have an hours lost due to traffic value higher than another identical city by a factor of about 1.000000277, a factor so minuscule that it is not even worth explaining. When dealing with such small (in magnitude) regression coefficients, it would make more sense to determine the effect of greater population differences to produce ratios that are more meaningful. In contrast, 1 "unit" for smaller-scale variables like people per household (and proportion statistics that go from 0 to 1) would be far more drastic and less realistic, which means there needs to be an optimal change in the explanatory variable that is based on the range of observations. Selecting a specific population difference requires that it fits the data relatively well; my estimate for a "good" number (in general terms) would be half of the interquartile range of the desired variable. The reason I chose this metric is because it represents the typical difference between two random observations, which is a 25-percentile difference. This avoids the 1 unit problem seen in large variables like population and small variables like people per household, while also removing any need to make arbitrary estimates for a "fair" change in the explanatory variable. I used this measure of spread instead of standard deviation since standard deviation can be skewed by outliers and can generate a number that serves as a less reasonable difference than IQR/2, which is not skewed by outliers in any way.

This method makes it possible to calculate a standardized ratio for each variable to determine the effect of changing each variable by IQR/2. Ratios above 1 indicate positive correlations, and ratios below 1 indicate negative correlations. Here is a table containing all of these results:

Table 5: Coefficient Table for Each Variable

Variable	Coeff	Ratio (x+1)	IQR/2	Ratio (x+IQR/2)
Population	$2.774*10^{-7}$	1.000000277	235624	1.067
People Per Household	$-6.451*10^{-1}$	.5246	.15625	.9041
Crude Prevalence of Binge Drinking	$1.338*10^{-1}$	1.143	1.514	1.225

Variable	Coeff	Ratio (x+1)	IQR/2	Ratio (x+IQR/2)
Population Change	$-1.167*10^{-1}$	.8899	.6074	.9316
Standard Deviation of Income (Wealth Disparity)	$-3.199*10^{-5}$	.9999680	3871	.8835
Percentage of Foreign-Born People	$1.966*10^{-2}$	1.0199	5.449	1.113
Public Transport Ridership Per Capita	$4.239*10^{-3}$	1.00425	20.84	1.092

The table shows (from left to right) each variable, the coefficients given by the model, the ratio based on change in 1 unit ($e^{coef.}$), half the IQR, and the adjusted ratio that takes the original ratio to the power of IQR/2. In terms of change in the explanatory variable by half the IQR, crude prevalence of binge drinking appears to have the greatest effect relative to its range of data; an increase of 1.514 percentage points causes hours lost due to traffic to increase by a factor of 1.225. Since the region variable (default: Midwest) is categorical, using the methods used in the other variables does not make much sense. The coefficient (or e raised to the coefficient) explains the ratio between hours lost due to traffic and each of the other regions, with all other variables kept constant. These ratios are:

$e^{.5542} = 1.741$ for Northeastern to Midwestern cities,

$e^{.6928} = 1.999$ for Southern to Midwestern cities, and

$e^{.5542} = 1.971$ for Western to Midwestern cities.

The fact that all of these ratios are considerably higher than the binge drinking IQR ratio suggests one of two things: a) either region truly is the most impactful factor in determining traffic congestion, or b) a change in region has an impact greater than a change in half the IQR for each variable. For reference, to get an increase by a factor of 1.999 (the largest change in region) from crude prevalence of binge drinking, it would take about 5.177 percentage points, which is almost 2 standard deviations; for the lowest change in region, it would take 4.144 percentage points, about 1.5 standard deviations. So, it can be concluded that the effect of changing region is somewhere between 1.5 to 2 standard deviations for the most impactful numerical variable, which suggests that a change in region can heavily impact hours lost due to traffic. This may be due to other factors such as how the roads were structured or how long the cities have been around; older cities are generally more densely-populated. Perhaps more

telling than the resulting IQR ratio for each of the variables is the sign of the coefficient, as it is much easier to explain and can more easily make reference to conventional wisdom or prior research to back it up.

Starting with population change, the coefficient sign being negative indicates that cities with a higher growth rate (or lower rate of population decline) tend to have fewer hours lost due to traffic after accounting for all other variables. This might suggest that traffic congestion is one of the main reasons why people are moving out of certain areas; while high-population areas usually have more traffic congestion, areas that are growing in population tend to have less traffic congestion. This might suggest that the cities experiencing the most growth are designed in such a way that ideally limits traffic congestion; this is backed up by the fact that density has a relatively strong negative correlation with population growth.

Next, wealth disparity has a negative coefficient, indicating that cities with more wealth disparity tend to have less traffic congestion, after all other variables are accounted for. One theory as to why this is the case is the fact that cities with higher wealth disparity tend to have higher proportions of people living in poverty, which indicates a high unemployment rate; less employment means that fewer people need to (or can afford to) drive, thus reducing traffic. This is backed up by an MIT study in 2009 that outlines the positive correlation between wealth inequality and unemployment using complex mathematical systems that are beyond the scope of this project (Cysne 2009).

The next variable is percentage of foreign-born people, which has a positive coefficient. It is conventional wisdom that foreign-born people generally immigrate to the US to find job opportunities, so they will most likely move to cities (or neighborhoods) with an extensive presence of businesses. Cities with more businesses will likely have more job activity, which means that more people will require some sort of transportation, leading to more cars on the road and thus greater traffic congestion.

However, the most interesting finding from this model is that public transport ridership per capita has a positive coefficient, meaning that cities in which people are more likely to use public transportation tend to have more hours lost due to traffic per person per year. While this is consistent with the scatterplot from the EDA, unlike the scatterplot, the model takes every other variable (notably, including population) into account when calculating the coefficient. The model simply adds more weight to the notion that cities with more developed public transportation systems have worse traffic congestion. However, this does not mean that public transportation causes traffic congestion; it is more reasonable to assume the inverse: areas that naturally have high traffic congestion are more likely to invest in public transportation as a solution. Furthermore, city planners believing in the positive impact of public transportation is also consistent with the scatterplots graphs shown at the end of the city-level EDA, which proved that growth in public transportation is negatively correlated with growth in traffic congestion. This is an example of how it is important to understand the realistic implications of the model results instead of blindly following the numbers without any reasonable context.

The second model focuses on change in traffic congestion due to COVID, according to INRIX, as the response variable. This model is expected to include more health-related variables than the first model, since prevalence of pre-existing conditions can determine how susceptible the population is to COVID, thus relating to change in traffic congestion due to COVID.

```
Call:
glm.nb(formula = covidChange ~ CHOLSCREEN_CrudePrev + PAPTEST_CrudePrev +
    COLON_SCREEN_CrudePrev + Average.Household.Size, data = model2keys,
    init.theta = 3.63666243, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1355  -0.9068  -0.2540   0.4049   3.2738

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             12.19745    2.35954   5.169 2.35e-07 ***
CHOLSCREEN_CrudePrev    -0.08431    0.03336  -2.527  0.01150 *
PAPTEST_CrudePrev       -0.08680    0.03426  -2.534  0.01129 *
COLON_SCREEN_CrudePrev   0.04741    0.02403   1.973  0.04853 *
Average.Household.Size   0.85772    0.27605   3.107  0.00189 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.6367) family taken to be 1)

    Null deviance: 154.278  on 75  degrees of freedom
Residual deviance:  78.205  on 71  degrees of freedom
AIC: 770.59

Number of Fisher Scoring iterations: 1

              Theta:  3.637
          Std. Err.:  0.590

 2 x log-likelihood:  -758.589
```

After removing all insignificant variables, the final model is a negative binomial model that includes crude prevalence of cholesterol screening, crude prevalence of pap smear use, crude prevalence of colonoscopy-related tests (ages 50-75), and people per household. According to the goodness-of-fit test, this model fits the data, with a p-value of .261; this again suggests

that the negative binomial model is clearly the best type of model for the data in question. Again, adding an interaction term, contrary to what the EDA shows, does not significantly improve the model. Just like in the first set of models, I will create a table showing the true value behind each coefficient relative to the data using the same IQR/2 strategy.

Variable	Coeff	Ratio (x+1)	IQR/2	Ratio (x+IQR/2)
Crude Prev. Cholesterol Screening	-.08431	.9191	1.662852	.8692
Crude Prev. Pap Smear Use	-.0868	.9169	1.93287	.8455
Crude Prev. Colonoscopy-Related Tests	.04741	1.049	3.43715	1.177
People Per Household	.85772	2.358	.15625	1.143

According to the coefficients shown in the model summary, it would appear that people per household has the greatest effect on change in traffic congestion due to COVID. However, this coefficient is based on an increase of one person per household, which is over 6 times larger than the IQR of this variable. Adjusting for IQR/2 yields much different results; in fact, by this metric, crude prevalence of colonoscopy-related tests has the greatest effect on change in traffic congestion due to COVID.

Crude prevalence of cholesterol screening and crude prevalence of pap smear use both have negative coefficients, which suggests they have negative correlations with change in traffic congestion due to COVID after accounting for all other variables. Both of these crude prevalence variables indicate the proportion of people who show up to their preventative health appointments, which may indicate access to health insurance. According to the correlation matrix, crude prevalence of access to health insurance is indeed negatively correlated to change in traffic congestion due to COVID, with a coefficient of -.420. However, the coefficient for colonoscopy-related tests is positive, suggesting that cities with a higher rate of colonoscopy tests within the ages of 50-75 had their traffic congestion decrease less or even increase during COVID lockdown. This implies that colonoscopies, which also indicate prevalence of access to health insurance, have opposing effects on change in traffic congestion due to COVID when compared to pap smear tests and cholesterol screening, which appears to have no reasonable explanation. There is a real possibility that the different coefficient signs in each of the variables could simply be a result of random variation, as the EDA discusses how a model using change in traffic congestion due to COVID may be relatively weak due to weak correlations in the scatterplots. Therefore, it is reasonable to suggest that the second model is the least useful model so far.

```
Call:
glm(formula = lastMile ~ density, family = "poisson", data = model3keys)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8880  -0.5982  -0.2444   0.5037   2.4289

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.205e+00  4.053e-02  79.067  < 2e-16 ***
density     -8.560e-05  1.844e-05  -4.642 3.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 78.471  on 72  degrees of freedom
Residual deviance: 54.069  on 71  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 413.67

Number of Fisher Scoring iterations: 4
```

The third model, which uses last-mile speed as the response variable, is by far the easiest to explain, with only one variable, being population density. Unlike the first two models, simply using Poisson regression is enough for the model to pass the goodness-of-fit test, meaning that a dispersion parameter is not required. Even so, the coefficient is interpreted in the same way as in the negative binomial model, which was explained in the first model analysis. Here is the IQR/2 table for Model 3:

Variable	Coeff	Ratio (x+1)	IQR/2	Ratio (x+IQR/2)
Population Density	$-8.560*10^{-5}$	.9999144	504.625	.9577

According to the resulting ratios, population density has a moderate effect on last-mile speed, although the effect is not as great as, for example, binge drinking on hours lost due to traffic. Even so, the negative correlation between population density and last-mile speed makes sense because denser cities have more congested city centers, compared to less dense cities, in which traffic congestion is more spread out throughout the entirety of the city. All in all, this model confirms what the EDA showed, which was the dominance of population density over all other potential explanatory variables when it comes to explaining variation in last-mile speed.

The final model intends to use highway fatality rate as the response variable, but since highway fatality rate uses decimal numbers, the amount of options for model types would be limited. So, the model will use total fatalities as the response variable and include total population as an offset term. Offset terms are mainly used for situations in which there exists a response variable that is known to be related to the response variable prior to any analysis; in general, total population (or log(total population)) is one of the most common offset terms.

```
Call:
glm.nb(formula = Fatalities ~ as.factor(region) + income + adjGas +
    adjMiles + prisonpct + primepct + incomeGrowth + offset(log(Total.population)),
    data = model4keys, init.theta = 95.76795514, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1422  -0.7043   0.0371   0.6412   2.2445

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -7.311e+00  8.541e-01  -8.560  < 2e-16 ***
as.factor(region)NE -1.697e-01  6.044e-02  -2.807 0.005001 **
as.factor(region)S   7.744e-02  5.141e-02   1.506 0.131942
as.factor(region)W   1.225e-01  5.290e-02   2.316 0.020553 *
income              -5.113e-06  2.606e-06  -1.962 0.049732 *
adjGas               1.715e-02  6.036e-03   2.842 0.004483 **
adjMiles             4.521e-01  2.011e-01   2.249 0.024525 *
prisonpct            4.524e-04  1.336e-04   3.387 0.000707 ***
primepct            -3.260e+00  1.105e+00  -2.950 0.003183 **
incomeGrowth         1.045e-02  3.872e-03   2.698 0.006976 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(95.768) family taken to be 1)

    Null deviance: 556.534  on 49  degrees of freedom
Residual deviance:  48.225  on 40  degrees of freedom
AIC: 563.77

Number of Fisher Scoring iterations: 1

              Theta:  95.8
          Std. Err.:  23.3
```

```
2 x log-likelihood:  -541.768
```

Using backwards regression on the negative binomial model yields a model with 7 variables that fits the data relatively well. The variables included are region, median income, gas used in gallons per person per year, miles driven per person per year, percentage of people in prison, percentage (values 0 to 1) of people in prime driving age (25-64), and change in median income from 2020 to 2021.

Once again, my IQR/2 strategy will be used to determine the effect of each variable relative to the data. Here is the table containing the results (with the exception of region, which is non-numeric):

Variable	Coeff	Ratio (x+1)	IQR/2	Ratio (x+IQR/2)
Median Income	$-5.113*10^{-6}$	.99999489	8829.125	.9559
Gas Used Per Person	$1.715*10^{-2}$	1.017	4.31136	1.077
Miles Driven Per Person	$4.521*10^{-1}$	1.572	.1172371	1.054
Prison Percent	$4.524*10^{-4}$	1.000453	146.3768	1.068
Prime Driving Age Percent	-3.260	.03839	.00896	.9712
Median Income Change	$1.045*10^{-2}$	1.0105	2.203699	1.023

Relative to their respective data ranges, the variable with the greatest effect on highway fatality rate is gas usage per person. This is interesting because gas usage and miles driven per person explain pretty much the same things about highway fatality rate, which might suggest that only one of these variables is useful, yet both variables show up in the final model. The best explanation for this is that fuel-efficiency (driving more miles while using less gas) has been proven to lessen the risk of traffic fatalities; "fuel efficiency has a statistically significant and negative effect on traffic fatalities at the state level; given that the results of all three models are the same, this result would be considered robust," (Gius 2009). However, this relationship must be taken with a grain of salt, as "this result runs counter to the results of most of the prior research," (Gius 2009). The fact that prior research exists to counter this explanation may suggest that the ratios shown in the table may be slightly misleading; conventional wisdom would first assume that miles driven can affect highway fatality rates before considering gas usage. Even so, both gas usage and miles driven show positive correlations with highway fatality rate, which makes sense because driving more miles increases one's risk of getting into an accident.

The negative coefficient for median income makes sense because states that are generally poorer might not have the resources to invest in major beneficial changes in traffic. This negative coefficient also helps explain the positive coefficient for prison population proportion, as median income is heavily negatively correlated with prison population proportion. However, prison population proportion must be significant regardless of whether the income variable is present, since it does not get removed from the model via the stepAIC function. This may be because states with more people in prison tend to have higher crime rates, which includes things such as reckless driving or driving under the influence, both of which can be deadly for both the driver and any nearby cars. Prime driving age percent has a negative coefficient, which also makes intuitive sense because states with more people in their driving prime have fewer people who are more prone to making poor decisions on the road. Finally, median income change has a positive coefficient, which suggests that states with growing incomes generally have higher fatality rates. Income inequality is a constant problem in the US that has only worsened over time, so states that are seeing recent growth in income are also more likely to have their income inequality increase. Furthermore, studies from the Canadian Economics Association show that income inequality is also a determining factor for highway fatality rate (Anbarci, Escaleras, and Register 2009), so, using income inequality, it can be logically proven that the positive coefficient for median income change makes sense in this context.

The effect of regions, however, appears to be by far the most drastic. The ratios in highway fatality rate when going between regions, accounting for all other variables in the model, are:

$e^{-.1624} = .8501$ for Northeastern to Midwestern cities,

$e^{.0655} = 1.068$ for Southern to Midwestern cities, and

$e^{.1429} = 1.154$ for Western to Midwestern cities.

This means that there are certain ways in which roads or highway systems are constructed in each region that affects highway fatality rate. The most interesting thing about this is that accounting for all other variables shows that Western states have a higher ratio to Midwestern states than Southern states, contradicting the EDA that portrays Southern states as having the worst highway fatality metrics. This is because variables that are associated with Southern states scoring the lowest or highest, such as gas usage, miles driven, prison population proportion, median income, etc. are all already present in the model. This is a good sign because it shows that there exists a combination of variables in the model that can at least somewhat explain Southern states' highway fatality problem.

Another thing that should be noted about the final model is the absence of population density and public transport ridership as variables. This does not suggest that these variables are useless for explaining highway fatality rate, it simply means that other variables that explain the same things are more significant. For example, the reason population density is not included can be attributed to the presence of miles driven per person in the model; as shown in the EDA, less dense states require drivers to drive more miles (and use more gas). The presence of miles driven (and gas used) per person also makes the inclusion of public transport ridership per person unnecessary since both variables are heavily negatively correlated with

public transport ridership per person, omitting what would seem like a very important variable from the final model. Furthermore, the absence of speed limit from the model is interesting since it contradicts the analysis done in the EDA, as well as preconceived notions that higher speed limits are more dangerous. Further research echoes these results; a study conducted by Lahore University of Management Sciences in conjunction with Texas A & M University concluded that "factors such as safety regulations, drunk driving, driving in the wrong lane etc. play a dominant role in fatal crashes. Thus, increasing the speed limit does not play a fundamental role in fatal crashes in [the states studied]," (Malik and Aftab 2017).

5. Discussion/Conclusions

In this paper, I used statistical modeling to determine what factors help explain traffic congestion in the US. I expected the most significant factor to be public transportation, followed closely by city density. For highway fatality rate, I also expected public transportation use, as well as speed limit, to play a big role in explaining the data. However, the results from the advanced models showed that public transportation use was not the most significant factor in explaining traffic congestion or highway fatality rate, and that speed limit has little to no effect on highway fatality rate. This suggests that two of the most instinctual "solutions" to traffic problems (investing in public transportation and lowering speed limits) are not cut-and-dry; there exist many other factors, most of them out of our control, that are potentially more significant and must be considered. However, this is not to say that either of these proposed solutions are entirely useless. For example, we found that cities that increase their public transportation usage per capita over a specific time period are likely to see their level of traffic congestion decrease over the same period. Even so, this is not sufficient proof that public transportation is good for reducing traffic congestion, since it is a simple scatterplot and not an advanced model in which all context is accounted for. After all, at the end of the day, this project is an observational study, so it is far more difficult to make conclusions of this caliber than if the project were an experimental study.

Despite the lack of statistical evidence to prove my prior hypotheses, the project still provides valuable information. The project shows that variables such as people per household, crude prevalence of binge drinking, and wealth disparity can have a sizable impact on traffic congestion, while variables such as miles driven, gas usage, and percentage of people in prison help explain highway fatality rate. More specifically, the combination of every variable in the final model is enough to at least somewhat explain why Southern states have higher fatality rates than every other region. Things like this can open the door for further research (statistical or not) on this topic; for example, historians can use the findings from this study and potentially add historical context to show a more theoretical explanation of discrepancies in traffic across the US. Overall, the results of this study show that there is not one, but many factors that can explain traffic congestion and highway fatality rate, most of which are beyond our control. This just goes to show that there is no easy solution to traffic problems. In short, despite the

real benefits of simply investing in more public transportation or lowering the speed limit, attempting to improve traffic goes much further than that, as it is important for traffic planners to account for all other variables before making a significant decision.

6. References

Anbarci, Nejat, Monica Escaleras, and Charles A. Register. 2009. "Traffic Fatalities: Does Income Inequality Create an Externality?" *The Canadian Journal of Economics / Revue Canadienne d'Economique* 42 (1): 244–66. https://www.jstor.org/stable/25478348.

Bureau, US Census. n.d. "City and Town Population Totals: 2020-2021." *Census.gov*. Accessed September 13, 2022. https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html.

Clapp, Joshua D., Denise M. Sloan, William Unger, Daniel J. Lee, Janie J. Jun, Scott D. Litwack, and J. Gayle Beck. 2019. "Problematic Driving in Former Service Members: An Evaluation of the Driving Behavior Survey in Veterans with Posttraumatic Stress Disorder." *J Anxiety Disord* 67 (October): 102134. https://doi.org/10.1016/j.janxdis.2019.102134.

Cysne, Rubens Penha. 2009. "On the Positive Correlation Between Income Inequality and Unemployment." *The Review of Economics and Statistics* 91 (1): 218–26. https://www.jstor.org/stable/25651330.

"Excerpt: Many Cities Have Transit. How Many Have Good Transit? | Kinder Institute for Urban Research." n.d. *Kinder Institute for Urban Research | Rice University*. Accessed January 23, 2023. https://kinder.rice.edu/urbanedge/excerpt-many-cities-have-transit-how-many-have-good-transit.

Gius, Mark. 2009. "Fuel Efficiency and the Determinants of Traffic Fatalities: A Comparison of Empirical Models." *New York Economic Review* 40 (January): 13–27.

Hummel, Daniel. 2020. "The Effects of Population and Housing Density in Urban Areas on Income in the United States." *Local Economy* 35 (1): 27–47. https://doi.org/10.1177/0269094220903265.

"'I Am Sick and Tired of This Congestion'_ Perceptions of Sanandaj Inhabitants on the Family Mental Health Impacts of Urban Traffic Jam | Elsevier Enhanced Reader." n.d. Accessed February 13, 2023. https://doi.org/10.1016/j.jth.2019.100587.

"Incarceration and Poverty in the United States." n.d. *AAF*. Accessed February 15, 2023. https://www.americanactionforum.org/research/incarceration-and-poverty-in-the-united-states/.

Jung, Gowoon, and Tse-Chuan Yang. 2016. "Household Structure and Suburbia Residence in U.S. Metropolitan Areas: Evidence from the American Housing Survey." *Soc Sci (Basel)* 5 (4): 74. https://doi.org/10.3390/socsci5040074.

Lyon, Craig, Dan Mayhew, Marie-Axelle Granié, Robyn Robertson, Ward Vanlaar, Heather Woods-Fry, Chloé Thevenet, Gerald Furian, and Aggelos Soteropoulos. 2020. "Age and Road Safety Performance: Focusing on Elderly and Young Drivers." *IATSS Research* 44 (3): 212–19. https://doi.org/10.1016/j.iatssr.2020.08.005.

Malik, Kashif Zaheer, and Ammar Aftab. 2017. "Speed Limit and Fatalities in the U.S.:

Implication for Transportation Policy." *Theoretical Economics Letters* 7 (5): 1398–1412. https://doi.org/10.4236/tel.2017.75094.

Malpezzi, Stephen. n.d. "Population Density: Some Facts and Some Predictions."

"Total Alcohol Consumption Per Capita by U.S. State 2020." n.d. *Statista*. Accessed January 11, 2023. https://www.statista.com/statistics/442848/per-capita-alcohol-consumption-of-all-beverages-in-the-us-by-state/.

Transportation. Bureau of Transportation Statistics, United States. Department of. 2019a. "National Transportation Statistics (NTS)." https://doi.org/10.21949/1503663.

———. 2019b. "State Transportation Statistics (STS)." https://doi.org/10.21949/1503664.

Vandegrift, Donald, and Vincent DiCaro. 1999. "Income, Population Density, and Residential Energy Use: An Analysis of U.s. States." *The Journal of Energy and Development* 25 (1): 37–45. https://www.jstor.org/stable/24808723.

Williams, A F. 2006. "Young Driver Risk Factors: Successful and Unsuccessful Approaches for Dealing with Them and an Agenda for the Future." *Inj Prev* 12 (Suppl 1): i4–8. https://doi.org/10.1136/ip.2006.011783.