Enhancing the Vehicle Data

Kafle, Lee, Siddique

- **Collecting Additional Attributes for Individual Cars:**

Sources used → edmunds

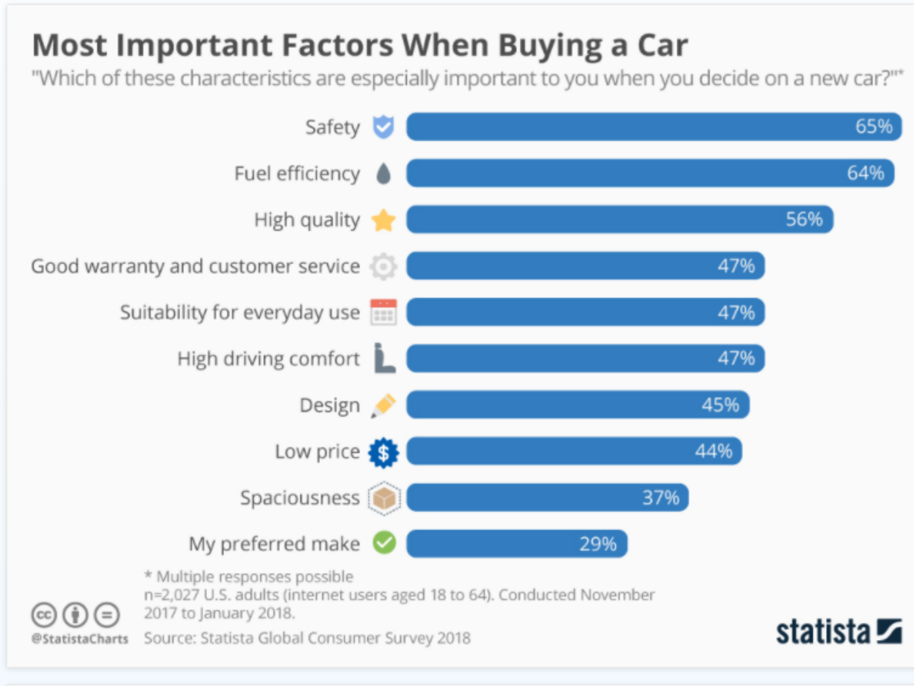
→ statista

- **Method Used for data collection:**

1) Python Web Scraping
2) Excel

- **Important Parameters Collected:**

1) Price
2) Seating Capacity
3) MPG
4) Warranty Year
5) Warranty Mileage
6) Cargo Capacity including seats
7) Fuel Capacity
8) Car Rating



Most Important Factors When Buying a Car

"Which of these characteristics are especially important to you when you decide on a new car?"*

* Multiple responses possible

n=2,027 U.S. adults (internet users aged 18 to 64). Conducted November 2017 to January 2018.

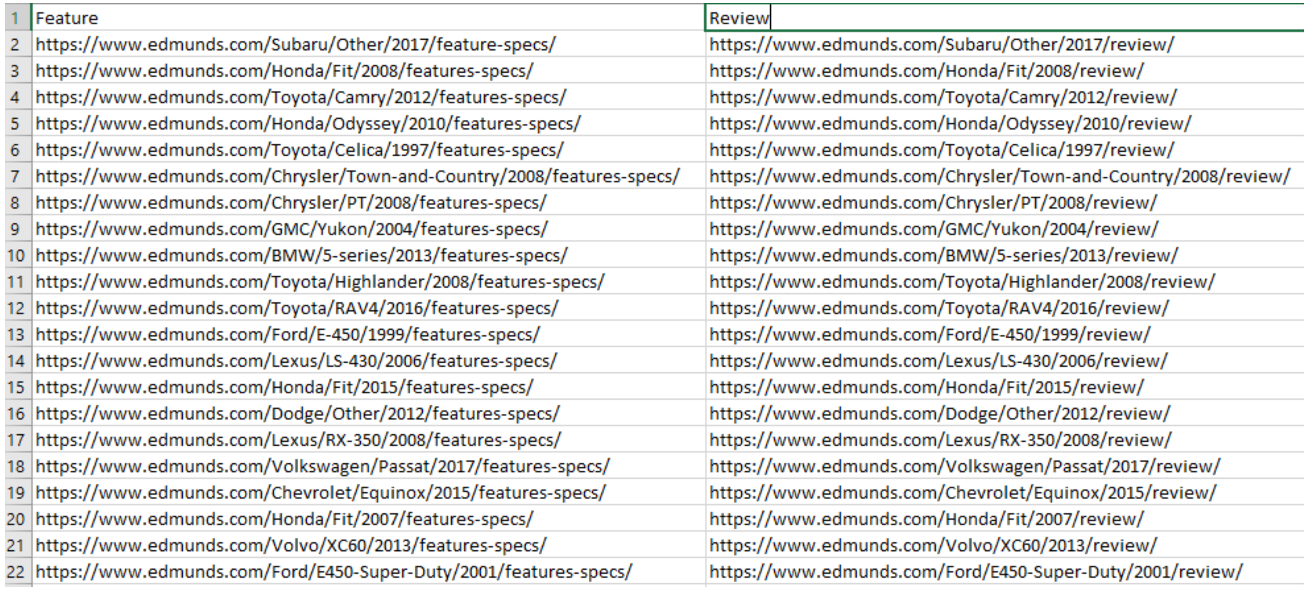
Source: Statista Global Consumer Survey 2018

Web Scrapping Vehicle Features & Review Data

Kafle, Lee, Siddique

- **Generating links for all the 765 vehicles using Python:**

  1) **Identifying unique vehicles - 623**
  2) **Generating two separate sets of links**
     a. Features
     b. Review



1	Feature	Review
2	https://www.edmunds.com/Subaru/Other/2017/feature-specs/	https://www.edmunds.com/Subaru/Other/2017/review/
3	https://www.edmunds.com/Honda/Fit/2008/features-specs/	https://www.edmunds.com/Honda/Fit/2008/review/
4	https://www.edmunds.com/Toyota/Camry/2012/features-specs/	https://www.edmunds.com/Toyota/Camry/2012/review/
5	https://www.edmunds.com/Honda/Odyssey/2010/features-specs/	https://www.edmunds.com/Honda/Odyssey/2010/review/
6	https://www.edmunds.com/Toyota/Celica/1997/features-specs/	https://www.edmunds.com/Toyota/Celica/1997/review/
7	https://www.edmunds.com/Chrysler/Town-and-Country/2008/features-specs/	https://www.edmunds.com/Chrysler/Town-and-Country/2008/review/
8	https://www.edmunds.com/Chrysler/PT/2008/features-specs/	https://www.edmunds.com/Chrysler/PT/2008/review/
9	https://www.edmunds.com/GMC/Yukon/2004/features-specs/	https://www.edmunds.com/GMC/Yukon/2004/review/
10	https://www.edmunds.com/BMW/5-series/2013/features-specs/	https://www.edmunds.com/BMW/5-series/2013/review/
11	https://www.edmunds.com/Toyota/Highlander/2008/features-specs/	https://www.edmunds.com/Toyota/Highlander/2008/review/
12	https://www.edmunds.com/Toyota/RAV4/2016/features-specs/	https://www.edmunds.com/Toyota/RAV4/2016/review/
13	https://www.edmunds.com/Ford/E-450/1999/features-specs/	https://www.edmunds.com/Ford/E-450/1999/review/
14	https://www.edmunds.com/Lexus/LS-430/2006/features-specs/	https://www.edmunds.com/Lexus/LS-430/2006/review/
15	https://www.edmunds.com/Honda/Fit/2015/features-specs/	https://www.edmunds.com/Honda/Fit/2015/review/
16	https://www.edmunds.com/Dodge/Other/2012/features-specs/	https://www.edmunds.com/Dodge/Other/2012/review/
17	https://www.edmunds.com/Lexus/RX-350/2008/features-specs/	https://www.edmunds.com/Lexus/RX-350/2008/review/
18	https://www.edmunds.com/Volkswagen/Passat/2017/features-specs/	https://www.edmunds.com/Volkswagen/Passat/2017/review/
19	https://www.edmunds.com/Chevrolet/Equinox/2015/features-specs/	https://www.edmunds.com/Chevrolet/Equinox/2015/review/
20	https://www.edmunds.com/Honda/Fit/2007/features-specs/	https://www.edmunds.com/Honda/Fit/2007/review/
21	https://www.edmunds.com/Volvo/XC60/2013/features-specs/	https://www.edmunds.com/Volvo/XC60/2013/review/
22	https://www.edmunds.com/Ford/E450-Super-Duty/2001/features-specs/	https://www.edmunds.com/Ford/E450-Super-Duty/2001/review/

Features
[1-7]

Revie
w
[8]

- **Important Parameters Collected:**

Attributes from Manufacturer

1) Price
2) Seating Capacity
3) MPG
4) Warranty Year
5) Warranty Mileage
6) Cargo Capacity including seats
7) Fuel Capacity

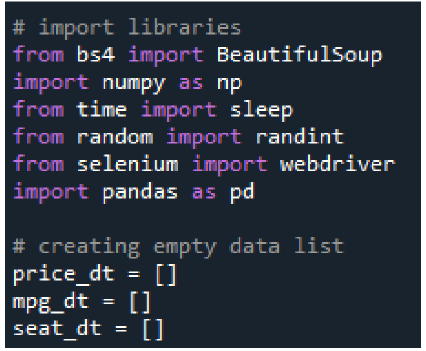
Pristine Real-Time Attributes

8) Car Rating

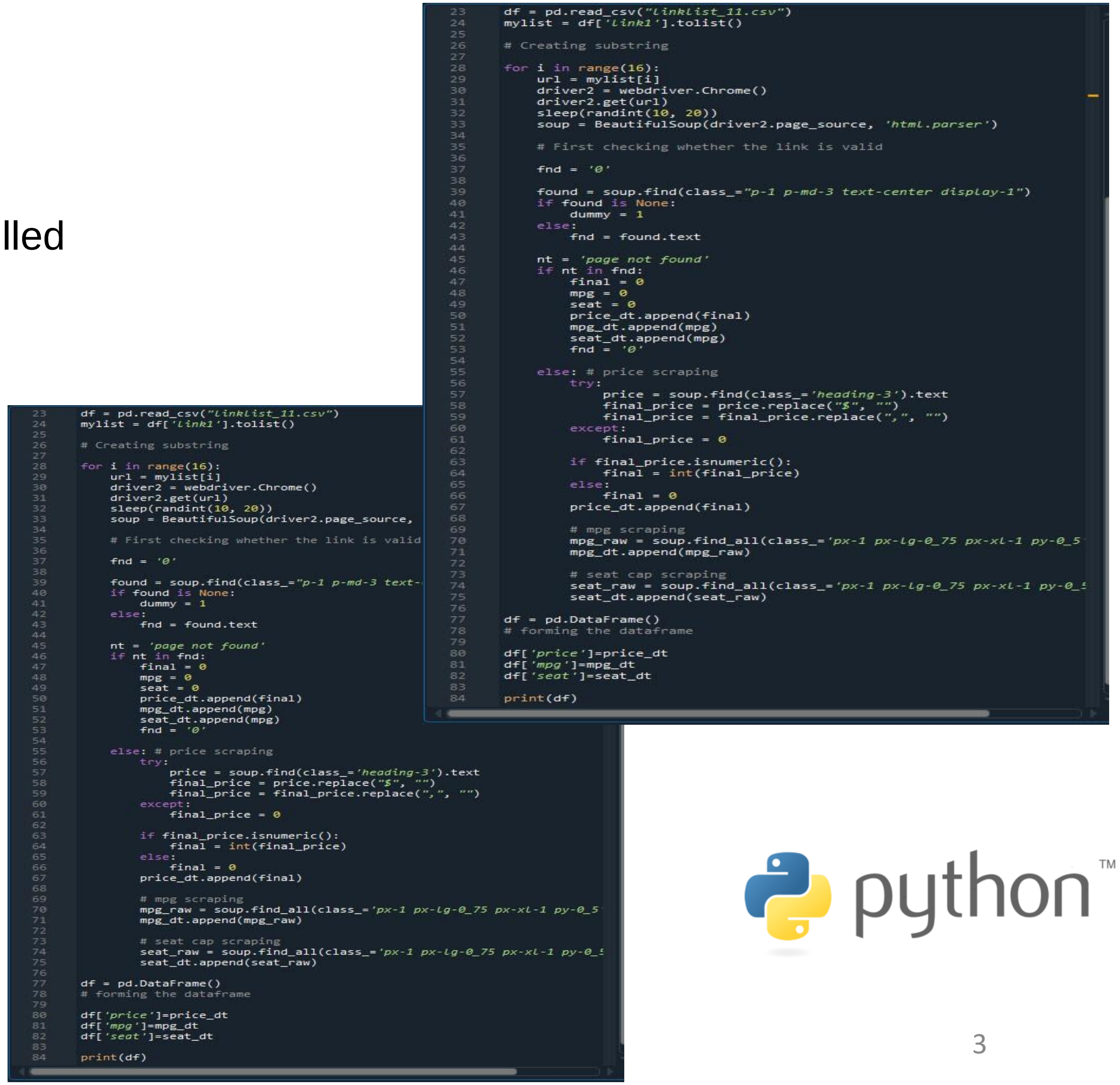
Web Scrapping Vehicle Features & Review Data

Kafle, Lee, Siddique

- **Packages Used in Python Web Scraping Method:**
  - BeautifulSoup - parsing HTML & Script files for websites
  - Selenium - pulling the web data
  - Chrome driver - launching & automated testing of the pulled website
  - Pandas - creating lists & dataframe
- **Methodology Used:**
  - Identifying HTML elements for each required data
  - Collecting the required numeric & text data
  - Cleaning the data
  - Storing the data in a dataframe
  - Bypassing the unfounded websites



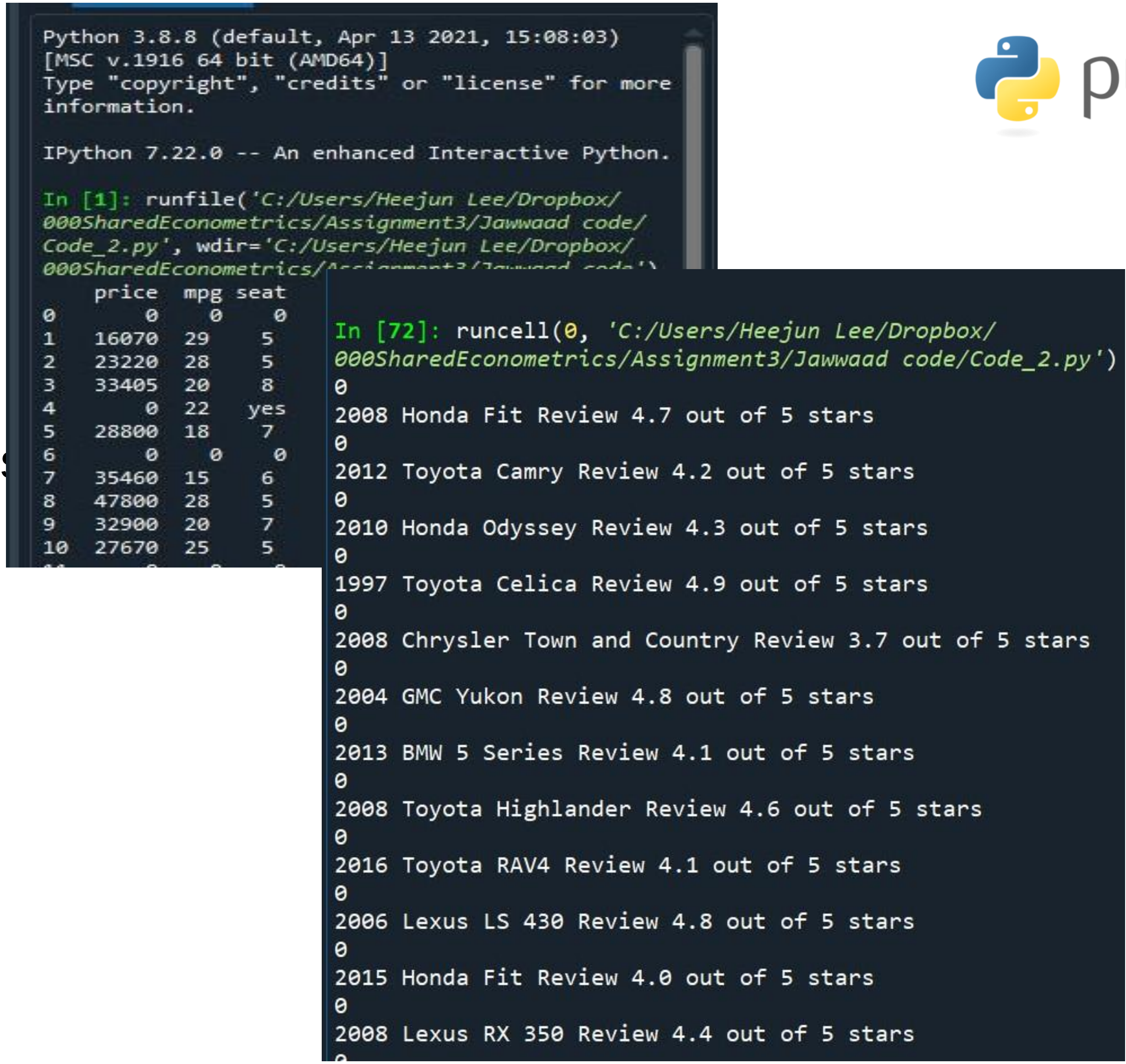
```python
# import libraries
from bs4 import BeautifulSoup
import numpy as np
from time import sleep
from random import randint
from selenium import webdriver
import pandas as pd

# creating empty data list
price_dt = []
mpg_dt = []
seat_dt = []
```



```python
df = pd.read_csv("linklist_11.csv")
mylist = df['link1'].tolist()

# Creating substring

for i in range(16):
    url = mylist[i]
    driver2 = webdriver.Chrome()
    driver2.get(url)
    sleep(randint(10, 20))
    soup = BeautifulSoup(driver2.page_source, 'html.parser')

    # First checking whether the link is valid

    fnd = '0'

    found = soup.find(class_="p-1 p-md-3 text-center display-1")
    if found is None:
        dummy = 1
    else:
        fnd = found.text

    nt = 'page not found'
    if nt in fnd:
        final = 0
        mpg = 0
        seat = 0
        price_dt.append(final)
        mpg_dt.append(mpg)
        seat_dt.append(mpg)
        fnd = '0'

    else: # price scraping
        try:
            price = soup.find(class_='heading-3').text
            final_price = price.replace("$", "")
            final_price = final_price.replace(",", "")
        except:
            final_price = 0

        if final_price.isnumeric():
            final = int(final_price)
        else:
            final = 0
        price_dt.append(final)

        # mpg scraping
        mpg_raw = soup.find_all(class_='px-1 px-lg-0_75 px-xl-1 py-0_5
        mpg_dt.append(mpg_raw)

        # seat cap scraping
        seat_raw = soup.find_all(class_='px-1 px-lg-0_75 px-xl-1 py-0_5
        seat_dt.append(seat_raw)

df = pd.DataFrame()
# forming the dataframe

df['price']=price_dt
df['mpg']=mpg_dt
df['seat']=seat_dt

print(df)
```

Web Scrapping Vehicle Features & Review Data

- **Parameters Collected:**
  - **From the Feature Set:**
    - Price
    - Seating Capacity
    - MPG
    - Warranty Year
    - Warranty Mileage
    - Cargo Capacity including seats
    - Fuel Capacity
  - **From the Review Set:**
    - Car Rating



```
Python 3.8.8 (default, Apr 13 2021, 15:08:03)
[MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more
information.

IPython 7.22.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Heejun Lee/Dropbox/
000SharedEconometrics/Assignment3/Jawwaad code/
Code_2.py', wdir='C:/Users/Heejun Lee/Dropbox/
000SharedEconometrics/Assignment3/Jawwaad code')
    price  mpg seat
0       0    0    0
1   16070  29    5
2   23220  28    5
3   33405  20    8
4       0  22  yes
5   28800  18    7
6       0    0    0
7   35460  15    6
8   47800  28    5
9   32900  20    7
10  27670  25    5
```

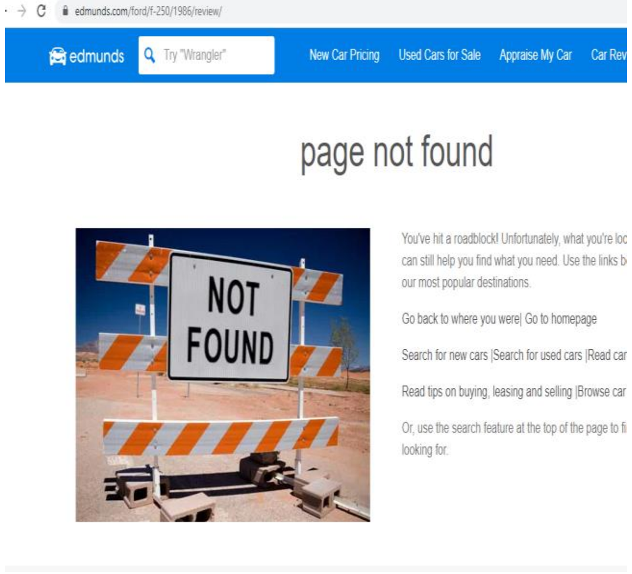
```
In [72]: runcell(0, 'C:/Users/Heejun Lee/Dropbox/
000SharedEconometrics/Assignment3/Jawwaad code/Code_2.py')
0
2008 Honda Fit Review 4.7 out of 5 stars
0
2012 Toyota Camry Review 4.2 out of 5 stars
0
2010 Honda Odyssey Review 4.3 out of 5 stars
0
1997 Toyota Celica Review 4.9 out of 5 stars
0
2008 Chrysler Town and Country Review 3.7 out of 5 stars
0
2004 GMC Yukon Review 4.8 out of 5 stars
0
2013 BMW 5 Series Review 4.1 out of 5 stars
0
2008 Toyota Highlander Review 4.6 out of 5 stars
0
2016 Toyota RAV4 Review 4.1 out of 5 stars
0
2006 Lexus LS 430 Review 4.8 out of 5 stars
0
2015 Honda Fit Review 4.0 out of 5 stars
0
2008 Lexus RX 350 Review 4.4 out of 5 stars
```

Web Scrapping Vehicle Features & Review Data

Kafle, Lee, Siddique

edmunds

- **Issues faced in web scraping:**
  - Dealing with unfounded web pages of unavailable car models
  - Dealing with specific car models like 'Other’
  - Time required for data pulling
  - Issues calculating safety & comfort points
  - Combining manual collect



page not found

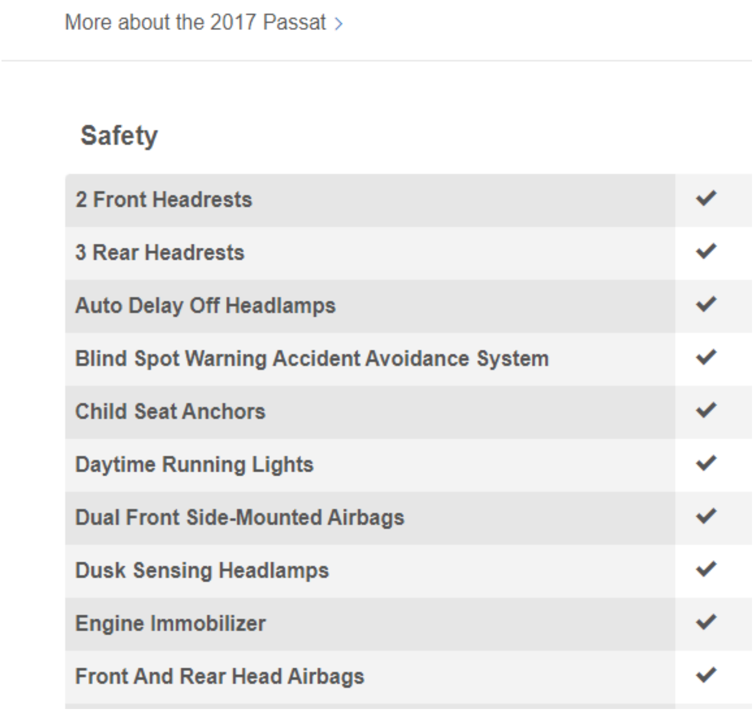
You've hit a roadblock! Unfortunately, what you're loc
can still help you find what you need. Use the links b
our most popular destinations.

Go back to where you were| Go to homepage

Search for new cars |Search for used cars |Read car

Read tips on buying, leasing and selling |Browse car

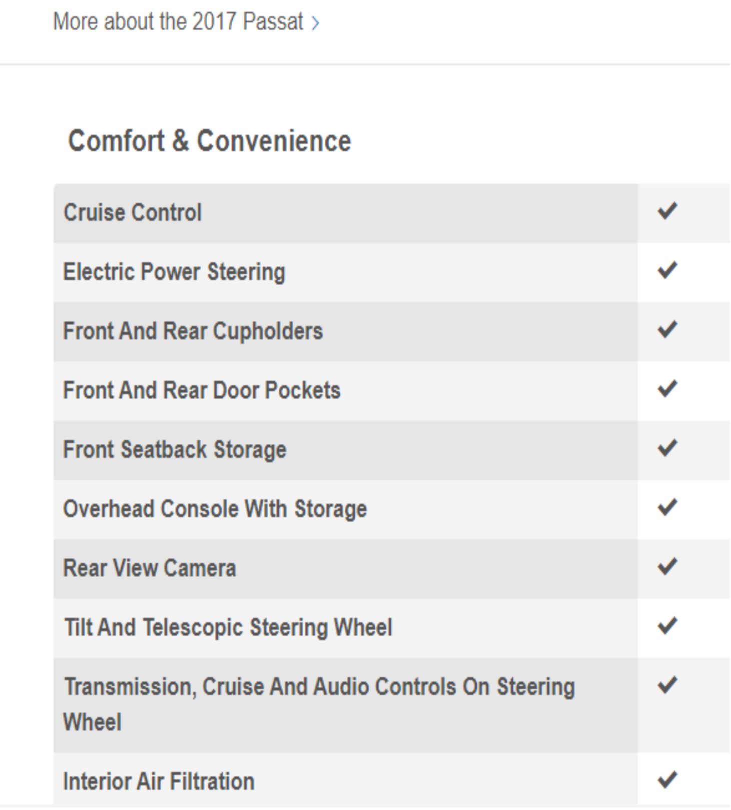
Or, use the search feature at the top of the page to fi
looking for.



More about the 2017 Passat ›

Safety

2 Front Headrests	✓
3 Rear Headrests	✓
Auto Delay Off Headlamps	✓
Blind Spot Warning Accident Avoidance System	✓
Child Seat Anchors	✓
Daytime Running Lights	✓
Dual Front Side-Mounted Airbags	✓
Dusk Sensing Headlamps	✓
Engine Immobilizer	✓
Front And Rear Head Airbags	✓



More about the 2017 Passat ›

Comfort & Convenience

Cruise Control	✓
Electric Power Steering	✓
Front And Rear Cupholders	✓
Front And Rear Door Pockets	✓
Front Seatback Storage	✓
Overhead Console With Storage	✓
Rear View Camera	✓
Tilt And Telescopic Steering Wheel	✓
Transmission, Cruise And Audio Controls On Steering Wheel	✓
Interior Air Filtration	✓