Gear

Used Car Price Analysis

Analyzing over 50,000 vehicles and building a price model

Furkan Demirdoven
Data Analyst

Business Problem

Gear

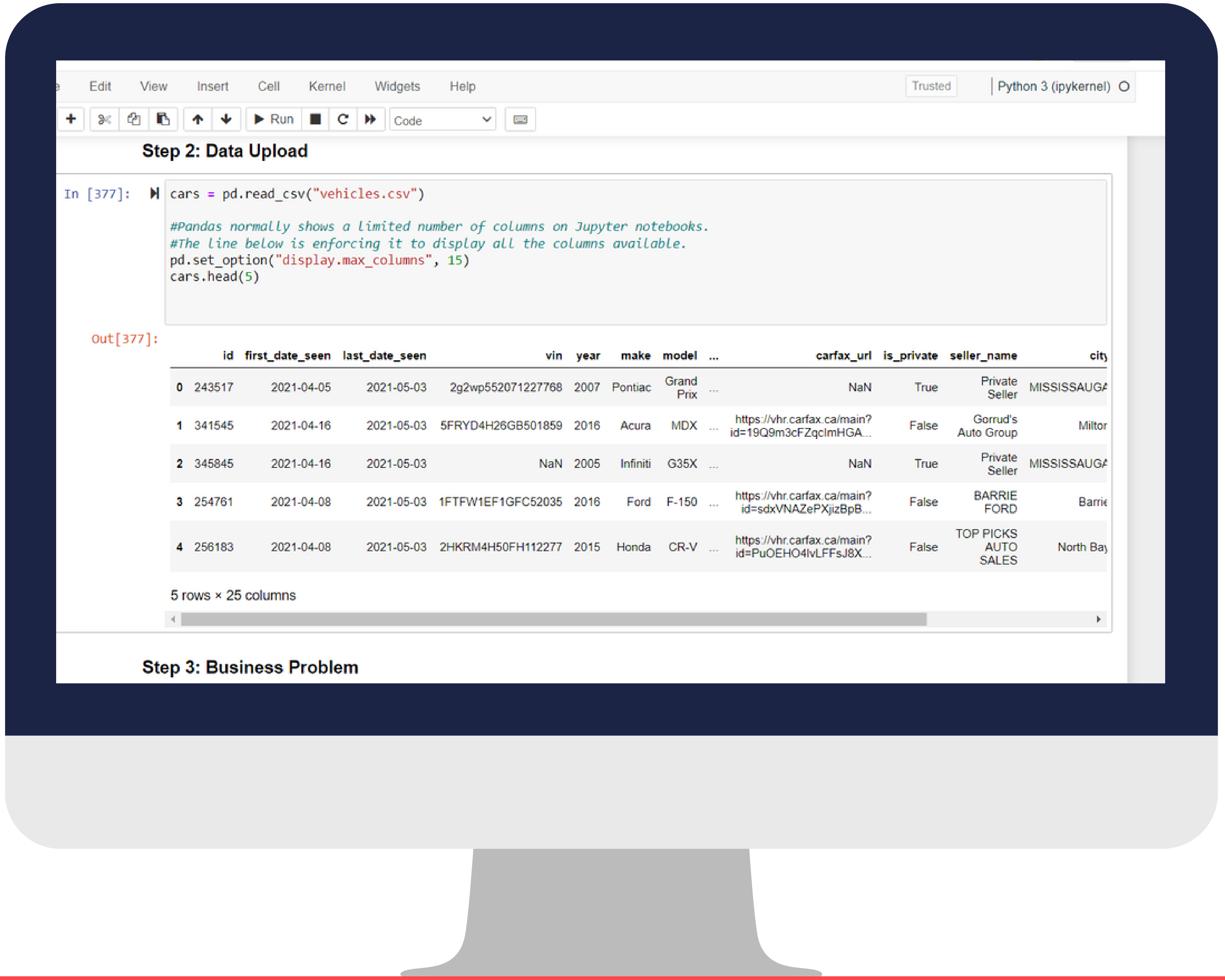
Online dealer **Gear** sits on rows of historic car data and would like to know how to get **data-driven price estimates** given that other attributes of vehicle are known.

1

Generate insights on past data

2

Leverage historic data to create predictive model



Gear

DATASET

- Over 50,000 **unique** used car listings
- April- May, 2021
- 25 columns

Steps followed

Gear

1. Exploratory Data Analysis (EDA)
2. Data Cleaning
3. Data Analysis
4. Modelling

EDA results

Gear

01

NULL values

- Actual NULL values that needed treatment in columns like **bodytype**, **drivetrain**
- NULL values are the actual values like the ones in **Carfax_url** field

02

Outliers

- Actual outliers like two rare listings from a US city or vehicles with over 3,000,000 km on them
- Outliers that signals a value for analysis like extremely old cars

03

Duplicate records

- No duplicate entries except for the two US listings

04

Data entry errors

- Many new vehicles with over a million km on them signal human errors like putting an extra digit by mistake
- A 2001 Jetta with 999,999 km
- Old cars with 0 km

05

Distribution of data

- Frequency distribution on categorical fields before data points. i.e. **Color** and **make** columns had many, insignificant unique values

Data Cleaning I

Gear

01

NULL values

- Fields like **carfax_url, vin** and **is_private** converted to boolean (1, 0)
- IMPUTATION. i.e. missing **body type** was imputed from vehicles with same model, make
- Missing values in numerical columns were filled with mean where applicable. i.e. NULLs under **mileage** field was imputed from vehicles with same **make** with same **year**

02

Data entry errors

- **Mileages** with erroneous digits are fixed. Simply divided by 10.
- **Mileage** for that 2001 Jetta with 999,999 km is rendered NULL and than filled with mean mileage of all the 2001 Jetta listings
- **City**='Richmond' converted to 'Richmond Hill' thanks to **longitude/latitude** data.

03

Distribution of data

- Captured all **color** shades under broader groups: Star White **->** White
- Tagged edge cases in fields like **color, make** under 'Other'

Data Cleaning II

Gear

04

New fields

- Engine field had lengthy strings not feasible for transformation. **Cylinder** information was captured and stored in a new column.
- **City** field had many unique values. New field, **toronto_gta** accounts for Toronto boroughs
- Old cars built more than 30 years ago is tagged **vintage**
- **Age** column instead of **year** makes more sense for analysis purposes
- No useful info from seller name. Dropped

05

Transformations

- Many values under categorical fields required proper case cleaning for better visibility.
- **Fuel type** is cleaned to have broader categories like Elektric, Gasoline, Diesel.
- Cars with asking price < 500$ and ad **description** including 'parts' tag removed.

Gear

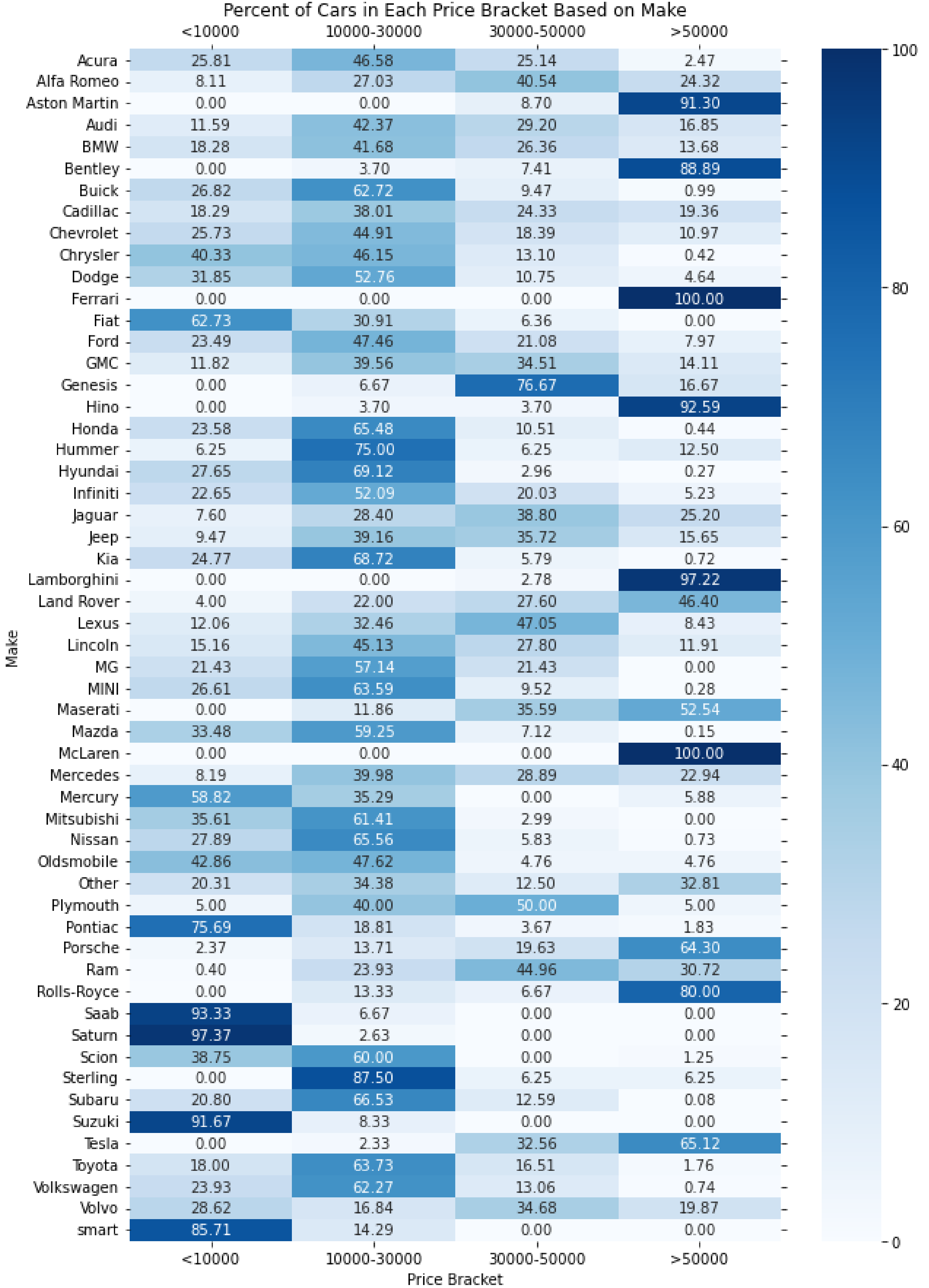
Data Analysis

Make vs Price

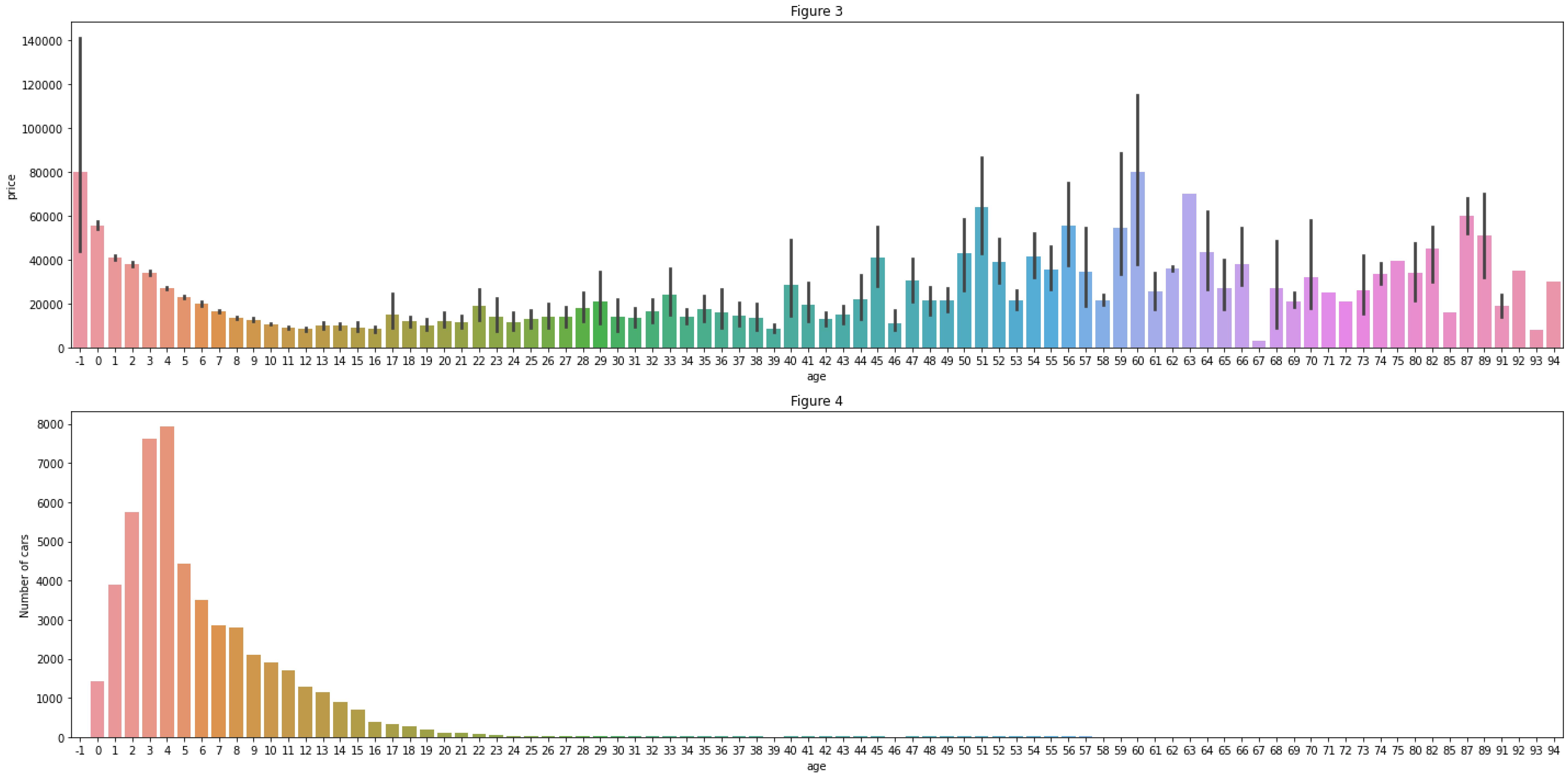
- Premium luxury vehicles such as Aston Martin, Bentley, Rolls-Royce, Ferrari, Tesla, Porsche, Maserati and Lamborghini have overwhelming majority of their cars listed upwards of **50,000$.**

- Brands like Toyota, Honda, Hyundai, Kia and Nissan have their cars mostly listed under the second tier,
 10,000$ - 30,000$.

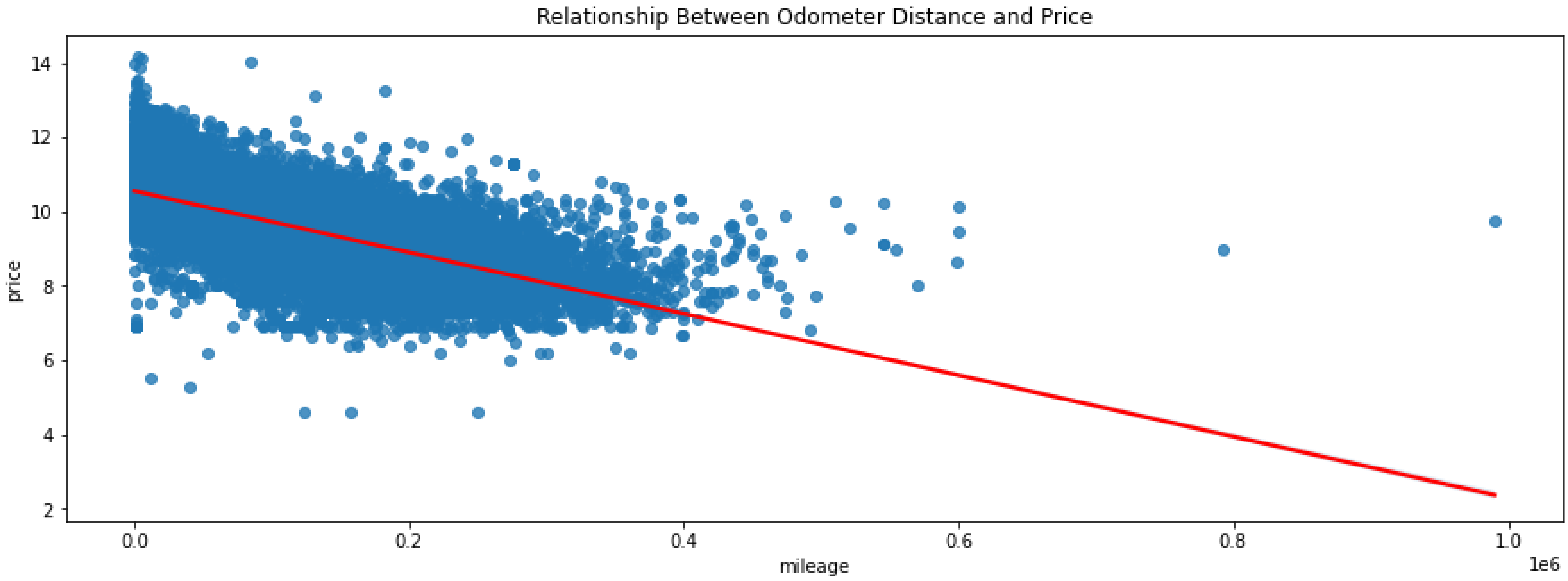
- The brands like Saturn, Saab, Suzuki, Smart and Fiat have almost all of their listings under **10,000$**. Could be due to the lower perception of brand quality in the Canadian used car market



Age vs Price

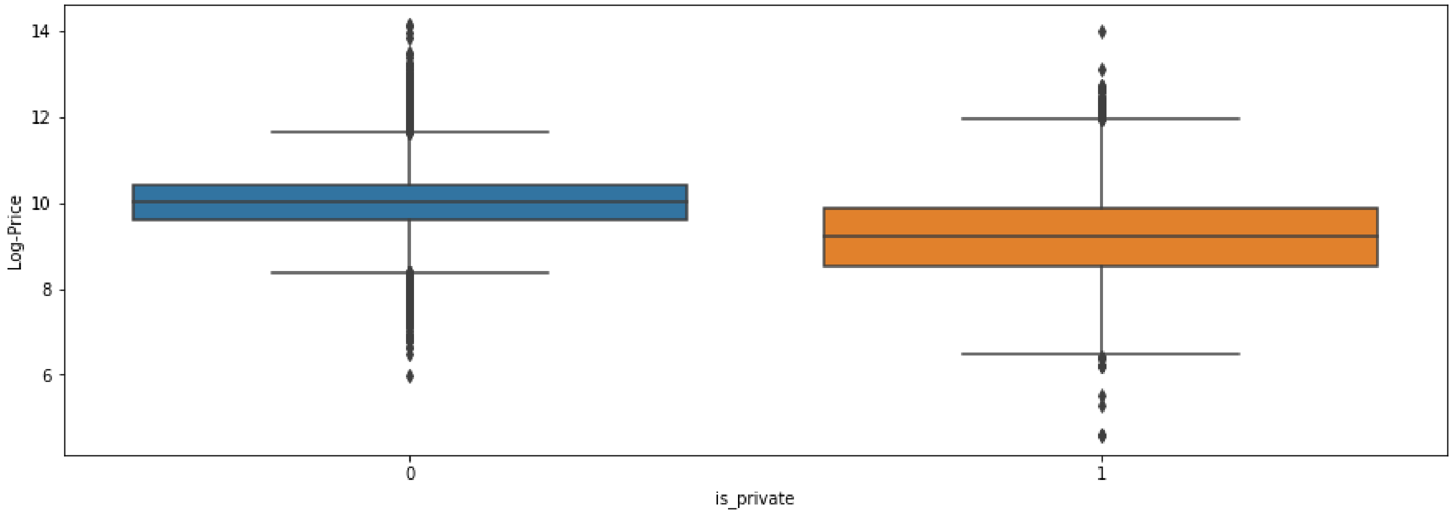


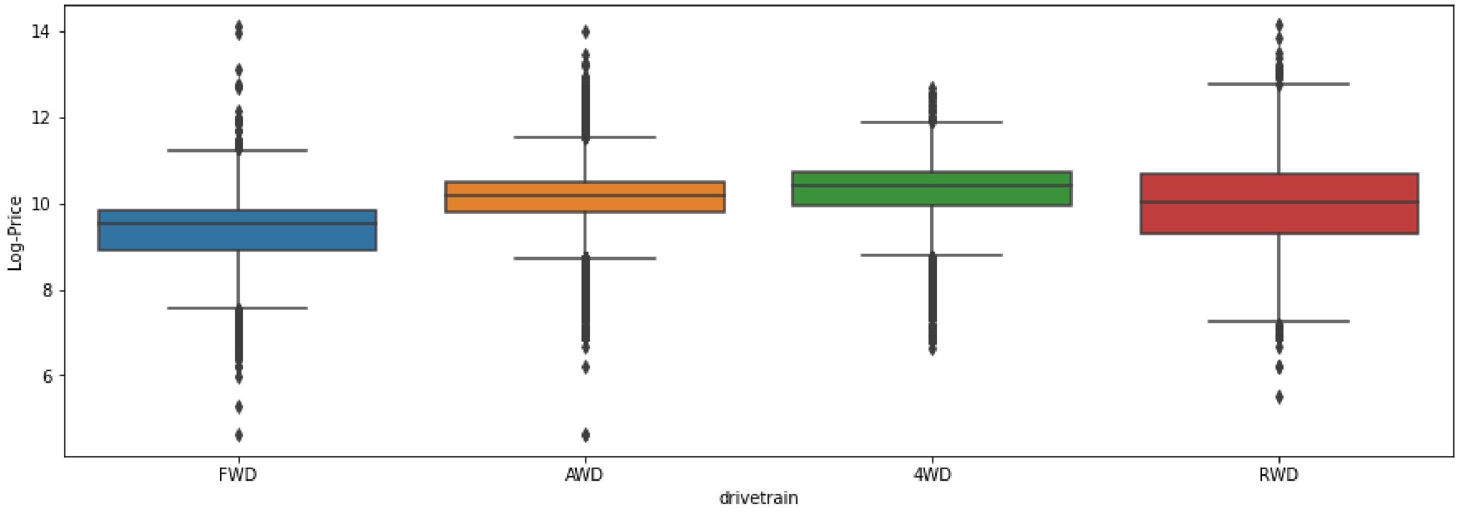
Mileage vs Price

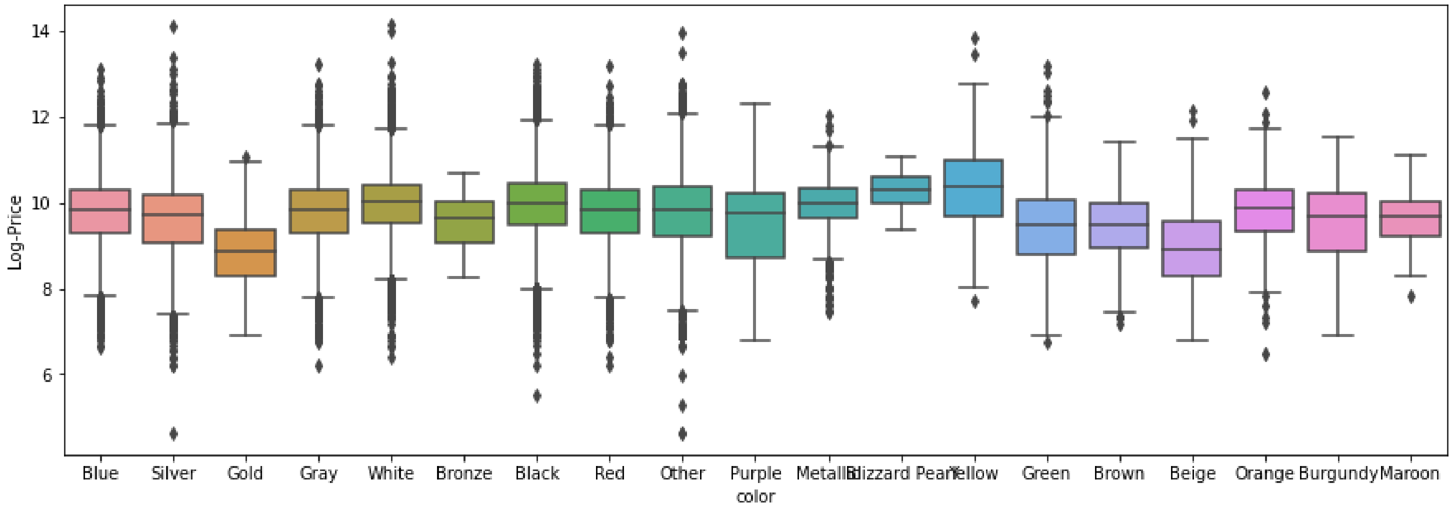


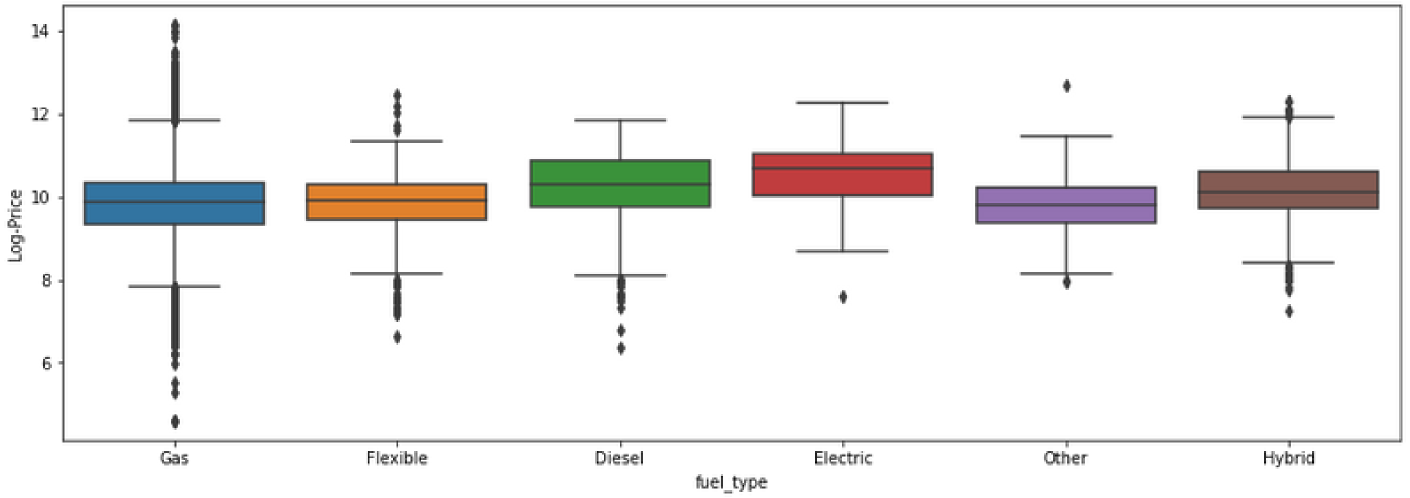
Various fields vs Price

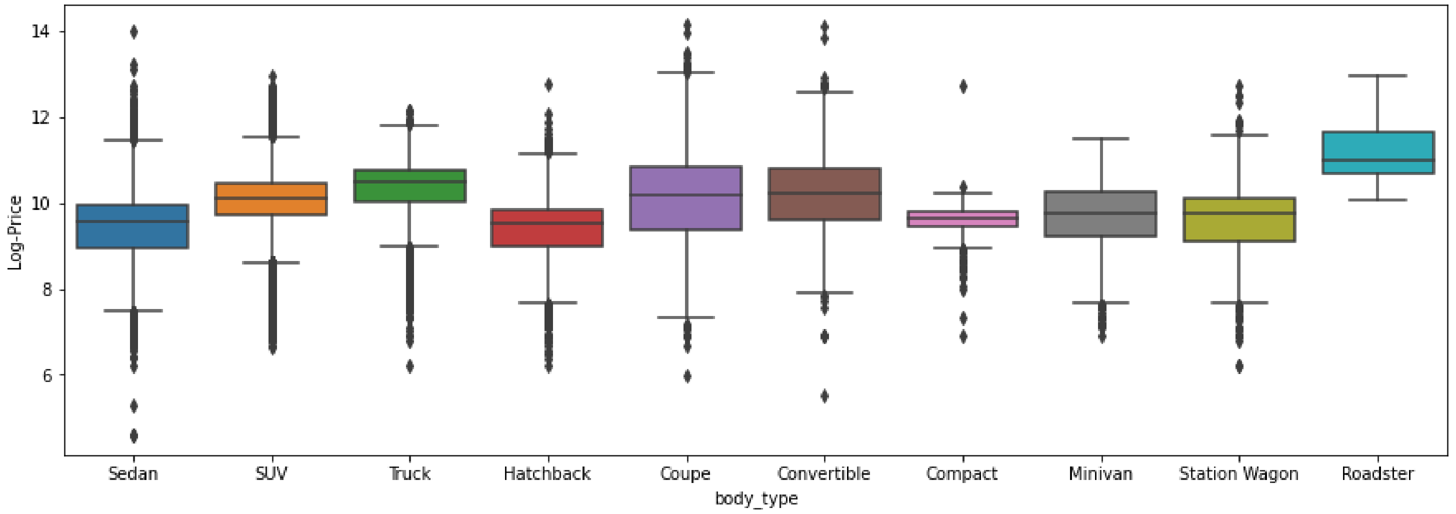
Gear

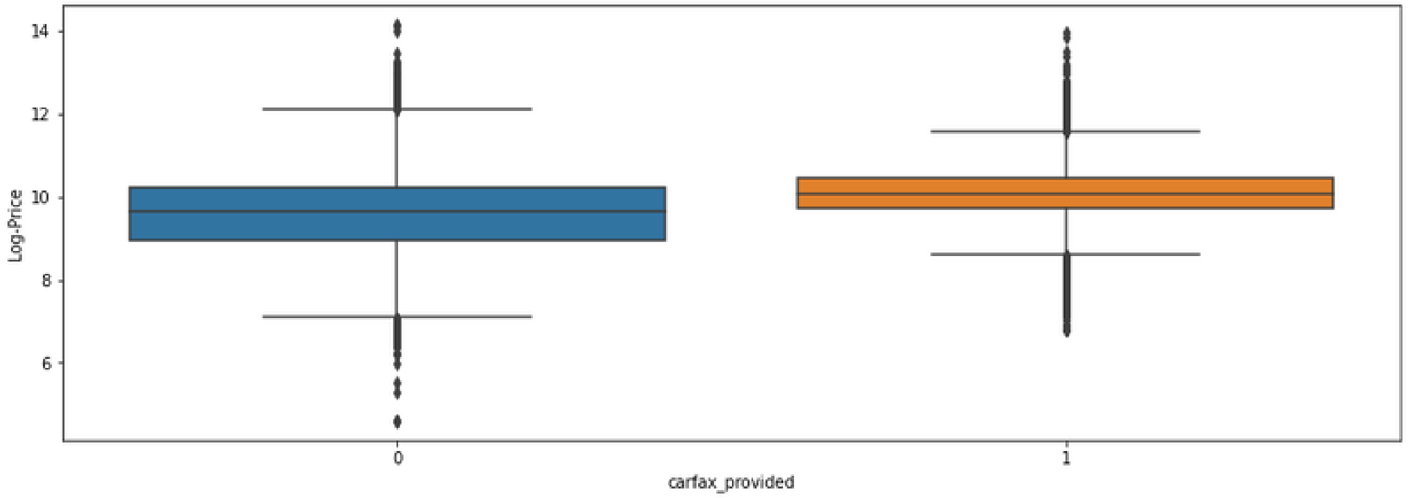






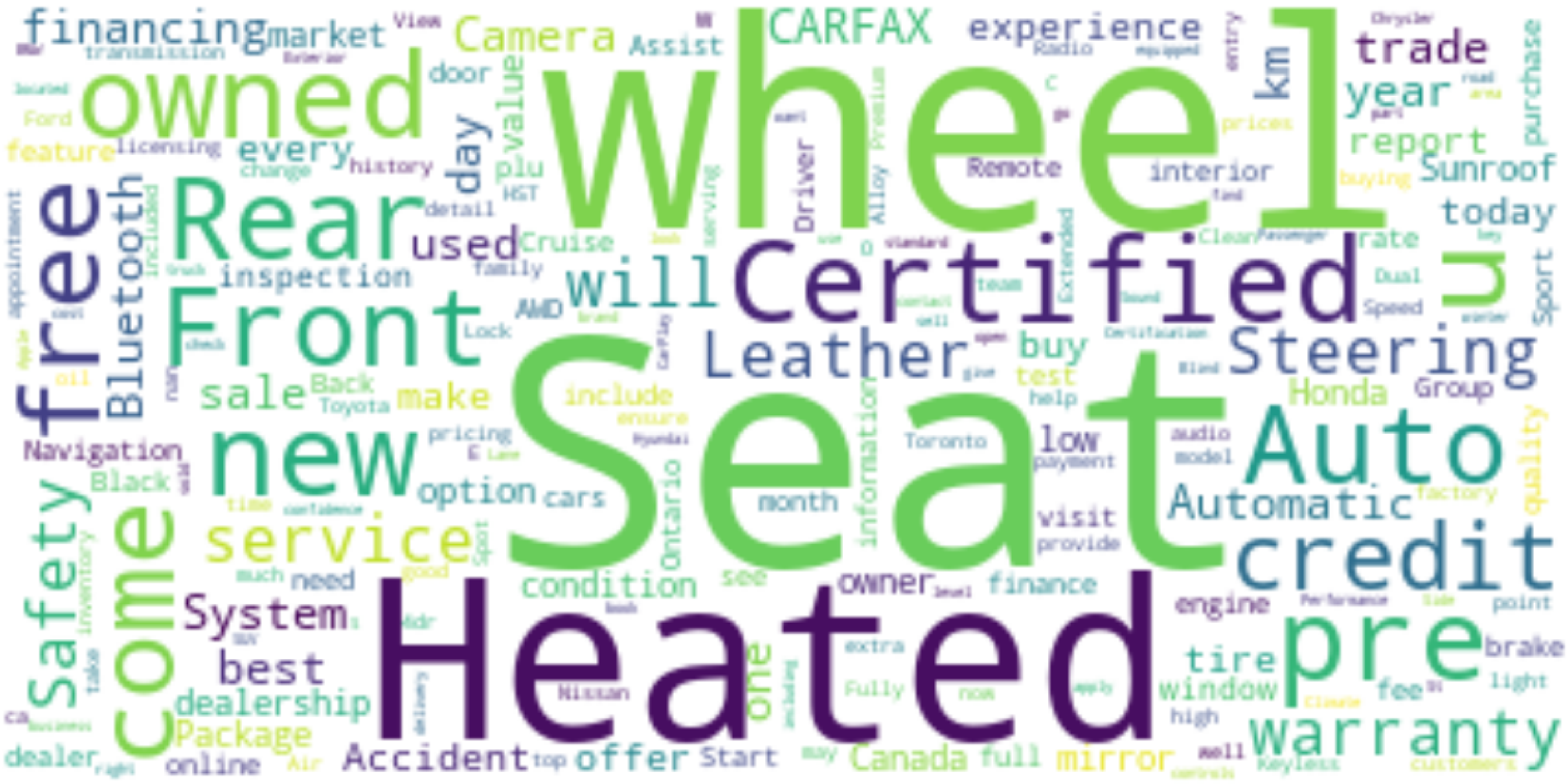




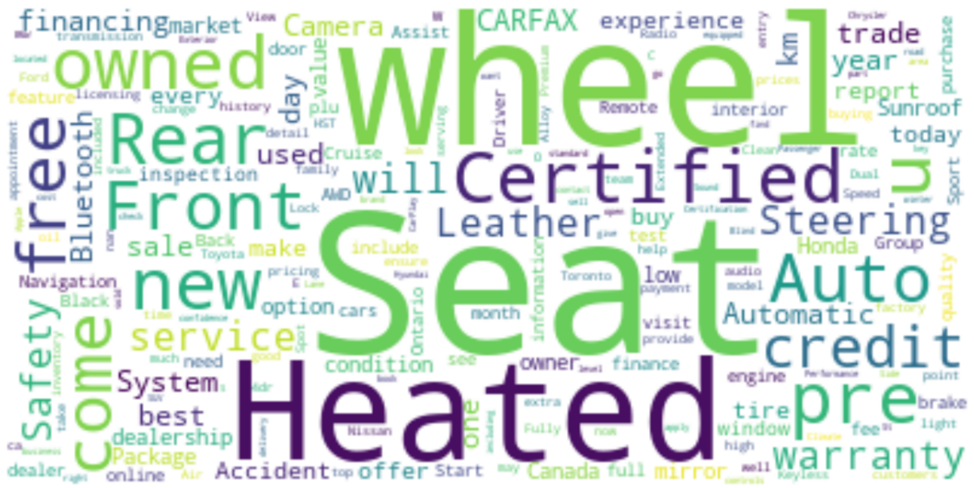


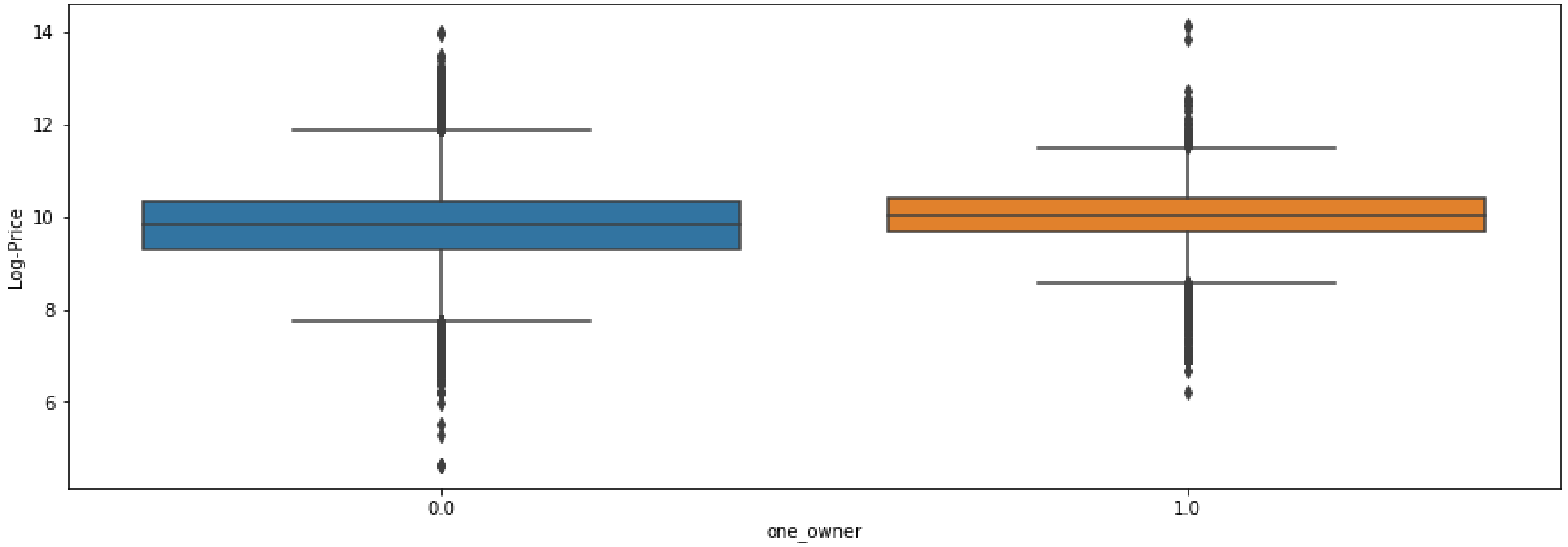
Improved description field

Gear

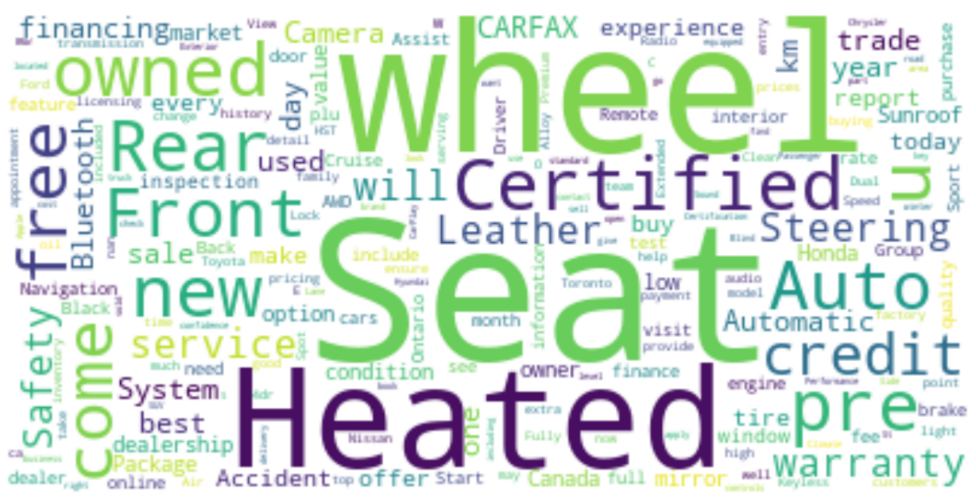


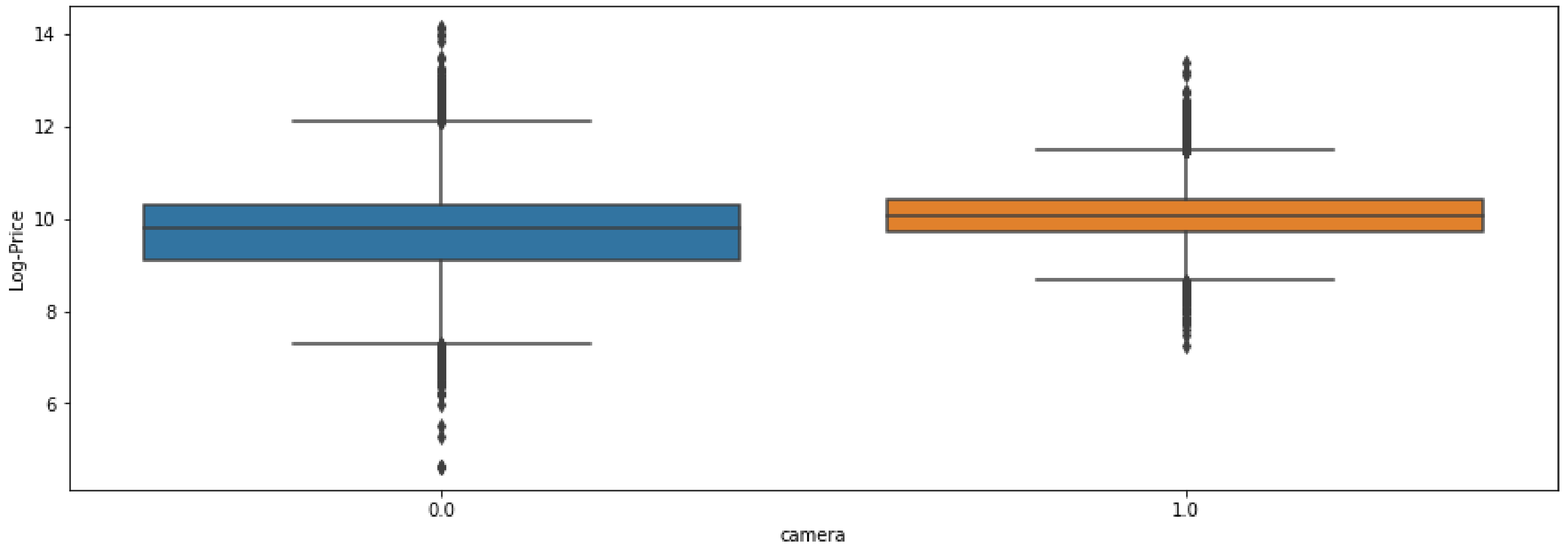
New column: One owner?



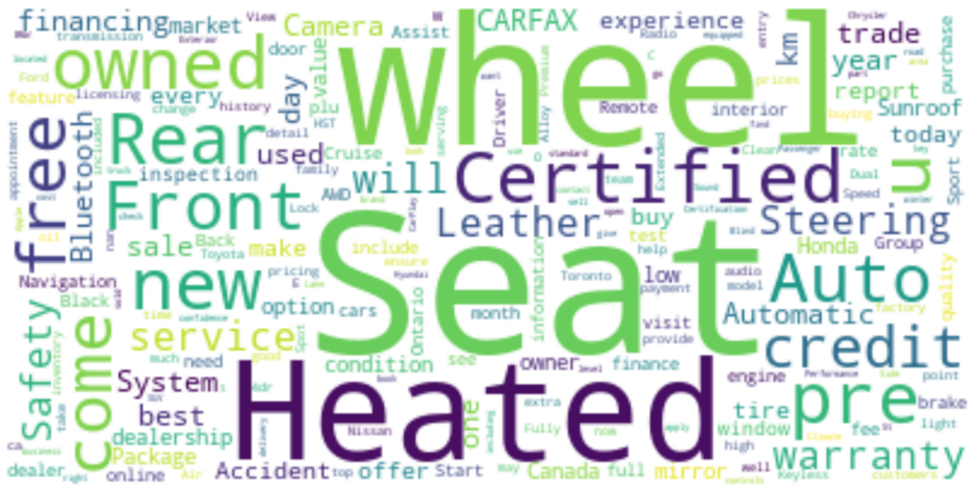


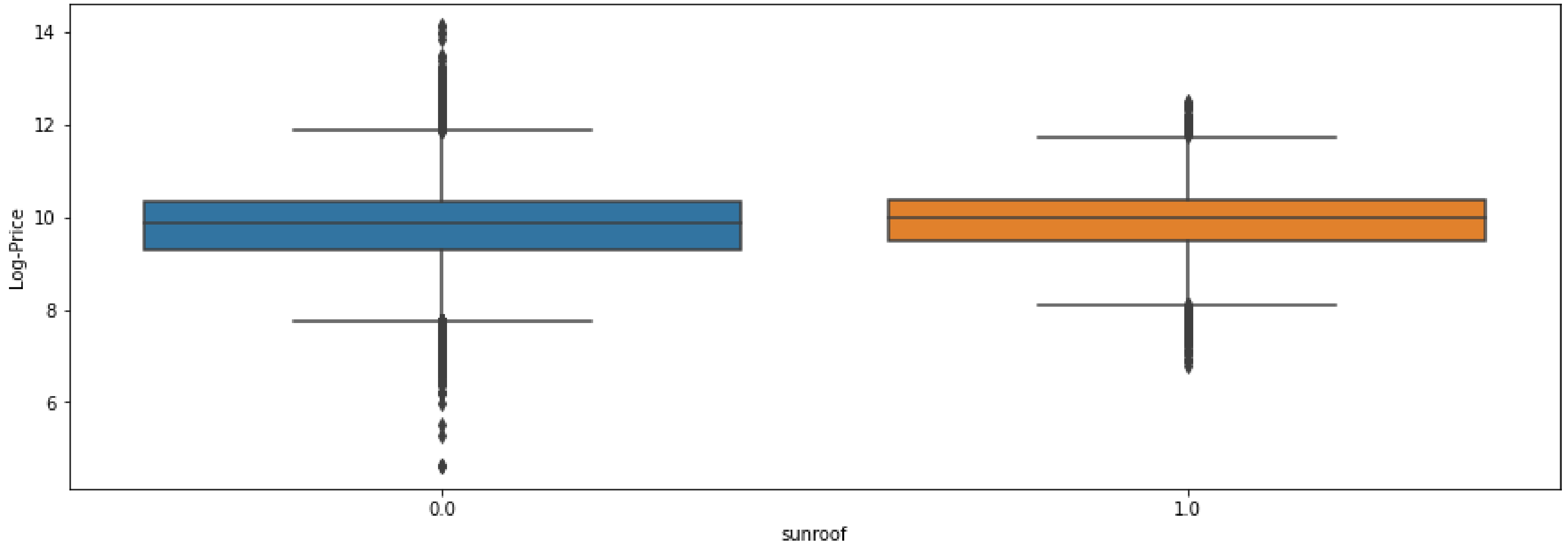
New column: Camera?



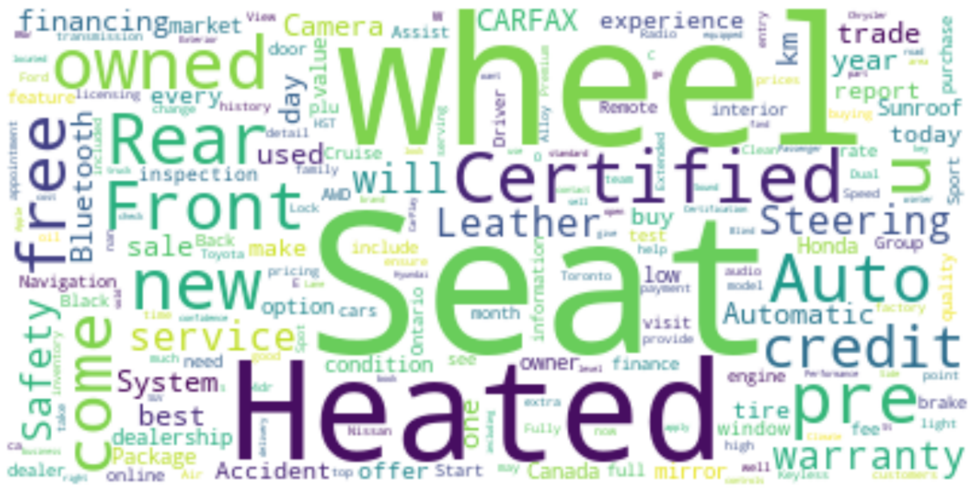


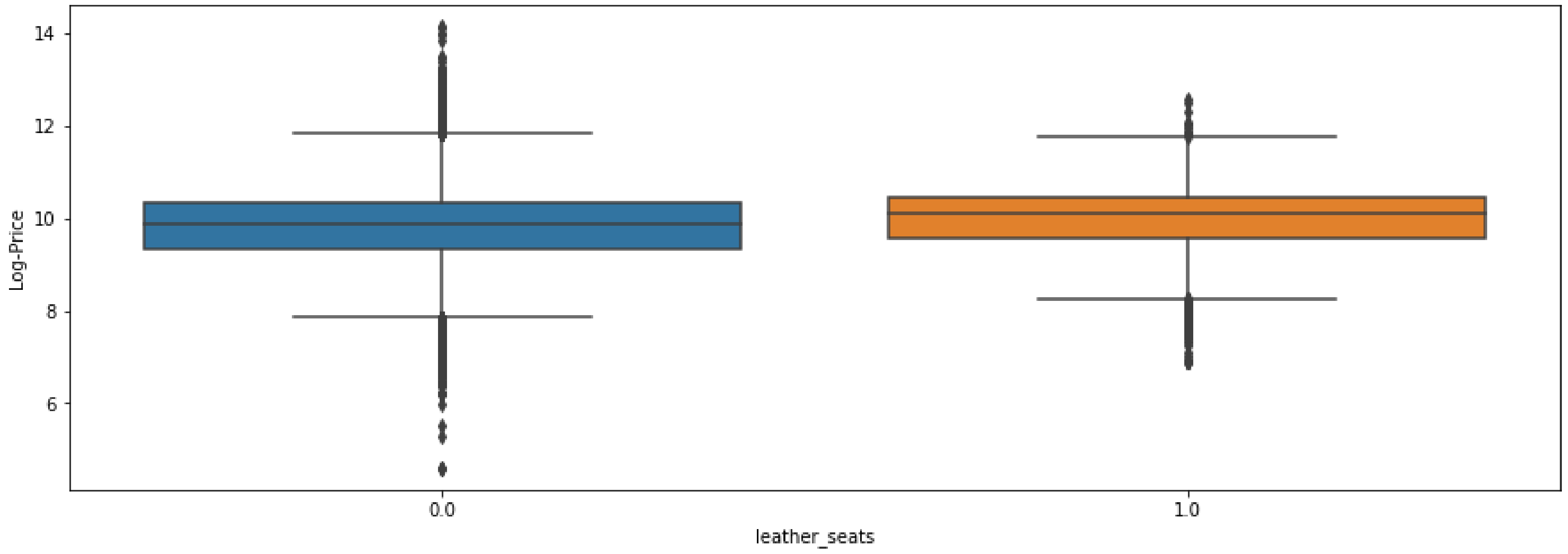
New column: Sunroof?





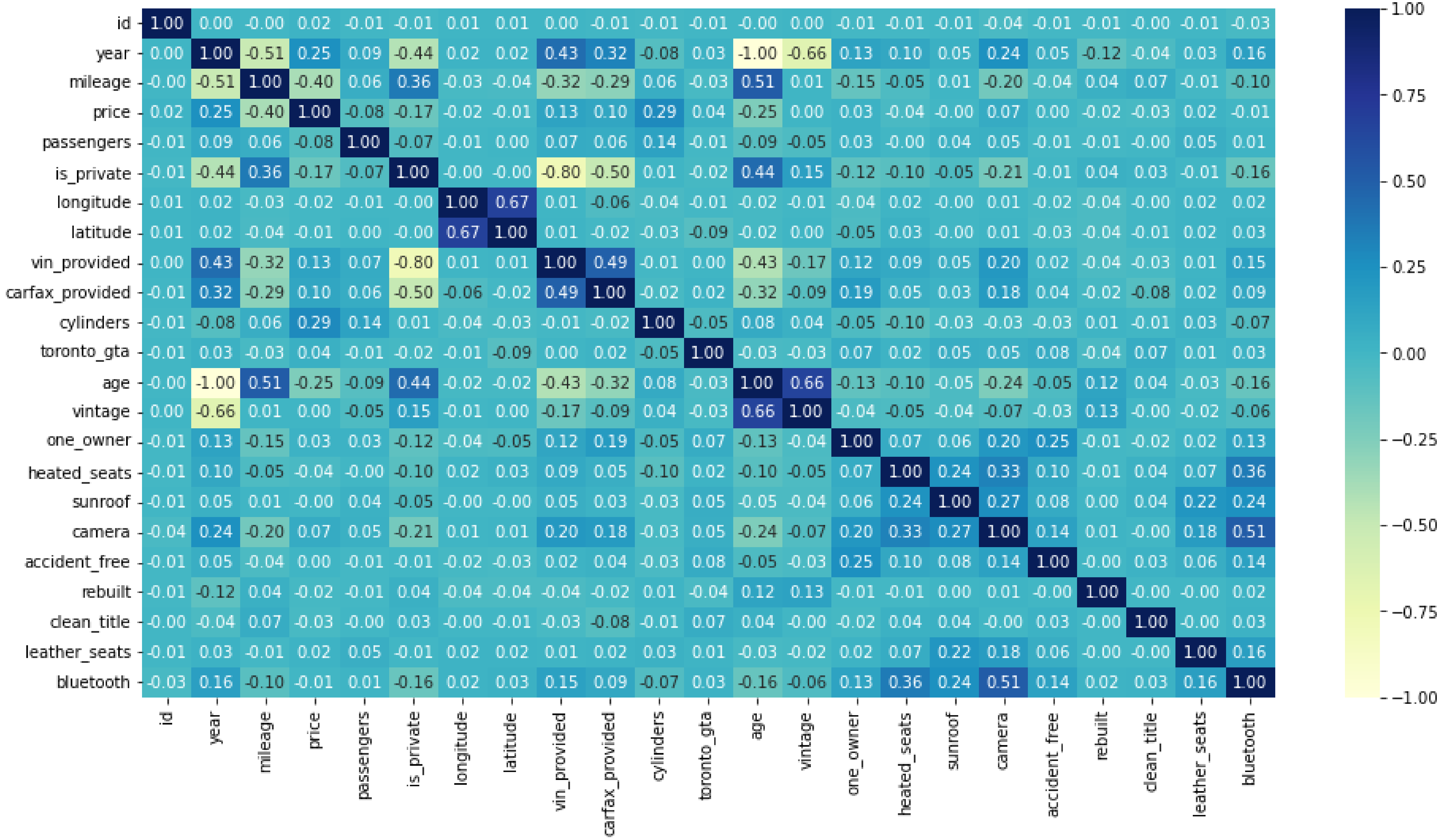
New column: Leather seats?





What fields are correlated?

Gear



Prediction modelling

Modelling: Preprocessing

Gear

01

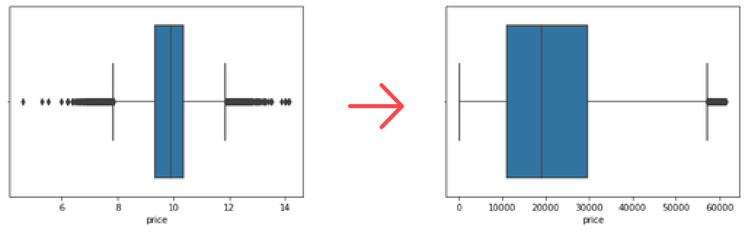
Unnecessary fields removed

- ID,
- first_date_seen
- last_date_seen
- year **(replaced by age)**
- model
- description
- longitude
- latitude
- classifications

02

Outliers removed

- Outlier removed from **price** and **mileage** fields by interquartile range filtering



03

Dummy* variables added (Total fields: 114)

- Make
- Color
- Body type
- Drive train
- Transmission
- Fuel Type

**This method assigns 1s and 0s for each class under variables in question.*

04

Variables scaled for better modelling

- When variables have different scales, it is always helpful to standardize them by subtracting the mean and then scaling to unit variance

05

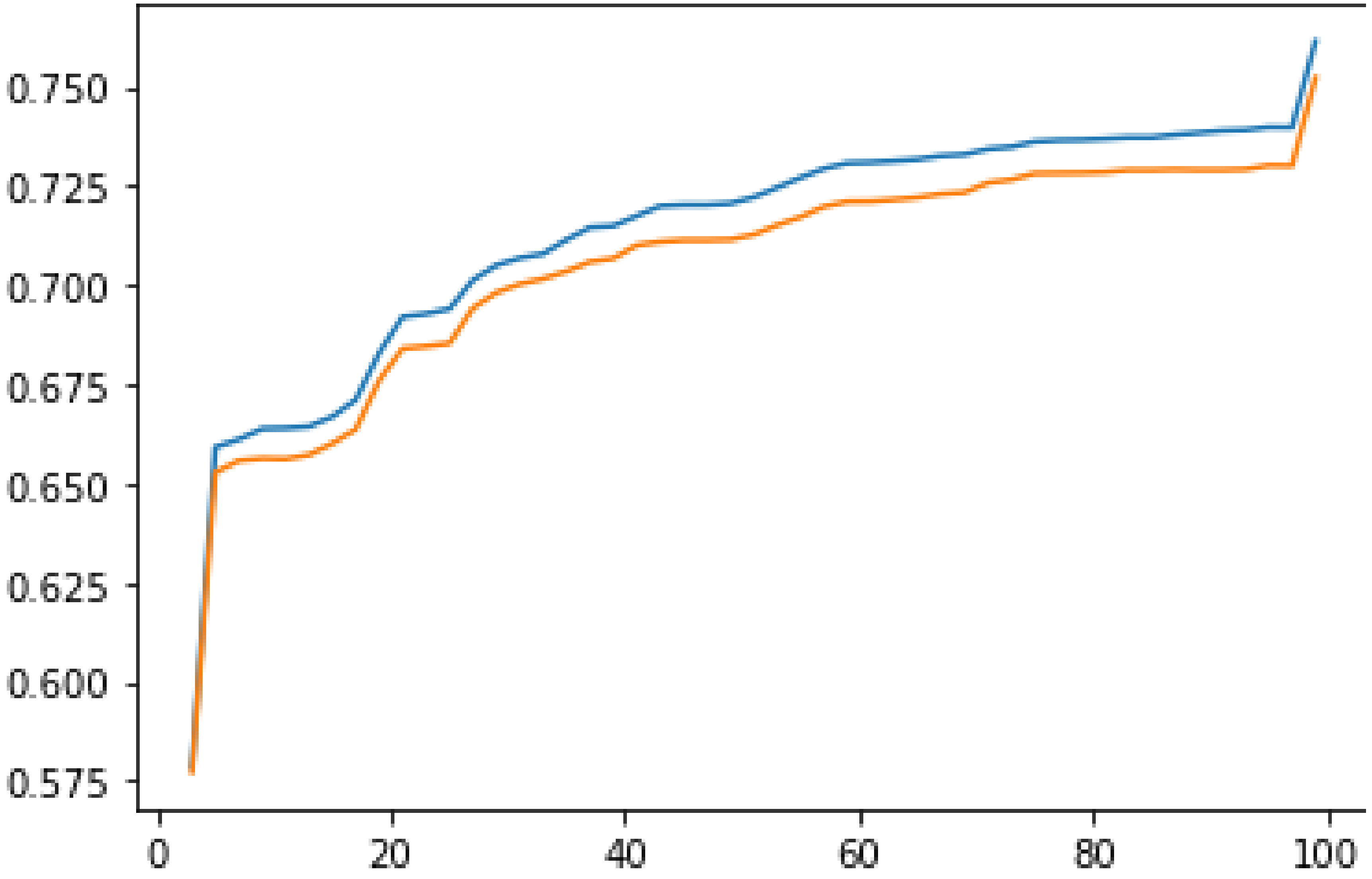
Dataset divided into test, train

- The dataset was divided into subsets, test and train by a ratio of 1/3

Modelling: Feature Selection

- **SelectKBest** from **sklearn** library is used to choose optimal number of variables

- Regression models reach an **R score** of 0.725 with around 60 variables.

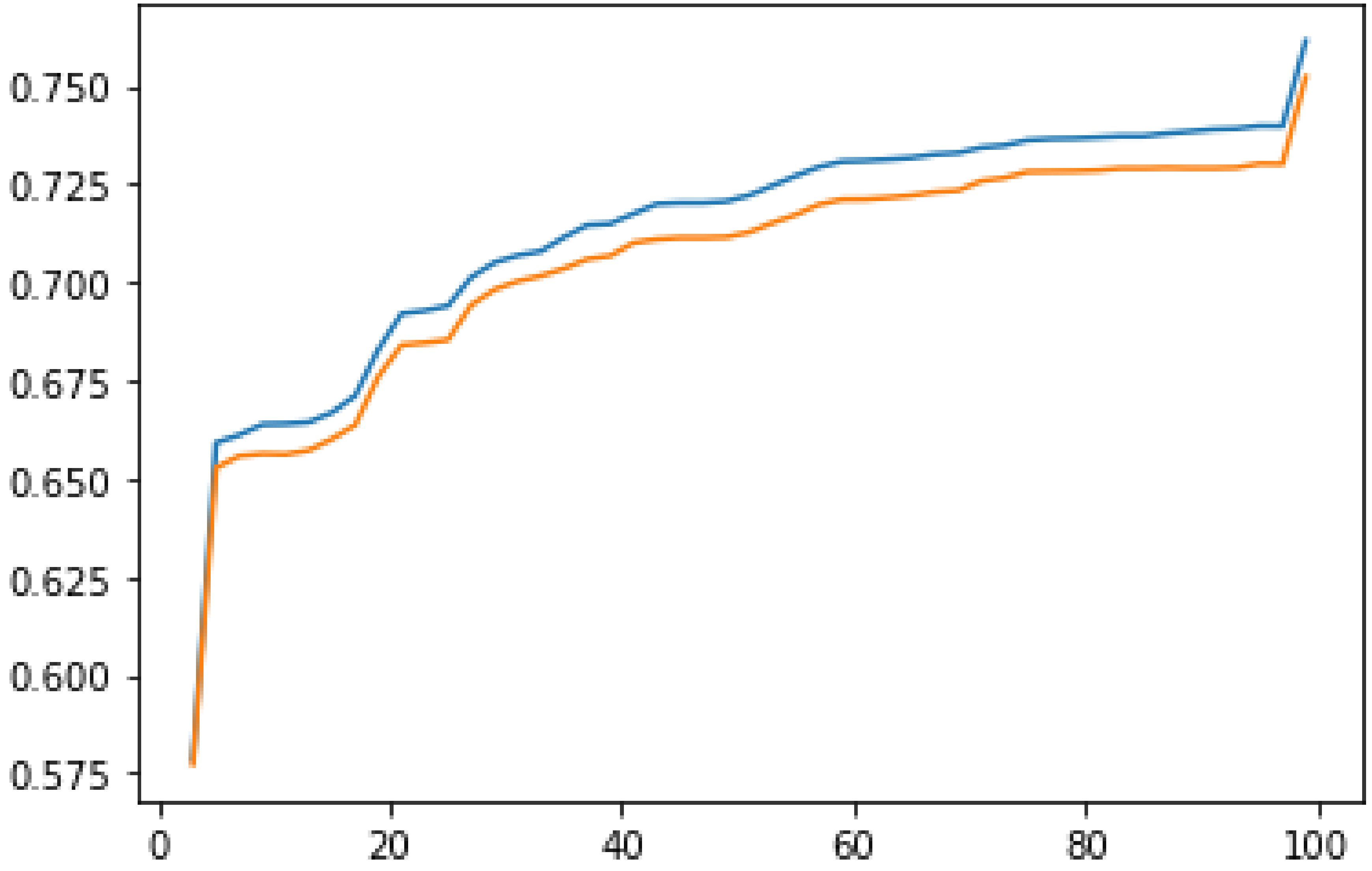


Modelling: What features are important?

Gear

- Mileage
- Passengers
- Is_private
- Vin_provided
- Carfax_provided
- Cylinders
- Age
- One_owner
- Sunroof
- Camera
- Clean title
- Leather seats
- Bluetooth
- Make (24/56)
- Color (8/19)
- Body type
- Drive train
- Transmission
- Fuel Type

- Toronto gta
- Vintage
- Heated_seats
- Accident_free
- Rebuilt
- Make (31/56)
- Color (10)



Modelling: Fitting regression

Gear

- With variables chosen in the earlier step, we fit all the available regression models to see that R score went even higher to 0.877

	Features	Model	Score
0	Linear	LinearRegression()	0.721019
1	Linear	Ridge()	0.721019
2	Linear	Lasso()	-0.000019
3	Linear	SVR()	0.864068
4	Linear	(DecisionTreeRegressor(max_features='auto', ra...	0.877071
5	Linear	MLPRegressor()	0.870174

Bringing all together

- An R score of .877 is a pretty good one. That means our model explains 88% of the price variation on used car prices. However, this is accomplished with a cleaned dataset. Real-life test would show performance better.

- If I had more time, I'd have
  - explored further transformations to increase performance;
  - deployed the model in a user interface with apps like Heroku;
  - been curious to have more historic data and account for COVID's impact on the used car market;
  - done some predictive analysis on Time to Sell;
  - looked for ways to get more recent data with non-expired CARFAX links;
  - explored opportunities to get broader geographical coverage
  - sought domain knowledge.

Gear

Q&A

Furkan Demirdoven