California Traffic and Demographic Trends

Link to Website:

https://sites.google.com/sdsu.edu/miguelsamigos/home

Link to Video:
https://sdsu.instructuremedia.com/embed/edd81f82-0f1d-4cf7-bd02-642415192601

BDA-594: Big Data Science and Analytics Platforms

San Diego State University

Authors: Eric Braga, Miguel Ángel Bravo Martínez del Valle, Xinyu Du, and Riley Rutan

Abstract

The state of California has been facing escalating traffic problems, with implications for economic vitality, environmental sustainability, and quality of life. This project seeks to elucidate the complex interplay between traffic volume, demographic shifts, and economic changes in California. By leveraging advanced data analytics and machine learning, we aim to understand historical trends and find economic and demographic factors that are associated with traffic volume. This endeavor holds significant importance for urban planning, infrastructure development, and policy-making, addressing the pressing need for informed decision-making in the face of rapid socio-economic transformations.

Introduction

With the ever-growing traffic problem in California, it becomes even more necessary to understand what affects future demands since it could be a reflection of underlying socio-economic patterns and burgeoning populations. We will rigorously analyze and identify correlations, exploring how shifts in economic dynamics and population growth across California's counties interplay with traffic volume. We aim to illuminate the nuanced factors influencing California's future transportation demons through comprehensive data analysis & visualization techniques. We will find the correlation and visualize the change in traffic in California based on economic and population growth. Through in-depth data analysis and insightful trend visualization, we seek to decode the intricate relationships between three pivotal variables: economic shifts, population dynamics, and traffic volume. Our objective is not just to find correlations but also to dive deep into the individual trajectories of each dataset.
In this project, we aim to understand the relationship between population dynamics, economic shifts, and traffic volume in California, trying to solve interesting questions focused on these three attributes:

- Is traffic getting worse? As California residents, we suspect that it has in the last 10 years. We plan to explore highway traffic volume data to see if this is true.
- Are trends in traffic associated with changes in California's population demographics? We plan to investigate the data of each county in California to find out.
- What role do economic factors have in California's traffic issues? We will analyze data from the US Census to dive deeper into this relationship

Literature Review

Some fundamental theories we took into consideration in our analysis and projects are highlighted here:

1. Theory of Urban Agglomeration:

The article "Transportation Issues in Developing China's urban agglomerations" (Huang et al. 1) talks about the theory of urban agglomeration. This theory suggests the increase in urbanization and population density found in China will lead to increased traffic due to higher demand for transportation services and infrastructure. As the population grows and economic activities intensify, traffic tends to increase.

2. Income Effect on Travel Behavior

In “Understanding Socioeconomic Disparities in Travel Behavior during the COVID-19 Pandemic”(Brough et al. 1) the author refers to how changes in income levels can impact travel behavior. This paper looks specifically at Covid 19 where traffic declined but not as much for low-income households. This can be because of fewer travel substitution options and working flexibility like remote work. Socioeconomic factors have the potential to affect traffic.

3. Land Use-Transportation Interaction

The site The Geography of Transport Systems states that “population growth is a vector for additional transportation demand, but rising incomes are as well”.Then goes on to say the relationship between land use patterns and transportation systems plays a significant role. The "Land Use-Transportation Feedback Cycle" theory suggests that variations in population density, income levels, and land-use patterns influence transportation choices and traffic flow.

Once these three attributes have been studied, we wanted to analyze the data to find trends in each dataset and utilize machine learning to find the factors that drive traffic the most.

Methods

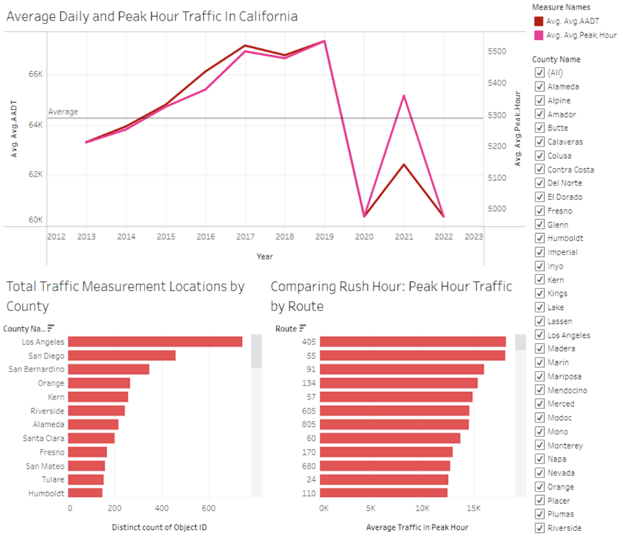
Traffic Data

We utilized data that was collected and distributed by the California Department of Transportation’s Traffic Census Program. Their website includes data on traffic counts and is broken up into four categories, one of which is “Traffic Volumes: Annual Average Daily Traffic (AADT)” which measures all vehicles on California State Highways and is the data we utilized for this project. The link to this data can be found here: https://dot.ca.gov/programs/traffic-operations/census

Historical data between 2013-2021 was collected from this source for later processing and analysis. While this data includes descriptions of the location in which the data was recorded, we wanted each data point to be associated with a latitude and longitude for more precise mapping and analysis. We located a dataset of records for the year 2022 in the form of a shapefile, with each measurement including a precise geographic location and unique location object ID. The link to that dataset can be found here:
https://gis.data.ca.gov/datasets/d8833219913c44358f2a9a71bda57f76/explore

The data from year to year had many inconsistencies, missing values, and errors. These had to be rectified, removed, and cleaned. Each row of data from the years 2013-2021 had to be matched with a location from the 2022 dataset that included latitude and longitude information. While our data cleaning script was able to fill the lat and long variables for most of the data, several entries did not find a matching input in the 2022 dataset. Location object IDs that did not have a record for every year between 2013 and 2022 were removed to ensure consistency and to avoid oversampling from a certain region. After cleaning, over 65,000 rows of traffic data remained. Variables that were not of interest were removed from the dataset, and front and back traffic values were aggregated into a single average AADT, peak hour AADT, and peak month AADT set of values. The data cleaning and preparation for analysis was performed using the Pandas package in Python, the dplyr package in R, and Excel.

For analysis, the data was imported into Tableau and several interactive dashboards were created. The average daily AADT and Peak Hour AADT were plotted over the years, as seen below. The total measurement locations by county and peak hour by route were also plotted below. Several other dashboards that visualize the traffic data can be found on our website for analysis. These dashboards show traffic by county or route over the years, as well as percentage change by county and region.



Population Data

The databases, information, and other data have been taken from the American Community Survey 1-Year from the official website of The United States Census Bureau (https://data.census.gov/).
The American Community Survey is conducted every year through sampling and provides estimates on a variety of demographic variables. Not every county in California is included in the 1-year survey due to their low populations.

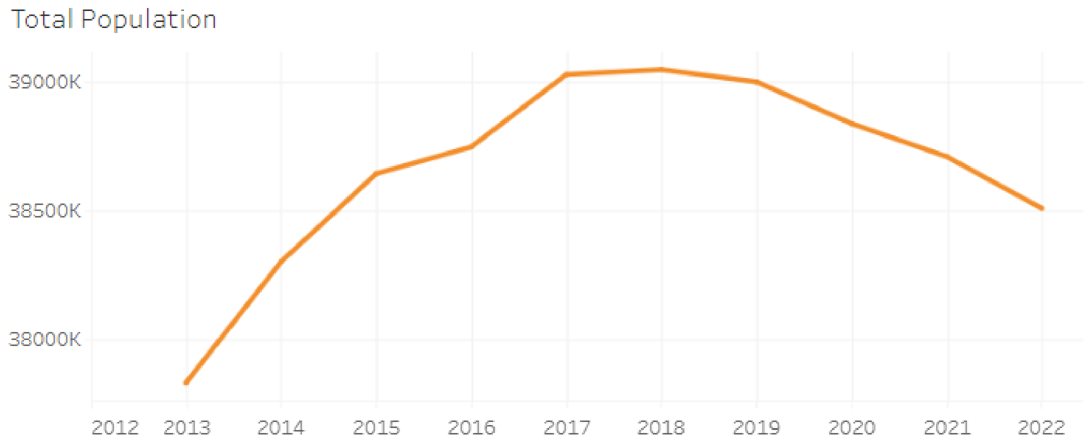
On this page, we were able to obtain data related to the population of the United States, from the national level to the county level. We found a database that gathered information about every county in the state of California, updated annually. There were several attributes that we did not need and were not necessary, so we focused on the following variables:

Under 5	5 To 9 Years	10 To 14 Years	15 To 19 Years
20 To 24 Years	25 to 29 Years	30 To 34 Years	35 To 39 Years
40 To 44 Years	45 To 49 Years	50 To 54 Years	55 to 59 Years
60 To 64 Years	65 To 69 Years	70 To 74 Years	75 to 79 Years
Total Population			

As you can see, we have many attributes and the groups are not really interesting, so we decided to regroup the ages to form the following age groups:

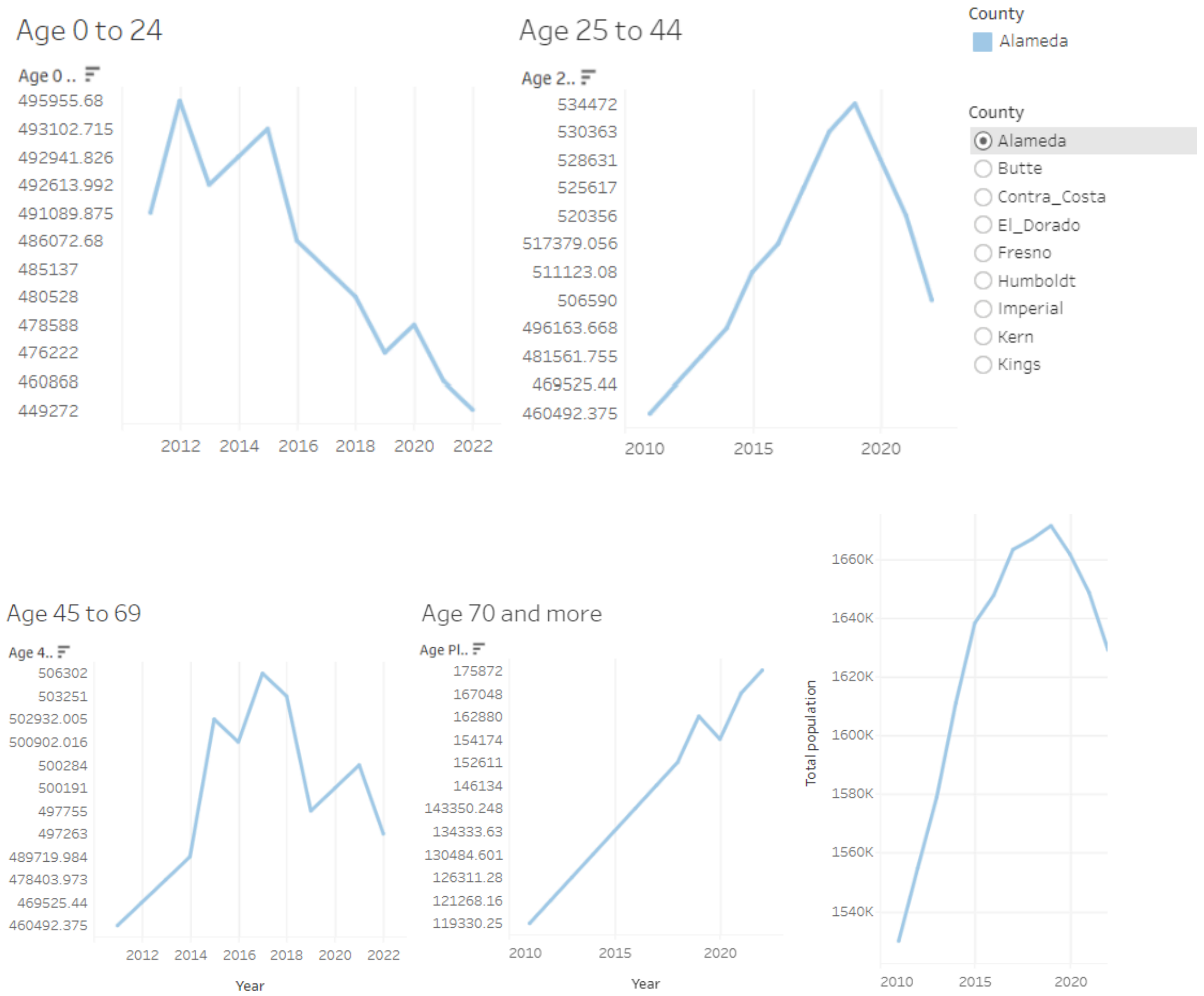
Age From 0 to 24	Age From 25 to 44	Age From 44 to 69	Age 70 and Over
Total Population			

Once the data is cleaned, we end up with a single dataset with all the information for each county. That is to say, the final dataset consists of the variables mentioned above, apart from the variables 'County' and 'Year'.



Until 2018, California's population demonstrated a consistent increase. However, from that point onwards, a gradual decline in demographic growth was observed, leading to a downward trend in the subsequent years until 2022. This shift marked a turning point in the state's population dynamics, signaling a change in the direction of demographic growth that had been steady until 2018.
This decline in population in California can be attributed to several factors as outlined in Hans Johnson's article called “What's Behind California's Recent Population Decline– and Why It Matters”.

- Over 100,000 deaths due to COVID-19
- Declining birthrate near record lows
- People moving out of state for a lower cost of living, often while working remotely
- Sharp drop in immigration due to the pandemic and federal policies

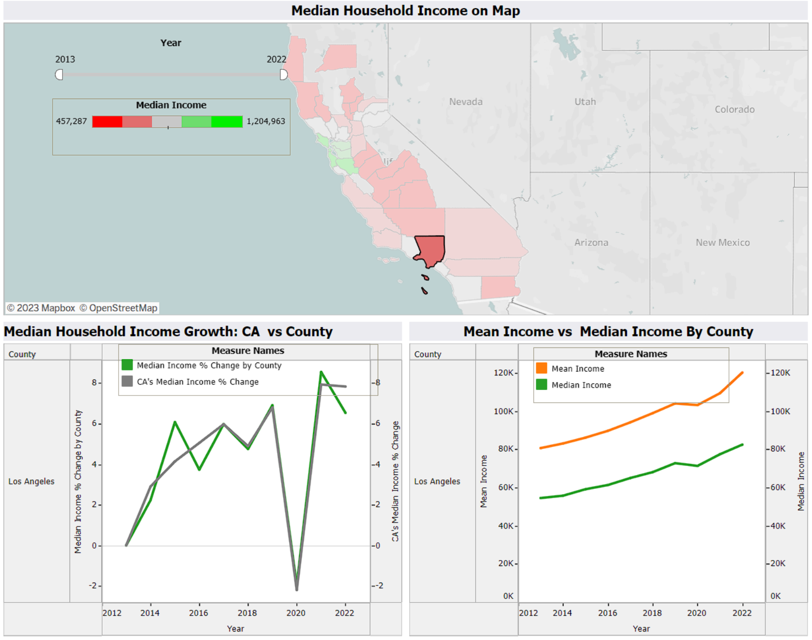


Income Data

We utilized data collected and distributed by the United States Census Bureau. The data spanned from 2013 to 2022, and the income data comprised various categories, including Household, Family, and Coupled-family. We specifically focused on household income data due to its greater inclusivity, enabling us to capture a broader spectrum of economic situations. The link to access this data is provided here: Income data link

We downloaded individual data files for each year from 2013 to 2022 and subsequently employed R Studio to consolidate the household data for each year into a unified CSV file. The dataset featured distinct ranges of household income, median income, and mean income. For a more reliable indicator of central tendency, particularly in the context of skewed distributions and to provide an overall economic overview, we predominantly relied on median income for our analyses. Additionally, we introduced a new column to calculate the percentage increase in median household income by county. To visualize the data, we used Tableau to create comprehensive visual representations.

The data visualization revealed distinct trends in household income data. Analysis of Median Household Income data revealed that areas with relatively high median household incomes are predominantly around the Bay Area, near San Francisco. A dramatic decrease in average median household income was observed in 2020, likely due to economic downturns triggered by coronavirus-related restrictions. Over the past decade, the gap between Mean and Median household income in each county widened, and it has the trends to keep diverging which points to an escalating issue of income inequality.



Machine Learning Models

With modeling our goal is to be able to derive the average AADT(average daily traffic) from other information to better understand the causes of traffic and to better predict future demand. Being able to predict or estimate demand is very important, especially if the process works on small roads where there aren't enough resources to measure every year. OLS Multivariable Linear Regression was used first and then Random Forest. These models were performed in Python with the help of scikit-learn.

Model 1 - All Data

The first model includes all the data. That was accomplished by adding the population demographic and income data to every measured traffic point. So, each measured traffic spot on the interstate has the associated county income and demographic data. For the set-up text columns were removed and blank (NaN) values were removed. The NaN values were present since the Census Bureau or specifically the American Community Survey does not collect population and income information for all the California counties annually. With the high values of some variables like population, the numbers were scaled with the scikit-learn standard scaler function. The model was then calculated in Python with the scikit-learn package.

The R-squared of this model was 0.993 which means the model could explain 99.3% of the Avg AADT value with the variables chosen. This is an incredibly good fit for a model. The model needed additional investigation. Interestingly, there were quite a few very statistically relevant variables that had P scores less than 0.05 specifically longitude was relevant, and latitude was not.

Model 2 - Filtered

This model had the same setup as model 1 but some variables were removed to leave only the economic and population information. These are hypothesized to be important factors.

The R-squared value was only 0.348 for this. Multiple iterations of this were tested and it is not that surprising that dropping the variables Avg Peak Hour and Avg Peak Month made the biggest difference. Still, some variables are found to be statistically relevant.

Model 3 – Filtered with Percent Change

Dealing with changes in large numbers is sometimes difficult to quantify as numbers in the model especially when working with data like population since the number it starts with is complex. Therefore, in this model, the numbers were changed to percent difference while grouped on ObjectID. No scalar was used since percent change is a way of scaling.

The R-square value for the model is 0.012. This did not perform as expected and potentially shows that the StandardScalar function in scikit-learn performs better than the percentage change.

Model 4a and 4b – Data by County

The small scope of points on the interstate was not very successful. Modeling the macro scope of whole counties may be more successful. The data frame used for this analysis was the same one used for model 1 but with some manipulation. First, unnecessary columns were removed and then everything was grouped by county. While grouping the AVG.AADT was aggregated as a mean (model 4a) and sum (model 4b). The StandardScalar function was used for this as well.

For the 4a model, the R-squared value was 0.823. This is an improvement over the smaller scope models. Interestingly the median income is more statistically significant than the mean income. Most population demographics were not relevant except for a few very specific age groups like ages 25-29. The 4b model performed even better with an R-squared value of 0.991. The importance of the variables is switched between the two models with model 4b finding population variables being statistically significant and income not being relevant.

Model 5 – Data by County with Percent Change

Percent did not perform well with the smaller scope, but it was still attempted at the larger scale. The setup was the same as model 4 except data was grouped by county and year to find the percent change over the years. No scalar was used.

The R-squared value was 0.439. This result was an improvement of the percentage change on the micro level but also performed worse than model 4. Again, the StandardScaler function in Scikit Learn seems to be better than manually using percent change.

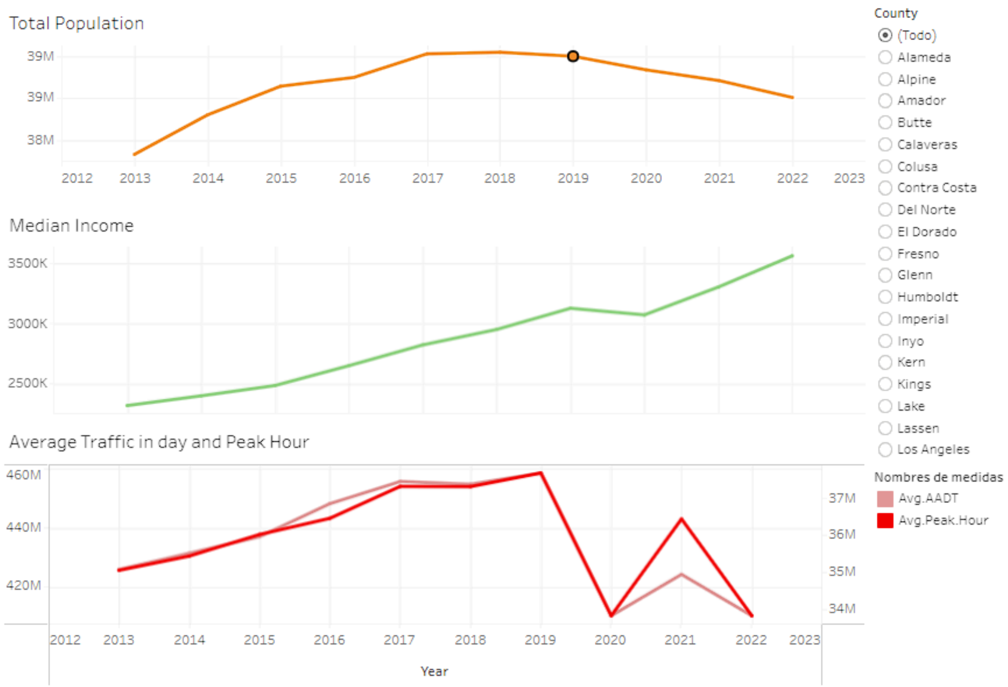
Model 6 – Random Forest Models

With the linear regression not being effective at the smallest scale a Random Forest model was tried as an alternative. The data setup used was the same as model 2 with some variables filtered out.

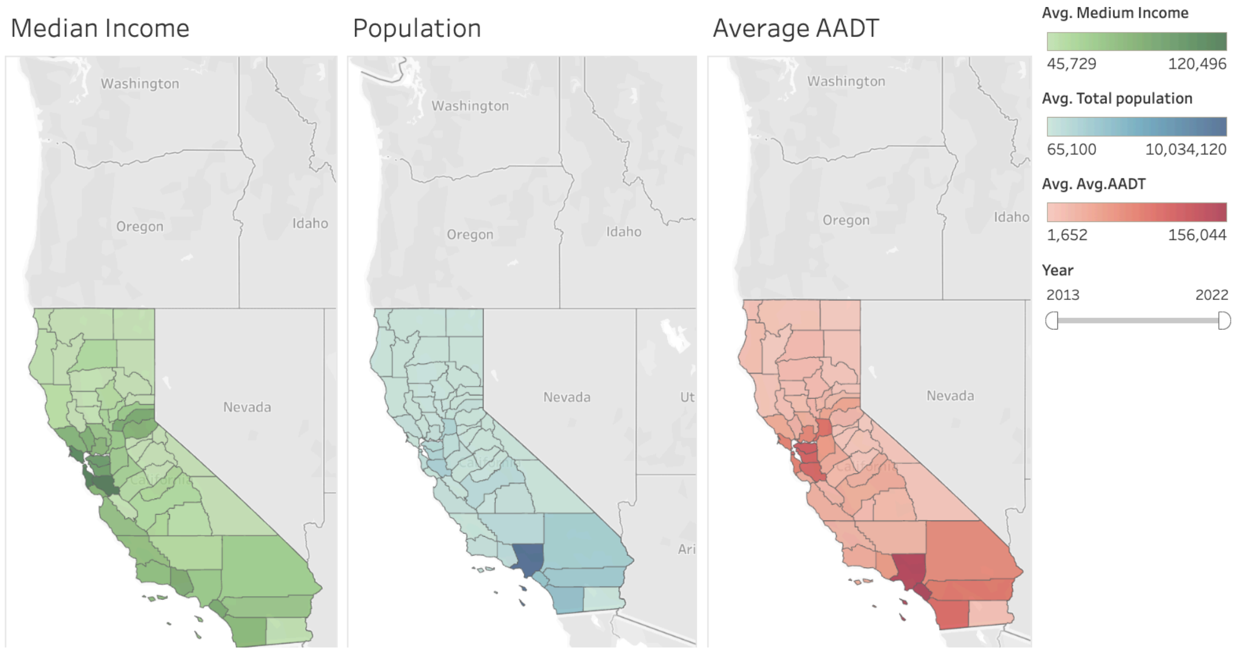
The R-squared value was 0.98. The model was very effective even without Avg Peak Hour and Avg Peak Month, unlike the OLS Regression. Without those variables latitude and longitude became the most important factors. Surprisingly the next most important was the age group of 85 years and older. However, when latitude and longitude are removed the model did not perform well with only an R-squared value of 0.37

Conclusion

The three datasets were combined, and the visualization of this data provided insight into their relationships over the years, and are displayed below. The traffic data from 2013 to 2022 revealed diverse trends. From 2013 to 2017, population income and traffic values grew at similar rates. However, between 2017 and 2019, while income growth continued, population growth stagnated, and traffic values declined. The most significant impact was seen in 2020, with all three measures experiencing a decline, particularly in traffic, due to the pandemic's effects on travel and economic activities. Post-2020, while income growth resumed, possibly due to inflation and federal interest rate changes, population decline persisted, and traffic volumes showed a rebound toward pre-COVID levels but with changes in peak-hour traffic patterns. We speculate that the changes in peak hours after the epidemic possibly reflect shifts to hybrid or remote work arrangements. The spreading of rush hour is described in Sam Ribakoff's article for The Courthouse News Service. We are unsure as to why the average daily AADT fell drastically in 2022.



All three datasets were used to create choropleth maps, so show counties with relatively low and high values over the years. A screenshot of this interactive dashboard can be seen below.



This map indicates a correlation between high traffic and population counts in Southern California, as well as a correlation between high traffic counts in median income and traffic counts in the Bay Area. The year slider on the right can be changed to see the change over time.

Much was learned through applying machine learning modeling to our datasets. The models indicate a correlation between average daily traffic, median income levels, and population. While these relationships are not always linear and quite complex, the results of the models greatly depend on how the model is built and how the data is preprocessed. The accuracy of some of the models was much better than expected. The assumption that a linear regression model makes of the variables being independent is a serious limitation, especially on the micro level. The complexities of the relationship between the variables need to have a more complicated model to reflect this. This was shown with the large condition number found with most models. Having a large condition number could mean there is a linear relationship between two or more independent variables. This makes it difficult for the model to estimate the individual effects of these variables on the dependent variable. There were also limitations on the data we collected with the traffic data being only on the interstates as well as the economic and population factors being only at the county level. This idea shows promise that would be worth pursuing by using a more complex model, having more data collected and more variables included.

This project illustrates the intricate relationships between economic shifts, population dynamics, and traffic volume in California. By visualizing this data, we have created comprehensive tools that can allow a user to explore three large datasets. These tools often break the data down for each county in California over the last 10 years and can be a valuable resource for those who are hoping to detect trends in this data and use these insights in a variety of fields. The combined analysis of these factors through data visualization and multivariable linear regression modeling provides a better understanding of how socioeconomic patterns and population changes will influence traffic trends. The insights from this study are critical for urban planning, infrastructure development, and strategic decision-making, offering a foundation for addressing California's evolving transportation challenges in the context of its changing economic and demographic landscape.

Works Cited

Brough, R., Freedman, M. L., & Phillips, D. (2021). Understanding socioeconomic disparities in travel behavior during the COVID-19 pandemic. Journal of Regional Science, 61(4), 753–774. https://doi.org/10.1111/jors.12527

Huang, Hai-Jun, et al. “Transportation Issues in Developing China’s Urban Agglomerations.” Transport Policy, vol. 85, Jan. 2020, pp. A1–22. https://doi.org/10.1016/j.tranpol.2019.09.007.

Johnson, Hans, et al. “What’s behind California’s Recent Population Decline-and Why It Matters.” *Public Policy Institute of California*, Public Policy Institute of California, 2 Oct. 2023, www.ppic.org/publication/whats-behind-californias-recent-population-decline-and-why-it-matters/#:~:text=Population%20change%20is%20determined%20by,birth%20rates%20continued%20to%20decline.

Ribakoff, Sam. “Study Shows California Traffic Improving as Rush Hour Peak Spreads out after Covid Restrictions.” *Courthouse News Service*, 13 Sept. 2023, www.courthousenews.com/study-shows-california-traffic-improving-as-rush-hour-peak-spreads-out-after-covid-restrictions/.

Rodrigue, Jean-Paul. “Transportation-Land Use Interactions: The Geography of Transport Systems.” *The Geography of Transport Systems | The Spatial Organization of Transportation and Mobility*, Routledge, 9 Oct. 2022, transportgeography.org/contents/chapter8/urban-land-use-transportation/transportation-land-use-interactions/.

About the Authors

Riley Rutan is currently a student in the Masters of Science in Big Data Analytics program at San Diego State University, with a planned graduation date of Spring 2025. He received a Bachelor of Science in Physics from the University of California, Santa Cruz, and most recently worked as a business analyst for a healthcare technology company in British Columbia, Canada. Riley has interests in big climate data analytics, business analytics, healthcare data science, machine learning engineering, and data visualization. His email is provided below.
Email: riley.rutan@gmail.com
Project Tasks: Traffic Data Wrangling and Cleaning, Data Visualization and Analysis for Traffic and Combined Datasets, Website Design, and Video Creation.

Eric Braga is a master's student in the Masters of Science Big Data Analytics program at San Diego State University set to graduate in spring of 2025. He previously graduated with a bachelor's degree in Economics from California State University Northridge. He has worked for various businesses as a business analyst including a chocolate company in the Los Angeles area. He has an interest in the fields of urban planning, transportation, machine learning engineering, and deep learning integration. He is working with The Smart Transportation Analytics Research (STAR) Lab. His email is provided below.
Email: ebraga890@gmail.com
Project Tasks: Coordinator, Traffic Data Analysis, Modeling.

Xinyu Du is a first-year master's in the Big Data Analytics program at San Diego State University and plans to graduate in Spring 2025. He received a Bachelor of Science degree in Aerospace Engineering from the University of California, San Diego in 2022. Xinyu has interests in the fields of data visualization, business analytics, image processing, and machine learning. His email is provided below.
Email: xinyu.xydu@gmail.com
Project Tasks: Household Income Data Cleaning, Data Visualization and Analysis for Household Income Datasets, Web Design and Content.

Miguel Bravo Martinez Del Valle is a first-year student at San Diego State University, in the Big Data Analytics Master's program, set to graduate in Spring of 2025. He took a bachelor's in Electronics, Robotics, and Mechatronics Engineering at the University of Málaga, back in Spain. Miguel has an interest in Data Analytics, Machine Learning, and Artificial Intelligence. He is currently in two Research Groups, one in Data Visualization working with an Ireland company and he is also in the AI4Businnes team, starting his research in Spring 2024.
Email: miguelangelbravo2000@gmail.com
Project Tasks: Population Data Cleaning and Analysis, Web Design and Content. Data Visualization and Analysis for Population and Combined Datasets.