

# Used Car Price Analysis

Analyzing over 50,000 vehicles and building a price model

Furkan Demirdoven Data Analyst



## Business Problem

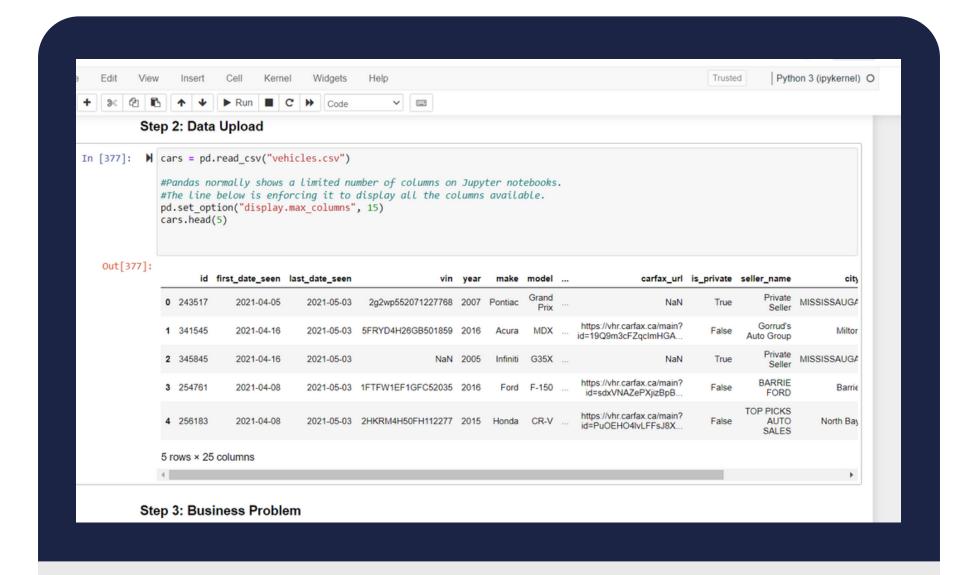
Online dealer **Gear** sits on rows of historic car data and would like to know how to get **data-driven price estimates** given that other attributes of vehicle are known.

1

Generate insights on past data

2

Leverage historic data to create predictive model





### DATASET

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



## Steps followed

Exploratory Data
Analysis (EDA)

2 Data Cleaning

3 Data Analysis

4 Modelling





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



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.

• City='Richmond'

converted to

data.

'Richmond Hill' thanks

to longitude/latitude

Mileage for that 2001
 Jetta with 999,999
 km is rendered NULL
 and than filled with
 mean mileage of all
 the 2001 Jetta listings

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



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.



## Data Analysis

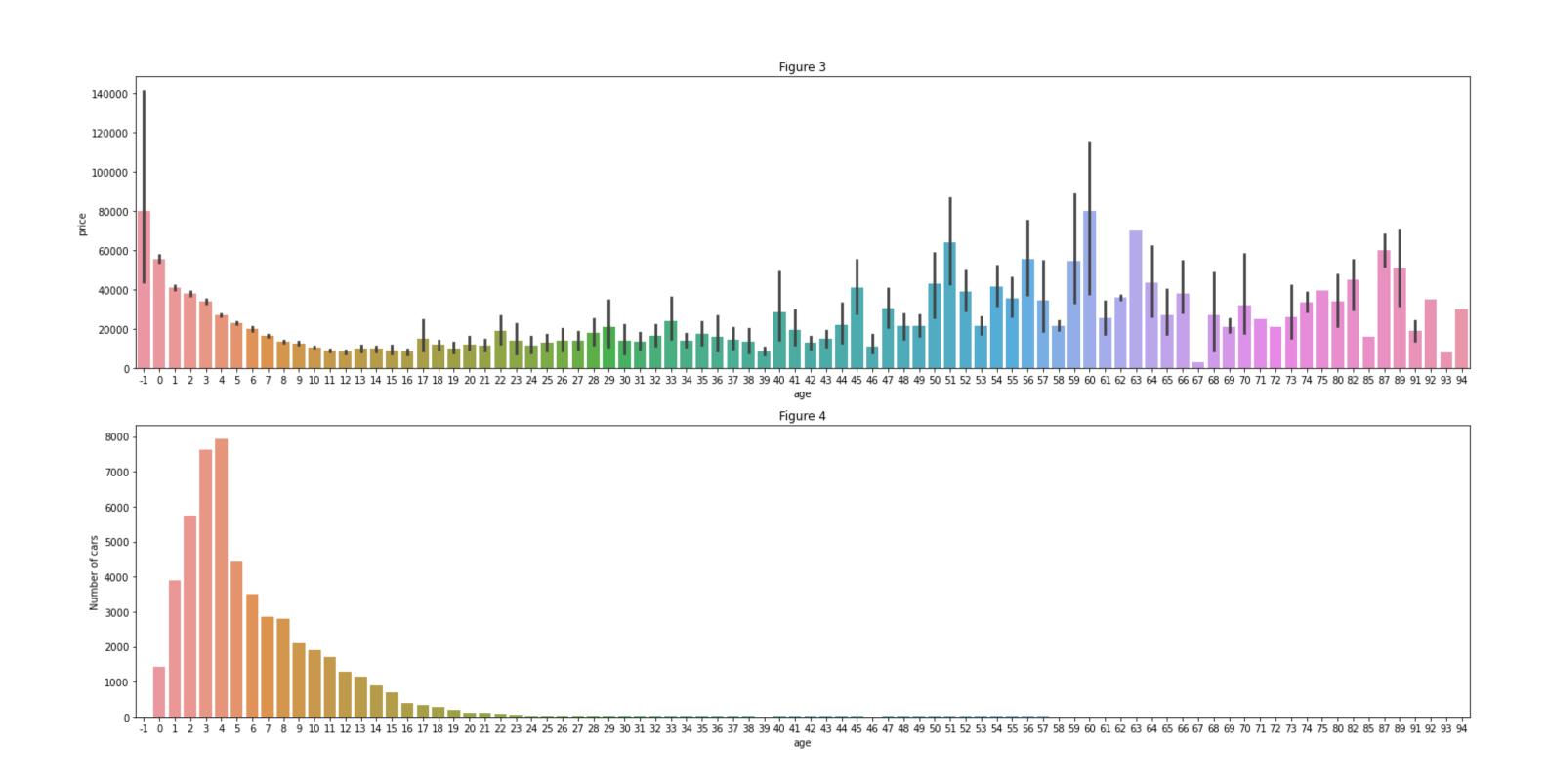
#### Make vs Price

- 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

Acura - 25.81 46.58 55.14 2.47 - Alta Rameo - 811 27.03 40.54 24.32 - Alta Rameo - 811 27.03 40.54 24.32 - Alta Martin - 0.00 0.00 87.0 91.30 - Alta Martin - 0.00 0.00 87.0 91.30 - Alta Martin - 0.00 3.70 7.41 88.89 - Bentiey - 0.00 3.70 7.41 88.89 - Bentiey - 0.00 3.70 7.41 88.89 - Bentiey - 0.00 3.70 7.41 88.89 - Buick - 25.73 44.91 18.39 10.97 - Chrysler - 40.33 46.15 13.10 0.42 - Dodge - 31.85 52.76 10.75 46.4 - Ferral - 0.00 0.00 0.00 100.00 - Ferral - 0.00 0.00 0.00 0.00 100.00 - Ferral - 0.00 0.00 0.00 0.00 100.00 - Ferral - 0.00 6.7 76.67 16.67 - Ferral - 0.00 0.00 6.67 76.67 16.67 - Ferral - 0.00 0.00 6.67 76.67 16.67 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 0.00 6.67 76.67 16.67 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 0.00 6.67 76.67 16.67 - Ferral - 22.49 47.45 21.08 7.97 - Ferral - 22.47 7.98 7.97 - Ferral - 22.55 5.50 9.50 9.50 9.50 9.50 9.50 9.50 9			of Cars in Each Pri				
Atla Rome				1	>50000		100
Aston Martin - 0.00	Acura	- 25.81	46.58	25.14	2.47	-	- 100
Audi	Alfa Romeo	- 8.11	27.03	40.54	24.32	-	
BMW   Bentley	Aston Martin	- 0.00	0.00	8.70	91.30	-	
Bentley	Audi	- 11.59	42.37	29.20	16.85	-	
Buick	BMW	- 18.28	41.68	26.36	13.68	-	
Cadillac   18.29   38.01   24.33   10.36   -   Chevrotet   25.73   44.91   18.39   10.97   -   Dodge   31.85   52.76   10.75   4.64   -   Ferral   0.00   0.00   0.00   100000   -   Fiat   62.73   30.91   6.36   0.00   -   Ford   23.49   47.46   21.08   7.97   -   GMC   11.82   39.56   34.51   14.11   -   Genesis   0.00   6.67   76.67   16.67   -   Hino   0.00   3.70   3.70   92.59   -   Honda   23.58   65.48   10.51   0.44   -   Hummer   6.25   75.00   6.25   12.50   -   Hyundai   27.65   69.12   2.96   0.27   -   Hepp   947   39.16   35.72   15.65   -   Jepp   947   39.16   35.72   15.65   -   Jepp   947   39.16   35.72   15.65   -   Lamborghini   0.00   0.00   2.78   97.22   -   Lamborghini   0.00   0.00   27.60   46.40   -   Land Rover   4.00   22.00   27.60   46.40   -   Lexus   12.06   32.46   47.05   8.43   -   Lincoln   15.16   45.13   27.80   11.91   -   MG   21.43   57.14   21.43   0.00   -   Mazerati   0.00   11.86   35.55   52.54   -   Mazda   33.48   59.25   7.12   0.15   -   Mactaren   0.00   0.00   0.00   0.00   0.00   -   Merceupy   58.82   35.29   0.00   5.88   -   Mercury   58.82   35.29   0.00   5.88   -   Mercury   58.82   35.29   0.00   5.00   -   Pontac   77.69   18.81   3.67   1.83   -   Plymouth   5.00   40.00   50.00   50.00   -   Pontac   77.69   18.81   3.67   1.83   -   Plymouth   5.00   40.00   50.00   0.00   -   Scion   38.75   60.00   0.00   0.00   -   Scion   56.71   14.29   0.00   0.00   -   Scion   56.71   14.29   0.00   0.00   -	Bentley	- 0.00	3.70	7.41	88.89	-	
Chevrolet - 25.73	Buick	- 26.82	62.72	9.47	0.99	-	
Chrysler	Cadillac	- 18.29	38.01	24.33	19.36	-	
Dodge	Chevrolet	- 25.73	44.91	18.39	10.97	-	
Dodge	Chrysler	- 40.33	46.15	13.10	0.42	-	
Ferrari - 0.00 0.00 0.00 100.00   180.0	-					-	
Fiat Ford	_			0.00		-	- 80
GMC 11.82 39.56 34.51 14.11   Genesis - 0.00 6.67 76.67 16.67 - 16.67						_	
GMC - 11.82   39.56   34.51   14.11   -   Genesis - 0.00   6.67   76.67   16.67   16.67   Hino - 0.00   3.70   3.70   92.59   -   Honda - 23.58   65.48   10.51   0.44   -   Hummer - 6.25   75.00   6.25   12.50   -   Hyundai - 27.65   69.12   2.96   0.27   -   Infiniti - 22.65   52.09   20.03   5.23   -   Jaguar - 7.60   28.40   38.80   25.20   -   Jeep - 9.47   39.16   35.72   15.65   -   Kia - 24.77   68.72   5.79   0.72   -   Lamborghini - 0.00   0.00   2.78   97.22     Lamborghini - 12.06   32.46   47.05   8.43   -   Lexus - 12.06   32.46   47.05   8.43   -   Lincoln - 15.16   45.13   27.80   11.91   -   MMNI   26.61   63.59   9.52   0.28   -   Maserati - 0.00   11.86   35.59   52.54   -   Mazarati - 0.00   11.86   35.59   52.54   -   McLaren - 0.00   0.00   0.00   100.00   Mercedes   8.19   39.98   28.89   22.94   -   Mercury   58.82   35.29   0.00   5.88   -   Mercury   58.82   35.29   0.00   5.88   -   Mercury   58.82   35.29   0.00   5.88   -   Oldsmobile   42.86   47.62   4.76   4.76   -   Cher   20.31   34.38   12.50   32.81   -   Plymouth - 5.00   40.00   50.00   5.00   -   Porsche   2.37   13.71   19.63   64.30   -   Ram   0.40   23.93   44.96   30.72   -   Ram   0.40   23.93   44.96   30.72   -   Ram   0.40   23.93   44.96   30.72   -   Sabaru   97.37   2.63   0.00   0.00   -   Sering   0.00   87.50   6.25   6.25   -   Subaru   20.80   66.53   12.59   0.08   -   Sering   0.00   23.93   32.56   65.12     Toyta   18.00   63.73   16.51   1.76   -   Volvo   28.62   16.84   34.68   19.87   -    Volvo   28.62   16.84   34.68   19.87   -						-	
Genesis - 0.00						-	
Hino - 0.00 3.70 3.70 92.59 - Honda - 23.58 65.48 10.51 0.44 - Honda - 23.58 65.48 10.51 0.44 - Honda - 23.58 65.48 10.51 0.44 - Honda - 27.65 65 69.12 2.96 0.27 - Infiniti - 22.65 52.09 20.03 5.23 - Jaguar - 7.60 28.40 38.80 52.20 - Jaguar - 7.60 28.40 38.80 25.20 - Jaguar - 7.60 27.2 - Jaguar - 7.60 27.60 46.40 27.60 2						-	
Honda						-	
Hummer - 6.25						_	
Hyundai - 27.65						_	
Infiniti						_	
Jaguar   7.60   28.40   38.80   25.20   -   -   -   -   -   -   -   -   -	-					_	
Sep   9.47   39.16   35.72   15.65						_	
Name						_	- 60
Lamborghini - 0.00						_	
Land Rover							
Lexus - 12.06	_						
Lincoln							
MINI - 26.61	U Lincoln						
MINI - 26.61	Σ MC						
Maserati         0.00         11.86         35.59         52.54         -           Mazda         -33.48         59.25         7.12         0.15         -           McLaren         0.00         0.00         0.00         100.00         -           Mercedes         8.19         39.98         28.89         22.94         -           Mercury         58.82         35.29         0.00         5.88         -           Mitsubishi         35.61         61.41         2.99         0.00         -           Nissan         27.89         65.56         5.83         0.73         -           Oldsmobile         42.86         47.62         4.76         4.76         -           Other         20.31         34.38         12.50         32.81         -           Plymouth         5.00         40.00         50.00         5.00         -           Portice         237         13.71         19.63         64.30         -           Ram         0.40         23.93         44.96         30.72         -           Rolls-Royce         0.00         13.33         6.67         0.00         0.00         -         -         20<	MO						
Mazda       33.48       59.25       7.12       0.15       -         McLaren       0.00       0.00       0.00       100.00       -         Mercedes       8.19       39.98       28.89       22.94       -         Mercury       58.82       35.29       0.00       5.88       -         Mitsubishi       35.61       61.41       2.99       0.00       -         Nissan       27.89       65.56       5.83       0.73       -         Oldsmobile       42.86       47.62       4.76       4.76       -         Other       20.31       34.38       12.50       32.81       -         Plymouth       5.00       40.00       50.00       5.00       -         Pontiac       75.69       18.81       3.67       1.83       -         Porsche       2.37       13.71       19.63       64.30       -         Ram       0.40       23.93       44.96       30.72       -         Rolls-Royce       0.00       13.33       6.67       80.00       -       -       -       20         Sab       93.33       6.67       0.00       0.00       -       - <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>							
McLaren - 0.00 0.00 0.00 100.00 40  Mercedes - 8.19 39.98 28.89 22.94 40  Mercury - 58.82 35.29 0.00 5.88							
Mercedes - 8.19         39.98         28.89         22.94         -           Mercury - 58.82         35.29         0.00         5.88         -           Mitsubishi - 35.61         61.41         2.99         0.00         -           Nissan - 27.89         65.56         5.83         0.73         -           Oldsmobile - 42.86         47.62         4.76         4.76         -           Other - 20.31         34.38         12.50         32.81         -           Plymouth - 5.00         40.00         50.00         5.00         -           Pontiac - Porsche - 2.37         13.71         19.63         64.30         -           Ram - 0.40         23.93         44.96         30.72         -           Rolls-Royce - 0.00         13.33         6.67         80.00         -         -         20           Saturn - 97.37         2.63         0.00         0.00         -         -         -         20           Scion - 38.75         60.00         0.00         1.25         -         -         -         -         -         -         -         -         -         -         -         -         -         -         -         -							
Mercury         58.82         35.29         0.00         5.88         -           Mitsubishi         35.61         61.41         2.99         0.00         -           Nissan         27.89         65.56         5.83         0.73         -           Oldsmobile         42.86         47.62         4.76         4.76         -           Other         20.31         34.38         12.50         32.81         -           Plymouth         5.00         40.00         50.00         5.00         -           Pontiac         75.69         18.81         3.67         1.83         -           Porsche         2.37         13.71         19.63         64.30         -           Ram         0.40         23.93         44.96         30.72         -           Rolls-Royce         0.00         13.33         6.67         80.00         -         -         20           Saturn         97.37         2.63         0.00         0.00         -         -         20           Sterling         0.00         87.50         6.25         6.25         -         -         -         -         20           Subaru         20							- 40
Mitsubishi - 35.61 61.41 2.99 0.00 - Nissan - 27.89 65.56 5.83 0.73 - Cldsmobile - 42.86 47.62 4.76 4.76 - Cther - 20.31 34.38 12.50 32.81 - Cther - 20.31 34.38 32.50 32.81 - Cther - 20.31 34.38 32.50 32.81 - Cther - 20.37 32.37 32.63 32.67 32.63 32.72 - Cther - 20.37 32.93 44.96 30.72 - Cther - 20.37 32.93 44.96 30.72 - Cther - 20.80 - 20.80 32.93 32.56 32.50							
Nissan - 27.89 65.56 5.83 0.73 - Oldsmobile - 42.86 47.62 4.76 4.76	-						
Oldsmobile - 42.86							
Other - 20.31							
Plymouth - 5.00							
Pontiac - 75.69 18.81 3.67 1.83 - Porsche - 2.37 13.71 19.63 64.30 - Ram - 0.40 23.93 44.96 30.72 - Rolls-Royce - 0.00 13.33 6.67 80.00 - Saab - 93.33 6.67 0.00 0.00 - Saturn - 97.37 2.63 0.00 0.00 - Scion - 38.75 60.00 0.00 1.25 - Sterling - 0.00 87.50 6.25 6.25 - Subaru - 20.80 66.53 12.59 0.08 - Suzuki - 91.67 8.33 0.00 0.00 - Tesla - 0.00 2.33 32.56 65.12 - Toyota - 18.00 63.73 16.51 1.76 - Volkswagen - 23.93 62.27 13.06 0.74 - Volvo - 28.62 16.84 34.68 19.87 - Smart - 85.71 14.29 0.00 0.00 - 50000							
Porsche - 2.37	•						
Ram - 0.40 23.93 44.96 30.72 -  Rolls-Royce - 0.00 13.33 6.67 80.00 -  Saab 93.33 6.67 0.00 0.00 -  Saturn 97.37 2.63 0.00 0.00 -  Scion 38.75 60.00 0.00 1.25 -  Sterling - 0.00 87.50 6.25 6.25 -  Subaru - 20.80 66.53 12.59 0.08 -  Suzuki 91.67 8.33 0.00 0.00 -  Tesla - 0.00 2.33 32.56 65.12 -  Toyota - 18.00 63.73 16.51 1.76 -  Volkswagen - 23.93 62.27 13.06 0.74 -  Volvo - 28.62 16.84 34.68 19.87 -  smart - 85.71 14.29 0.00 - <a href="#">&lt; 19.87</a>							
Rolls-Royce - 0.00 13.33 6.67 80.00 - 20  Saab 93.33 6.67 0.00 0.00 - 20  Saturn 97.37 2.63 0.00 0.00 - 25  Scion 38.75 60.00 0.00 1.25 - 25  Sterling - 0.00 87.50 6.25 6.25 - 25  Subaru - 20.80 66.53 12.59 0.08 - 25  Suzuki 91.67 8.33 0.00 0.00 - 25  Tesla - 0.00 2.33 32.56 65.12 - 25  Toyota - 18.00 63.73 16.51 1.76 - 25  Volkswagen - 23.93 62.27 13.06 0.74 - 25  Volvo - 28.62 16.84 34.68 19.87 - 25  smart - 85.71 14.29 0.00 - 20.00							
Saab - 93.33 6.67 0.00 0.00 20  Saturn - 97.37 2.63 0.00 0.00							
Saturn - 97.37 2.63 0.00 0.00 - 5.50 - 38.75 60.00 0.00 0.00 - 5.50 - 5.	-						- 20
Scion - Sterling - O.00     87.50     6.25     -       Subaru - Suzuki - Suzuki - Tesla - O.00     8.33     0.00     0.00     -       Tesla - O.00     2.33     32.56     65.12     -       Toyota - Volkswagen - Volvo - Smart - Volvo - Smart						-	
Sterling -     0.00     87.50     6.25     6.25     -       Subaru -     20.80     66.53     12.59     0.08     -       Suzuki -     91.67     8.33     0.00     0.00     -       Tesla -     0.00     2.33     32.56     65.12     -       Toyota -     18.00     63.73     16.51     1.76     -       Volkswagen -     23.93     62.27     13.06     0.74     -       Volvo -     28.62     16.84     34.68     19.87     -       smart -     85.71     14.29     0.00     0.00     -       <0						-	
Subaru -     20.80     66.53     12.59     0.08     -       Suzuki -     91.67     8.33     0.00     0.00     -       Tesla -     0.00     2.33     32.56     65.12     -       Toyota -     18.00     63.73     16.51     1.76     -       Volkswagen -     23.93     62.27     13.06     0.74     -       Volvo -     28.62     16.84     34.68     19.87     -       smart -     85.71     14.29     0.00     0.00     -       <10000						-	
Suzuki     91.67     8.33     0.00     0.00     -       Tesla     0.00     2.33     32.56     65.12     -       Toyota     18.00     63.73     16.51     1.76     -       Volkswagen     23.93     62.27     13.06     0.74     -       Volvo     28.62     16.84     34.68     19.87     -       smart     85.71     14.29     0.00     0.00     -       <10000	_					-	
Tesla - 0.00 2.33 32.56 65.12 - Toyota - 18.00 63.73 16.51 1.76 - Volkswagen - 23.93 62.27 13.06 0.74 - Volvo - 28.62 16.84 34.68 19.87 - smart - 85.71 14.29 0.00 0.00 - <10000 10000-30000 30000-50000 >50000						-	
Toyota - 18.00 63.73 16.51 1.76 - Volkswagen - 23.93 62.27 13.06 0.74 - Volvo - 28.62 16.84 34.68 19.87 - smart - 85.71 14.29 0.00 0.00 - <10000 10000-30000 30000-50000 >50000						-	
Volkswagen - 23.93 62.27 13.06 0.74 - Volvo - 28.62 16.84 34.68 19.87 - Smart - 85.71 14.29 0.00 0.00 - <10000 10000-30000 30000-50000 >50000						-	
Volvo - 28.62 16.84 34.68 19.87 - smart - 85.71 14.29 0.00 0.00 - <10000 10000-30000 30000-50000 >50000	-					-	
smart - 85.71 14.29 0.00 0.00 0						-	
<10000 10000-30000 30000-50000 >50000						-	
<10000 10000-30000 30000-50000 >50000	smart	85.71	14.29	0.00	0.00	-	-0
Price Bracket		<10000			>50000		Ū
			Price B	racket			

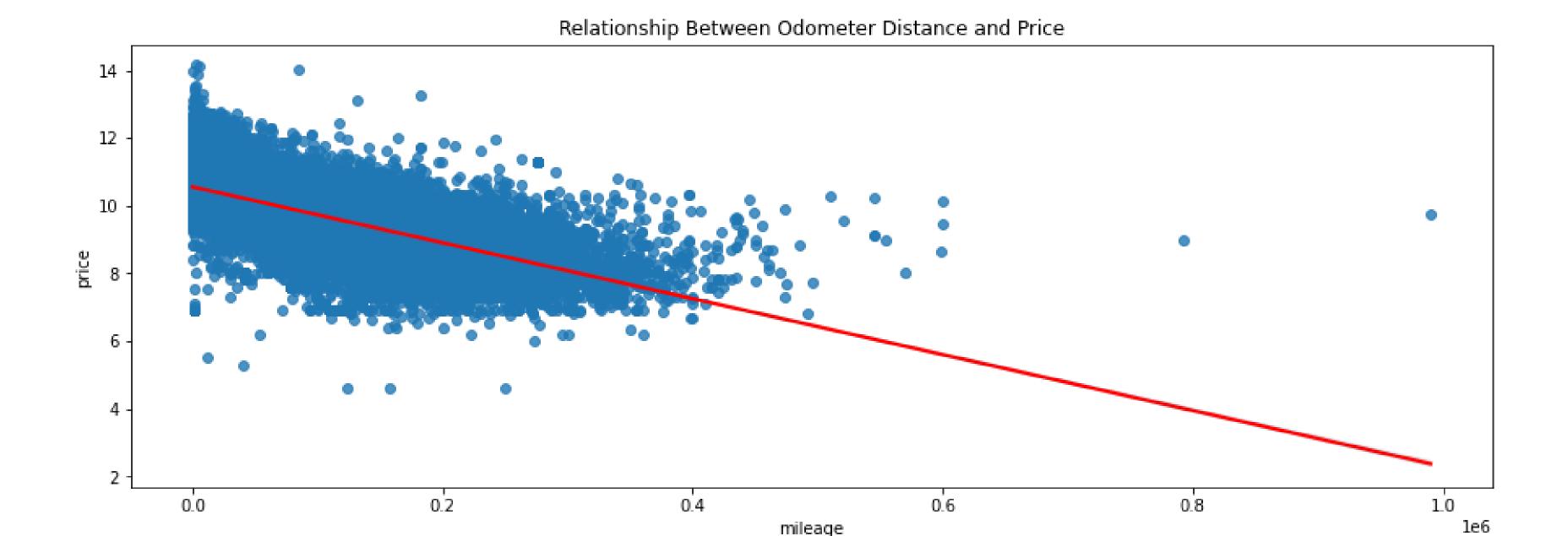
### Age vs Price





### Mileage vs Price

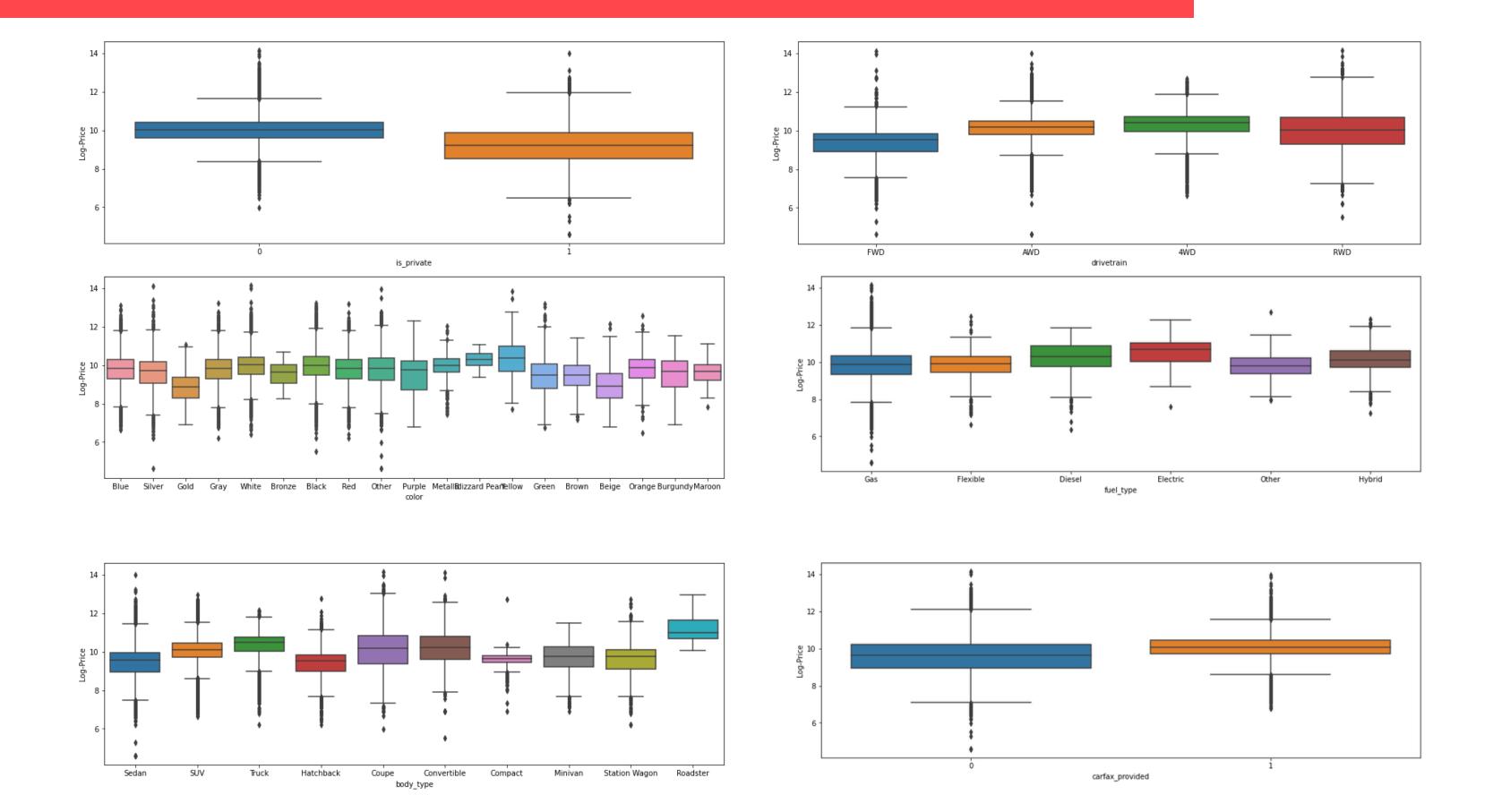




mileage

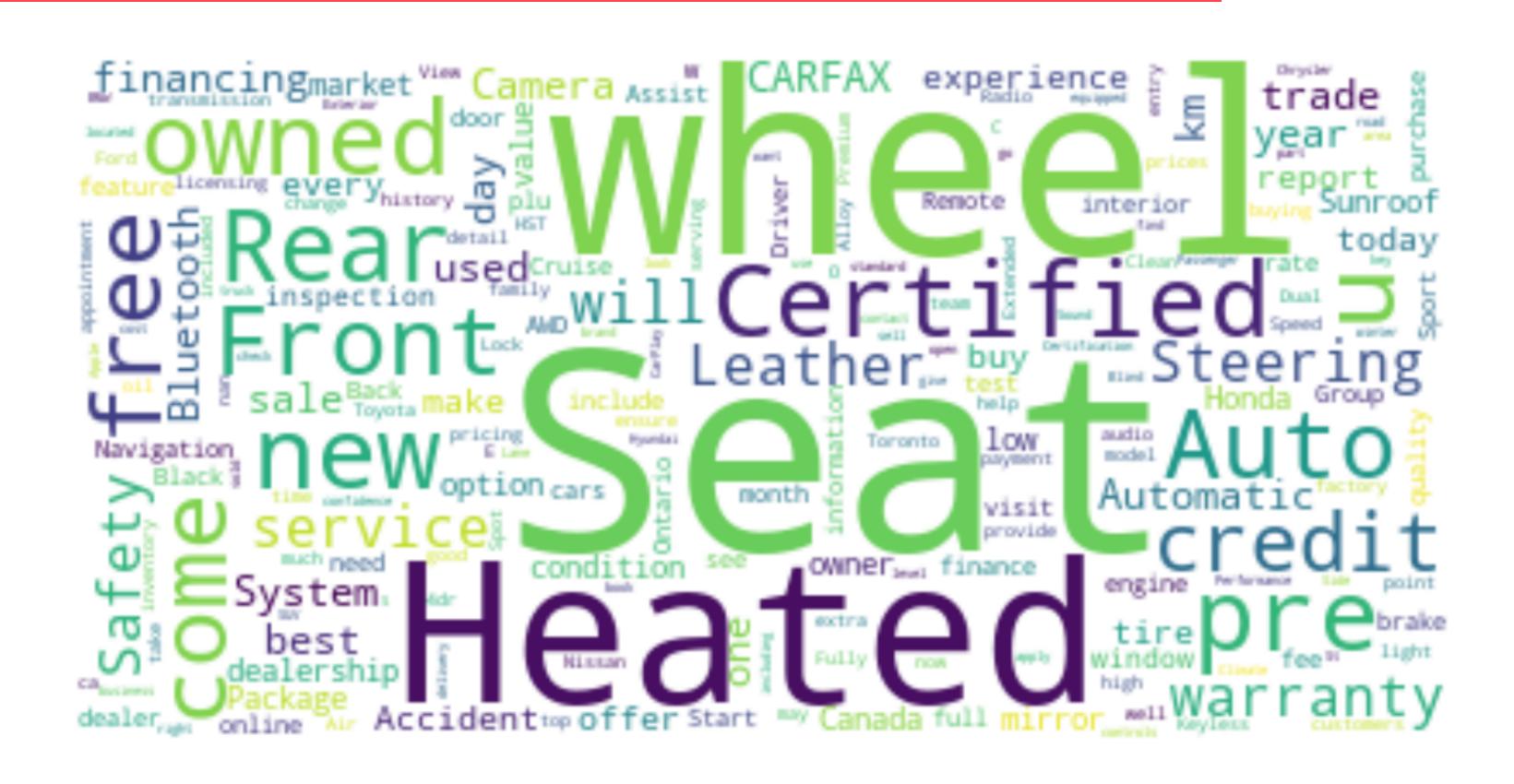
#### Various fields vs Price





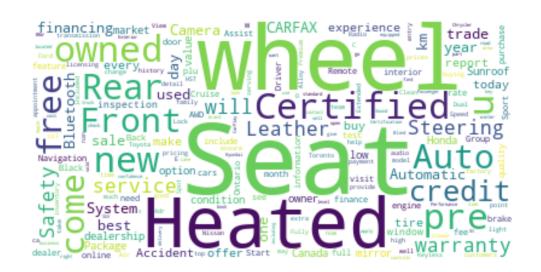
#### Improved description field



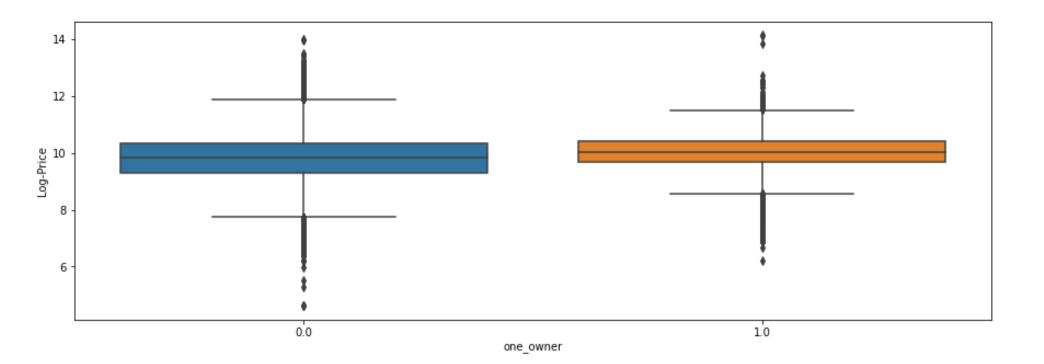


#### New column: One owner?







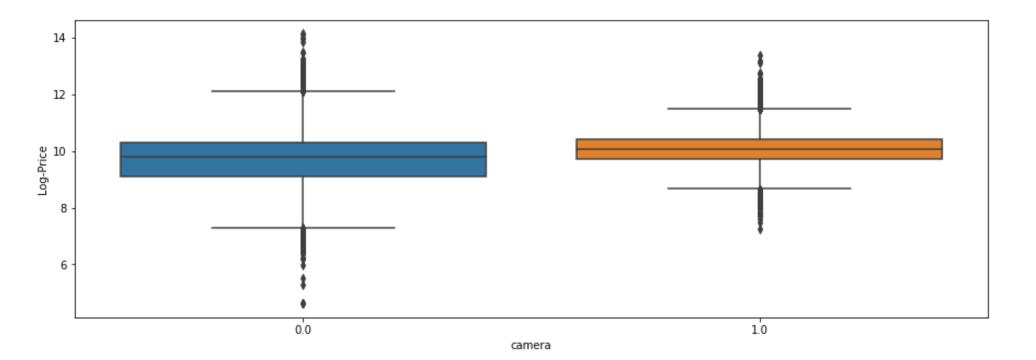


#### New column: Camera?







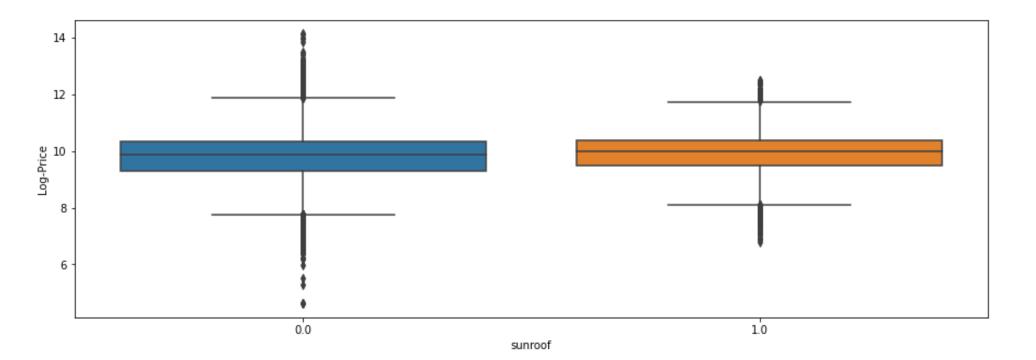


#### New column: Sunroof?



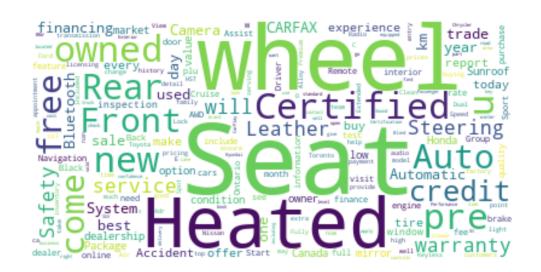




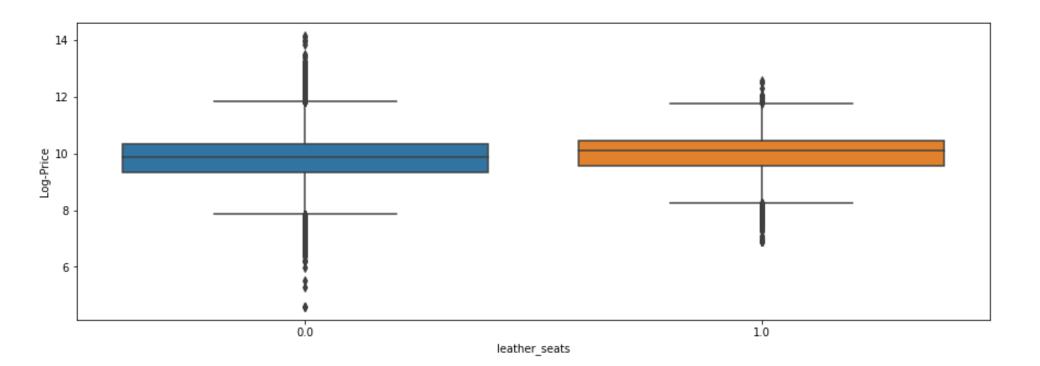


#### New column: Leather seats?



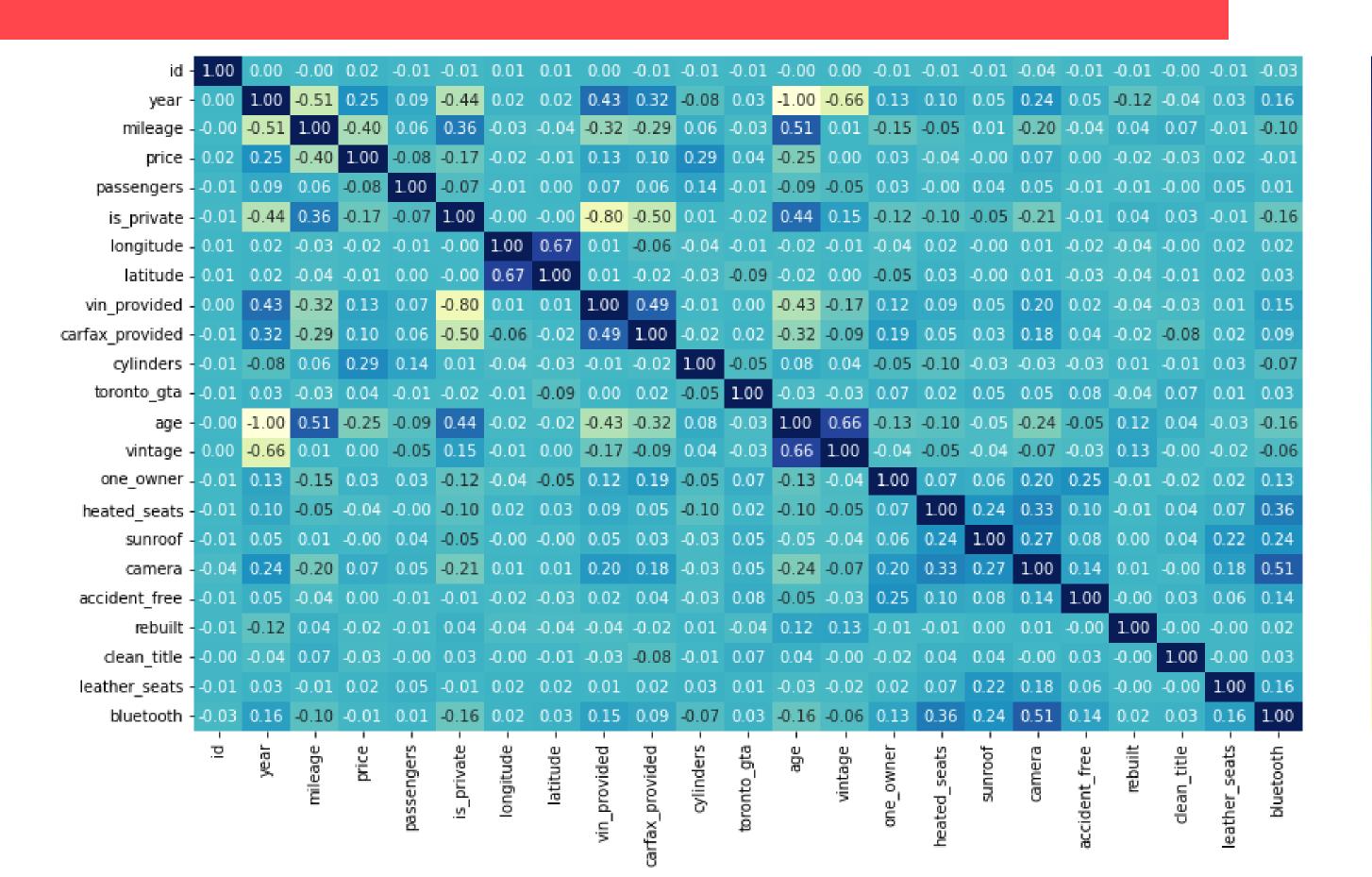






#### What fields are correlated?





- 1.00

- 0.75

- 0.50

- 0.25

- 0.00

- -0.25

- -0.50

- -0.75

- -1.00



# Prediction modelling

## Modelling: Preprocessing



01

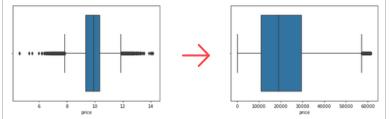
# Unnecessary fields removed

- 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

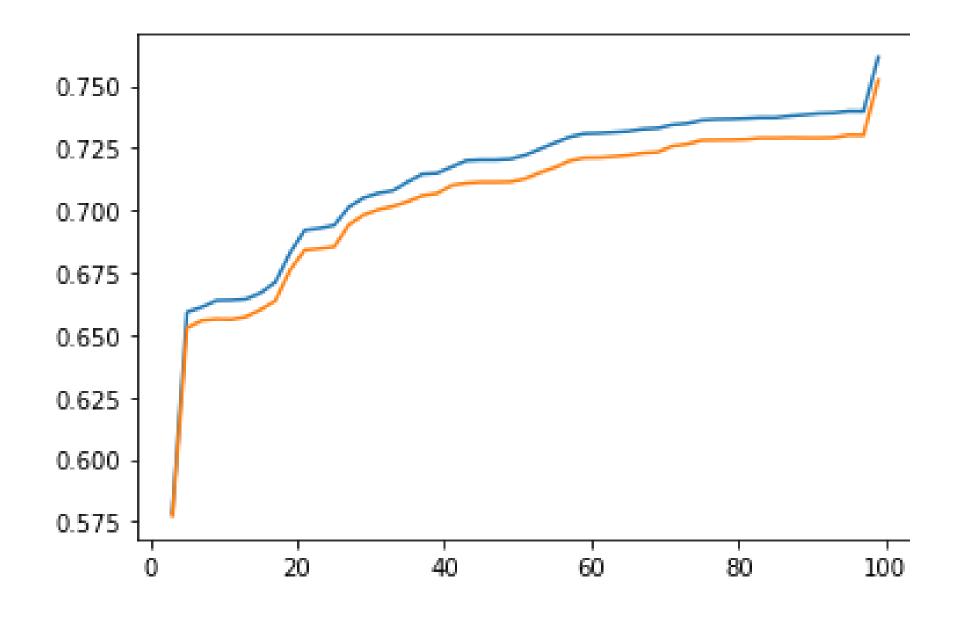
## Dataset divided into test, train

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



- 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?

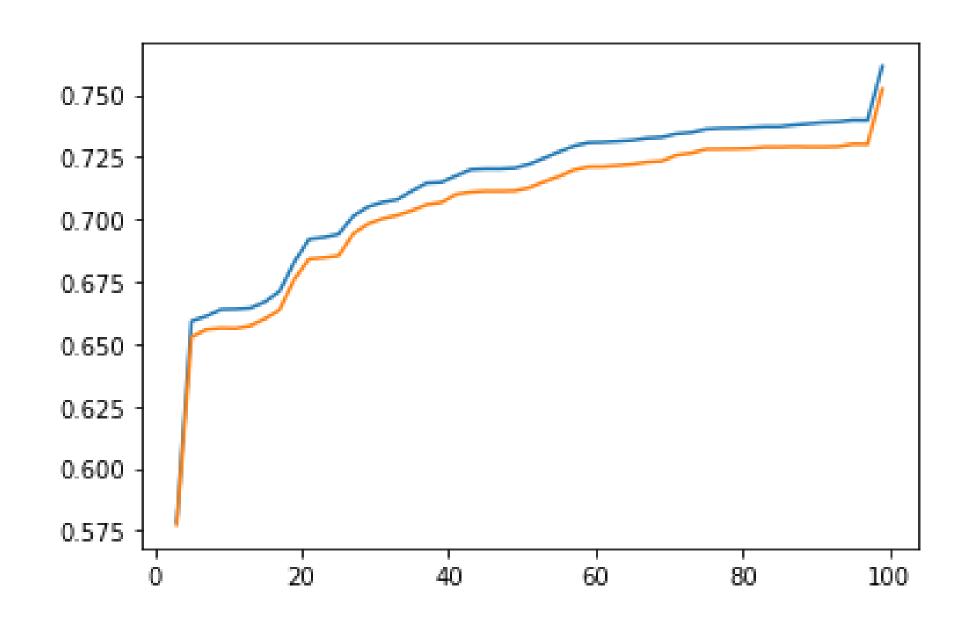




- Mileage
- Passengers
- Is\_private
- Vin\_provided
- Carfax\_provided
- Cylinders
- Age
- One\_owner
- Sunroof
- Camera
- Clean title
- Leather seats
- Bluetooth
- Make (24/56)
- Color (8/19)
- Body type
- Drive train
- Transmission
- Fuel Type



- Toronto gta
- Vintage
- Heated\_seats
- Accident\_free
- Rebuilt
- Make (31/56)
- Color (10)



## Modelling: Fitting regression

• 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.





## Q&A

**Furkan Demirdoven**