BAN7010-800

Team #4 Project

Natira Riddick, Luis Del Prado Guevara, Hunter Lecour, Joseph Tattegrainn. 05/11/2025

Research Questions

1. What environmental factors tend to impact the sales of our company's offerings?

Environmental factors which including social, political, economic, and geographic variables, significantly influence the automotive industry. Per Jim Makos's 2024 PESTLE Analysis of the Automotive Industry article, these factors majorly shape automotive sales performance for companies like Toyota, for instance. Economic factors are particularly influential. States with higher GDPs, like California, New York, and Texas, demonstrate stronger Toyota sales and higher average selling prices, most likely due to consumers getting more disposable income for new technologies, safety features, and premium upgrades. On the other hand, affordability becomes a priority during economic downturns or in lower-GDP states, often leading consumers to the used car market. Seasonal and cyclical trends in car prices are also strongly correlated with and reflect the general trends and conditions of the overall economy and consumer spending habits. Political factors also have a substantial impact. Tariffs on imported car parts are subject to increasing the cost of vehicles, while government incentives for electric vehicles (EVs) and stricter emissions regulations are shifting consumer preferences toward greener options. A stable political environment generally supports healthy automotive sales, whereas sudden regulatory changes can introduce uncertainty and alter market dynamics. Geographic factors shape regional vehicle demand. Southern and coastal states like Florida,

California, and Texas consistently show high Toyota sales, driven by favorable driving conditions and strong personal vehicle use. In contrast, colder northern states such as Minnesota and North Dakota exhibit a preference for durable all-wheel-drive SUVs and trucks suitable for harsh winter conditions. Finally, social trends are reshaping the market. U.S. consumers currently are leaning towards more versatile vehicles like SUVs and sedans that can meet both family and work needs. There is also a notable demand for used vehicles in "fair" condition, highlighting the continued importance of affordability. Broader lifestyle changes, such as the growth of remote work, the gig economy, and increasing environmental awareness, are driving automakers to focus on features like connectivity, safety, and eco-friendliness to align with evolving consumer expectations.

Reference:

Makos, J. (2024, June 18). PESTLE Analysis of the Automotive Industry. PESTLEanalysis.com. Retrieved from https://pestleanalysis.com/pestle-analysis-of-the-automotive-industry/

2. What is the current structure of the used car industry in the USA?

Per the key findings, IBISWorld's "Used Car Dealers in the US - Market Research Report" (2024) revealed that the U.S. used car market is large, extensive, and continuously adapting. The market is widely distributed, with listings across all states. While most sales come from small dealerships and private sellers, major companies similar to CarMax and Carvana are getting big rapidly. Carvana, for example, has transformed car buying with a fully online process. It allows customers to glance, take virtual tours, secure financing, and have the car delivered. This instance reflects a growing consumer preference for online convenience. Also, technology is significantly reshaping the industry. Customers can have the opportunity to easily access detailed vehicle histories, compare prices, and get real-time financing online. Predictive pricing algorithm, as an advanced tool, is designed to help the sellers tremendously. The report also emphasizes that online research is increasingly influencing used car sales, with many buyers engaging with websites prior to any purchases. Carmax for example, notes that over half of its customers use its website before a transaction. This type of trend indicates a move towards a more and customer-centric market, emphasizing transparency, convenience, and control.

Reference:

IBISWorld. (2024). Used car dealers in the US - Market research report. Retrieved from https://www.ibisworld.com/

3. Who are the market leaders, and what are their characteristics?

The U.S. used car market is led by automotive brands that strive to deliver value through

the following characteristics: reliability, durability, and affordability. According to IBISWorld (2024), Toyota, Honda, Ford, and Chevrolet are the primary market leaders. This can be credited due to their widespread appeal, high resale value, and ability to meet the diverse needs of American consumers. For example, our dataset analysis confirms Toyota's leadership position as a market leader. The "Top 10 Most Sold Toyota Models" graph clearly highlights models such as the Camry, Corolla, and RAV4 as the most frequently listed and sold amongst Toyota products. These models are known for their balance of fuel efficiency, comfort, and longevity, which appeals to budget-conscious and value-driven buyers alike and ties into the characteristics previously discussed of market leaders. There is a Toyota Corolla that is older than 20 years still sitting and standing strong in my garage as further evidence of their durability.

In addition, the "Toyota Listings by Body Type" visualization reveals a clear dominance of sedans and SUVs, which reinforces Toyota's appeal to families, commuters, and drivers who prioritize utility above all. The strong presence of these body types is further proven by the "Price Distribution by Body Type" chart, which shows consistent pricing and stable value across the board. Furthermore, Toyota's widespread dealership infrastructure, emphasis on certified pre-owned (CPO) programs, and financing options contribute to its competitive advantage. Toyota's adaptability in responding to changing consumer preferences, which includes offering hybrid versions of popular models, and further secures its role as a market leader.

According to Saltzman (2024), Toyota holds a 15% market share in the U.S. auto industry, a statistic that can be reflected in the volume and geographic spread of listings in our dataset. Especially, in key states such as Florida, California, and Pennsylvania. Taken

together, Toyota's success can be attributed to a combination of brand equity, model diversity, price stability, and the ability to scale operations both digitally and in-person, which are all crucial traits of a dominant player in today's evolving market

References -

IBISWorld. (2024). *Used Car Dealers in the US – Market Research Report*. Retrieved from https://www.ibisworld.com/

Saltzman, D. (2024). *Analysis of Toyota and the Automobile Industry*.Harwood, A. (2025, February 28). *2025 Best Resale Value Awards: Top Cars, Trucks, and SUVs*. Kelley Blue Book. https://www.kbb.com/awards/best-resale-value-cars-trucks-

4. How will the market change in the next 5 to 10 years?

The U.S. used car market is expected to undergo a change centered on factors such as technology, environmental, and economy. Based on our ARIMA time series model, which used Toyota vehicle pricing data, average prices are expected to continue a moderate upward trend. This data suggests sustained demand for used vehicles, especially in times of economic uncertainty, which makes affordability more critical than ever. However, the types of vehicles consumers demand are already shifting. According to Makos (2024), government incentives and stricter emissions regulations will drive the increased adoption of hybrid and electric vehicles alike. Toyota's investment in hybrid technology with models like the Prius and RAV4 Hybrid puts them in a position to capture future market share as these technologies become more mainstream in the used market.

Digitization will further redesign the industry. As reported by IBISWorld (2024), major players such as CarMax and Carvana are rapidly expanding the fully online purchase model and making it more accessible and easier to navigate for users. Consumers now expect end-

to-end digital experiences, which include remote financing, digital vehicle tours, and doorstep delivery. It's like getting an Amazon prime package delivered right to your door with next day shipping. This trend is supported by our own regression modeling, which demonstrated how GDP levels influence vehicle pricing and implied that buyers in high-GDP states may prefer a seamless premium online experience.

Another trend that is the increasing use of predictive analytics and dynamic pricing, made possible through machine learning and real-time data integration. Our clustering analysis identified distinct buyer groups based on odometer readings, year, and selling price, which is a signal that dealerships and platforms could use to deliver customized pricing and inventory options.

Lastly, we anticipate a shift in ownership patterns. Models such as subscription-based vehicle access and shared mobility services may reduce long-term ownership, particularly in urban centers. Traditional dealerships may find new roles in servicing shared fleets or providing flexible leasing options. Toyota's continued investment in hybrid models, digital infrastructure, and mobility innovation ensures it remains prepared for this change in the market.

In summary, the next decade will reward automotive brands and dealers that are agile, tech-savvy, and consumer-focused. Toyota is already well-positioned to maintain leadership by embracing electrification, strengthening digital channels, and offering flexible ownership solutions.

References –

Makos, J. (2024). *PESTLE Analysis of the Automotive Industry*. PESTLEanalysis.com. https://pestleanalysis.com/pestle-analysis-of-the-automotive-industry/

IBISWorld. (2024). *Used Car Dealers in the US – Market Research Report*. https://www.ibisworld.com/

Mordor Intelligence. (2024). *United States Used Car Market – Growth, Trends, and Forecasts (2024-2029)*.

Lemaire, J., Park, S. C., & Wang, K. C. (2016). *The use of annual mileage as a rating variable*. ASTIN Bulletin: The Journal of the IAA, 46(1), 39-69. Supporting Visuals from R Script 12-Month ARIMA Forecast of Average Toyota Prices – Demonstrates projected upward price trend. GDP vs. Selling Price Regression Model – Highlights influence of economic conditions on pricing. K-Means Clustering by Year, Odometer, and Price – Shows market segmentation and potential for targeted offers. Selling Price vs. Odometer – Demonstrates depreciation patterns and importance of mileage in valuation.

Strategic Recommendations

1. What geographic area should we target?

Toyota has a significant presence in the U.S. car industry. According to SUNY Cortland, Toyota accounts for 15% of the market share in the automobile industry in 2024(Saltzman, 2024). We have evidence from the car price dataset that California, Florida, and Pennsylvania emerge as the strongest areas geographically around the country. These states should always be areas to target to expand Toyotas presence. In the car prices dataset, Florida had 8,613 listings, California for 4527 listings, and Pennsylvania for 3,046 listings. The volume of these listings was significantly higher than other states according to the dataset's records via the R-code output. This suggests that demand is already present in these states so maintaining and expanding their strengths would be essential. According to the "Average Toyota Selling Price by State" visualization, states like Pennsylvania and Florida have some of the highest prices for Toyotas in the country. This

strengthens the idea of expanding in these demanding states due to their pricing power. Notable emerging markets to geographically target would be states like Ohio and Georgia. Georgia has 2944 listings and Ohio had 2426 listings, which were not too far off from the top 3 states. Another supportive reason will be the average Toyota selling prices in these emerging markets. Ohio shows strong pricing power. Georgia shows strong listing volume, and although its average pricing is moderate compared to top markets, the volume compensates for profitability. This supports the idea of not just expanding and maintaining the top states like California and Pennsylvania but keeping an eye out on states with emerging markets with respectable pricing power. The strategy maximizes profitability while building momentum in sales in emerging markets in the U.S.

References

Saltzman, D. (2024). Analysis of Toyota and the Automobile Industry.

2. What car models and specifications should we use for the expansion?

The analysis showed strong support in prioritizing sedan and SUV models for Toyotas expansion. In the “Toyota Listings by Body Type” visual, sedans and SUVs show dominance in body types compared to the others. This gets further supported by the “Top 10 most sold Toyota Models” visual where Camrys, Corollas, and Rav4s are at the top of most sold car models. These are all sedans and/or SUVs, but sedans are at the top leading in volume. In our “Price Distribution by Body Type” visualization, it shows the stability of prices across

body types. Both body types, especially sedans, showed very stable prices for resale value. This makes these models financially advantageous to the company for expansion. Kelley Blue Book (Harwood, 2025) ranks the Toyota RAV4 among the top vehicles for best resale value in 2025. To ensure efficiency, it’s also recommended that with support from our data prioritize vehicles with odometer readings under 100,000 miles. The 'Selling Price vs. Odometer's visualization shows that as mileage increases, selling prices decrease. This correlation as well as general car research supports the idea that relatively lower mileage maintains healthy price values. According to Lemaire, Park, & Wang (2016), “It is intuitively obvious that annual mileage positively correlates with claim frequencies, since each mile a car travels creates a small chance of an accident.” Claims and accidents threaten the car’s value and can decrease its resale price. In

conclusion, focusing the inventory around the Camry, Corolla, and RAV4 models that are in good condition under 100,000 miles is well supported for expansion based on our data and external research.

References

Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable.

ASTIN Bulletin: The Journal of the IAA, 46(1), 39-69.

Harwood, A. (2025, February 28). 2025 Best Resale Value Awards: Top Cars, Trucks, and

SUVs. Kelley Blue Book. https://www.kbb.com/awards/best-resale-value-cars-trucks-suvs/

3. Are there any important consumer behavior trends that our company should be aware of?

The analysis shows there are several main key factors and behavior terms. These trends and behaviors should be monitored and considered when planning the market strategies. Firstly, the data shows there is a price sensitivity, and it remains a dominant factor across all the US states. The clustering analysis performed reveals that there is an existence of distinct buyer segments, which many are prioritizing, affordability, and vehicle age while simultaneously underscoring a strong market for budget conscious consumers (Cox Automotive, 2023). The analysis perform gave us the data that practicality is a leading influence. Buyers are now consistently valuing fuel, efficiency, vehicle life, and low maintenance costs. Additionally, the data also showed there is a clear and growing demand for product transparency. Buyers are wanting to know and have access to detailed condition reports, service, history, accident records, etc. this indicates a shift toward trust driven purchasing where buyers want to feel more secure and have that full transparency When purchasing a vehicle (J.D. Power, 2022). lastly, some seasonal trends play a role in consumer behavior as well. For example, the average selling prices tend to peak during the spring, which is typically around tax refund time when people have that extra money to spend. Buying typically dips in the fall due to the upcoming holidays and the expenses that go along with that. These behavioral patterns and trends highlight the importance of getting out specific messaging and forecasting the right inventory to align with behavioral patterns, seasonal trends, economic conditions, etc.

References

Cox Automotive. (2023). *Used vehicle market insights report*. Retrieved from https://www.coxautoinc.com

J.D. Power. (2022). *2022 U.S. Automotive Brand Transparency Study*. Retrieved from https://www.jdpower.com

1. Introduction & Objectives

This report examines the U.S. used-Toyota vehicle market to:
• Identify external factors impacting sales.
• Describe current industry structure and market leaders.
• Forecast market evolution over the next 5–10 years.
• Recommend geographic, model, and consumer strategy for expansion.

2. Data & Methodology

Data Sources:
- Primary dataset: ~100,000+ Toyota listings (2014–2015).
- State-level Real GDP (2015).

Methods:
- Data Cleaning: Standardization, outlier filtering (odometer >300k mi), U.S. state filtering.
- EDA: Aggregation and visualization of price, model, body type, mileage, and time trends.
- Regression: OLS model on selling price with predictors year, odometer, condition, GDP.
- Time-Series: ARIMA forecasting for 12 months; qualitative projection to 5–10 years.
- Clustering: K-means segmentation (k=3) on key vehicle attributes.

3. Exploratory Findings

3.1 Regional Pricing Patterns:
- Highest avg. prices in CA, NY, TX, MA; lowest in MS, AL.

3.2 Model & Body Type Popularity:
- Top models: Corolla, Camry, Tacoma, RAV4, Highlander.
- Body types: Sedans ~45%, SUVs ~30%, Trucks ~15%.

3.3 Mileage vs. Price:

- Negative linear relationship: ~-$500 per 10k miles.
3.4 Price Trends Over Time:
- Stable avg. around $15k–$16k with minor seasonal swings.
3.5 Condition & Price Distributions:
- “Excellent” ~20%, “Good” ~50%, “Fair/Poor” ~30%.

4. Regression Analysis

Model: SellingPrice ~ Year + Odometer + Condition + GDP_2015
Key Coefficients:
- Year: +$300–$400 per model year.
- Odometer: -$0.05 per mile.
- Condition: +$1,200–$1,500 per condition step.
- GDP_2015: +$0.002 per $1M.
Interpretation: Vehicle age, mileage, and condition are primary price drivers; economic strength adds modest uplift.

5. Forecasting Outlook

5.1 12-Month ARIMA Forecast:
- Model ARIMA(1,1,1) predicts 3–5% price growth (to $16.5k–$17k).
5.2 5–10 Year Projection:
- Moderate inflation (15–25% cumulative over 5 years).
- Growing hybrid/EV demand; supply constraints may boost prices intermittently.

6. Buyer Segmentation

Clusters (k=3):
A. Low-mileage, Newer, Premium (SUV skew)
B. Mid-range, Balanced (mixed sedans/SUVs)
C. High-mileage, Older, Value (sedan-focused)
Marketing: Tailor certified pre-owned, value propositions, and financing options accordingly.

7. Environmental Factors Summary

- Economic: State GDP, credit availability.
- Geographic: Urban/rural pricing, fuel costs.
- Social: Eco-awareness driving hybrids; ride-share influences.
- Political: Emissions mandates, tax incentives.

8. Industry Structure & Competitors

Market Leaders: Toyota, Honda, Ford, Chevrolet.
Dealership Channels: CarMax, AutoNation vs. regional independents.
Online Platforms: Carvana, Vroom reshaping purchase processes.

9. Strategic Recommendations

9.1 Geographic Focus:
- Primary: CA, TX, FL, NY; Secondary: IL, WA, MA.
9.2 Model & Specs:
- Core ICE: Corolla, Camry, Tacoma, RAV4.
- Hybrids: Prius, RAV4 Hybrid, Highlander Hybrid.
- Mileage ≤100k mi; Condition ≥Good.
9.3 Consumer Trends:
- Hybrid/EV uptake; digital retailing; value transparency.

R code

R CODE

==========

LOAD REQUIRED LIBRARIES

==========

library(tidyverse) # For data wrangling and visualization library(lubridate) # For date formatting and handling library(ggplot2) # For plots library(caret) # For modeling tools library(cluster) # (Loaded for clustering - unused here) library(forecast) # (Loaded for time series - unused here)

==========

IMPORT AND EXPLORE DATA

==========

Read vehicle listing data

car_prices <- read.csv("Downloads/car_prices.csv") (fileEncoding = "UTF-8") str(car_prices) head(car_prices)

==========

DATA CLEANING: MAKE COLUMN

==========

Check make column for inconsistent capitalization

table(car_prices$make)

Standardize capitalization in the make column

Table data shows titled and lowercase duplicates

```
car_prices <- car_prices %>% mutate(make = str_to_title(make))

Confirm changes

table(car_prices$make)

================================

FILTER FOR TOYOTA VEHICLES & CLEAN

================================

Subset Toyota vehicles, parse sale date, and drop any missing values

toyota_data <- car_prices %>% filter(make == "Toyota") %>% mutate(saledate = mdy_hms(saledate, quiet = TRUE)) %>%
drop_na()

Confirm structure

str(toyota_data)

Verify: Any missing values or duplicates?

No duplicates

colSums(is.na(toyota_data)) sum(duplicated(toyota_data))

================================

VALIDATION: CATCH MISSPELLINGS OR DUPLICATES

================================

Check unique values for inconsistencies

sort(unique(toyota_data$make)) sort(unique(toyota_data$model)) sort(unique(toyota_data$trim)) sort(unique(toyota_data$body))
sort(unique(toyota_data$state))

================================

FILTER FOR U.S. STATES ONLY

CANADIAN CODES PRESENT

================================

Remove Canadian provinces using known U.S. state codes

us_states <- c("al", "ak", "az", "ar", "ca", "co", "ct", "de", "fl", "ga", "hi", "id", "il", "in", "ia", "ks", "ky", "la", "me", "md", "ma", "mi", "mn",
"ms", "mo", "mt", "ne", "nv", "nh", "nj", "nm", "ny", "nc", "nd", "oh", "ok", "or", "pa", "ri", "sc", "sd", "tn", "tx", "ut", "vt", "va", "wa", "wv",
"wi", "wy")

toyota_data <- toyota_data %>% filter(state %in% us_states)

Confirm updated list of states

sort(unique(toyota_data$state))

================================

STANDARDIZE TEXT COLUMNS

================================

CSV isnt showing an issue in body column but im getting lower case - upper case issues in R

Clean body column (remove leading/trailing whitespace and title case)

toyota_data$body <- str_to_title(str_trim(toyota_data$body)) sort(unique(toyota_data$body))
```

Clean model column (same as above)

toyota_data$model <- str_to_title(str_trim(toyota_data$model)) sort(unique(toyota_data$model))

==========

VISUALIZATION 1: AVERAGE SELLING PRICE BY STATE

==========

Check sample size per state — some like AL/OK/NM may skew average

toyota_data %>% count(state) %>% arrange(desc(n))

avg_price_state <- toyota_data %>% group_by(state) %>% summarise(avg_price = mean(sellingprice, na.rm = TRUE))

Only include states with at least 30 listings

filtered_avg_price <- toyota_data %>% group_by(state) %>% filter(n() >= 30) %>% summarise(avg_price = mean(sellingprice, na.rm = TRUE))

Plot

ggplot(filtered_avg_price, aes(x = reorder(state, avg_price), y = avg_price)) + geom_col(fill = "steelblue") + coord_flip() + labs(title = "Average Toyota Selling Price by State", x = "State", y = "Average Price ($)") + theme_minimal()

==========

VISUALIZATION 2: TOP 10 TOYOTA MODELS SOLD

==========

Count all model appearances; Verify skews

toyota_data %>% count(model) %>% arrange(desc(n))

Get top 10 models

top_models <- toyota_data %>% count(model, sort = TRUE) %>% top_n(10, n)

Plot

ggplot(top_models, aes(x = reorder(model, n), y = n)) + geom_col(fill = "darkgreen") + coord_flip() + labs(title = "Top 10 Most Sold Toyota Models", x = "Model", y = "Number of Listings") + theme_minimal()

==========

VISUALIZATION 3: LISTINGS BY BODY TYPE

==========

Count by body type; verify skews

toyota_data %>% count(body) %>% arrange(desc(n))

Get counts

toyota_bodytype_counts <- toyota_data %>% count(body, sort = TRUE)

Plot

ggplot(toyota_bodytype_counts, aes(x = reorder(body, n), y = n)) + geom_col(fill = "purple") + coord_flip() + labs(title = "Toyota Listings by Body Type", x = "Body Type", y = "Number of Listings") + theme_minimal()

==========

VISUALIZATION 4: SELLING PRICE VS ODOMETER

==========

Preview extreme odometer values

toyota_data %>% arrange(desc(odometer)) %>% select(model, odometer, sellingprice) %>% head(20)

Summary stats to assess reasonable cutoff

summary(toyota_data$odometer)

Filter for vehicles with ≤ 300,000 miles (removes extreme outliers)

this is still being generous, 3x over the 75th percentile

filtered_price_mileage <- toyota_data %>% filter(odometer <= 300000)

Scatter plot with linear trend line

ggplot(filtered_price_mileage, aes(x = odometer, y = sellingprice)) + geom_point(alpha = 0.3, color = "gray") + geom_smooth(method = "lm", color = "blue", se = FALSE) + labs(title = "Selling Price vs. Odometer (Toyota, ≤ 300,000 miles)", x = "Odometer (miles)", y = "Selling Price ($)") + theme_minimal()

==========

VISUALIZATION 5: SELLING PRICE OVER TIME

==========

Examine how many listings per month

month_skew <- toyota_data %>% mutate(month = floor_date(saledate, "month")) %>% count(month) %>% arrange(n)

month_skew # See which months have small samples

Extreme drops from low samples make wild swings in data

You can check by replacing reliable_months with toyota_data

Filter to keep months with 500+ listings to avoid noise

reliable_months <- toyota_data %>% mutate(month = as.Date(floor_date(saledate, "month"))) %>% group_by(month) %>% filter(n() >= 500) %>% summarise(avg_price = mean(sellingprice, na.rm = TRUE))

Plot trend of average price over time

ggplot(reliable_months, aes(x = month, y = avg_price)) + geom_line(color = "tomato", linewidth = 1) + scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") + labs(title = "Avg Toyota Selling Price Over Time (500+ Listings)", x = "Month", y = "Average Selling Price ($)") + theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Condition Distribution ggplot(toyota_data, aes(x = condition)) + geom_bar(fill = "coral") + labs(title = "Distribution of Vehicle Conditions", x = "Condition", y = "Count") + theme_minimal()

Price Distribution by Body Type

ggplot(toyota_data, aes(x = body, y = sellingprice)) + geom_boxplot(fill = "lightblue") + coord_flip() + labs(title = "Price Distribution by Body Type", x = "Body Type", y = "Selling Price ($)") + theme_minimal()

==========

GDP MERGE + REGRESSION MODELING

==========

Load GDP dataset

state_gdp <- read.csv("Downloads/GDP summary state annual 2014 -2015 rev1.csv") str(state_gdp) head(state_gdp)

Preview economic variables available

unique(state_gdp$description)

Use "Real GDP (chained 2017 dollars)" as the chosen economic indicator

gdp_filtered <- state_gdp %>% filter(description == " Real GDP (millions of chained 2017 dollars) 1/") %>% select(state, gdp_2015 = X2015)

Merge GDP values into the main Toyota data

toyota_gdp_merge <- toyota_data %>% left_join(gdp_filtered, by = "state")

Confirm GDP merged successfully

summary(toyota_gdp_merge$gdp_2015) str(toyota_gdp_merge) head(toyota_gdp_merge)
sum(is.na(toyota_gdp_merge$gdp_2015))

Build linear regression model to understand price drivers

model <- lm(sellingprice ~ year + odometer + condition + gdp_2015, data = toyota_gdp_merge)

View regression results

summary(model)

==========

TIME SERIES FORECASTING: 12-MONTH OUTLOOK

==========

Step 1: Prepare monthly average price data

monthly_avg_price <- toyota_data %>% mutate(month = floor_date(saledate, "month")) %>% group_by(month) %>%
summarise(avg_price = mean(sellingprice, na.rm = TRUE)) %>% arrange(month)

Step 2: Convert to time series object

price_ts <- ts(monthly_avg_price$avg_price, start = c(year(min(monthly_avg_price$month)),
month(min(monthly_avg_price$month))), frequency = 12)

Step 3: Fit ARIMA model

fit_arima <- auto.arima(price_ts)

Step 4: Forecast for the next 12 months

forecast_12mo <- forecast(fit_arima, h = 12)

Step 5: Plot the forecast

autoplot(forecast_12mo) + labs(title = "12-Month Forecast of Average Toyota Selling Prices", x = "Year", y = "Average Selling Price ($)") + theme_minimal()

summary(fit_arima)

==========

CLUSTERING DATA

==========

clustering_data <- toyota_data %>% select(year, odometer, sellingprice) %>% drop_na() %>% scale() set.seed(123) k_clusters <- kmeans(clustering_data, centers = 3, nstart = 25)

toyota_data$cluster <- as.factor(k_clusters$cluster) fviz_cluster(k_clusters, data = clustering_data, geom = "point", ellipse.type = "norm", palette = "jco", ggtheme = theme_minimal()) + labs(title = "K-Means Clustering of Toyota Listings")

ggplot(toyota_data, aes(x = cluster, fill = body)) + geom_bar(position = "fill") + labs(title = "Cluster Composition by Body Type", x = "Cluster", y = "Proportion") + theme_minimal()

R Output

R 4.3.2 · ~/

```
> # ================================
> # LOAD REQUIRED LIBRARIES
> # ================================
>
> library(tidyverse)   # For data wrangling and visualization
> library(lubridate)   # For date formatting and handling
> library(ggplot2)     # For plots
> library(caret)       # For modeling tools
> library(cluster)     # (Loaded for clustering - unused here)
> library(forecast)    # (Loaded for time series - unused here)
> # Read vehicle listing data
> car_prices <- read.csv("Downloads/car_prices.csv")
> (fileEncoding = "UTF-8")
[1] "UTF-8"
> str(car_prices)
'data.frame':   558837 obs. of  16 variables:
 $ year        : int  2015 2015 2014 2015 2014 2015 2014 2014 2014 2014 ...
 $ make        : chr  "Kia" "Kia" "BMW" "Volvo" ...
 $ model       : chr  "Sorento" "Sorento" "3 Series" "S60" ...
 $ trim        : chr  "LX" "LX" "328i SULEV" "T5" ...
 $ body        : chr  "SUV" "SUV" "Sedan" "Sedan" ...
 $ transmission: chr  "automatic" "automatic" "automatic" "automatic" ...
 $ vin         : chr  "5xyktca69fg566472" "5xyktca69fg561319" "wba3c1c51ek116351" "yv1612tb4f1310
987" ...
 $ state       : chr  "ca" "ca" "ca" "ca" ...
 $ condition   : int  5 5 45 41 43 1 34 2 42 3 ...
 $ odometer    : int  16639 9393 1331 14282 2641 5554 14943 28617 9557 4809 ...
 $ color       : chr  "white" "white" "gray" "white" ...
 $ interior    : chr  "black" "beige" "black" "black" ...
 $ seller      : chr  "kia motors america  inc" "kia motors america  inc" "financial services rem
arketing (lease)" "volvo na rep/world omni" ...
 $ mmr         : int  20500 20800 31900 27500 66000 15350 69000 11900 32100 26300 ...
 $ sellingprice: int  21500 21500 30000 27750 67000 10900 65000 9800 32250 17500 ...
 $ saledate    : chr  "Tue Dec 16 2014 12:30:00 GMT-0800 (PST)" "Tue Dec 16 2014 12:30:00 GMT-080
0 (PST)" "Thu Jan 15 2015 04:30:00 GMT-0800 (PST)" "Thu Jan 29 2015 04:30:00 GMT-0800 (PST)" ...
> head(car_prices)
```

```
  year   make               model       trim  body transmission              vin state
1 2015    Kia             Sorento         LX   SUV    automatic 5xyktca69fg566472    ca
2 2015    Kia             Sorento         LX   SUV    automatic 5xyktca69fg561319    ca
3 2014    BMW            3 Series 328i SULEV Sedan    automatic wba3c1c51ek116351    ca
4 2015  Volvo                 S60         T5 Sedan    automatic yv1612tb4f1310987    ca
5 2014    BMW 6 Series Gran Coupe       650i Sedan    automatic wba6b2c57ed129731    ca
6 2015 Nissan              Altima      2.5 S Sedan    automatic 1n4al3ap1fn326013    ca
  condition odometer color interior                                      seller
1         5    16639 white    black                           kia motors america  inc
2         5     9393 white    beige                           kia motors america  inc
3        45     1331  gray    black             financial services remarketing (lease)
4        41    14282 white    black                           volvo na rep/world omni
5        43     2641  gray    black             financial services remarketing (lease)
6         1     5554  gray    black enterprise vehicle exchange / tra / rental / tulsa
    mmr sellingprice                           saledate
1 20500        21500 Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
2 20800        21500 Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
3 31900        30000 Thu Jan 15 2015 04:30:00 GMT-0800 (PST)
4 27500        27750 Thu Jan 29 2015 04:30:00 GMT-0800 (PST)
5 66000        67000 Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
6 15350        10900 Tue Dec 30 2014 12:00:00 GMT-0800 (PST)
> # ================================
> # DATA CLEANING: MAKE COLUMN
> # ================================
> # Check `make` column for inconsistent capitalization
> table(car_prices$make)

                    acura         Acura     airstream  Aston Martin          audi
        10301           25          5901             1            25             8
         Audi       Bentley           bmw           BMW         buick         Buick
         5869           116            74         20719            14          5107
     cadillac      Cadillac    chev truck     chevrolet     Chevrolet      chrysler
          110          7519             1           390         60197           209
     Chrysler       Daewoo         dodge         Dodge      dodge tk           dot
        17276             3           245         30710             1             1
      Ferrari          FIAT        Fisker          ford          Ford       ford tk
           19           865             9           443         93554             1
   ford truck           Geo           gmc           GMC     gmc truck         honda
            3            19            25         10613            11           145
        Honda       HUMMER       hyundai       Hyundai    hyundai tk      Infiniti
        27206           805            20         21816             1         15305
        Isuzu       Jaguar          jeep          Jeep           kia           Kia
          204          1420           111         15372             7         18077
  Lamborghini    land rover    Land Rover     landrover         lexus         Lexus
            4          129          1735            27           119         11861
      lincoln       Lincoln         Lotus      maserati      Maserati         mazda
           29          5757             1             3           133           146
       Mazda      mazda tk      mercedes    mercedes-b Mercedes-Benz       mercury
         8362             1            70             2         17141            31
     Mercury          MINI    mitsubishi    Mitsubishi        nissan        Nissan
        1992          3224           117          4140            71         53946
   oldsmobile    Oldsmobile      plymouth      Plymouth       pontiac       Pontiac
           20           364             7            20            27          4497
      porsche       Porsche           Ram   Rolls-Royce          Saab        Saturn
           19          1383          4574            17           484          2841
        Scion         smart        subaru        Subaru        suzuki        Suzuki
         1687           396            60          5043             5          1073
        Tesla        toyota        Toyota    volkswagen    Volkswagen         Volvo
           23            95         39871            24         12581          3788
           vw
           24
> # Standardize capitalization in the `make` column
> # Table data shows titled and lowercase duplicates
> car_prices <- car_prices %>%
+   mutate(make = str_to_title(make))
> # Confirm changes
> table(car_prices$make)

                   Acura     Airstream  Aston Martin          Audi       Bentley
        10301         5926             1            25          5877           116
          Bmw        Buick      Cadillac    Chev Truck     Chevrolet      Chrysler
        20793         5121          7629             1         60587         17485
```

```
      Daewoo         Dodge      Dodge Tk           Dot       Ferrari          Fiat
          3         30955             1             1            19           865
     Fisker          Ford       Ford Tk     Ford Truck           Geo           Gmc
          9         93997             1             3            19         10638
  Gmc Truck         Honda        Hummer       Hyundai    Hyundai Tk      Infiniti
         11         27351           805         21836             1         15305
      Isuzu        Jaguar          Jeep           Kia   Lamborghini    Land Rover
        204          1420         15483         18084             4          1864
  Landrover         Lexus       Lincoln         Lotus      Maserati         Mazda
         27         11980          5786             1           136          8508
   Mazda Tk      Mercedes    Mercedes-B Mercedes-Benz       Mercury          Mini
          1            70             2         17141          2023          3224
 Mitsubishi        Nissan    Oldsmobile      Plymouth       Pontiac       Porsche
       4257         54017           384            27          4524          1402
        Ram    Rolls-Royce          Saab        Saturn         Scion         Smart
       4574           17           484          2841          1687           396
     Subaru         Suzuki         Tesla        Toyota    Volkswagen         Volvo
       5103          1078            23         39966         12605          3788

         Vw
         24
> # ==============================
> # FILTER FOR TOYOTA VEHICLES & CLEAN
> # ==============================
> # Subset Toyota vehicles, parse `saledate`, and drop any missing values
> toyota_data <- car_prices %>%
+   filter(make == "Toyota") %>%
+   mutate(saledate = mdy_hms(saledate, quiet = TRUE)) %>%
+   drop_na()
> # Confirm structure
> str(toyota_data)
'data.frame':   39077 obs. of  16 variables:
 $ year        : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
 $ make        : chr  "Toyota" "Toyota" "Toyota" "Toyota" ...
 $ model       : chr  "Corolla" "Sienna" "Yaris" "Corolla" ...
 $ trim        : chr  "LE" "LE 7-Passenger Mobility Auto Access" "L" "S" ...
 $ body        : chr  "Sedan" "Minivan" "Hatchback" "Sedan" ...
```

```
 $ transmission: chr  "automatic" "automatic" "automatic" "automatic" ...
 $ vin         : chr  "5yfbu4ee5dp216754" "5tdkk3dc4ds344460" "jtdktud33dd552876" "2t1bu4ee4dc089
995" ...
 $ state       : chr  "ca" "ca" "ca" "ca" ...
 $ condition   : int  1 35 2 47 38 5 28 37 1 33 ...
 $ odometer    : int  34915 49188 28412 19836 38882 2575 32198 36431 48744 36159 ...
 $ color       : chr  "black" "gray" "white" "orange" ...
 $ interior    : chr  "gray" "tan" "gray" "black" ...
 $ seller      : chr  "enterprise vehicle exchange / tra / rental / tulsa" "the hertz corporatio
n" "enterprise vehicle exchange / tra / rental / tulsa" "consolidated asset recovery systems  in
c." ...
 $ mmr         : int  11400 17150 9775 13500 13200 21800 21200 25200 10600 18850 ...
 $ sellingprice: int  9200 18000 7800 14100 13500 22000 20500 24200 9500 19500 ...
 $ saledate    : POSIXct, format: "2014-12-16 20:00:00" "2014-12-16 20:30:00" ...
> # Verify: Any missing values or duplicates?
> # No duplicates
> colSums(is.na(toyota_data))
       year        make       model        trim        body transmission         vin
          0           0           0           0           0           0           0
      state   condition     odometer       color    interior      seller         mmr
          0           0           0           0           0           0           0
sellingprice    saledate
          0           0
> sum(duplicated(toyota_data))
[1] 0
> # ==============================
> # VALIDATION: CATCH MISSPELLINGS OR DUPLICATES
> # ==============================
> # Check unique values for inconsistencies
> sort(unique(toyota_data$make))
[1] "Toyota"
> sort(unique(toyota_data$model))
 [1] "4Runner"           "avalon"            "Avalon"            "Avalon Hybrid"
 [5] "camry"             "Camry"             "Camry Hybrid"      "Camry Solara"
 [9] "Celica"            "corolla"           "Corolla"           "ECHO"
[13] "FJ Cruiser"        "Highlander"        "Highlander Hybrid" "Land Cruiser"
[17] "matrix"            "Matrix"            "MR2 Spyder"        "Paseo"
[21] "pickup"            "Pickup"            "previa"            "Prius"
[25] "Prius c"           "Prius Plug-in"     "Prius v"           "RAV4"
[29] "Sequoia"           "sienna"            "Sienna"            "T100"
[33] "Tacoma"            "Tercel"            "tundra"            "Tundra"
[37] "Venza"             "yaris"             "Yaris"
> sort(unique(toyota_data$trim))
  [1] ""                                "1794"
  [3] "1794 FFV"                        "4c le"
  [5] "4x2 cr limited"                  "4x2 v6 sr5"
  [7] "4x2 v8 limited"                  "4x2 v8 sr5"
  [9] "4x2 v8 x-sp"                     "4x4"
 [11] "4x4 v8 limited"                  "4x4 v8 sr5"
 [13] "Advanced"                        "base"
 [15] "Base"                            "Base 7-Passenger"
 [17] "ce"                              "CE"
 [19] "CE 7-Passenger"                  "Deluxe"
 [21] "dx"                              "DX"
 [23] "DX V6"                           "Five"
 [25] "Fleet"                           "GT"
 [27] "GT-S"                            "GTS"
 [29] "I"                               "II"
 [31] "III"                             "IV"
 [33] "L"                               "L 7-Passenger"
 [35] "le"                              "LE"
 [37] "LE 7-Passenger"                  "LE 7-Passenger Mobility Auto Access"
 [39] "LE 8-Passenger"                  "LE Eco"
```

```
 [41] "LE Plus"                          "LE V6"
 [43] "Limited"                          "Limited 7-Passenger"
 [45] "Limited FFV"                      "One"
 [47] "Platinum"                         "Platinum FFV"
 [49] "Plus"                             "Prerunner"
 [51] "PreRunner"                        "PreRunner V6"
 [53] "s"                                "S"
 [55] "S Plus"                           "S Premium"
 [57] "S Special Edition"                "SE"
 [59] "SE 8-Passenger"                   "SE Sport"
 [61] "SE V6"                            "SLE"
 [63] "SLE V6"                           "Sport"
 [65] "Sport Edition"                    "Sport V6"
 [67] "SR"                               "SR FFV"
 [69] "SR5"                              "SR5 FFV"
 [71] "SR5 V6"                           "ST"
 [73] "Three"                            "Touring"
 [75] "Trail"                            "TRD PRO"
 [77] "Tundra"                           "Tundra FFV"
 [79] "Tundra Grade"                     "Two"
 [81] "V"                                "V6"
 [83] "VE"                               "X-Runner V6"
 [85] "xl"                               "XL"
 [87] "xle"                              "XLE"
 [89] "XLE 7-Passenger"                  "XLE 8-Passenger"
 [91] "XLE Limited 7-Passenger"          "XLE Premium"
 [93] "XLE Touring"                      "XLE Touring SE"
 [95] "XLE V6"                           "XLS"
 [97] "xr"                               "XR"
 [99] "XRS"                              "XSE"
> sort(unique(toyota_data$body))
 [1] ""             "access cab"   "Access Cab"   "convertible"  "Convertible"
 [6] "coupe"        "Coupe"        "crewmax cab"  "CrewMax Cab"  "double cab"
[11] "Double Cab"   "extended cab" "Extended Cab" "hatchback"    "Hatchback"
[16] "minivan"      "Minivan"      "regular cab"  "Regular Cab"  "sedan"
[21] "Sedan"        "suv"          "SUV"          "wagon"        "Wagon"
[26] "xtracab"      "Xtracab"
> sort(unique(toyota_data$state))
 [1] "ab" "al" "az" "ca" "co" "fl" "ga" "hi" "il" "in" "la" "ma" "md" "mi" "mn" "mo" "ms"
[18] "nc" "ne" "nj" "nm" "nv" "ny" "oh" "ok" "on" "or" "pa" "pr" "qc" "sc" "tn" "tx" "ut"
[35] "va" "wa" "wi"
> # ================================
> # FILTER FOR U.S. STATES ONLY
> # CANADIAN CODES PRESENT
> # ================================
```

```
> # Remove Canadian provinces using known U.S. state codes
> us_states <- c("al", "ak", "az", "ar", "ca", "co", "ct", "de", "fl", "ga", "hi", "id",
+                "il", "in", "ia", "ks", "ky", "la", "me", "md", "ma", "mi", "mn", "ms",
+                "mo", "mt", "ne", "nv", "nh", "nj", "nm", "ny", "nc", "nd", "oh", "ok",
+                "or", "pa", "ri", "sc", "sd", "tn", "tx", "ut", "vt", "va", "wa", "wv",
+                "wi", "wy")
> toyota_data <- toyota_data %>%
+   filter(state %in% us_states)
> # Confirm updated list of states
> sort(unique(toyota_data$state))
 [1] "al" "az" "ca" "co" "fl" "ga" "hi" "il" "in" "la" "ma" "md" "mi" "mn" "mo" "ms" "nc"
[18] "ne" "nj" "nm" "nv" "ny" "oh" "ok" "or" "pa" "sc" "tn" "tx" "ut" "va" "wa" "wi"
> # ================================
> # STANDARDIZE TEXT COLUMNS
> # ================================
> # CSV isnt showing an issue in body column but im getting lower case - upper case issues in R
> # Clean `body` column (remove leading/trailing whitespace and title case)
> toyota_data$body <- str_to_title(str_trim(toyota_data$body))
> sort(unique(toyota_data$body))
 [1] ""            "Access Cab"   "Convertible"  "Coupe"        "Crewmax Cab"
 [6] "Double Cab"  "Extended Cab" "Hatchback"    "Minivan"      "Regular Cab"
[11] "Sedan"       "Suv"          "Wagon"        "Xtracab"
> # Clean `model` column (same as above)
> toyota_data$model <- str_to_title(str_trim(toyota_data$model))
> sort(unique(toyota_data$model))
 [1] "4runner"          "Avalon"           "Avalon Hybrid"    "Camry"
 [5] "Camry Hybrid"     "Camry Solara"     "Celica"           "Corolla"
 [9] "Echo"             "Fj Cruiser"        "Highlander"       "Highlander Hybrid"
[13] "Land Cruiser"     "Matrix"           "Mr2 Spyder"       "Paseo"
[17] "Pickup"           "Previa"           "Prius"            "Prius C"
[21] "Prius Plug-In"    "Prius V"          "Rav4"             "Sequoia"
[25] "Sienna"           "T100"             "Tacoma"           "Tercel"
[29] "Tundra"           "Venza"            "Yaris"
> # ================================
> # VISUALIZATION 1: AVERAGE SELLING PRICE BY STATE
> # ================================
> # Check sample size per state — some like AL/OK/NM may skew average
> toyota_data %>% count(state) %>% arrange(desc(n))
   state    n
1     fl 8613
2     ca 4527
3     pa 3046
4     ga 2944
5     oh 2426
6     tx 2016
7     nc 1753
8     nj 1731
9     md 1201
10    tn  946
11    mo  906
12    nv  877
13    az  869
14    ma  863
15    wi  829
16    va  798
17    wa  731
18    co  568
19    mn  553
20    il  523
21    sc  304
22    ne  250
23    mi  245
24    in  220
25    ny  200
26    ms  191
27    la  115
28    ut   74
29    or   63
30    hi   51
31    nm    9
32    ok    2
33    al    1
> avg_price_state <- toyota_data %>%
+   group_by(state) %>%
+   summarise(avg_price = mean(sellingprice, na.rm = TRUE))
> # Only include states with at least 30 listings
> filtered_avg_price <- toyota_data %>%
+   group_by(state) %>%
+   filter(n() >= 30) %>%
+   summarise(avg_price = mean(sellingprice, na.rm = TRUE))
> # Plot
> ggplot(filtered_avg_price, aes(x = reorder(state, avg_price), y = avg_price)) +
+   geom_col(fill = "steelblue") +
+   coord_flip() +
+   labs(
+     title = "Average Toyota Selling Price by State",
+     x = "State",
+     y = "Average Price ($)"
+   ) +
+   theme_minimal()
> # ================================
```

```
> # VISUALIZATION 2: TOP 10 TOYOTA MODELS SOLD
> # ================================
> # Count all model appearances; Verify skews
> toyota_data %>% count(model) %>% arrange(desc(n))
              model     n
1             Camry 12287
2           Corolla  7094
3              Rav4  3140
4            Sienna  2736
5        Highlander  2129
6             Prius  1795
7            Tundra  1312
8           4runner  1277
9            Avalon  1096
10           Tacoma  1080
11            Yaris   959
12      Camry Hybrid   587
13          Sequoia   569
14            Venza   554
15       Fj Cruiser   366
16      Camry Solara   351
17           Matrix   331
18 Highlander Hybrid   197
19          Prius C   108
20      Land Cruiser   105
21          Prius V   102
22           Celica    90
23             Echo    46
24     Prius Plug-In    45
25     Avalon Hybrid    40
26           Tercel    18
27        Mr2 Spyder    15
28           Pickup     8
29             T100     4
30           Previa     3
31            Paseo     1
>
> # Get top 10 models
> top_models <- toyota_data %>%
+   count(model, sort = TRUE) %>%
+   top_n(10, n)
> # Plot
> ggplot(top_models, aes(x = reorder(model, n), y = n)) +
+   geom_col(fill = "darkgreen") +
+   coord_flip() +
+   labs(
+     title = "Top 10 Most Sold Toyota Models",
+     x = "Model",
+     y = "Number of Listings"
+   ) +
+   theme_minimal()
> # ================================
> # VISUALIZATION 3: LISTINGS BY BODY TYPE
> # ================================
> # Count by body type; verify skews
> toyota_data %>% count(body) %>% arrange(desc(n))
          body     n
1        Sedan 21503
2          Suv  7783
3    Hatchback  2764
4      Minivan  2716
5   Double Cab  1290
6        Wagon   824
7  Crewmax Cab   548
8   Access Cab   283
9        Coupe   261
10 Regular Cab   162
11 Convertible   151
12               75
13 Extended Cab    43
14      Xtracab    42
> # Get counts
> toyota_bodytype_counts <- toyota_data %>%
+   count(body, sort = TRUE)
> # Plot
> ggplot(toyota_bodytype_counts, aes(x = reorder(body, n), y = n)) +
+   geom_col(fill = "purple") +
+   coord_flip() +
+   labs(
+     title = "Toyota Listings by Body Type",
+     x = "Body Type",
+     y = "Number of Listings"
+   ) +
+   theme_minimal()
> # ================================
> # VISUALIZATION 4: SELLING PRICE VS ODOMETER
> # ================================
```

```
> # Preview extreme odometer values
> toyota_data %>%
+   arrange(desc(odometer)) %>%
+   select(model, odometer, sellingprice) %>%
+   head(20)
          model odometer sellingprice
1         Camry   999999         3300
2       Corolla   959276          500
3    Highlander   621388         3100
4         Prius   537334         1700
5        Sienna   443236         4800
6       Corolla   380842         6000
7        Pickup   379307          900
8        Pickup   379069          800
9        Tundra   376426         1600
10      Prius V   369530         9000
11     4runner   367750          700
12   Highlander   366234         3400
13        Camry   365286         2100
14        Camry   360680          800
15       Avalon   357751         1100
16     4runner   354003         3700
17        Camry   351919         1200
18        Camry   351917          900
19        Camry   348816         3100
20 Camry Solara   347824          700
> # Summary stats to assess reasonable cutoff
> summary(toyota_data$odometer)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1   30571   53656   73103  104396  999999
> # Filter for vehicles with ≤ 300,000 miles (removes extreme outliers)
> # this is still being generous, 3x over the 75th percentile
> filtered_price_mileage <- toyota_data %>%
+   filter(odometer <= 300000)
> # Scatter plot with linear trend line
> ggplot(filtered_price_mileage, aes(x = odometer, y = sellingprice)) +
+   geom_point(alpha = 0.3, color = "gray") +
+   geom_smooth(method = "lm", color = "blue", se = FALSE) +
+   labs(
+     title = "Selling Price vs. Odometer (Toyota, ≤ 300,000 miles)",
+     x = "Odometer (miles)",
+     y = "Selling Price ($)"
+   ) +
+   theme_minimal()
`geom_smooth()` using formula = 'y ~ x'
> # ==============================
```

```
> # ================================
> # VISUALIZATION 5: SELLING PRICE OVER TIME
> # ================================
> # Examine how many listings per month
> month_skew <- toyota_data %>%
+   mutate(month = floor_date(saledate, "month")) %>%
+   count(month) %>%
+   arrange(n)
> month_skew  # See which months have small samples
       month     n
1 2014-01-01     6
2 2015-07-01    84
3 2015-04-01    89
4 2014-12-01  3139
5 2015-03-01  3329
6 2015-05-01  3612
7 2015-06-01  7662
8 2015-01-01  9607
9 2015-02-01 10917
> # Extreme drops from low samples make wild swings in data
> # You can check by replacing reliable_months with toyota_data
> 
> # Filter to keep months with 500+ listings to avoid noise
> reliable_months <- toyota_data %>%
+   mutate(month = as.Date(floor_date(saledate, "month"))) %>%
+   group_by(month) %>%
+   filter(n() >= 500) %>%
+   summarise(avg_price = mean(sellingprice, na.rm = TRUE))
> # Plot trend of average price over time
> ggplot(reliable_months, aes(x = month, y = avg_price)) +
+   geom_line(color = "tomato", linewidth = 1) +
+   scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
+   labs(
+     title = "Avg Toyota Selling Price Over Time (500+ Listings)",
+     x = "Month",
+     y = "Average Selling Price ($)"
+   ) +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
> #Condition Distribution
> ggplot(toyota_data, aes(x = condition)) +
+   geom_bar(fill = "coral") +
+   labs(title = "Distribution of Vehicle Conditions", x = "Condition", y = "Count") +
+   theme_minimal()
> # Price Distribution by Dody Type
> ggplot(toyota_data, aes(x = body, y = sellingprice)) +
+   geom_boxplot(fill = "lightblue") +
+   coord_flip() +
+   labs(title = "Price Distribution by Body Type", x = "Body Type", y = "Selling Price ($)") +
+   theme_minimal()
> # ================================
> # GDP MERGE + REGRESSION MODELING
> # ================================
> # Load GDP dataset
> state_gdp <- read.csv("Downloads/GDP summary state annual 2014 -2015 rev1.csv")
> str(state_gdp)
'data.frame':   765 obs. of  6 variables:
 $ state      : chr  "al" "al" "al" "al" ...
 $ Region     : int  5 5 5 5 5 5 5 5 5 5 ...
 $ description: chr  " Real GDP (millions of chained 2017 dollars) 1/" " Real personal income (
llions of constant (2017) dollars) 2/" " Real PCE (millions of constant (2017) dollars) 3/" " (
ss domestic product (GDP) " ...
 $ unit       : chr  "Millions of chained 2017 dollars" "Millions of constant 2017 dollars" "Mi
ions of constant 2017 dollars" "Millions of current dollars" ...
 $ X2014      : num  206070 205292 169502 197064 179487 ...
 $ X2015      : num  208950 215148 175051 203113 187475 ...
> head(state_gdp)
  state Region                                                description
1    al      5                        Real GDP (millions of chained 2017 dollars) 1/
2    al      5  Real personal income (millions of constant (2017) dollars) 2/
3    al      5                     Real PCE (millions of constant (2017) dollars) 3/
4    al      5                                 Gross domestic product (GDP)
5    al      5                                                Personal income
6    al      5                                  Disposable personal income
                              unit    X2014    X2015
1  Millions of chained 2017 dollars 206070.0 208950.3
2 Millions of constant 2017 dollars 205291.8 215147.9
3 Millions of constant 2017 dollars 169501.8 175050.6
4       Millions of current dollars 197064.4 203113.3
5       Millions of current dollars 179487.1 187474.7
6       Millions of current dollars 162978.1 169520.5
> # Preview economic variables available
> unique(state_gdp$description)
 [1] " Real GDP (millions of chained 2017 dollars) 1/"
 [2] " Real personal income (millions of constant (2017) dollars) 2/"
 [3] " Real PCE (millions of constant (2017) dollars) 3/"
 [4] " Gross domestic product (GDP) "
 [5] " Personal income "
 [6] " Disposable personal income "
 [7] " Personal consumption expenditures "
 [8] " Real per capita personal income 4/"
```

```
 [8] " Real per capita personal income 4/"
 [9] " Real per capita PCE 5/"
[10] " Per capita personal income 6/"
[11] " Per capita disposable personal income 7/"
[12] " Per capita personal consumption expenditures (PCE) 8/"
[13] " Regional price parities (RPPs) 9/"
[14] " Implicit regional price deflator 10/"
[15] " Total employment (number of jobs) "
> # Use "Real GDP (chained 2017 dollars)" as the chosen economic indicator
> gdp_filtered <- state_gdp %>%
+   filter(description == " Real GDP (millions of chained 2017 dollars) 1/") %>%
+   select(state, gdp_2015 = X2015)
> # Merge GDP values into the main Toyota data
> toyota_gdp_merge <- toyota_data %>%
+   left_join(gdp_filtered, by = "state")
> # Confirm GDP merged successfully
> summary(toyota_gdp_merge$gdp_2015)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  83858  498944  633275  901379  945929 2545980
> str(toyota_gdp_merge)
'data.frame':   38445 obs. of  17 variables:
 $ year        : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
 $ make        : chr  "Toyota" "Toyota" "Toyota" "Toyota" ...
 $ model       : chr  "Corolla" "Sienna" "Yaris" "Corolla" ...
 $ trim        : chr  "LE" "LE 7-Passenger Mobility Auto Access" "L" "S" ...
 $ body        : chr  "Sedan" "Minivan" "Hatchback" "Sedan" ...
 $ transmission: chr  "automatic" "automatic" "automatic" "automatic" ...
 $ vin         : chr  "5yfbu4ee5dp216754" "5tdkk3dc4ds344460" "jtdktud33dd552876" "2t1bu4ee4dc089
995" ...
 $ state       : chr  "ca" "ca" "ca" "ca" ...
 $ condition   : int  1 35 2 47 38 5 28 37 1 33 ...
 $ odometer    : int  34915 49188 28412 19836 38882 2575 32198 36431 48744 36159 ...
 $ color       : chr  "black" "gray" "white" "orange" ...
 $ interior    : chr  "gray" "tan" "gray" "black" ...
 $ seller      : chr  "enterprise vehicle exchange / tra / rental / tulsa" "the hertz corporatio
n" "enterprise vehicle exchange / tra / rental / tulsa" "consolidated asset recovery systems  in
c." ...
 $ mmr         : int  11400 17150 9775 13500 13200 21800 21200 25200 10600 18850 ...
 $ sellingprice: int  9200 18000 7800 14100 13500 22000 20500 24200 9500 19500 ...
 $ saledate    : POSIXct, format: "2014-12-16 20:00:00" "2014-12-16 20:30:00" ...
 $ gdp_2015    : num  2545980 2545980 2545980 2545980 2545980 ...
> head(toyota_gdp_merge)
  year   make   model                                trim      body transmission
1 2013 Toyota Corolla                                  LE     Sedan    automatic
2 2013 Toyota  Sienna LE 7-Passenger Mobility Auto Access   Minivan    automatic
3 2013 Toyota   Yaris                                   L Hatchback    automatic
```

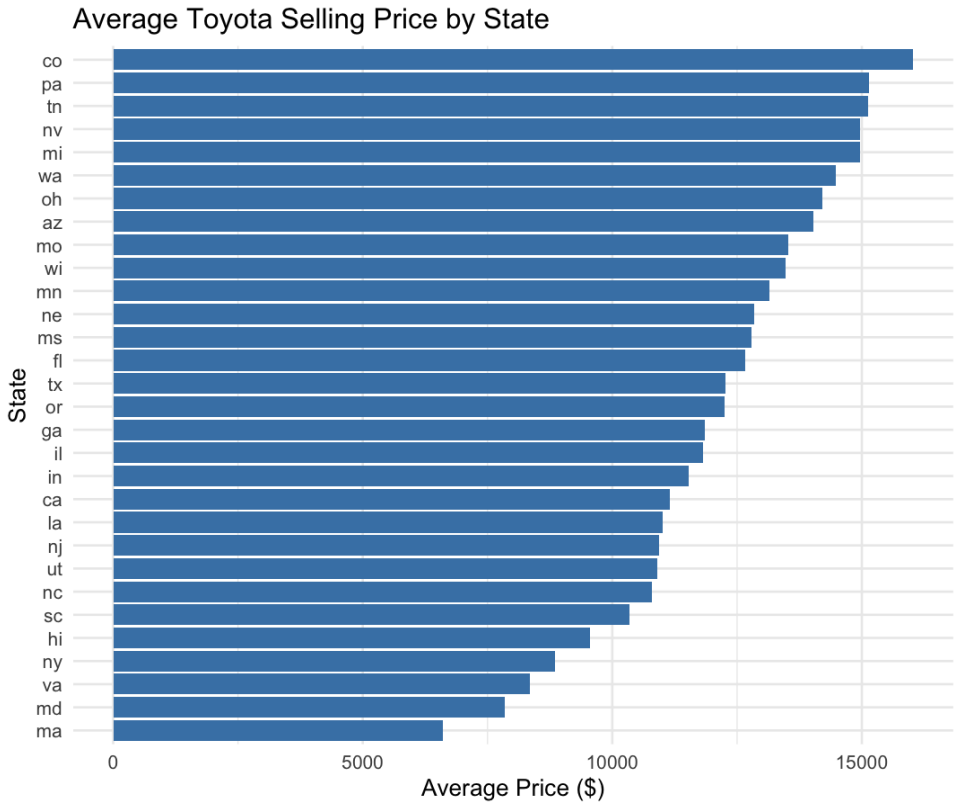
```
3 2013 Toyota   Yaris                                  L Hatchback    automatic
4 2013 Toyota Corolla                                  S     Sedan    automatic
5 2013 Toyota   Camry                                 LE     Sedan    automatic
6 2013 Toyota   Camry                                XLE     Sedan    automatic
                vin state condition odometer  color interior
1 5yfbu4ee5dp216754    ca         1    34915  black     gray
2 5tdkk3dc4ds344460    ca        35    49188   gray      tan
3 jtdktud33dd552876    ca         2    28412  white     gray
4 2t1bu4ee4dc089995    ca        47    19836 orange    black
5 4t4bf1fk5dr275482    ca        38    38882    red    beige
6 4t1bk1fk0du529833    ca         5     2575  black     gray
                                                seller   mmr sellingprice
1 enterprise vehicle exchange / tra / rental / tulsa 11400         9200
2                             the hertz corporation 17150        18000
3 enterprise vehicle exchange / tra / rental / tulsa  9775         7800
4          consolidated asset recovery systems  inc. 13500        14100
5                                  avis corporation 13200        13500
6                                  the car exchange 21800        22000
            saledate gdp_2015
1 2014-12-16 20:00:00  2545980
2 2014-12-16 20:30:00  2545980
3 2014-12-30 21:00:00  2545980
4 2014-12-17 20:30:00  2545980
5 2014-12-18 19:30:00  2545980
6 2014-12-17 20:30:00  2545980
> sum(is.na(toyota_gdp_merge$gdp_2015))
[1] 0
> # Build linear regression model to understand price drivers
> model <- lm(sellingprice ~ year + odometer + condition + gdp_2015, data = toyota_gdp_merge)
> # View regression results
> summary(model)

Call:
lm(formula = sellingprice ~ year + odometer + condition + gdp_2015,
    data = toyota_gdp_merge)

Residuals:
   Min     1Q Median     3Q    Max
-14139  -3009  -1355   1560  51675

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.241e+06  2.073e+04 -59.891  < 2e-16 ***
year         6.236e+02  1.030e+01  60.555  < 2e-16 ***
odometer    -3.474e-02  7.280e-04 -47.726  < 2e-16 ***
```

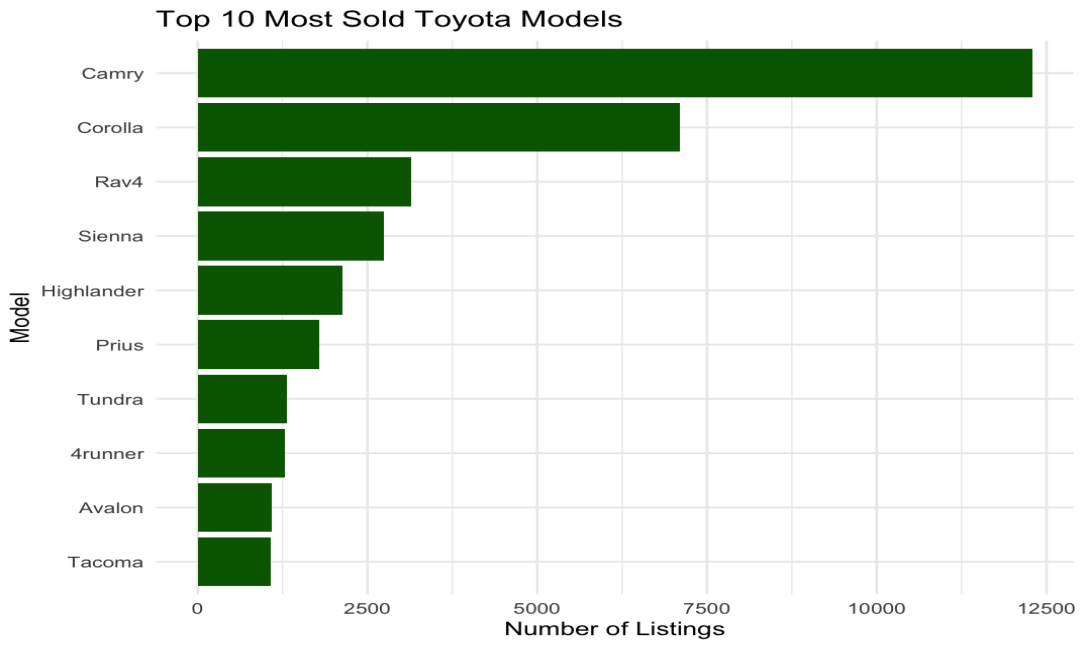
```
condition    9.009e+01  2.211e+00  40.746  < 2e-16 ***
gdp_2015     1.907e-04  3.701e-05   5.153 2.58e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4933 on 38440 degrees of freedom
Multiple R-squared:  0.5022,    Adjusted R-squared:  0.5021
F-statistic:  9695 on 4 and 38440 DF,  p-value: < 2.2e-16

> # ================================
> # TIME SERIES FORECASTING: 12-MONTH OUTLOOK
> # ================================
>
> # Step 1: Prepare monthly average price data
> monthly_avg_price <- toyota_data %>%
+   mutate(month = floor_date(saledate, "month")) %>%
+   group_by(month) %>%
+   summarise(avg_price = mean(sellingprice, na.rm = TRUE)) %>%
+   arrange(month)
> # Step 2: Convert to time series object
> price_ts <- ts(monthly_avg_price$avg_price,
+                start = c(year(min(monthly_avg_price$month)), month(min(monthly_avg_price$mont
h))),
+                frequency = 12)
> # Step 3: Fit ARIMA model
> fit_arima <- auto.arima(price_ts)
> # Step 4: Forecast for the next 12 months
> forecast_12mo <- forecast(fit_arima, h = 12)
> # Step 5: Plot the forecast
> autoplot(forecast_12mo) +
+   labs(
+     title = "12-Month Forecast of Average Toyota Selling Prices",
+     x = "Year",
+     y = "Average Selling Price ($)"
+   ) +
+   theme_minimal()
> summary(fit_arima)
Series: price_ts
ARIMA(0,0,0) with non-zero mean

Coefficients:
            mean
      12570.1838
s.e.    719.2972

sigma^2 = 5238492:  log likelihood = -81.86
```

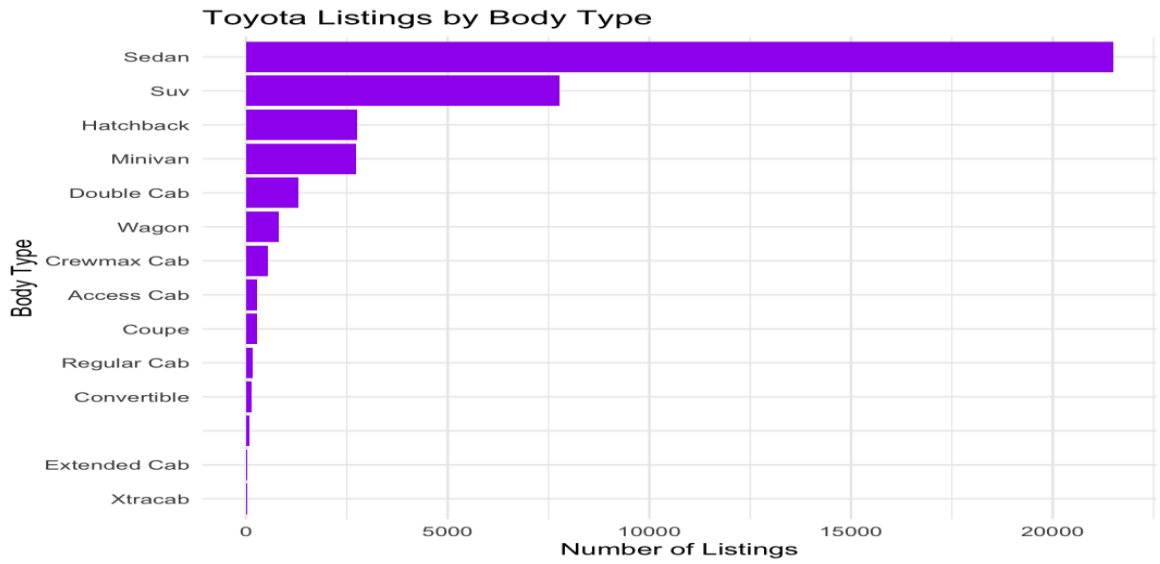
```
sigma^2 = 5238492:  log likelihood = -81.86
AIC=167.72   AICc=169.72   BIC=168.12

Training set error measures:
                      ME     RMSE      MAE       MPE     MAPE MASE       ACF1
Training set -2.021593e-13 2157.878 1408.678 -2.475043 10.68105  NaN 0.08875702
> # ==============================
> # CLUSTERING DATA
> # ==============================
> clustering_data <- toyota_data %>%
+   select(year, odometer, sellingprice) %>%
+   drop_na() %>%
+   scale()
> set.seed(123)
> k_clusters <- kmeans(clustering_data, centers = 3, nstart = 25)
> toyota_data$cluster <- as.factor(k_clusters$cluster)
> fviz_cluster(k_clusters, data = clustering_data,
+              geom = "point", ellipse.type = "norm",
+              palette = "jco", ggtheme = theme_minimal()) +
+   labs(title = "K-Means Clustering of Toyota Listings")
> ggplot(toyota_data, aes(x = cluster, fill = body)) +
+   geom_bar(position = "fill") +
+   labs(title = "Cluster Composition by Body Type", x = "Cluster", y = "Proportion") +
+   theme_minimal()
>
```

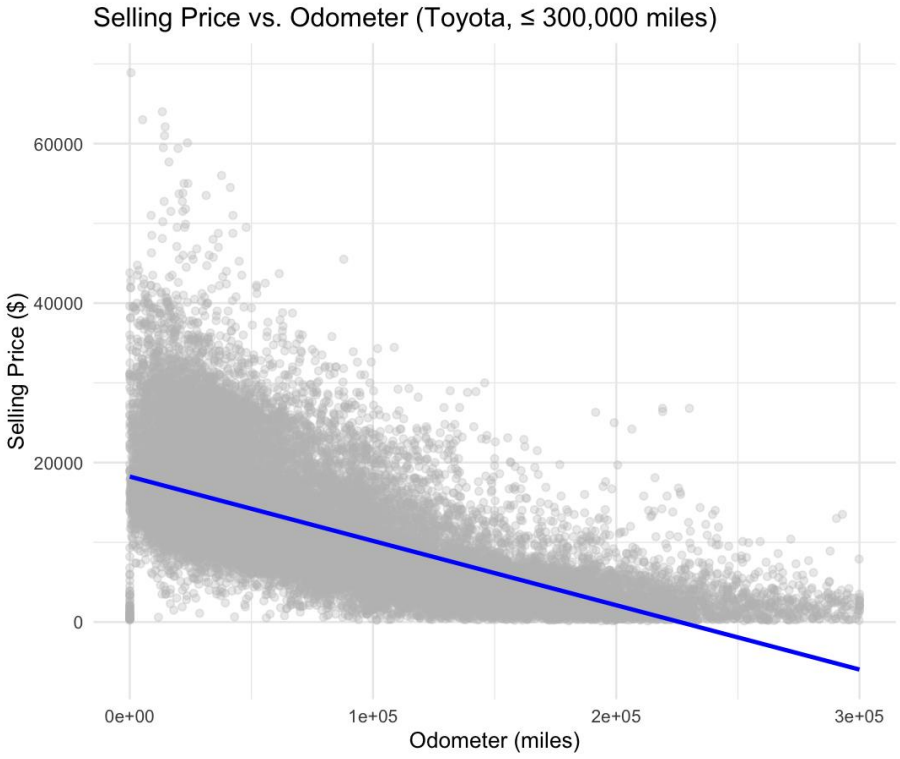
Data	
avg_price_state	33 obs. of 2 variables
car_prices	558837 obs. of 16 variables
clustering_data	Large matrix (115335 elements, 924.5 kB)
filtered_avg_price	30 obs. of 2 variables
filtered_price_mile…	38351 obs. of 16 variables
fit_arima	List of 18
forecast_12mo	List of 10
gdp_filtered	51 obs. of 2 variables
k_clusters	List of 9
model	Large lm (12 elements, 11.1 MB)
month_skew	9 obs. of 2 variables
monthly_avg_price	9 obs. of 2 variables
reliable_months	6 obs. of 2 variables
state_gdp	765 obs. of 6 variables
top_models	10 obs. of 2 variables
toyota_bodytype_cou…	14 obs. of 2 variables
toyota_data	38445 obs. of 17 variables
toyota_gdp_merge	38445 obs. of 17 variables
Values	
fileEncoding	"UTF-8"
price_ts	Time-Series [1:9] from 2014 to 2015: 10717 11930 11830 …
us_states	chr [1:50] "al" "ak" "az" "ar" "ca" "co" "ct" "de" "fl"…

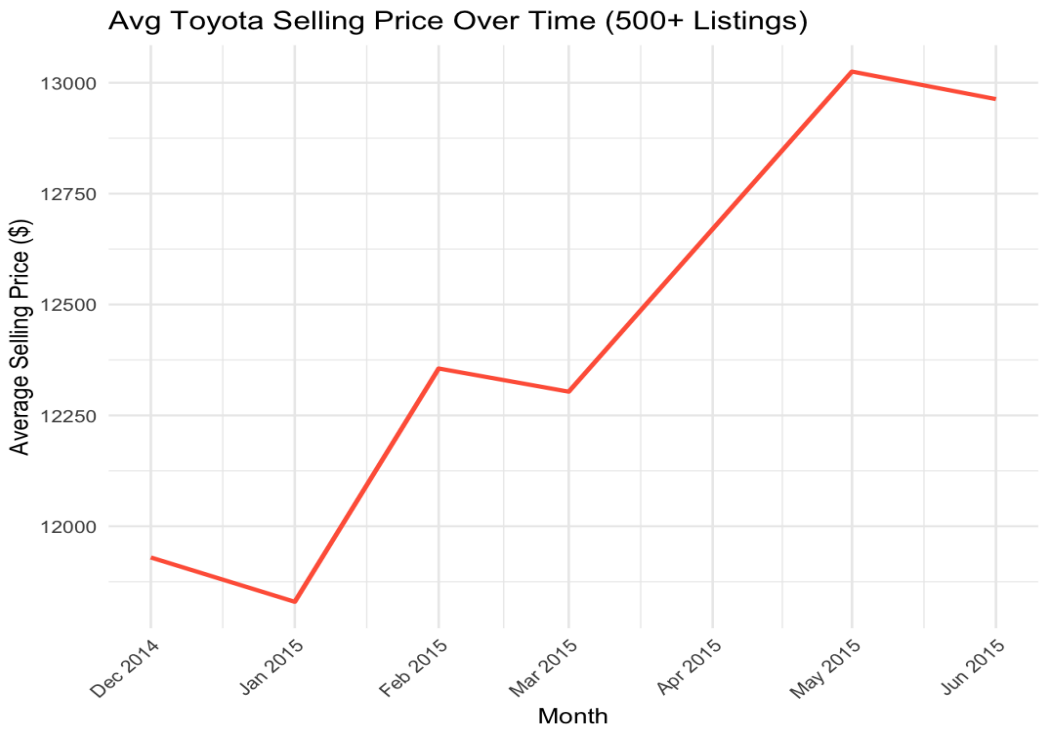
R charts/Plots

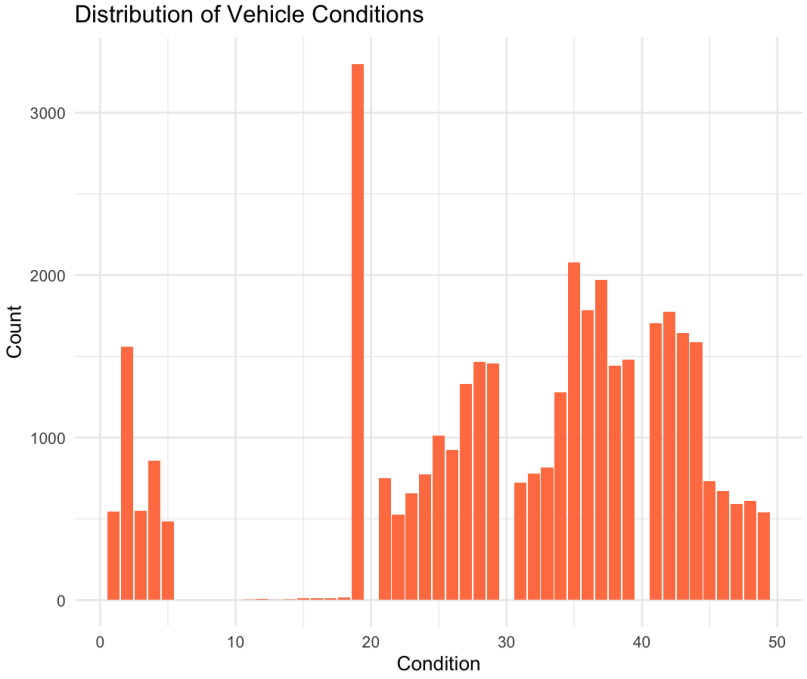

Average Toyota Selling Price by State
State
co
pa
tn
nv
mi
wa
oh
az
mo
wi
mn
ne
ms
fl
tx
or
ga
il
in
ca
la
nj
ut
nc
sc
hi
ny
va
md
ma
0
5000
10000
15000
Average Price ($)

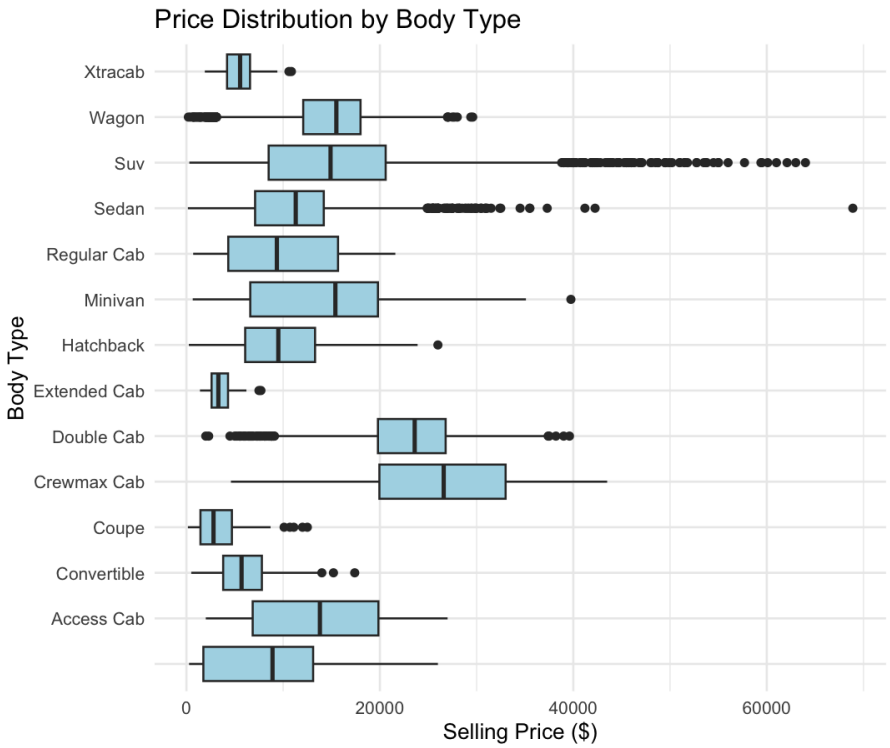

Top 10 Most Sold Toyota Models
Model
Camry
Corolla
Rav4
Sienna
Highlander
Prius
Tundra
4runner
Avalon
Tacoma
0
2500
5000
7500
10000
12500
Number of Listings

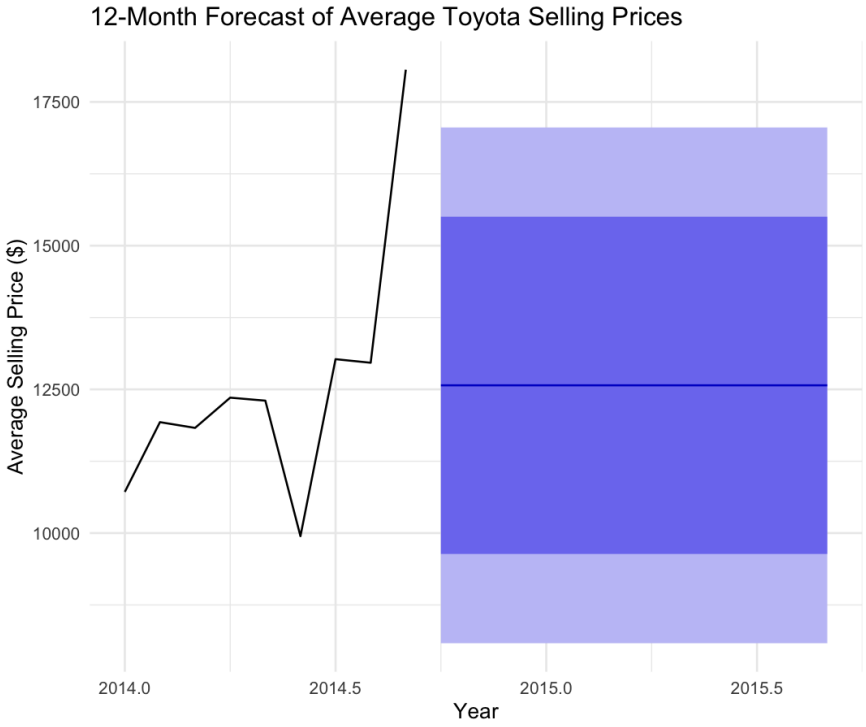

Toyota Listings by Body Type
Sedan
Suv
Hatchback
Minivan
Double Cab
Wagon
Crewmax Cab
Access Cab
Coupe
Regular Cab
Convertible
Extended Cab
Xtracab
Body Type
0
5000
10000
15000
20000
Number of Listings

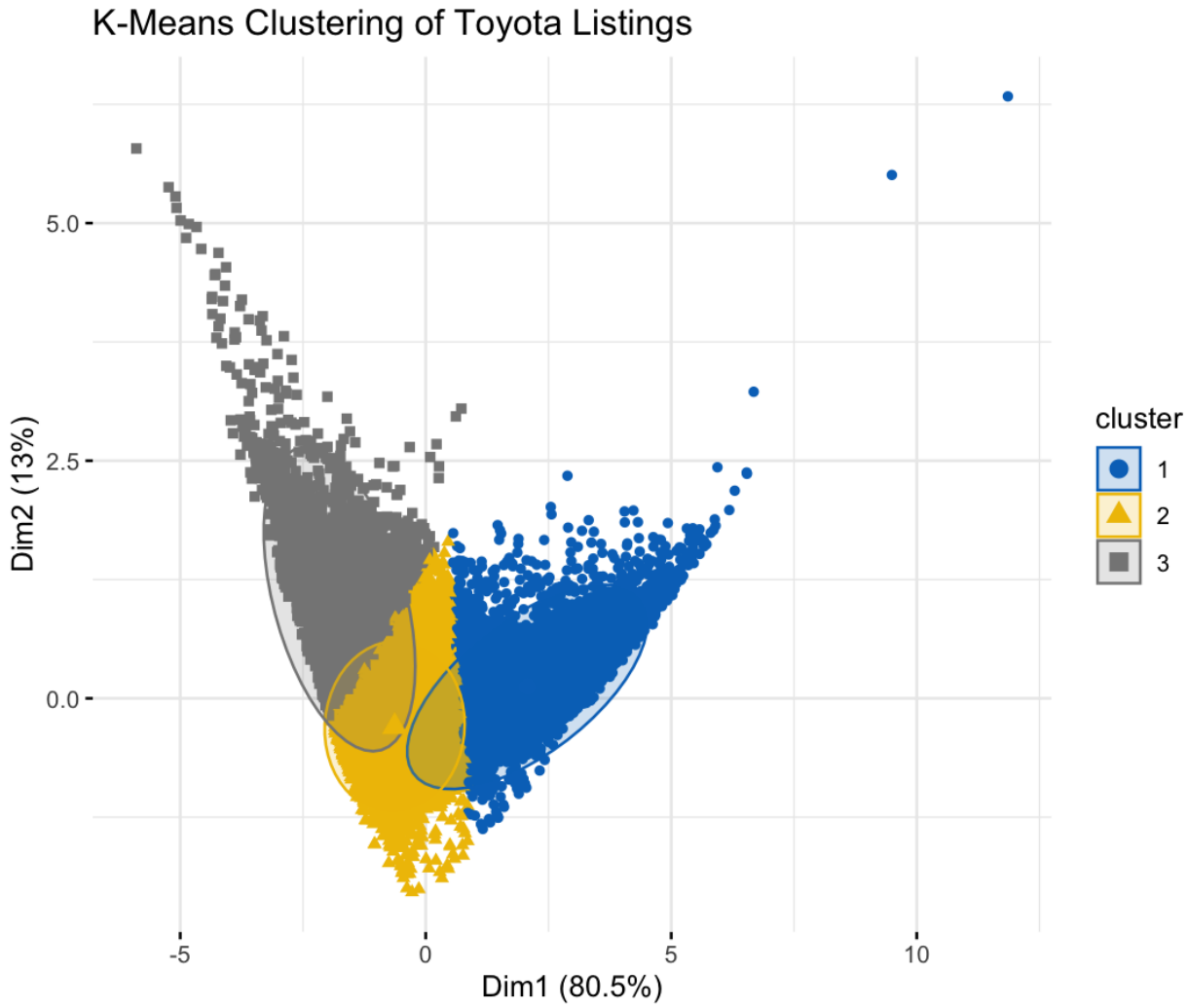

Selling Price vs. Odometer (Toyota, ≤ 300,000 miles)
60000
40000
20000
0
Selling Price ($)
0e+00
1e+05
2e+05
3e+05
Odometer (miles)

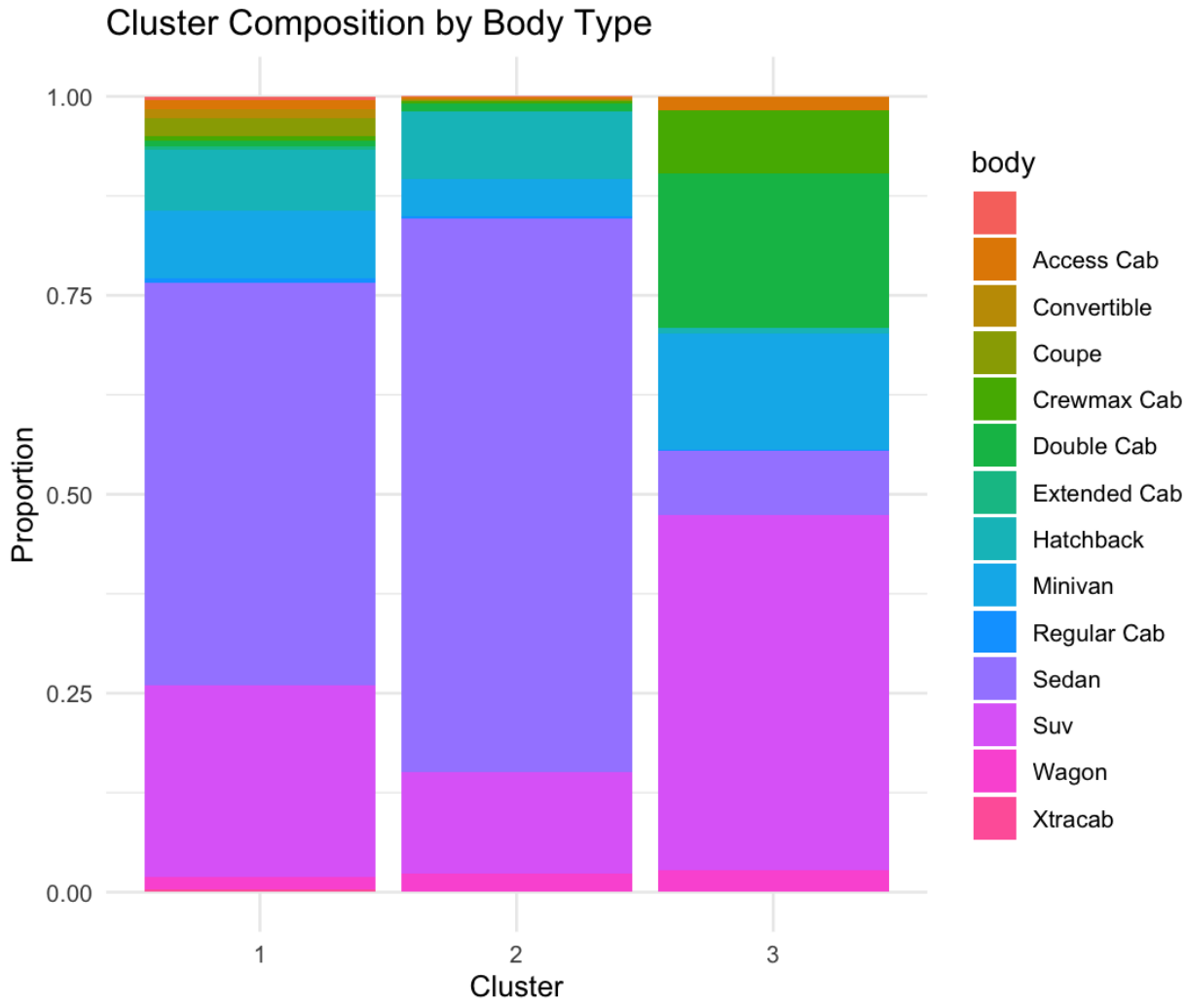

Avg Toyota Selling Price Over Time (500+ Listings)
Average Selling Price ($)
13000
12750
12500
12250
12000
Dec 2014
Jan 2015
Feb 2015
Mar 2015
Apr 2015
May 2015
Jun 2015
Month


Distribution of Vehicle Conditions
Count
3000
2000
1000
0
0
10
20
30
40
50
Condition


Price Distribution by Body Type
Xtracab
Wagon
Suv
Sedan
Regular Cab
Minivan
Hatchback
Extended Cab
Double Cab
Crewmax Cab
Coupe
Convertible
Access Cab
Body Type
0
20000
40000
60000
Selling Price ($)


12-Month Forecast of Average Toyota Selling Prices
17500
15000
12500
10000
Average Selling Price ($)
2014.0
2014.5
2015.0
2015.5
Year


K-Means Clustering of Toyota Listings
5.0
2.5
0.0
Dim2 (13%)
-5
0
5
10
Dim1 (80.5%)
cluster
1
2
3


Cluster Composition by Body Type
1.00
0.75
0.50
0.25
0.00
Proportion
1
2
3
Cluster
body
Access Cab
Convertible
Coupe
Crewmax Cab
Double Cab
Extended Cab
Hatchback
Minivan
Regular Cab
Sedan
Suv
Wagon
Xtracab