Web Scraping Project

Autotrader.ca

•••

Zack Chen / Jack Hu

Oct 20th, 2020

Overview

- Motivation
- Website
- Web Scraping
- Data Cleaning
- Stats/Trend with Data Visualization
- Challenges
- What’s next

Motivations

A great place to find out used car’s information

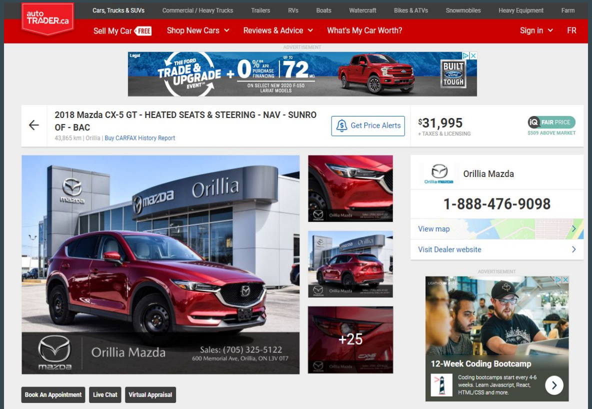
- The largest online automotive advertising website in Canada
- Over 130,000 listings for both new and used cars
- Detailed and comparable information for each listing

After web scraping class - How about build an app to notify me a great deal

After Pandas class - What insight can we find out in the market

Website

- Search filters:
  - Location: Ontario
  - Condition: Used
  - Year: >2013



2018 Mazda CX-5 GT - HEATED SEATS & STEERING - NAV - SUNROOF - BAC

43,865 km | Orillia | Buy CARFAX History Report

Get Price Alerts

$31,995
+ TAXES & LICENSING

FAIR PRICE
$509 ABOVE MARKET

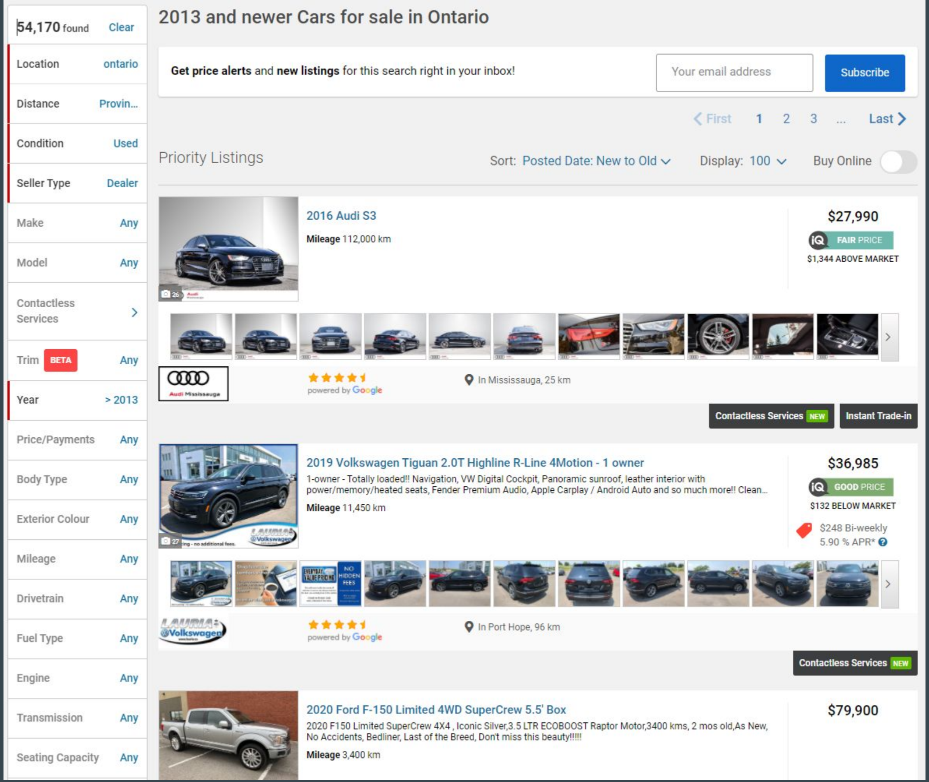
Orillia Mazda

1-888-476-9098

View map

Visit Dealer website

Book An Appointment | Live Chat | Virtual Appraisal



54,170 found | Clear

Location	ontario
Distance	Provin...
Condition	Used
Seller Type	Dealer
Make	Any
Model	Any
Contactless Services	›
Trim BETA	Any
Year	> 2013
Price/Payments	Any
Body Type	Any
Exterior Colour	Any
Mileage	Any
Drivetrain	Any
Fuel Type	Any
Engine	Any
Transmission	Any
Seating Capacity	Any

2013 and newer Cars for sale in Ontario

Get price alerts and **new listings** for this search right in your inbox!

Your email address | Subscribe

‹ First 1 2 3 ... Last ›

Priority Listings

Sort: Posted Date: New to Old | Display: 100 | Buy Online

2016 Audi S3
Mileage 112,000 km
$27,990
FAIR PRICE
$1,344 ABOVE MARKET
★★★★½ powered by Google
In Mississauga, 25 km
Contactless Services NEW | Instant Trade-in

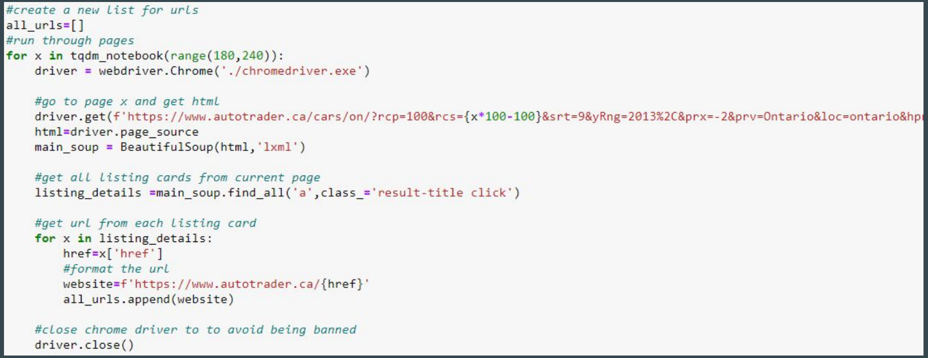
2019 Volkswagen Tiguan 2.0T Highline R-Line 4Motion - 1 owner
1-owner - Totally loaded!! Navigation, VW Digital Cockpit, Panoramic sunroof, leather interior with power/memory/heated seats, Fender Premium Audio, Apple Carplay / Android Auto and so much more!! Clean...
Mileage 11,450 km
$36,985
GOOD PRICE
$132 BELOW MARKET
$248 Bi-weekly
5.90 % APR*
★★★★½ powered by Google
In Port Hope, 96 km
Contactless Services NEW

2020 Ford F-150 Limited 4WD SuperCrew 5.5' Box
2020 F150 Limited SuperCrew 4X4 , Iconic Silver,3.5 LTR ECOBOOST Raptor Motor,3400 kms, 2 mos old,As New, No Accidents, Bedliner, Last of the Breed, Dont miss this beauty!!!!!
Mileage 3,400 km
$79,900

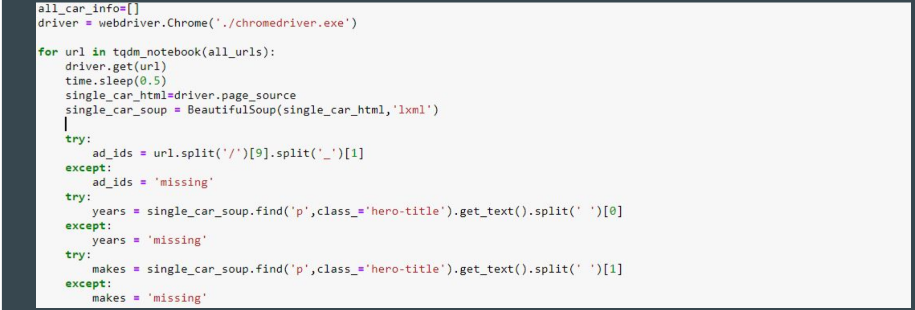
Data Scraping Code

Web Scraping

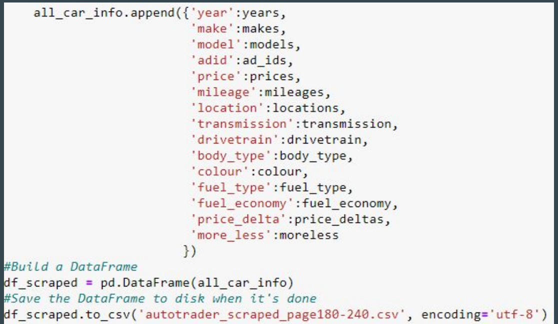
1. Selenium/BS4
2. Went through pages of listings and scraped the url of each listing
3. Used selenium to open each url and scraped detailed listing information
4. Formed scraped listing information into DataFrame and saved to disk(.csv) in partitions to prevent data loss



```python
#create a new list for urls
all_urls=[]
#run through pages
for x in tqdm_notebook(range(180,240)):
    driver = webdriver.Chrome('./chromedriver.exe')

    #go to page x and get html
    driver.get(f'https://www.autotrader.ca/cars/on/?rcp=100&rcs={x*100-100}&srt=9&yRng=2013%2C&prx=-2&prv=Ontario&loc=ontario&hp
    html=driver.page_source
    main_soup = BeautifulSoup(html,'lxml')

    #get all listing cards from current page
    listing_details =main_soup.find_all('a',class_='result-title click')

    #get url from each listing card
    for x in listing_details:
        href=x['href']
        #format the url
        website=f'https://www.autotrader.ca/{href}'
        all_urls.append(website)

    #close chrome driver to to avoid being banned
    driver.close()
```



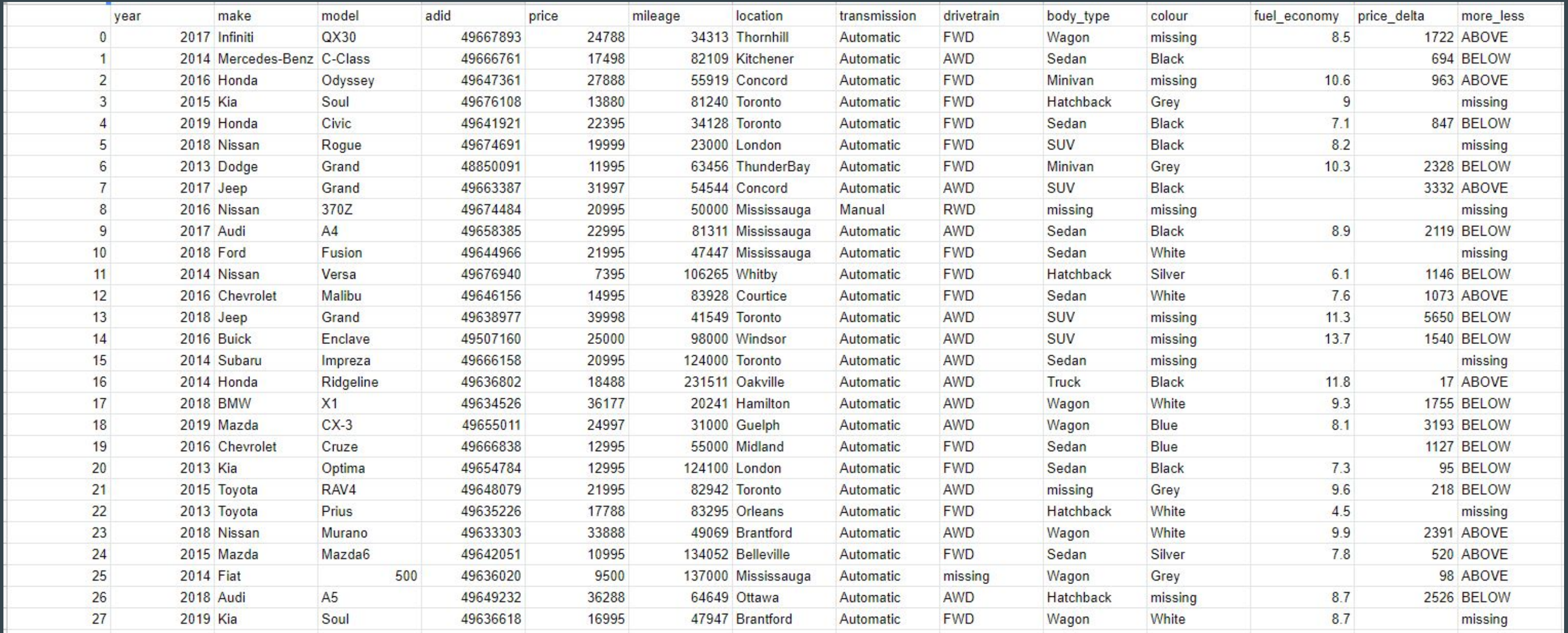
```python
all_car_info=[]
driver = webdriver.Chrome('./chromedriver.exe')

for url in tqdm_notebook(all_urls):
    driver.get(url)
    time.sleep(0.5)
    single_car_html=driver.page_source
    single_car_soup = BeautifulSoup(single_car_html,'lxml')

    try:
        ad_ids = url.split('/')[9].split('_')[1]
    except:
        ad_ids = 'missing'
    try:
        years = single_car_soup.find('p',class_='hero-title').get_text().split(' ')[0]
    except:
        years = 'missing'
    try:
        makes = single_car_soup.find('p',class_='hero-title').get_text().split(' ')[1]
    except:
        makes = 'missing'
```



```python
    all_car_info.append({'year':years,
                         'make':makes,
                         'model':models,
                         'adid':ad_ids,
                         'price':prices,
                         'mileage':mileages,
                         'location':locations,
                         'transmission':transmission,
                         'drivetrain':drivetrain,
                         'body_type':body_type,
                         'colour':colour,
                         'fuel_type':fuel_type,
                         'fuel_economy':fuel_economy,
                         'price_delta':price_deltas,
                         'more_less':moreless
                        })
#Build a DataFrame
df_scraped = pd.DataFrame(all_car_info)
#Save the DataFrame to disk when it's done
df_scraped.to_csv('autotrader_scraped_page180-240.csv', encoding='utf-8')
```

Data Collected

- Columns: 14 Rows: 11686

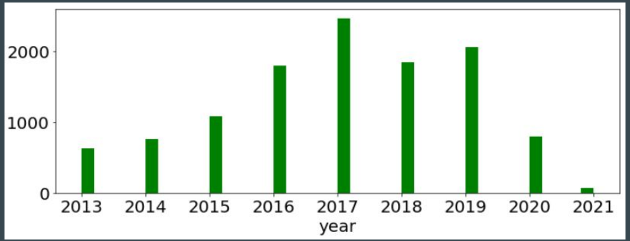


	year	make	model	adid	price	mileage	location	transmission	drivetrain	body_type	colour	fuel_economy	price_delta	more_less
0	2017	Infiniti	QX30	49667893	24788	34313	Thornhill	Automatic	FWD	Wagon	missing	8.5	1722	ABOVE
1	2014	Mercedes-Benz	C-Class	49666761	17498	82109	Kitchener	Automatic	AWD	Sedan	Black		694	BELOW
2	2016	Honda	Odyssey	49647361	27888	55919	Concord	Automatic	FWD	Minivan	missing	10.6	963	ABOVE
3	2015	Kia	Soul	49676108	13880	81240	Toronto	Automatic	FWD	Hatchback	Grey	9		missing
4	2019	Honda	Civic	49641921	22395	34128	Toronto	Automatic	FWD	Sedan	Black	7.1	847	BELOW
5	2018	Nissan	Rogue	49674691	19999	23000	London	Automatic	FWD	SUV	Black	8.2		missing
6	2013	Dodge	Grand	48850091	11995	63456	ThunderBay	Automatic	FWD	Minivan	Grey	10.3	2328	BELOW
7	2017	Jeep	Grand	49663387	31997	54544	Concord	Automatic	AWD	SUV	Black		3332	ABOVE
8	2016	Nissan	370Z	49674484	20995	50000	Mississauga	Manual	RWD	missing	missing			missing
9	2017	Audi	A4	49658385	22995	81311	Mississauga	Automatic	AWD	Sedan	Black	8.9	2119	BELOW
10	2018	Ford	Fusion	49644966	21995	47447	Mississauga	Automatic	FWD	Sedan	White			missing
11	2014	Nissan	Versa	49676940	7395	106265	Whitby	Automatic	FWD	Hatchback	Silver	6.1	1146	BELOW
12	2016	Chevrolet	Malibu	49646156	14995	83928	Courtice	Automatic	FWD	Sedan	White	7.6	1073	ABOVE
13	2018	Jeep	Grand	49638977	39998	41549	Toronto	Automatic	AWD	SUV	missing	11.3	5650	BELOW
14	2016	Buick	Enclave	49507160	25000	98000	Windsor	Automatic	AWD	SUV	missing	13.7	1540	BELOW
15	2014	Subaru	Impreza	49666158	20995	124000	Toronto	Automatic	AWD	Sedan	missing			missing
16	2014	Honda	Ridgeline	49636802	18488	231511	Oakville	Automatic	AWD	Truck	Black	11.8	17	ABOVE
17	2018	BMW	X1	49634526	36177	20241	Hamilton	Automatic	AWD	Wagon	White	9.3	1755	BELOW
18	2019	Mazda	CX-3	49655011	24997	31000	Guelph	Automatic	AWD	Wagon	Blue	8.1	3193	BELOW
19	2016	Chevrolet	Cruze	49666838	12995	55000	Midland	Automatic	FWD	Sedan	Blue		1127	BELOW
20	2013	Kia	Optima	49654784	12995	124100	London	Automatic	FWD	Sedan	Black	7.3	95	BELOW
21	2015	Toyota	RAV4	49648079	21995	82942	Toronto	Automatic	AWD	missing	Grey	9.6	218	BELOW
22	2013	Toyota	Prius	49635226	17788	83295	Orleans	Automatic	FWD	Hatchback	White	4.5		missing
23	2018	Nissan	Murano	49633303	33888	49069	Brantford	Automatic	AWD	Wagon	White	9.9	2391	ABOVE
24	2015	Mazda	Mazda6	49642051	10995	134052	Belleville	Automatic	FWD	Sedan	Silver	7.8	520	ABOVE
25	2014	Fiat	500	49636020	9500	137000	Mississauga	Automatic	missing	Wagon	Grey		98	ABOVE
26	2018	Audi	A5	49649232	36288	64649	Ottawa	Automatic	AWD	Hatchback	missing	8.7	2526	BELOW
27	2019	Kia	Soul	49636618	16995	47947	Brantford	Automatic	FWD	Wagon	White	8.7		missing

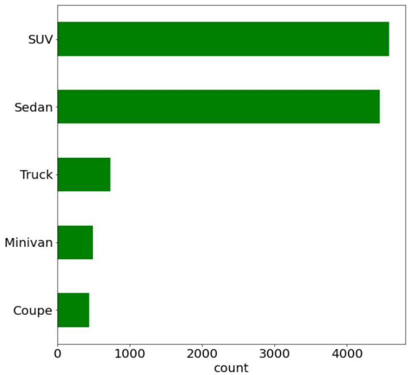
Understanding the market

Used Car Market Overview

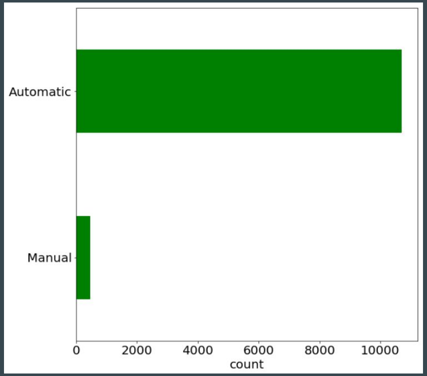
By Year



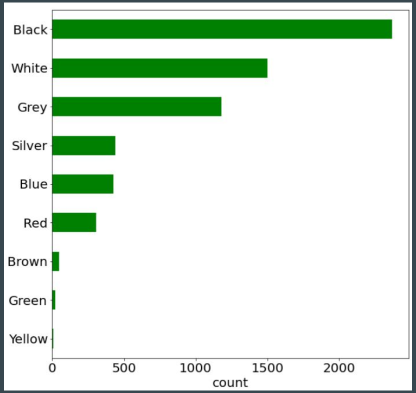
By Body Type



By Transmission

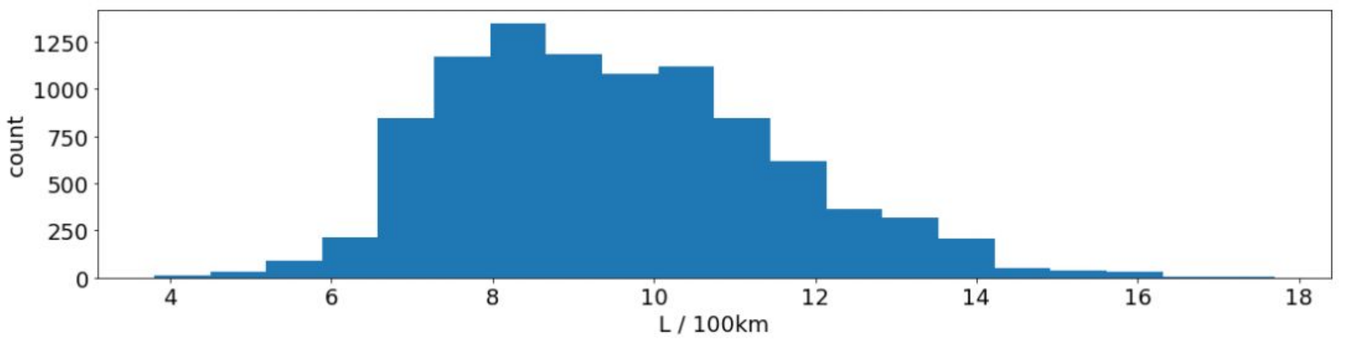


By Colors

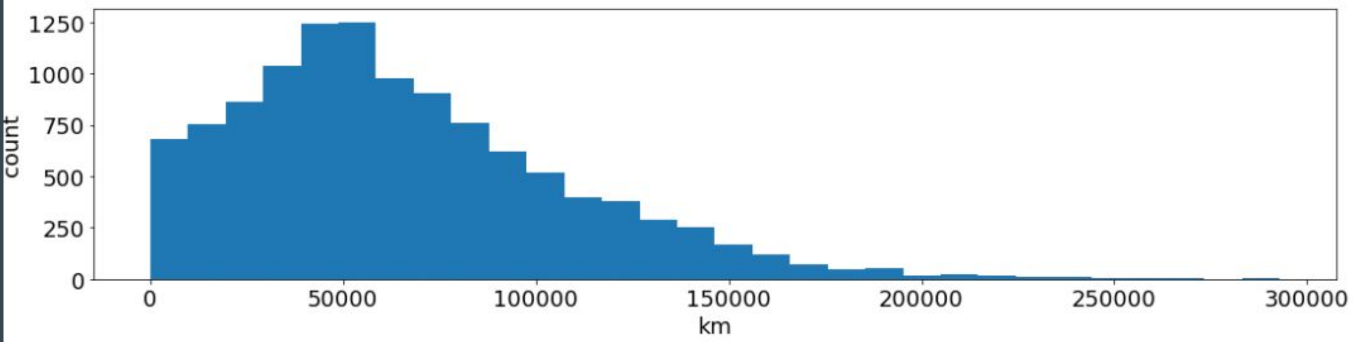


Used Car Market Overview

By Combined Fuel Economy

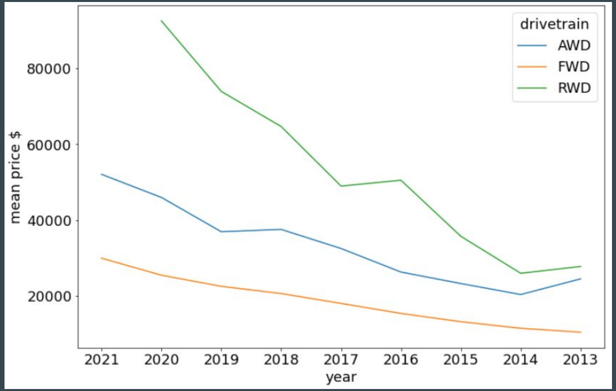


By Mileage

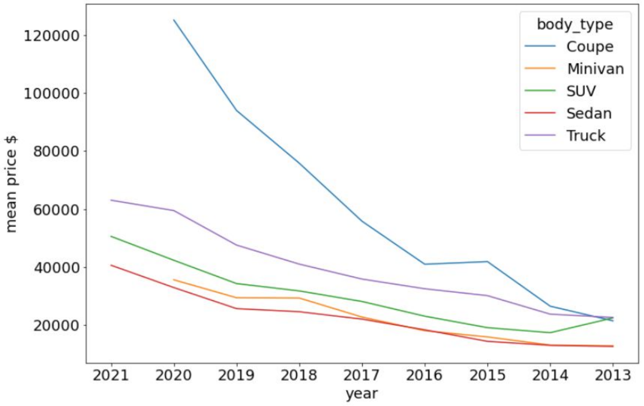


Mean Price vs. Year

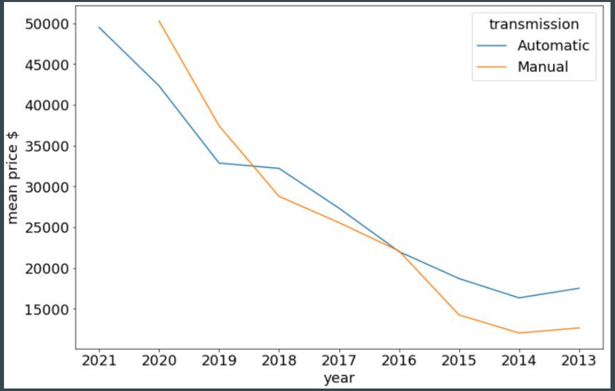
By Drivetrain



By Body Type

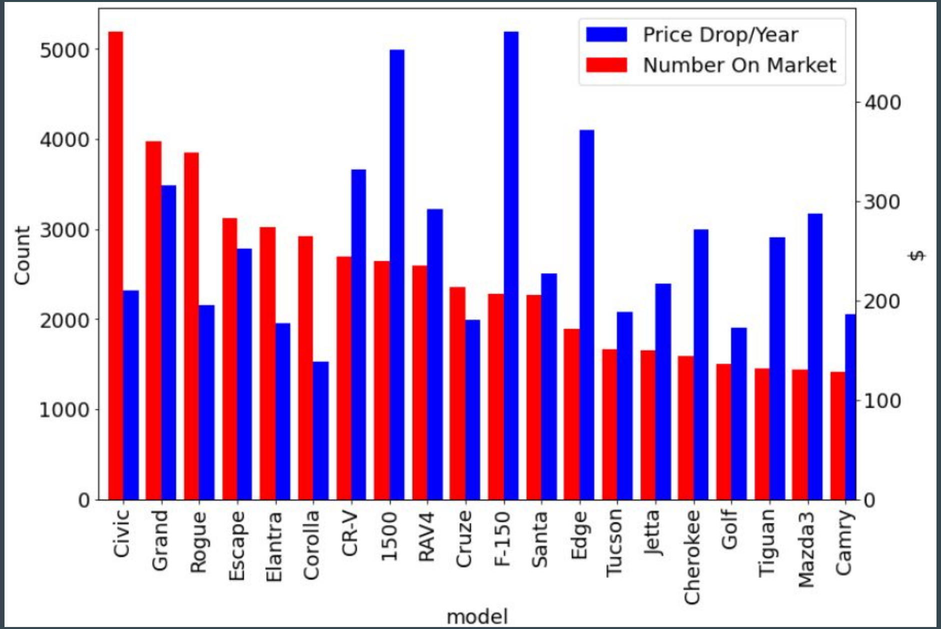


By Transmission



Depreciation Rate Among Popular Models

By Year



Best Cars That Hold Their Value:

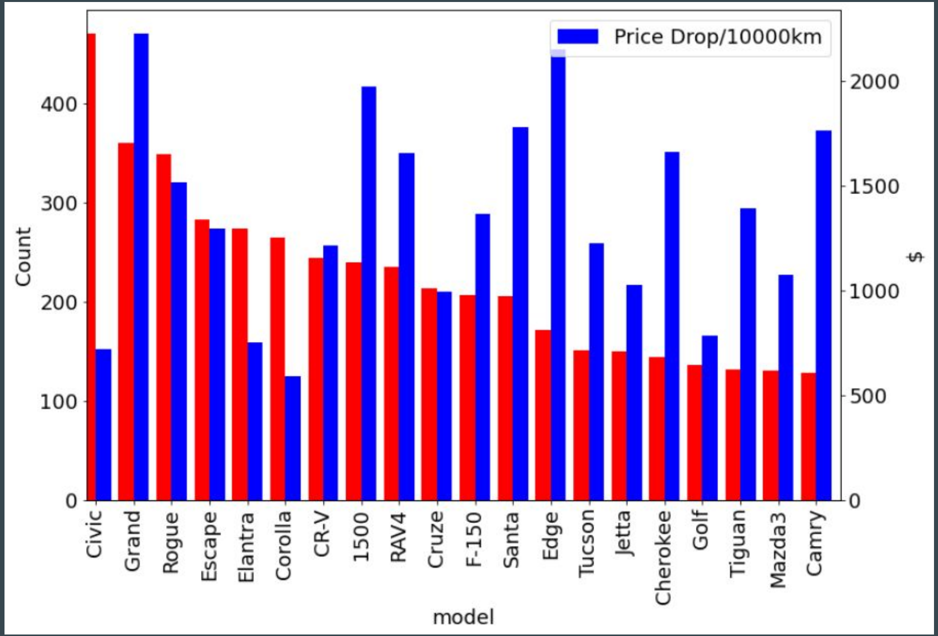
1. Toyota Corolla
2. Volkswagen Golf
3. Hyundai Elantra

Worst Cars At Holding Their Value:

4. Ford F-150
5. Ram 1500
6. Ford Edge

Depreciation Rate Among Popular Models

By Mileage



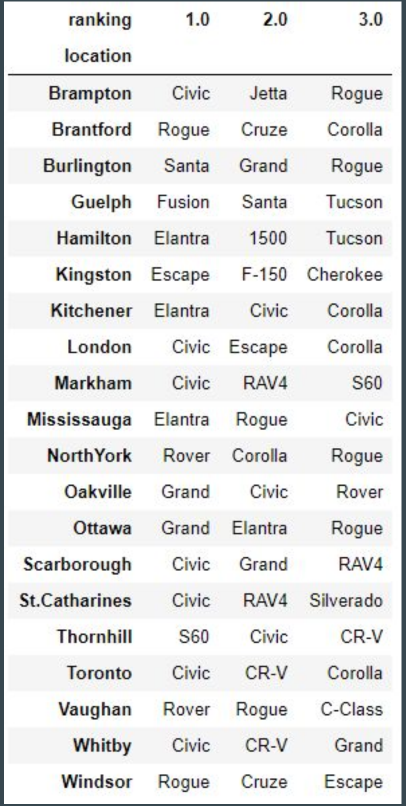
Best Cars That Hold Their Value:

1. Toyota Corolla
2. Honda Civic
3. Hyundai Elantra

Worst Cars At Holding Their Value:

4. Dodge Caravan
5. Ford Edge
6. Ram 1500

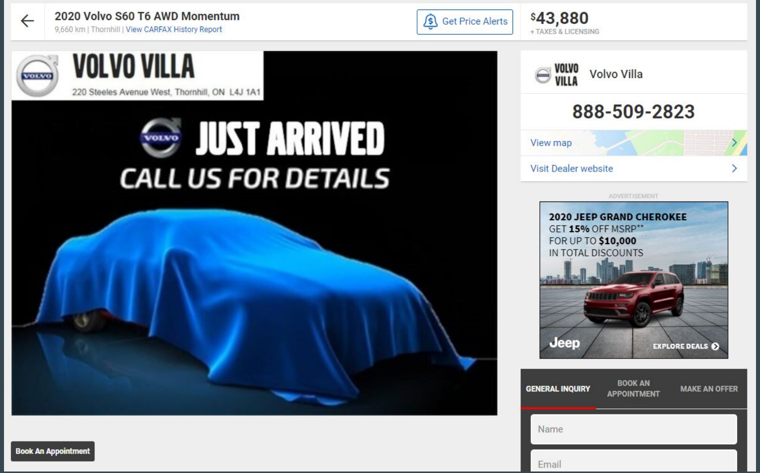
Most Models Being Sold



ranking	1.0	2.0	3.0
location			
Brampton	Civic	Jetta	Rogue
Brantford	Rogue	Cruze	Corolla
Burlington	Santa	Grand	Rogue
Guelph	Fusion	Santa	Tucson
Hamilton	Elantra	1500	Tucson
Kingston	Escape	F-150	Cherokee
Kitchener	Elantra	Civic	Corolla
London	Civic	Escape	Corolla
Markham	Civic	RAV4	S60
Mississauga	Elantra	Rogue	Civic
NorthYork	Rover	Corolla	Rogue
Oakville	Grand	Civic	Rover
Ottawa	Grand	Elantra	Rogue
Scarborough	Civic	Grand	RAV4
St.Catharines	Civic	RAV4	Silverado
Thornhill	S60	Civic	CR-V
Toronto	Civic	CR-V	Corolla
Vaughan	Rover	Rogue	C-Class
Whitby	Civic	CR-V	Grand
Windsor	Rogue	Cruze	Escape

Interesting Finding

- A Volvo dealership is selling lots of used 2019/2020 S60 model



2020 Volvo S60 T6 AWD Momentum

9,660 km | Thornhill | View CARFAX History Report

$43,880

+ TAXES & LICENSING

VOLVO VILLA

220 Steeles Avenue West, Thornhill, ON L4J 1A1

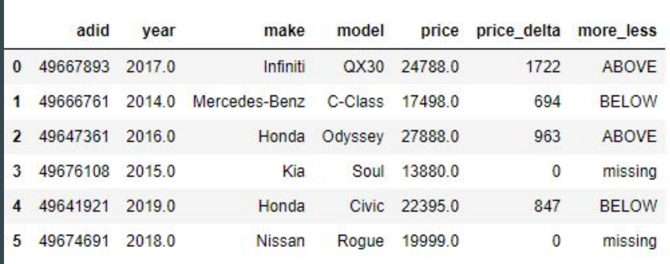
JUST ARRIVED

CALL US FOR DETAILS

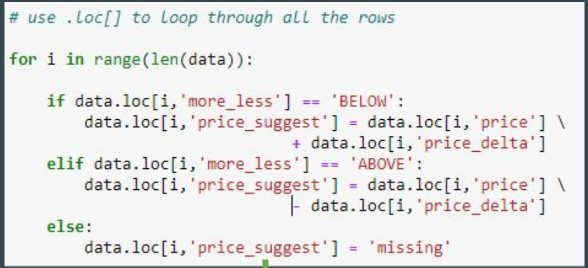
Volvo Villa

888-509-2823

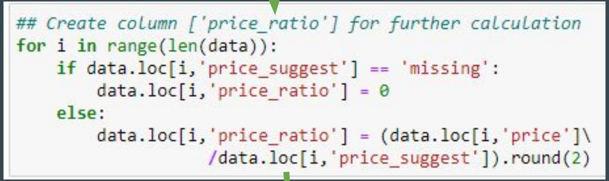
Price Comment



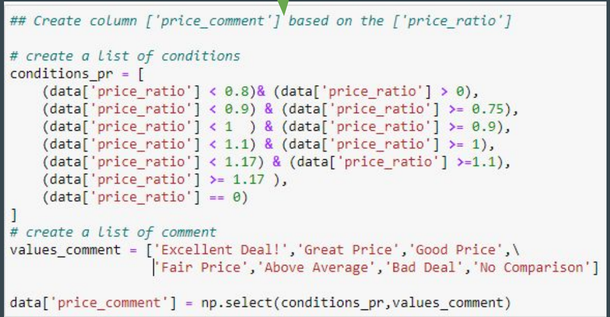
	adid	year	make	model	price	price_delta	more_less
0	49667893	2017.0	Infiniti	QX30	24788.0	1722	ABOVE
1	49666761	2014.0	Mercedes-Benz	C-Class	17498.0	694	BELOW
2	49647361	2016.0	Honda	Odyssey	27888.0	963	ABOVE
3	49676108	2015.0	Kia	Soul	13880.0	0	missing
4	49641921	2019.0	Honda	Civic	22395.0	847	BELOW
5	49674691	2018.0	Nissan	Rogue	19999.0	0	missing



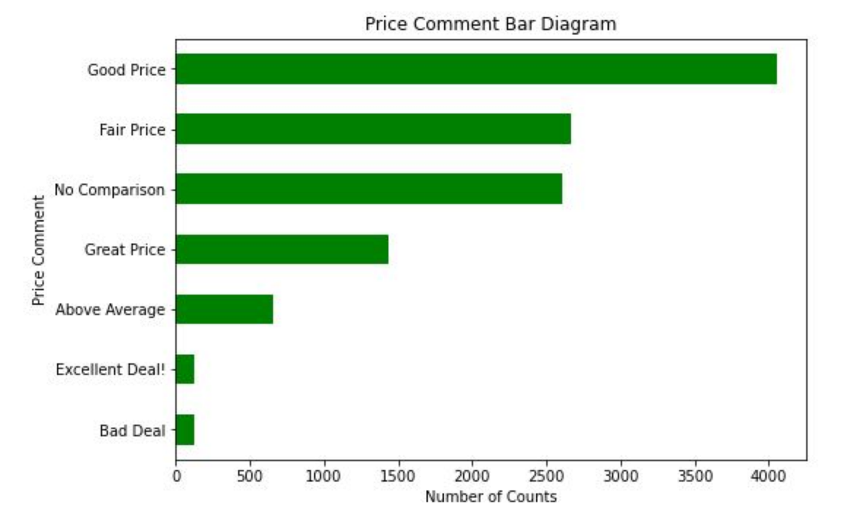
```
# use .loc[] to loop through all the rows

for i in range(len(data)):

    if data.loc[i,'more_less'] == 'BELOW':
        data.loc[i,'price_suggest'] = data.loc[i,'price'] \
                                    + data.loc[i,'price_delta']
    elif data.loc[i,'more_less'] == 'ABOVE':
        data.loc[i,'price_suggest'] = data.loc[i,'price'] \
                                    - data.loc[i,'price_delta']
    else:
        data.loc[i,'price_suggest'] = 'missing'
```



```
## Create column ['price_ratio'] for further calculation
for i in range(len(data)):
    if data.loc[i,'price_suggest'] == 'missing':
        data.loc[i,'price_ratio'] = 0
    else:
        data.loc[i,'price_ratio'] = (data.loc[i,'price']\
                                /data.loc[i,'price_suggest']).round(2)
```

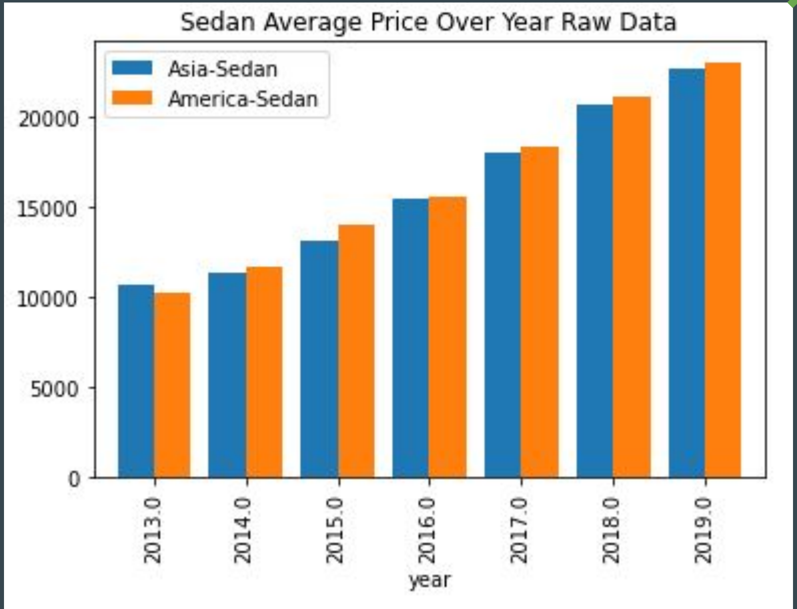


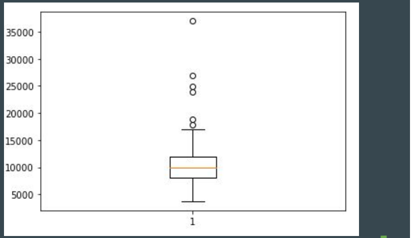
```
## Create column ['price_comment'] based on the ['price_ratio']

# create a list of conditions
conditions_pr = [
    (data['price_ratio'] < 0.8)& (data['price_ratio'] > 0),
    (data['price_ratio'] < 0.9) & (data['price_ratio'] >= 0.75),
    (data['price_ratio'] < 1   ) & (data['price_ratio'] >= 0.9),
    (data['price_ratio'] < 1.1) & (data['price_ratio'] >= 1),
    (data['price_ratio'] < 1.17) & (data['price_ratio'] >=1.1),
    (data['price_ratio'] >= 1.17 ),
    (data['price_ratio'] == 0)
]
# create a list of comment
values_comment = ['Excellent Deal!','Great Price','Good Price',\
                 'Fair Price','Above Average','Bad Deal','No Comparison']

data['price_comment'] = np.select(conditions_pr,values_comment)
```

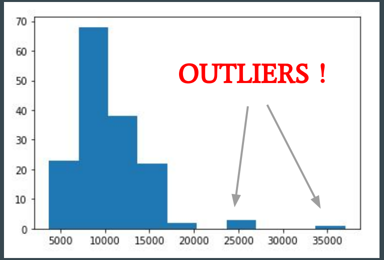


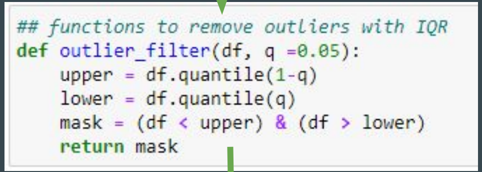
Sedan Segment Investigation

Have over 3,500 lists/rows falls under these two categories BUT no clear trend observed ?! WHY ?!

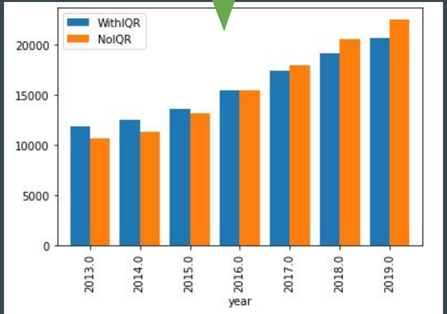

Sedan Average Price Over Year Raw Data



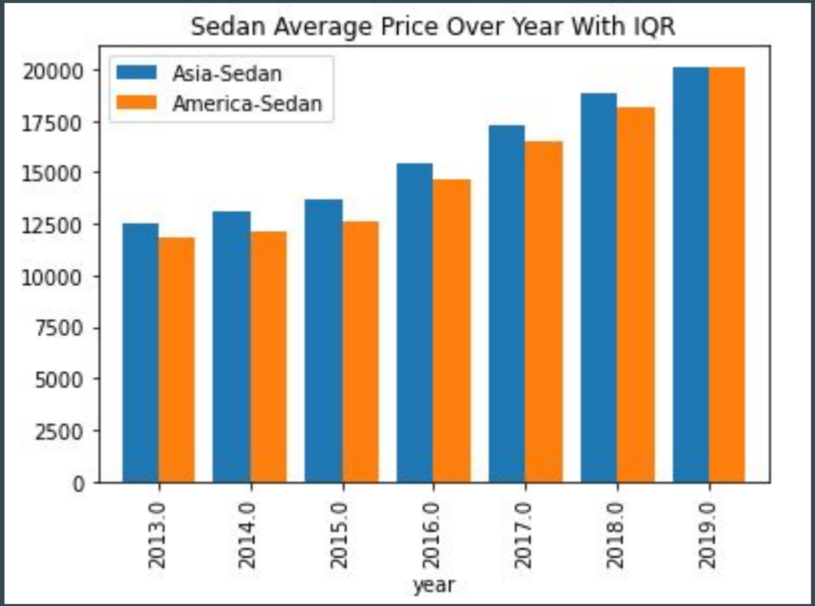

OUTLIERS !



```
## functions to remove outliers with IQR
def outlier_filter(df, q =0.05):
    upper = df.quantile(1-q)
    lower = df.quantile(q)
    mask = (df < upper) & (df > lower)
    return mask
```



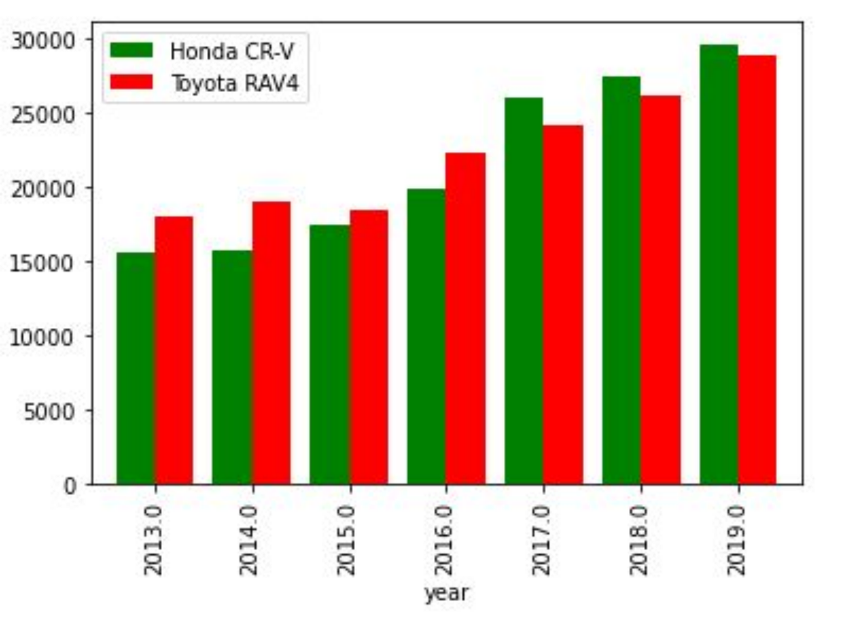
Sedan Segment Investigation Continued



The Trend now is MUCH clear!!

- Both segments starts off very closely
- Japanese/Korean car retains much value over the time
- The max price difference appears at 2015(5 year-old car)
- The older the car gets, the less price difference among the two segments

SUV Segment Investigation - Showcase CR-V vs. RAV4



The Trend

- CR-V retains better value than RAV4 between 2017 - 2019
- BIG price dip for CR-V in 2016 models. WHY?
- RAV4 retains better value than CR-V between 2013 - 2016
- Two of the best models on retaining its values

Challenges

BS4 & Selenium

- Bypass the bot
- Automation & progress tracking

Data Munging

- Data type convert and fillna() with what value
- Effectively loop through the entire dataframe (iterrow() vs. np.select vs. iloc[])

Find out what data is try to tell us

- First what features to create to gain more insight
- Outlier
- You actually need understand the market to know where to look

What is NEXT?

❏ Scrap more data!

❏ Develop an user interactive front-end app

❏ Find out more insight with more data accumulated over the time

❏ Develop some machine learning model to help us predict any used car market worth

❏ Get better at Data Science :)