Used Car Listings Dataset (U.S.A.) Name：Wei-Xuan Hu

I. Introduction

Owning a car is helpful for living in the US due to its vast size. For international students, getting around can be challenging without a car, often relying on friends or public transportation. Buying a used car may offer a feasible solution due to its lower price. However, to choose a suitable used car that meets criteria such as affordability, acceptable condition, and age, comprehensive comparisons are necessary. Additionally, differences between states further complicate the decision-making process. Factors like climate variations, terrain ruggedness, and urban-rural divide also play significant roles in determining the ideal car for different regions. In summary, gaining a deeper understanding of the market and considering regional factors is essential for individuals unfamiliar with car ownership in the US.

A. Research Topic

The research focuses on the challenges and considerations involved in purchasing a used car in the United States, particularly for international students.

B. Motivation

The motivation behind this research stems from the recognition of the pivotal role cars play in navigating the vast expanse of the United States. International students, often without access to personal vehicles, face unique obstacles in transportation, which can significantly impact their mobility and overall experience in the country. By delving into the intricacies of used car ownership, we aim to provide valuable insights and guidance to help mitigate these challenges.

C. Goals

- To examine the feasibility of purchasing a used car as a solution for international students navigating transportation challenges in the US.
- To identify and analyze the key criteria, such as affordability, condition, age, and regional factors, that influence the selection of a suitable used car.
- To explore the variations in car ownership needs and preferences across different states, considering factors such as climate, terrain, and urban-rural disparities.
- To provide comprehensive comparisons and guidelines to assist international students and other individuals unfamiliar with car ownership in making informed decisions when purchasing a used car in the US.

D. Research Questions

- What are the most popular makes and models of used cars in the dataset, and do their prices vary significantly by region?
- Do regions with high latitude and low temperatures prefer certain brands or types?
- Does the quantity and density of major cities have an impact on the prices of used cars?

II. Methodology

A. Data Collection

In this study, we will primarily utilize two datasets. The first dataset, obtained from Kaggle(https://www.kaggle.com/datasets/tushar5harma/signate-career-up-2023-dataset?resource=download), titled final_data.csv, provides detailed information about a diverse collection of used cars available for sale in various regions across the United States. With a total of approximately 27,500 entries, this dataset offers a rich resource for analyzing and modeling the factors that influence used car prices and market trends.

The second dataset, sourced from the PRISM Climate Group (https://prism.oregonstate.edu/), consists of global land temperatures by state. Specifically, we have obtained the dataset titled 'GlobalLandTemperaturesByState.csv'.

Table 1 ：Field Overview of final_data.csv

Attributes	Description
id	A unique identifier for each car listing.
region	The location where the car is listed for sale (e.g., North, South, East).
Year	The year the car was manufactured.
Manufacturer	The manufacturer of the car.
condition	The condition of the car (e.g., excellent, good, fair).
cylinders	The number of cylinders in the car's engine.
fuel	The type of fuel the car uses (e.g., gas, diesel).
odometer	The mileage on the car.
title_status	The status of the car's title (e.g., clean, rebuilt).
transmission	The type of transmission in the car (e.g., manual, automatic).
drive	The type of drive in the car (e.g., fwd, rwd, 4wd).
size	The size category of the car (e.g., compact, mid-size, full-size).
type	The body type of the car (e.g., sedan, SUV, convertible).
paint_color	The color of the car.
state	The state where the car is listed for sale.
price	The listing price of the car.

Table 2 ：Field Overview of GlobalLandTemperaturesByState.csv

Attributes	Description
dt	Date in the format "YYYY-MM-DD" representing the observation date of the temperature measurement.
AverageTempe-rature	Average temperature measured in degrees Celsius for the specified date, state, and country.
AverageTempe-ratureUncerta-inty	Uncertainty associated with the average temperature measurement, typically provided in degrees Celsius.
State	Name of the state where the temperature measurement was recorded.
Country	Name of the country where the temperature measurement was recorded.

In this project, for the sake of data consistency and to mitigate the influence of the Financial Crisis (2007-2008), we will focus solely on exploring the trends from 2010 to 2013.

I have obtained supplementary data from the PRISM Climate Group, specifically sourced from the dataset titled 'Global Land Temperatures By State.csv'.

Secondly, we will also use the "us_states" dataset (spData package) to obtain the geometry of the states and boundaries, then overlay the histograms.

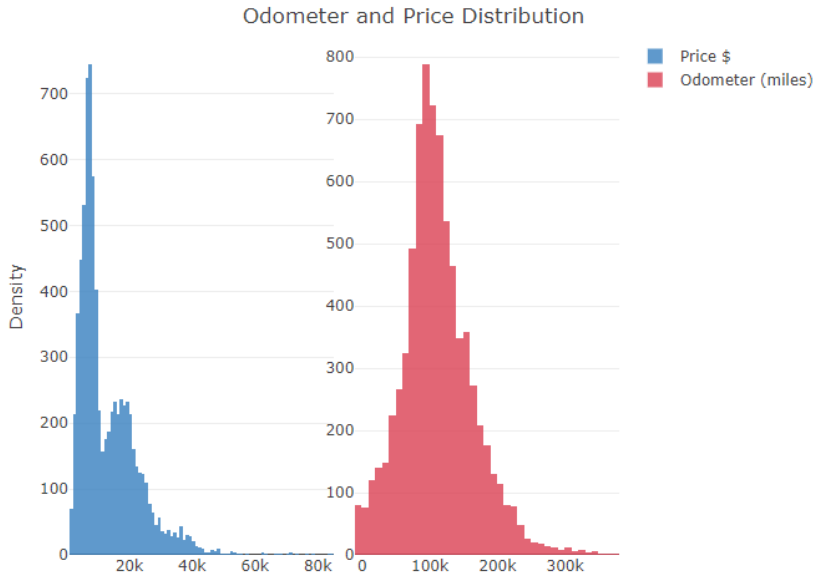
Finally, because this project aims to investigate the impact of distance on the prices of used cars, we need to import the "USA Major Cities" dataset (https://hub.arcgis.com/datasets/esri::usa-major-cities/explore) to determine the locations of urban areas. Additionally, the "state" data in the used car dataset is abbreviated, so we need to import the "List-of-US-States" dataset (https://github.com/jasonong/List-of-US-States/tree/master) to convert these state abbreviations into full names for better integration with the "us_states" dataset to display this information on the map of the United States.

B. Data Preparation

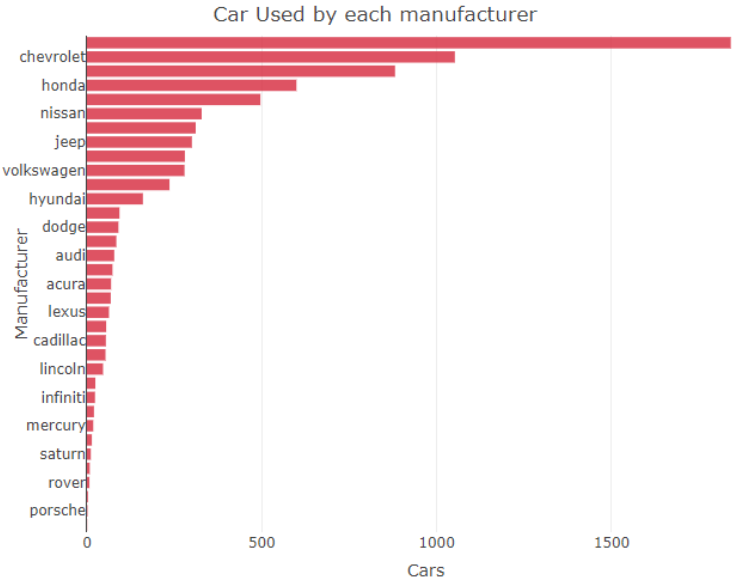
Before importing the data, it is necessary to synchronize the input of various brands in Excel, as some original data entries for manufacturers are in full-width characters, and inconsistencies in capitalization may occur for some identical brands. **Prior to data organization, there are 125 brand categories. Using the Replace function, the categories within the data were unified into 37 categories.**

C. Exploratory Data Analysis

1. The histograms for both price and odometer display right-skewed distributions, indicating that vehicles with higher odometer readings are less frequently listed for sale, and similarly, cars with higher prices are less commonly available for purchase.



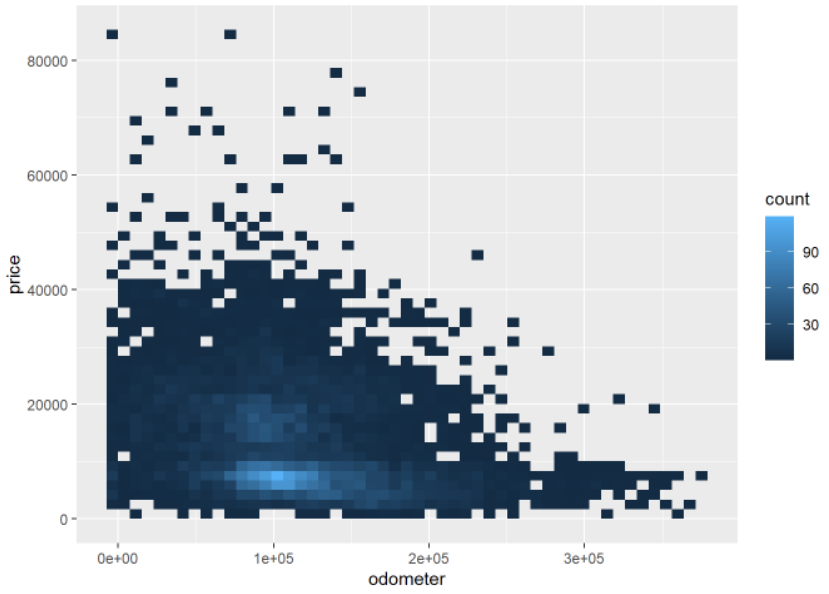
2. For this graph, it is obvious that Ford, Chevrolet, and BMW have the most vehicles.



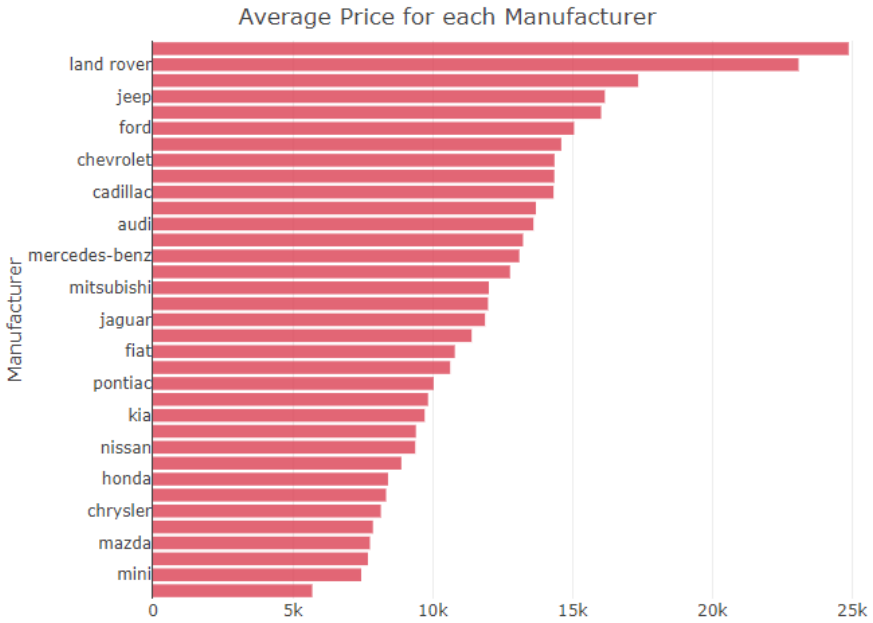
3. From the data, we can observe that the top two brands are from the United States, while the third-ranked brand, BMW, is from Germany. The subsequent three brands are all from Japan. Additionally, Japanese brands have average prices approximately $5,000 lower than European and American brands.

	manufacturer	avg_price
acura	acura	8336.54
audi	audi	13604.95
bmw	bmw	14601.64
buick	buick	11396.08
cadillac	cadillac	14323.94
chevrolet	chevrolet	14358.07
chrysler	chrysler	8148.92
dodge	dodge	14352.72
fiat	fiat	10794.86
ford	ford	15058.45
gmc	gmc	17351.43
honda	honda	8406.72
hyundai	hyundai	7687.96
infiniti	infiniti	10627.70
jaguar	jaguar	11875.33
jeep	jeep	16159.65
kia	kia	9716.39
land rover	land rover	23088.00
lexus	lexus	11982.30
lincoln	lincoln	9836.50
mazda	mazda	7759.22
mercedes-benz	mercedes-benz	13100.55
mercury	mercury	5696.89
mini	mini	7446.50
mitsubishi	mitsubishi	12011.58
nissan	nissan	9379.56
pontiac	pontiac	10035.75
porsche	porsche	16026.00
ram	ram	24878.32
rover	rover	13234.14
saturn	saturn	8882.45
subaru	subaru	12767.76
toyota	toyota	9402.93
volkswagen	volkswagen	7866.45
volvo	volvo	13694.25

4. From the above chart, we can observe that the popular odometer is around 100,000. And, there appears to be a negative relationship between the mileage of a car and its price. This is not surprising, as cars with lower mileage typically command higher prices.



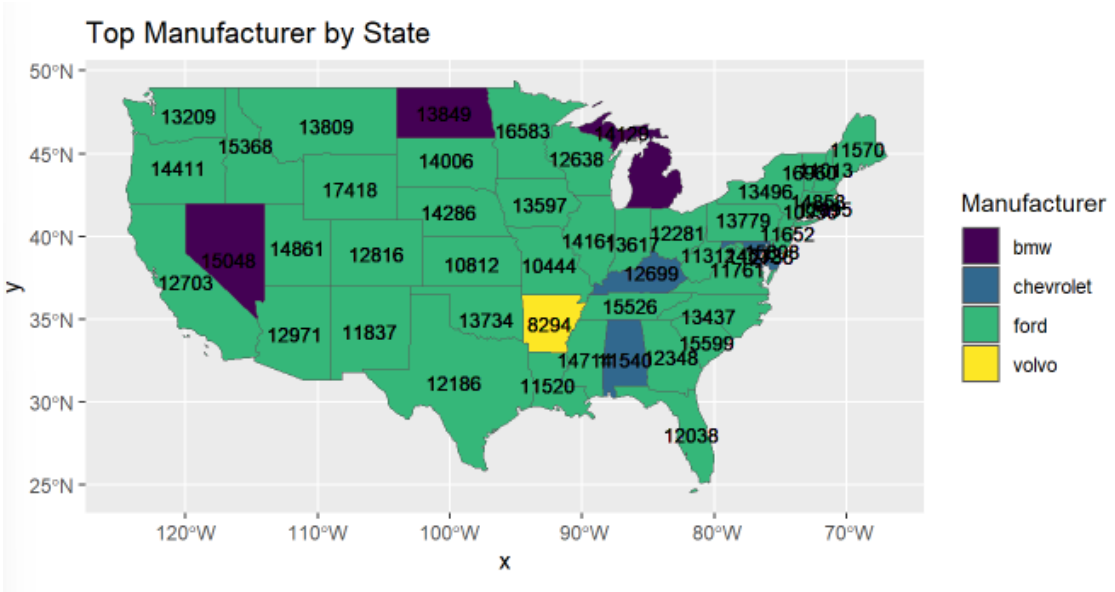
5. Observing the top six best-selling manufacturers mentioned earlier, we can see that the prices for Ford, Chevrolet, and BMW cars are approximately in the range of 14-15k, while the prices for Honda, Toyota, and Nissan cars fall around 8.4-9k. From this, we can infer that these two price ranges are desirable in the market, and consumers are willing to pay for cars within these price ranges.



III. Founds and Results

A. What are the most popular makes of used cars in the dataset, and do their prices vary significantly by region?

- As seen from the exploratory analysis in the previous section, we found that Ford, Chevrolet, and BMW are among the most popular vehicle manufacturers. Based on the summary of filtered_data, the mean price is $13,126, with the third quartile at $18,373. This suggests that residents who prefer to purchase luxury cars like BMW tend to reside in states with higher average car prices. Conversely, if the average car price in a state is lower, it is more likely that residents lean towards purchasing domestic brands like Ford. The region where Volvo sells best is Arkansas, where the average car price is the lowest among states, just slightly higher than the first quartile at $6,430, with an average price of $8,294. Furthermore, from a geographical perspective, this project speculates that areas with dry and low-temperature climates are more inclined towards European brands like BMW. This is because the climate in those areas resembles that of inland European countries, which is conducive to the use of automobiles.

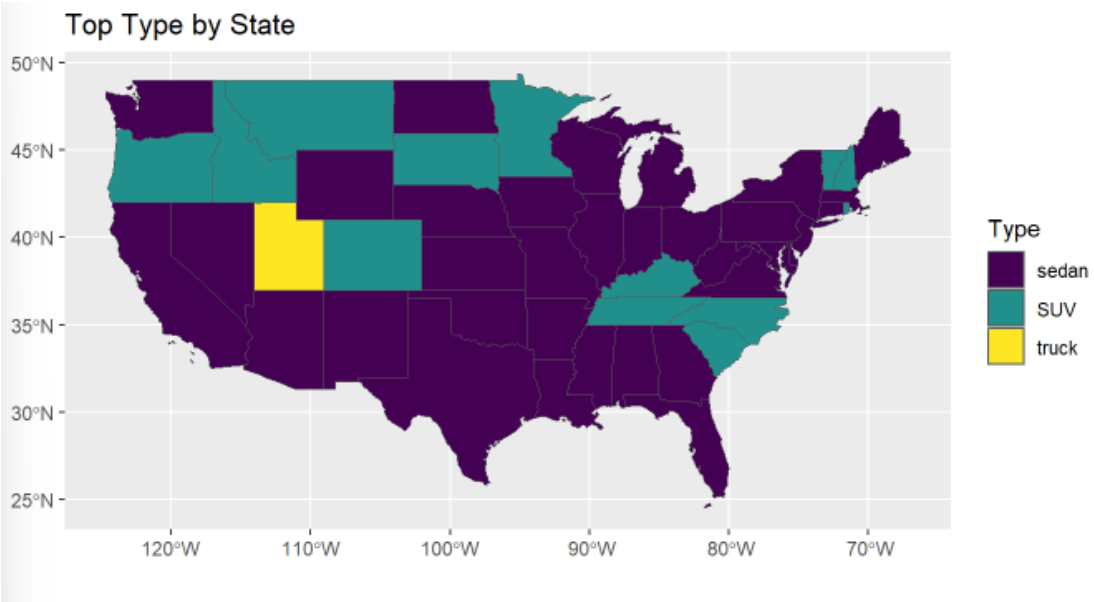


B. Do regions with high latitude and low temperatures prefer certain brands or types?

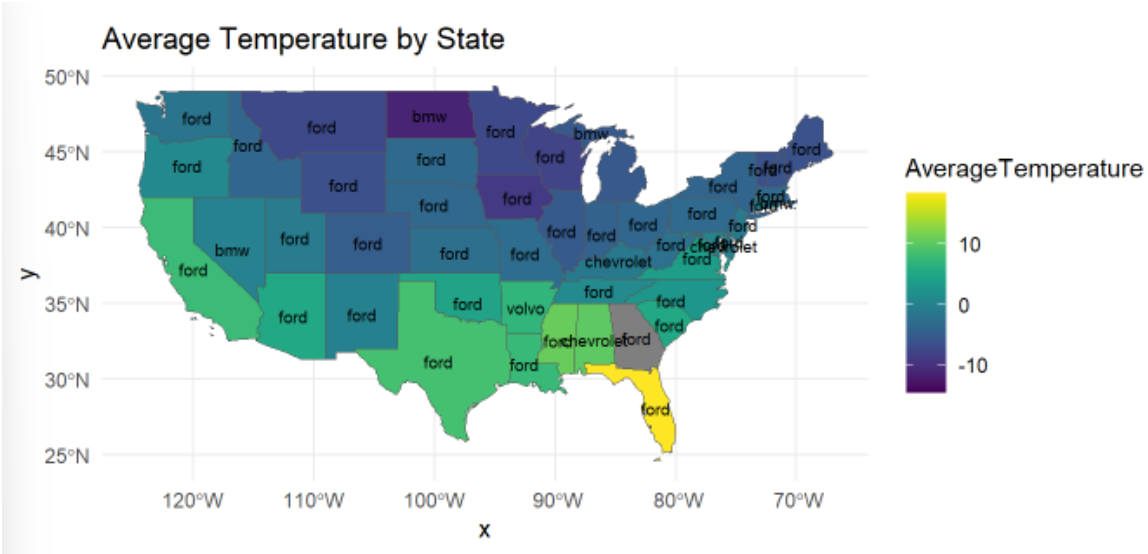
- States with significant mountain ranges include Colorado, California, Washington, Oregon, Montana, and Wyoming. From the data presented, it is evident that many of the regions where SUVs are preferred are characterized by abundant mountainous terrain. Additionally, Kentucky, Tennessee, North Carolina, and South Carolina stand out for their notable preference for SUVs, driven by several factors. These states feature diverse landscapes encompassing mountains, hills, and plains, making SUVs an appealing choice due to their higher ground clearance and four-wheel drive capabilities, which are well-suited for navigating varied road conditions. Furthermore, the prevalence of outdoor recreational activities such as hiking, camping, and fishing in these regions enhances the desirability of SUVs, known for their spacious interiors and off-road performance. Moreover, the humid climates and seasonal extreme weather conditions in these states underscore the importance of SUVs, providing stability and safety for drivers facing challenging weather

conditions. Additionally, the presence of rural areas within these states necessitates vehicles capable of traversing uneven and muddy rural roads, further fueling the demand for SUVs among residents.

➢

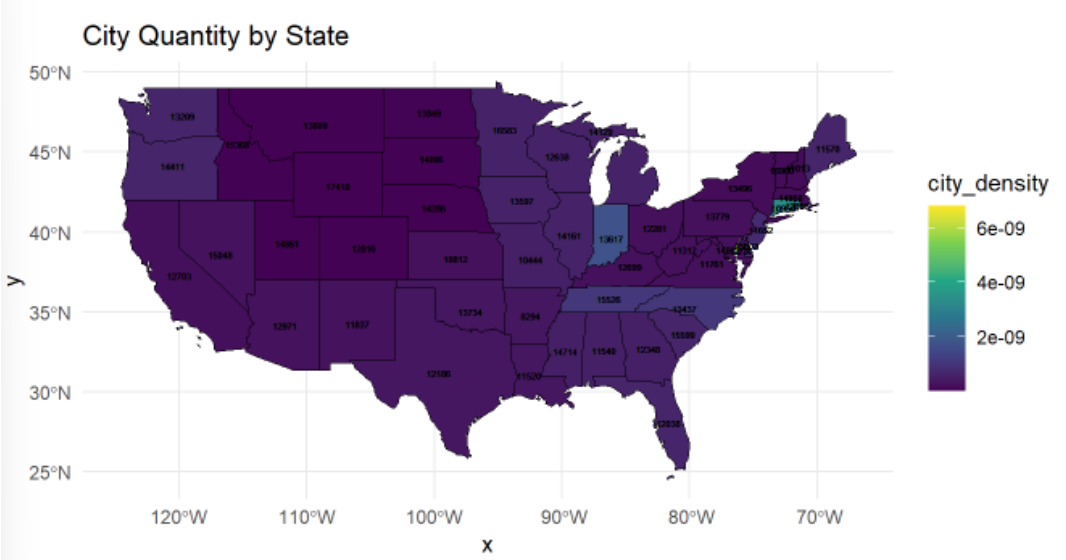


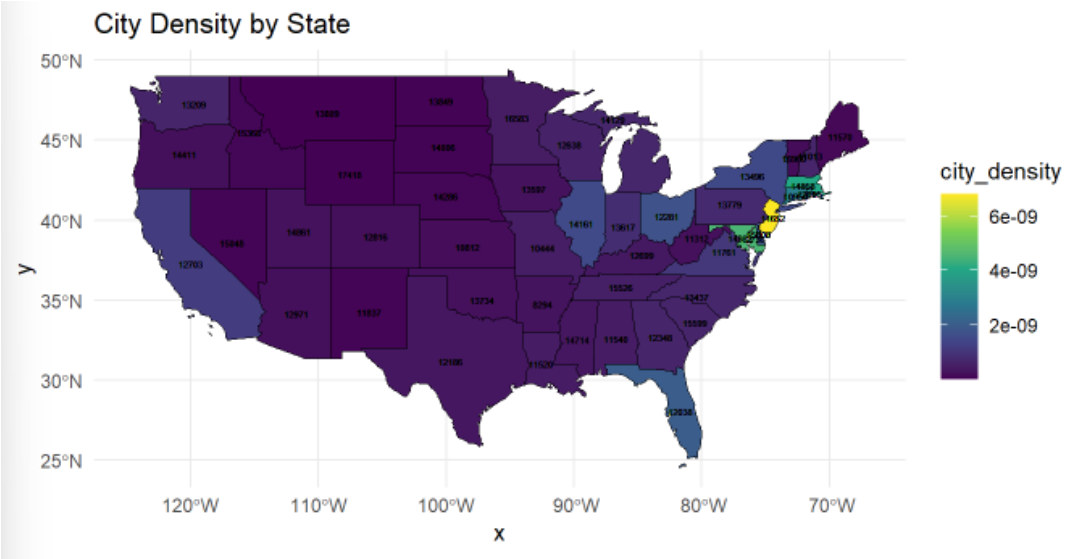
➢ From the average temperature by state chart, it's evident that the average January temperatures tend to decrease as you move further north. Conversely, regions located more towards the south and closer to the coast experience relatively higher temperatures, notably Florida. However, upon closer inspection, it becomes apparent that the most popular car brands do not show a clear correlation with the temperature of the region. Ford remains the top-selling brand in most states, with a few states showing preferences for BMW, Volvo, or Chevrolet. Interestingly, among these states, there are no significant common geographical features.



C. Do regions with high latitude and low temperatures prefer certain brands or types?

- In theory, the more numerous the big cities in a state, the more developed and well-constructed it is, leading people to rely less on personal vehicle ownership and more on public transportation. Looking at states like California and New York, the average car price is around $13,000. Especially on the East Coast, excluding the capital city areas, including New York, Delaware, New Jersey, Maryland, and Connecticut. However, in other regions, based on the average car price, there does appear to be a positive correlation between the quantity of cities and the average price.





IV. Summary

In this project, we investigated various factors influencing the purchase of used cars in the United States, with a particular focus on the considerations for international students. The exploration involved data processing, exploratory analysis, and addressing specific research questions.

A. Exploratory Analysis

1. The exploratory analysis encompassed visualizations and statistical summaries to gain insights into the dataset:
2. Popular Car Manufacturers: We identified the most popular car manufacturers, with Ford, Chevrolet, and BMW emerging as the top contenders. Japanese brands such as Honda, Toyota, and Nissan also had a significant presence in the market.
3. Price Distribution and Odometer Mileage: Histograms were utilized to visualize the distribution of car prices and odometer mileage. The data indicated right-skewed distributions for both variables, suggesting that cars with higher prices and lower mileage were less frequently listed for sale.
4. Average Price by Manufacturer: We calculated the average prices for different car manufacturers, revealing varying price ranges across brands. European and American brands like Ford, Chevrolet, and BMW had higher average prices compared to Japanese brands like Honda, Toyota, and Nissan.
5. Car Type Preference: Analysis of car type preferences indicated that front-wheel drive (fwd) vehicles were the most common, followed by rear-wheel drive (rwd) and four-wheel drive (4wd) options.
6. Geographical Preferences: Geographical analysis revealed regional preferences for certain car types and manufacturers. States with significant mountain ranges tended to prefer SUVs, while other factors such as climate and terrain also influenced car preferences.
7. Average Temperature by State: The relationship between car preferences and average temperatures by state was explored, revealing that temperature variations did not necessarily correlate with the popularity of specific car brands.
8. City Quantity and Density: Analysis of city quantity and density suggested a positive correlation between the number of major cities and the average price of used cars. States with higher city densities tended to have higher average car prices.

B. Research Questions

Finally, specific research questions were addressed, including the impact of geographical factors on car preferences, the influence of temperature on brand selection, and the relationship between city quantity/density and car prices.

In summary, this project provides valuable insights into the complexities of purchasing used cars in the United States, offering guidance for international students and other individuals navigating the car market in the country. Further research could delve deeper into specific regional preferences and factors influencing car ownership decisions.