EAGLE FANG DATA CONSULTING

Team 36: Final Presentation

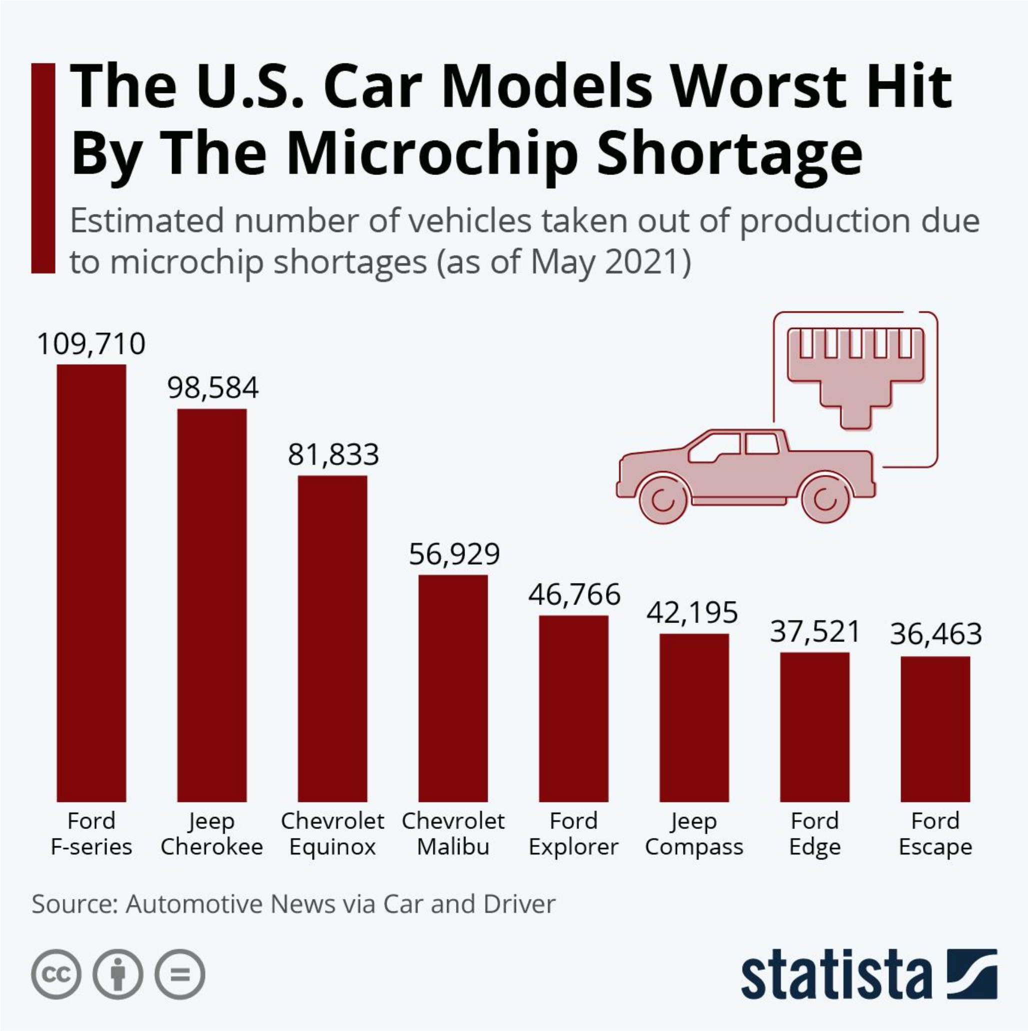
Authors: Andrew Bartels, Jason Young, Matt Palmer, Ngoc Nguyen, Shawn Azzu

Where have all the cheap cars gone?

- Customers no longer have an upper-hand
- In the past, the moment a car was driven off the lot, the price would immediately depreciate.
- A car's price would be cut in half after just 3 years.
- This is no longer the case.

Problem/Background

- Huge Microchip Shortage over the past few years has resulted in decreased car production for new models
- This has increased the price of new and used cars, but has turned more buyers to look at the secondhand market



Project Task and Introduction

- Penske Motorgroup, LLC has hired Eagle Fang Data Consulting to investigate arbitrage opportunities in the used car market
- Our task is to create models that will predict used car prices in various regions in the U.S
- We have built in app that will enable our clients to easily identify and compare prices

Research and Background

- Our Research has shown us that when keeping the mileage, model, and year constant, location has an impact on car price in two ways:
  - Impact on car based on environment: weather and terrain
  - Car Price fluctuation based on regional economics:
    - Regional Economic Variables indirectly influences cars' value and the overall state of the used-car market
  - With this knowledge we utilize multiple linear regressions, regressing price on regional and economic variables

- The Cost of Driving comes from gas and general maintenance and car price depreciation is exponential within the first 2 to 3 years
  - From then, decline in price is seen to be more gradual
  - The depreciation is related to miles on the car, we have this variable as Odometer

Approach and Methodology

Data Preparation and Cleaning

- Primary Cars Dataset contains 426,880 observations and 22 explanatory variables.
- The complementary datasets that we combined with our car table include economic data pulled from the 2020 census as well as outdoor recreation data by state from 2021.
- For the first step of the cleaning process, we wrote the tables to an SQLite3 database for lightweight storage and efficient use between team members.

Approach and Methodology

Cleaning Challenges

- Handling NA Values (1,688,231 NA values!)
- Mapping income to cities – extract ZCTA codes
- Extracting missing information like Manufacturers & Models from description
- Discrepancies (ex. old vehicles with low mileage)

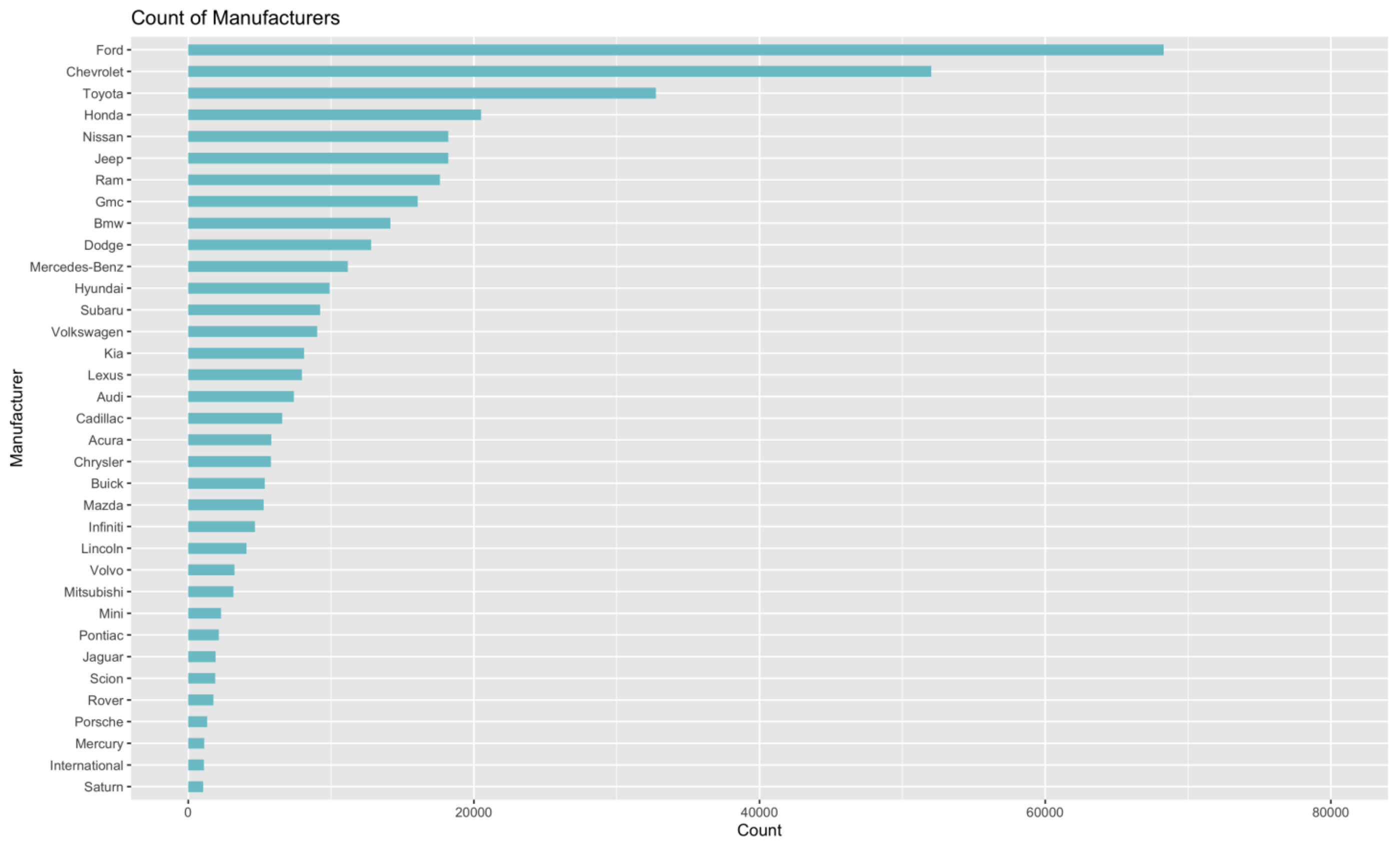
Variable	# NA Values	% of Observations	Amt. Recovered
Year	1,205	28%	1,184
Manufacturer	17,646	4.13%	11,975
Paint color	130,203	30.5%	37,758

Variable	# NA Values	% of Observations	Amt. Recovered
Type	92,858	21.75%	23,214
Drive	130,567	30.59%	12,998

Exploratory Data Analysis

Strategy: Select top 5 Manufacturers

Focus on cleaning the data so that we could build clean visualizations and have accurate models



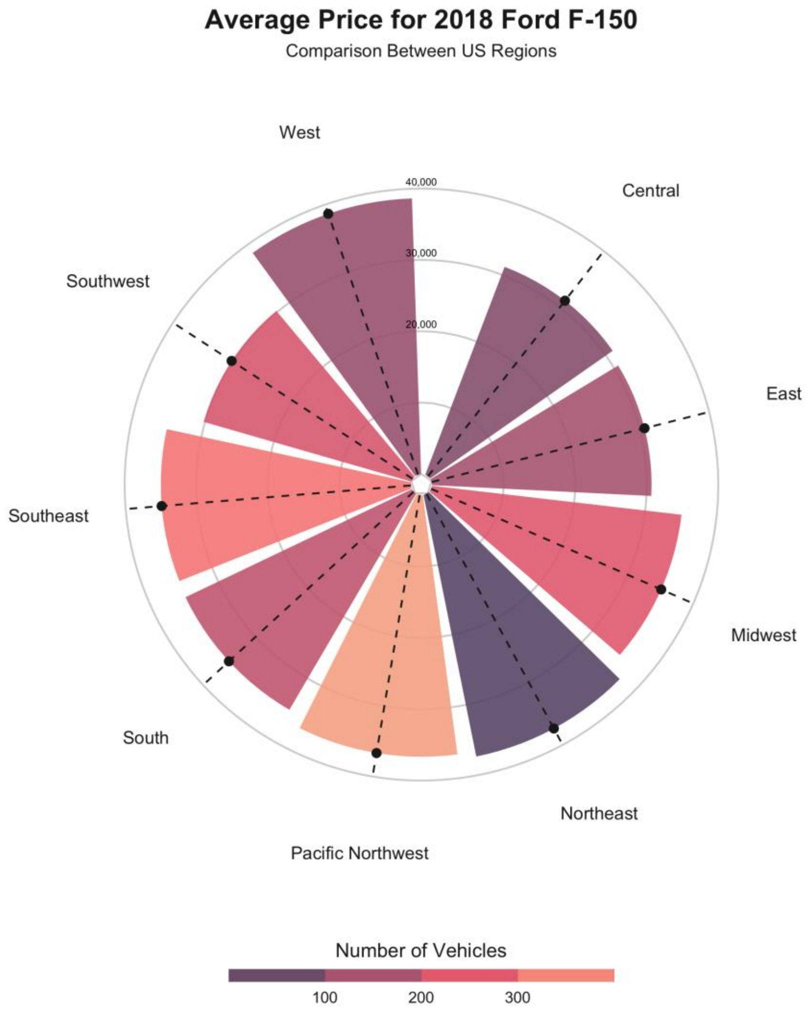
Analysis and Visualizations

Strategy: Create Regional Comparison

Ultimately, we wanted our client to be able to compare vehicle prices between different parts of the country

We segmented the dataset into 9 key regions:

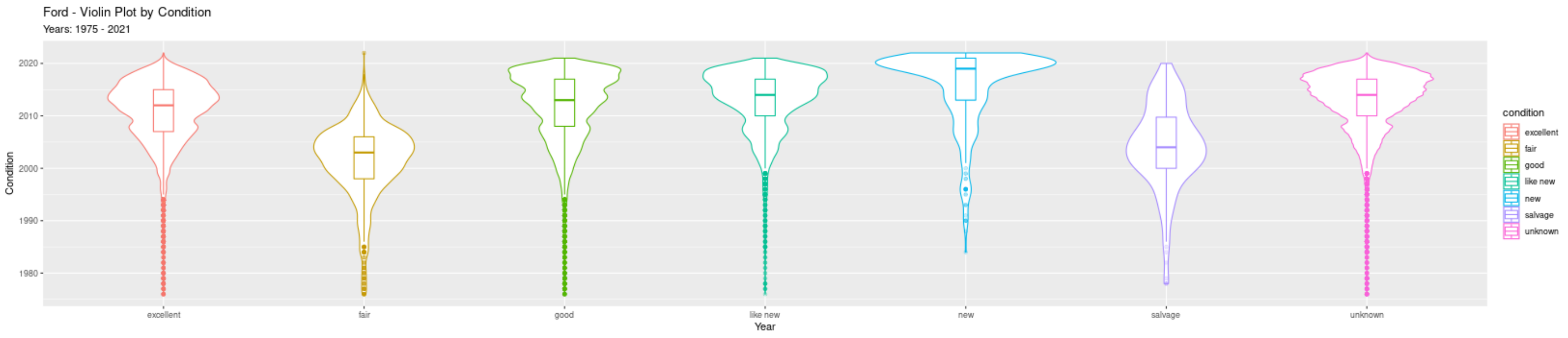
1. West
2. Southwest
3. Central
4. Pacific Northwest
5. Northeast
6. Southeast
7. Midwest
8. South
9. East



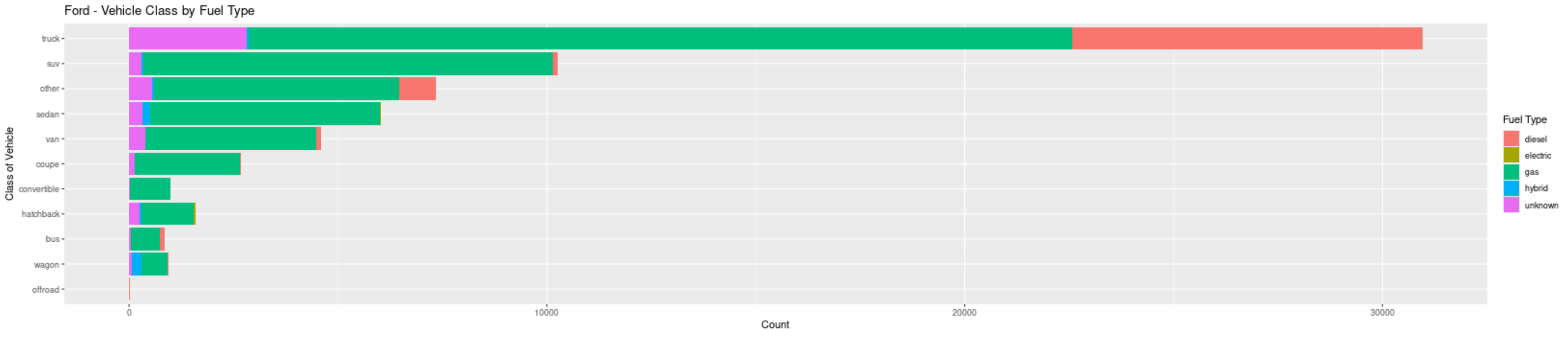
Analysis and Visualizations

Strategy: Build a tool that can compare vehicle attributes

Condition Comparison per Manufacturer



Fuel Comparison per Manufacturer



Comparing Regression Models

Model_Comparison():

```
Model_Comparison <- function(df, list_of_tuples)
{
  # split dataframe into training, validation, and test sets (60-20-20% rule)

  train_size = round(0.6 * nrow(df), 0)
  valid_size = round(0.2 * nrow(df), 0)
  test_size = round(0.2 * nrow(df), 0)
  shuffled_rows <- sample(nrow(df))
  df = df[shuffled_rows, ]
  train_data = df[1:train_size, ]
  valid_data = df[(train_size + 1):(train_size + valid_size), ]
  test_data = df[(train_size + valid_size + 1):nrow(df), ]

  # Create the empty output dataframe for comparison
  output_df <- data.frame(matrix(ncol = 4, nrow = 0))
  colnames(output_df) <- c("Model", "R^2 on Validation Data", "Best Model?",
"R^2 on Test Data")

  # Create empty lists to collect R^2 values
  R_2_list = c()
  best_R_2_list = c()
```

Input:

```
tuple_list =
  list(
    list("Standard LM (car info only)", standard_lm_function_car_only),
    list("Standard LM plus local economic data", standard_lm_function),
    list("Log Age (All Columns)", standard_lm_log_age),
    list("Log Age Log Odometer (All Columns)", log_age_log_odometer)
  )


Model_Comparison(cars, tuple_list)
```

Sample Output:

	Model <chr>	R^2 on Validation Data <chr>	Best Model? <chr>	R^2 on Test Data <chr>
1	Model 1	0.272363884414753	0	NA
2	Model 2	0.369859962701729	0	NA
3	Model 3	0.383760232034281	1	0.389529506010875

Results of Regression Model Comparisons

	Model <chr>	R^2 on Validation Data <chr>	Best Model? <chr>	R^2 on Test Data <chr>
1	Standard LM (car info only)	0.423	0	*NA*
2	Standard LM plus local economic data	0.459	0	*NA*
3	Log Age (All Columns)	0.537	0	*NA*
4	Log Age Log Odometer (All Columns)	0.549	1	0.548

Results of Regression Model Comparisons

	Model <chr>	R^2 on Validation Data <chr>	Best Model? <chr>	R^2 on Test Data <chr>
1	Standard LM (car info only)	0.423	0	NA
2	Standard LM plus local economic data	0.459	0	NA
3	Log Age (All Columns)	0.537	0	NA
4	Log Age Log Odometer (All Columns)	0.549	1	0.548

outdoor_rec_by_state

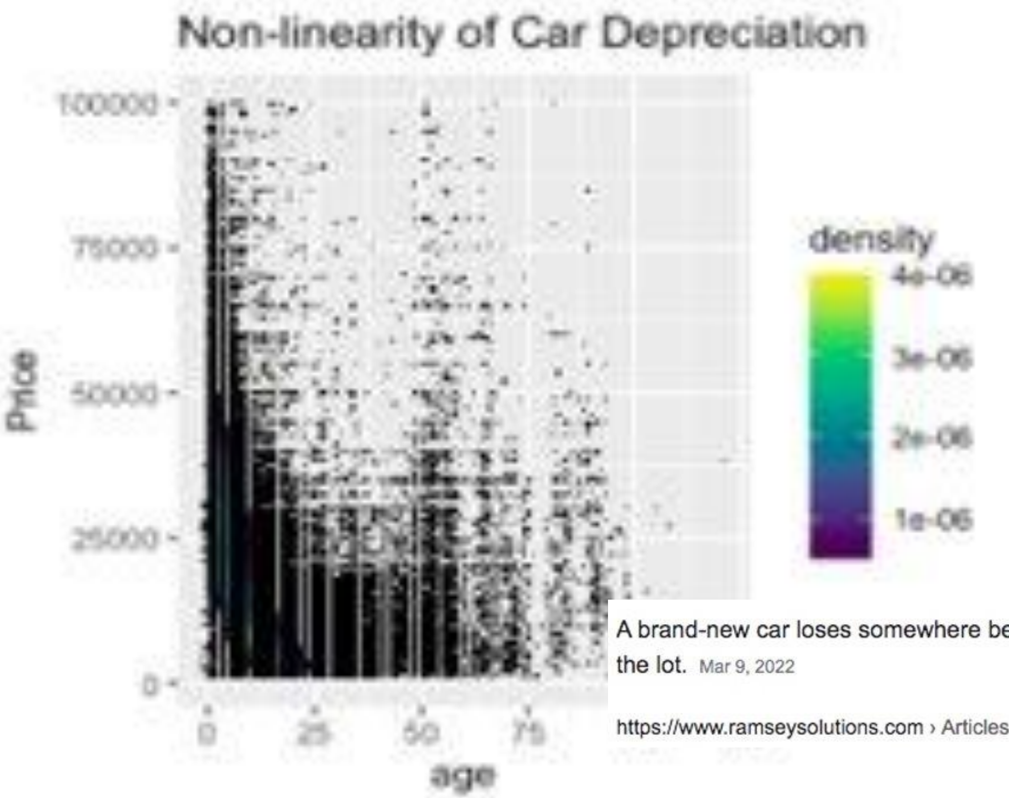
index	INTEGER
State	TEXT
Total outdoor recreation value a...	TEXT
Percent of total value added1	REAL
Total outdoor recreation employ...	TEXT
Percent of total wage and salary...	REAL
Total outdoor recreation compe...	TEXT
Percent of total compensation1	REAL

Results of Regression Model Comparisons

	Model <chr>	R^2 on Validation Data <chr>	Best Model? <chr>	R^2 on Test Data <chr>
1	Standard LM (car info only)	0.423	0	*NA*
2	Standard LM plus local economic data	0.459	0	*NA*
3	Log Age (All Columns)	0.537	0	*NA*
4	Log Age Log Odometer (All Columns)	0.549	1	0.548

outdoor_rec_by_state

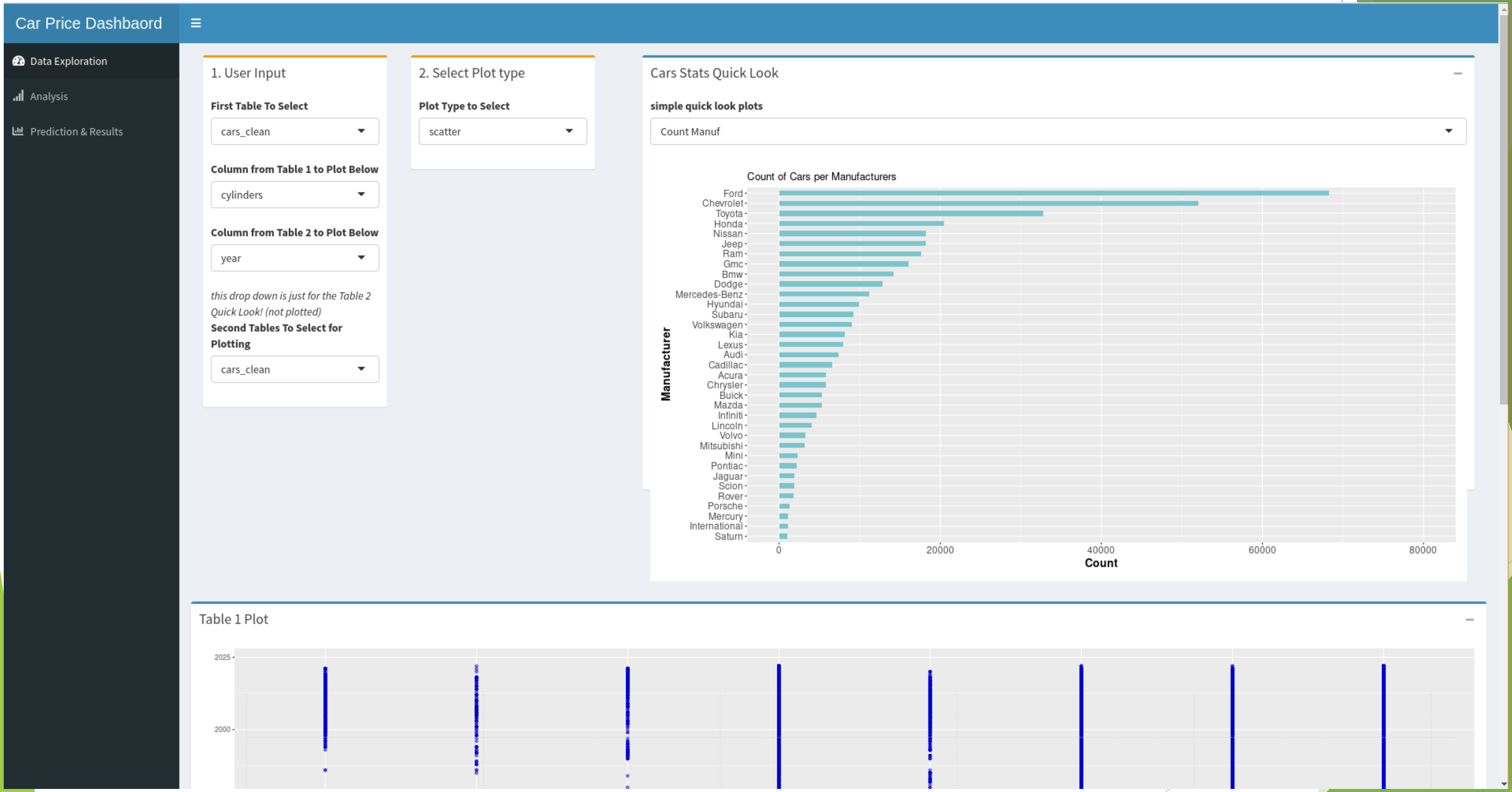
index	INTEGER
State	TEXT
Total outdoor recreation value a...	TEXT
Percent of total value added1	REAL
Total outdoor recreation employ...	TEXT
Percent of total wage and salary...	REAL
Total outdoor recreation compe...	TEXT
Percent of total compensation1	REAL



A brand-new car loses somewhere between **9–11% of its value** the moment you drive off the lot. Mar 9, 2022

https://www.ramseysolutions.com › Articles

Results – Demo & Visualization



Results – Demo & Visualization

Table 1 Plot

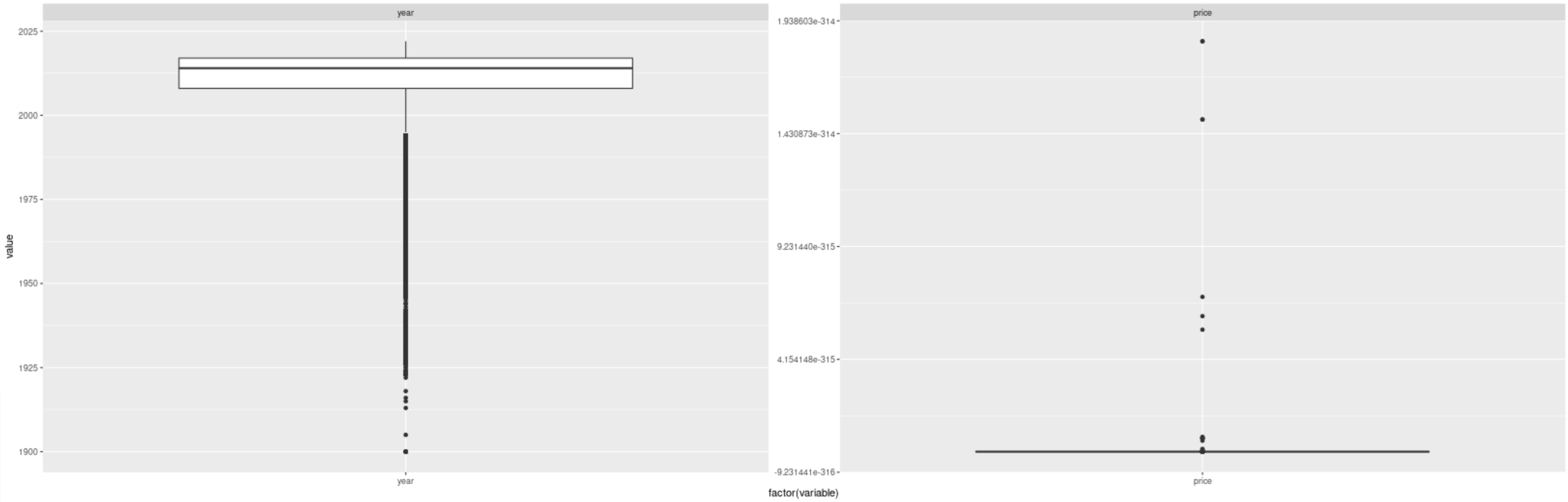


Table 1 Quick Look

Show 10 entries

Search:

	id	region	city	state	year	manufacturer	model	condition	cylinders
1	7316814884	auburn	Auburn	al	2014	Gmc	Sierra 1500 Crew Cab Slt	good	8 cylinders
2	7316814758	auburn	Auburn	al	2010	Chevrolet	Silverado 1500	good	8 cylinders
3	7316814989	auburn	Auburn	al	2020	Chevrolet	Silverado 1500 Crew	good	8 cylinders

Table 2 Quick Look

Show 10 entries

Search:

	id	region	city	state	year	manufacturer	model	condition	cylinders
1	7316814884	auburn	Auburn	al	2014	Gmc	Sierra 1500 Crew Cab Slt	good	8 cylinders
2	7316814758	auburn	Auburn	al	2010	Chevrolet	Silverado 1500	good	8 cylinders
3	7316814989	auburn	Auburn	al	2020	Chevrolet	Silverado 1500 Crew	good	8 cylinders

Results - Demo & Visualization

Analysis tab contents

This tab contains some more complex visualizations for a specific Manufacturer to ensure the most value and consistent app experience.

Select Manufacturer to Analyze

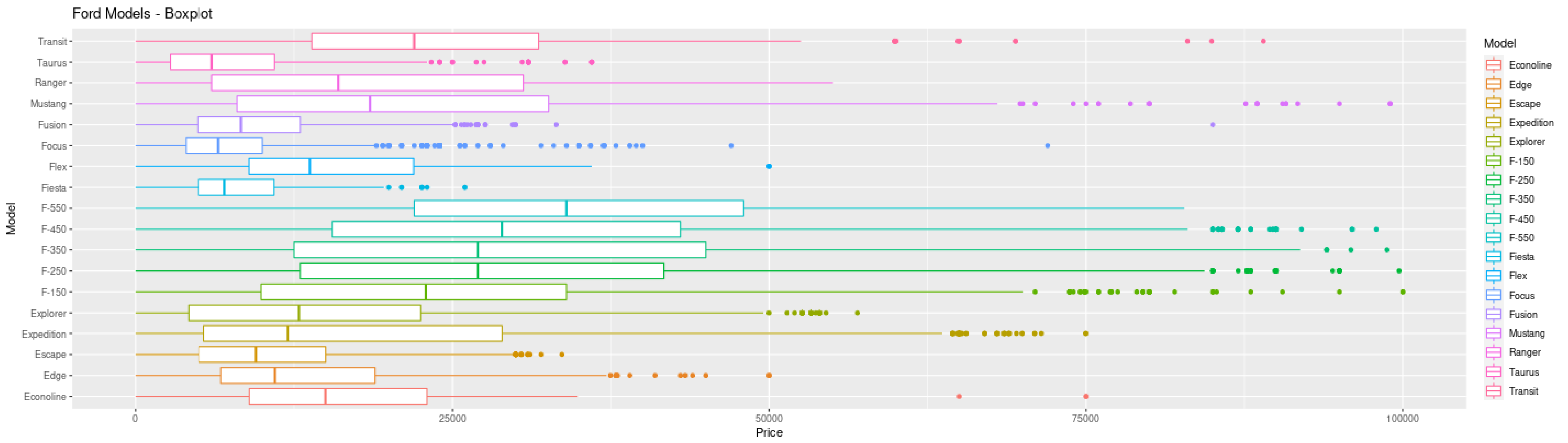
Select Manufacturer

Ford

Select the Make/Model

Mustang

Per Make/Model Box



Results – Demo & Visualization

2. City

Bakersfield

3. Manufacturer

Ford

4. Make/Model

Mustang

5. Year

2015

6. Condition

like new

7. Mileage

500

8. Drive

Rear Wheel Drive

9. Number of Cylinders

4

Predicting Car Price!

Model Predicted Price and Region

```
You are now going to get a price prediction for 2015 Ford Mustang RWD 4 cylinders, in  Bakersfield , CA  -  like new condition with 500 miles in the Boxes below!
```

28169
State Prediction Estimate

38175
National Prediction Estimate

19795
KNN Prediction Estimate

If the models return 0/NA, there was not enough data for a good prediction!

Conclusion and Final Thoughts

- Overall, this research project for the Penske Motorgroup was a large success. The visualization for observing and predicting car prices are now available to a larger subset of Analysts to generate business decisions
- Visualizations are informative, responsive and generate value
- Infrastructure for the app, backend, and frontend are simple, sleek and reproducible