

Web Scraping & PowerBI Visualisation – DMV Autonomous Vehicle Incident Reports

This document describes the Python script used to automate the scraping of Autonomous Vehicle (AV) Collision Reports on the State of California Department of Motor Vehicles (DMV) website - <https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports>.

The Collision Reports come in the form of individual PDF files. To scrape the data, we will need to iterate through the 400 (as of 8 Dec 2022) reports and extract the PDF fields and checkboxes using Python.

In the later part of this document, I will also share some of the visualisations created and insights gathered by ingesting the scraped data into PowerBI.

DMV Webpage ([with each Collision Report saved in a PDF](#))

AUTONOMOUS VEHICLE COLLISION REPORTS

Manufacturers who are testing autonomous vehicles need to report any collision that resulted in property damage, bodily injury, or death within 10 days of the incident.

As of December 2, 2022, the DMV has received 533 Autonomous Vehicle Collision Reports. Collision reports prior to January 1, 2019 have been archived by DMV and are available upon request. Please email AVarchive@dmv.ca.gov to request a digital copy of an archived report. Requests must include the manufacturer and the date of the collision. Please do not include any sensitive personal information such as your social security number, driver license number, or financial account number on the request.

2022

- [Zoox November 20, 2022 \(PDF\)](#)
- [Mercedes Benz November 19, 2022 \(PDF\)](#)
- [Mercedes Benz November 17, 2022 \(PDF\)](#)
- [Cruise November 16, 2022 \(PDF\)](#)

HTML Text ([showing the URL for the Collision Report](#))

```
<ul>
  <li>
    ::marker
    <a href="/portal/file/zoox_112022-pdf/">...</a> == $0
  </li>
</ul>
```

Sample of DMV AV Collision Report



REPORT OF TRAFFIC COLLISION INVOLVING AN AUTONOMOUS VEHICLE

DMV USE ONLY	
A/T NUMBER	
NAME	

Instructions: Please print within the spaces and boxes on this form. If you need to provide additional information on a separate piece of paper(s) or you include a copy of any law enforcement agency report, please check the box to indicate "Additional Information Attached."

- Write **unk (for unknown)** or **none** in any space or box when you do not have the information on the other party involved.
- Give insurance information that is complete and which correctly and **fully** identifies the **company** that issued the insurance policy or surety bond, or whether there is a certificate of self-insurance.
- Place the National Association of Insurance Commissioners (NAIC) number for your Insurance or Surety Company in the boxes provided. The NAIC number should be located on the proof of insurance provided by you company or you can contact your insurer for that information.
- Identify any person involved in the accident (driver, passenger, bicyclist, pedestrian, etc) that you saw was injured or complained of bodily injury or know to be deceased.
- Record in the PROPERTY DAMAGE line any damage to telephone poles, fences, street signs, guard post, trees, livestock, dogs, buildings, parked vehicles, etc., including a description of the damage.
- Once you have completed this report, please mail to: Department of Motor Vehicles, Occupational Licensing Branch, P.O. Box 932342, MS: L224, Sacramento, CA 94232-3420

SECTION 1 — MANUFACTURER'S INFORMATION

MANUFACTURER'S NAME GM Cruise LLC		A/T NUMBER	
BUSINESS NAME Cruise		TELEPHONE NUMBER ()	
STREET ADDRESS	CITY	STATE	ZIP CODE

SECTION 2 — ACCIDENT INFORMATION/VEHICLE 1

DATE OF ACCIDENT 01/07/2019	TIME OF ACCIDENT 06:54 <input type="checkbox"/> AM <input checked="" type="checkbox"/> PM	VEHICLE YEAR 2019	MAKE Chevrolet	MODEL Bolt
LICENSE PLATE NUMBER		VEHICLE IDENTIFICATION NUMBER		STATE VEHICLE IS REGISTERED IN CA
ADDRESS/LOCATION OF ACCIDENT Folsom St. and 11th St.		CITY San Francisco	COUNTY San Francisco	STATE ZIP CODE CA 94103
Vehicle was: <input checked="" type="checkbox"/> Moving <input type="checkbox"/> Stopped in Traffic		Involved in the Accident: <input type="checkbox"/> Pedestrian <input type="checkbox"/> Bicyclist <input type="checkbox"/> Other		NUMBER OF VEHICLES INVOLVED 2
DRIVER'S FULL NAME (FIRST, MIDDLE, LAST)		DRIVER LICENSE NUMBER		STATE DATE OF BIRTH
INSURANCE COMPANY NAME OR SURETY COMPANY AT TIME OF ACCIDENT		POLICY NUMBER		
COMPANY NAIC NUMBER		POLICY PERIOD FROM TO		

Describe Vehicle Damage

☐ UNK ☐ NONE ☐ MINOR
☒ MOD ☐ MAJOR

Shade in Damaged Area



Go to Page 2



SECTION 3 — OTHER PARTY'S INFORMATION/VEHICLE 2

VEHICLE YEAR 2018	MODEL Pex150		
LICENSE PLATE NUMBER	VEHICLE IDENTIFICATION NUMBER	STATE VEHICLE IS REGISTERED IN CA	
Vehicle was: <input checked="" type="checkbox"/> Moving <input type="checkbox"/> Stopped in Traffic	Involved in the Accident: <input type="checkbox"/> Pedestrian <input type="checkbox"/> Bicyclist <input type="checkbox"/> Other	NUMBER OF VEHICLES INVOLVED 2	
DRIVER'S FULL NAME (FIRST, MIDDLE, LAST)		DRIVER LICENSE NUMBER unk	STATE un
INSURANCE COMPANY NAME OR SURETY COMPANY AT TIME OF ACCIDENT		DATE OF BIRTH	
COMPANY NAIC NUMBER		POLICY NUMBER	
POLICY PERIOD FROM		TO	

☐ Additional information attached.**SECTION 4 — INJURY/DEATH, PROPERTY DAMAGE**

NAME (FIRST, MIDDLE, LAST)			
ADDRESS unk	CITY unk	STATE un	ZIP CODE unk
CHECK ALL THAT APPLY <input checked="" type="checkbox"/> Injured <input type="checkbox"/> Deceased <input checked="" type="checkbox"/> Driver <input type="checkbox"/> Passenger <input type="checkbox"/> Bicyclist <input type="checkbox"/> Property			
NAME (FIRST, MIDDLE, LAST)			
ADDRESS	CITY	STATE	ZIP CODE
CHECK ALL THAT APPLY <input type="checkbox"/> Injured <input type="checkbox"/> Deceased <input type="checkbox"/> Driver <input type="checkbox"/> Passenger <input type="checkbox"/> Bicyclist <input type="checkbox"/> Property			
PROPERTY DAMAGE			
PROPERTY OWNER'S NAME		TELEPHONE NUMBER ()	
STREET ADDRESS	CITY	STATE	ZIP CODE
WITNESS NAME		TELEPHONE NUMBER ()	
STREET ADDRESS	CITY	STATE	ZIP CODE
WITNESS NAME		TELEPHONE NUMBER ()	
STREET ADDRESS	CITY	STATE	ZIP CODE

☐ Additional information attached.**SECTION 5 — ACCIDENT DETAILS - DESCRIPTION**☐ Autonomous Mode ☒ Conventional Mode

A Cruise autonomous vehicle ("Cruise AV"), operating in conventional mode, was making a left turn from northeast bound Folsom Street onto northwest bound 11th Street when a scooterist, attempting to pass the Cruise AV on the left, made contact with the front left side of the Cruise AV, damaging the front left fender, radar, and wheel well of the Cruise AV. The scooterist reported injuries and emergency services and the police arrived at the scene, but the scooterist declined medical treatment. No police report was available at the time of the filing of this report.

☐ Additional information attached.[Go to Page 3](#)

ITEMS MARKED BELOW FOLLOWED BY AN ASTERISK (*) SHOULD BE EXPLAINED IN THE NARRATIVE								
WEATHER (MARK 1 to 2 ITEMS)		VEH 1	VEH 2	MOVEMENT PRECEDING COLLISION		VEH 1	VEH 2	OTHER ASSOCIATED FACTOR(s) (MARK ALL APPLICABLE)
A. CLEAR		X	X	A. STOPPED				A. CVC SECTIONS VIOLATED CITED <input type="checkbox"/> YES <input type="checkbox"/> NO
B. CLOUDY				B. PROCEEDING STRAIGHT				
C. RAINING				C. RAN OFF ROAD				
D. SNOWING				D. MAKING RIGHT TURN				
E. FOG/VISIBILITY				E. MAKING LEFT TURN		X		
F. OTHER				F. MAKING U TURN				B. VISION OBSCUREMENT <input type="checkbox"/>
G. WIND				G. BACKING				C. INATTENTION* <input type="checkbox"/>
LIGHTING				H. SLOWING/STOPPING				D. STOP & GO TRAFFIC <input type="checkbox"/>
A. DAYLIGHT				I. PASSING OTHER VEHICLE				E. ENTERING/LEAVING RAMP <input type="checkbox"/>
B. DUSK – DAWN				J. CHANGING LANES			X	F. PREVIOUS COLLISION <input type="checkbox"/>
C. DARK –STREET LIGHTS		X	X	K. PARKING MANUEVER				G. UNFAMILIAR WITH ROAD <input type="checkbox"/>
D. DARK – NO STREET LIGHTS				L. ENTERING TRAFFIC				H. DEFECTIVE WEH EQUIP CITED <input type="checkbox"/> YES <input type="checkbox"/> NO
E. DARK –STREET LIGHTS NOT FUNCTIONING*				M. OTHER UNSAFE TURNING				
ROADWAY SURFACE				N. XING INTO OPPOSING LANE				
A. DRY		X	X	O. PARKED				I. UNINVOLVED VEHICLE <input type="checkbox"/>
B. WET				P. MERGING				J. OTHER* <input type="checkbox"/>
C. SNOWY – ICY				Q. TRAVELING WRONG WAY				K. NONE APPARENT <input type="checkbox"/>
D. SLIPPERY (MUDDY, OILY, ETC.)				R. OTHER*				L. RUNAWAY VEHICLE <input type="checkbox"/>
ROADWAY CONDITIONS (MARK 1 TO 2 ITEMS)				TYPE OF COLLISION				
A. HOLES, DEEP RUT*				A. HEAD-ON				
B. LOOSE MATERIAL ON ROADWAY				B. SIDE SWIPE			X	
C. OBSTRUCTION ON ROADWAY*				C. REAR END				
D. CONSTRUCTION – REPAIR ZONE				D. BROADSIDE				
E. REDUCED ROADWAY WIDTH				E. HIT OBJECT				
F. FLOODED*				F. OVERTURNED				
G. OTHER*				G. VEHICLE/PEDESTRIAN				
H. NO UNUSUAL CONDITIONS		X	X	H. OTHER*				

SECTION 6 — CERTIFICATION

I certify (or declare) under penalty of perjury under the laws of the State of California that the foregoing is true and correct.

I further certify that I am the authorized Administrator of the program for the above named employer.

PROGRAM DIRECTOR/AUTHORIZED REPRESENTATIVE PRINTED NAME AND TITLE Kevin Chu, Director of AV Engineering	TELEPHONE NUMBER ()
SIGNATURE X	DATE SIGNED 03/13/2019

First, we import the relevant Python libraries. Next, send a **GET** request to the DMV webpage and create a response object that stores the request response.

Using the BeautifulSoup library, we can parse the html to get the URLs of the various AV Collision Reports.

Import Python Libraries and Parsing Webpage HTML

```
1 import requests
2 import PyPDF2
3 import json
4 import fitz
5 import pandas as pd
6 import io
7 from io import BytesIO
8 from bs4 import BeautifulSoup

1 response = requests.get("https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-coll")
<

1 dmv_web_page = response.text

1 soup = BeautifulSoup(dmv_web_page, "html.parser")

1 # Get all spans with the word file
2 dmv_report = soup.select('a[href*=file]')
3 print(dmv_report)

[<a href="/portal/file/zoox_112022-pdf/">Zoox November 20, 2022</a>, <a href="/portal/file/mercedes-benz_111922-pdf/">Mercede
s Benz November 19, 2022</a>, <a href="/portal/file/mercedes-benz_111722-pdf/">Mercedes Benz November 17, 2022</a>, <a href
="/portal/file/cruise_111622-pdf/">Cruise November 16, 2022</a>, <a href="/portal/file/zoox_111422-pdf/">Zoox November 14, 20
22</a>, <a href="/portal/file/waymo_110822-pdf/">Waymo November 8, 2022</a>, <a href="/portal/file/waymo_110322-pdf/">Waymo N
ovember 3, 2022</a>, <a href="/portal/file/zoox_110422-pdf/">Zoox November 4, 2022</a>, <a href="/portal/file/waymo_102122-pd
f/">Waymo October 21, 2022</a>, <a href="/portal/file/apple_102022-pdf/">Apple October 20, 2022</a>, <a href="/portal/file/ap
ple_101022-pdf/">Apple October 10, 2022</a>, <a href="/portal/file/zoox_101422-pdf/">Zoox October 14, 2022</a>, <a href="/por
tal/file/zoox_100322-pdf/">Zoox October 3, 2022</a>, <a href="/portal/file/ghost_100722-pdf/">Ghost Autonomy Inc October 7, 2
022</a>, <a href="/portal/file/argo_092222-pdf/">Argo AI September 22, 2022</a>, <a href="/portal/file/waymo_091722-pdf/">Way
mo September 17, 2022</a>, <a href="/portal/file/waymo_091522-pdf/">Waymo September 15, 2022</a>, <a href="/portal/file/zoox_
091422-pdf/">Zoox September 14, 2022</a>, <a href="/portal/file/waymo_091022-pdf/">Waymo September 10, 2022</a>, <a href="/po
rtal/file/cruise_090522-pdf/">Cruise September 5, 2022</a>, <a href="/portal/file/cruise_090422-pdf/">Cruise September 4, 202
2</a>, <a href="/portal/file/zoox_090122-pdf/">Zoox September 1, 2022</a>, <a href="/portal/file/cruise_090122-pdf/">Cruise S
eptember 1, 2022</a>, <a href="/portal/file/zoox_083022-pdf/">Zoox August 30, 2022</a>, <a href="/portal/file/zoox_082322-pd
f/">Zoox August 23, 2022</a>, <a href="/portal/file/waymo_082522-pdf/">Waymo August 25, 2022</a>, <a href="/portal/file/cruis
e_081622-pdf/">Cruise August 16, 2022</a>, <a href="/portal/file/waymo_081622_1-pdf/">Waymo August 16, 2022 (1)</a>, <a href
="/portal/file/waymo_081622_2-pdf/">Waymo August 16, 2022 (2)</a>, <a href="/portal/file/waymo_081422-pdf/">Waymo August 14,
2022</a>, <a href="/portal/file/waymo_081322-pdf/">Waymo August 13, 2022</a>, <a href="/portal/file/waymo_081022-pdf/">Waymo
August 10, 2022</a>, <a href="/portal/file/woven_planet_080822-pdf/">Woven Planet August 8, 2022</a>, <a href="/portal/file/c
```

Extract URL from HTML text

```
1 dmV_report_links = []
2
3 for link in dmV_report:
4     dmV_report_links.append(link.get('href'))
5
6 print(dmV_report_links)
```

```
['/portal/file/zoox_112022-pdf/', '/portal/file/mercedes-benz_111922-pdf/', '/portal/file/mercedes-benz_111722-pdf/', '/portal/
file/cruise_111622-pdf/', '/portal/file/zoox_111422-pdf/', '/portal/file/waymo_110822-pdf/', '/portal/file/waymo_110322-pd
f/', '/portal/file/zoox_110422-pdf/', '/portal/file/waymo_102122-pdf/', '/portal/file/apple_102022-pdf/', '/portal/file/apple
_101022-pdf/', '/portal/file/zoox_101422-pdf/', '/portal/file/zoox_100322-pdf/', '/portal/file/ghost_100722-pdf/', '/portal/f
ile/argo_092222-pdf/', '/portal/file/waymo_091722-pdf/', '/portal/file/waymo_091522-pdf/', '/portal/file/zoox_091422-pdf/',
'/portal/file/waymo_091022-pdf/', '/portal/file/cruise_090522-pdf/', '/portal/file/cruise_090422-pdf/', '/portal/file/zoox_09
0122-pdf/', '/portal/file/cruise_090122-pdf/', '/portal/file/zoox_083022-pdf/', '/portal/file/zoox_082322-pdf/', '/portal/fil
e/waymo_082522-pdf/', '/portal/file/cruise_081622-pdf/', '/portal/file/waymo_081622_1-pdf/', '/portal/file/waymo_081622_2-pd
f/', '/portal/file/waymo_081422-pdf/', '/portal/file/waymo_081322-pdf/', '/portal/file/waymo_081022-pdf/', '/portal/file/wove
n_planet_080822-pdf/', '/portal/file/cruise_080222-pdf/', '/portal/file/waymo_072422-pdf/', '/portal/file/zoox_072222-pdf/',
'/portal/file/cruise_071822-pdf/', '/portal/file/cruise_071622-pdf/', '/portal/file/cruise_071422-pdf/', '/portal/file/zoox_0
70822-pdf/', '/portal/file/waymo_070522-pdf/', '/portal/file/waymo_070322-pdf/', '/portal/file/waymo_070222_1-pdf/', '/porta
l/file/waymo_070222_2-pdf/', '/portal/file/waymo_070122-pdf/', '/portal/file/cruise_062922-pdf/', '/portal/file/zoox_062822-p
df/', '/portal/file/waymo_062122-pdf/', '/portal/file/apple_061422-pdf/', '/portal/file/cruise_061722-pdf/', '/portal/file/zo
ox_061122-pdf/', '/portal/file/zoox_060822-pdf/', '/portal/file/waymo_060522-pdf/', '/portal/file/cruise_060322-pdf/', '/port
al/file/cruise_060222-pdf/', '/portal/file/waymo_060122-pdf/', '/portal/file/zoox_060122-1-pdf/', '/portal/file/zoox_060122-2
-pdf/', '/portal/file/waymo_052722-pdf/', '/portal/file/cruise_052522-pdf/', '/portal/file/zoox_053122-pdf/', '/portal/file/w
aymo_052122-pdf/', '/portal/file/zoox_052022-pdf/', '/portal/file/mercedes-benz_051922-pdf/', '/portal/file/waymo_051722-pd
f/', '/portal/file/pony-ai_051722-pdf/', '/portal/file/motional_051722-pdf/', '/portal/file/cruise_051322-pdf/', '/portal/fil
```

As of 8 Dec 2022, there are close to 400 AV Collision Reports. Downloading them one by one and inputting the data in the PDF files manually would be time consuming. Let's make use of Python to automate this effort.

Create an Empty Dataframe with the Columns that You Want

```

1 # create empty dataframe
2 df = pd.DataFrame({
3     'BUSINESS NAME': [],
4     'DATE OF ACCIDENT': [],
5     'Time of Accident': [],
6     'VEHICLE YEAR' : [],
7     'MAKE' : [],
8     'MODEL' : [],
9     'NUMBER OF VEHICLES INVOLVED' : []
10 })
11
12 # print dataframe
13 print("\n *** Original DataFrames ** \n")
14 print(df)
15
16
17 # keys that you want to retain (for filtering dictionary later before appending to dataframe)
18 keys = ['BUSINESS NAME', 'DATE OF ACCIDENT', 'Time of Accident',
19         'VEHICLE YEAR', 'MAKE', 'MODEL', 'NUMBER OF VEHICLES INVOLVED']

```


After creating an empty dataframe, let's iterate over the 400 AV Collision Report URL, and extract the data from the PDF. Save the PDF data in the variable 'dictionary'.

```

1 for link in dmv_report_links:
2     url = 'https://www.dmv.ca.gov' + link
3
4     response = requests.get(url)
5     f = io.BytesIO(response.content)
6
7     with f as data:
8         #read it and get the pages number
9         pdfreader=PyPDF2.PdfFileReader(data)
10        x=pdfreader.numPages
11        pageobj=pdfreader.getPage(0)
12
13        # extract the pdf to text
14        text=pageobj.extractText()
15
16        # extract pdf form fields
17        dictionary = pdfreader.getFormTextFields()

```

Note that data from checkboxes such as the one below cannot be extracted using the above method.

ITEMS MARKED BELOW FOLLOWED BY AN ASTERISK (*) SHOULD BE EXPLAINED IN THE NARRATIVE						
WEATHER (MARK 1 to 2 ITEMS)	VEH 1	VEH 2	MOVEMENT PRECEDING COLLISION	VEH 1	VEH 2	OTHER ASSOCIATED FACTOR(s) (MARK ALL APPLICABLE)
A. CLEAR	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	A. STOPPED	<input checked="" type="checkbox"/>		A. CVC SECTIONS VIOLATED CITED <input type="checkbox"/> YES <input type="checkbox"/> NO
B. CLOUDY			B. PROCEEDING STRAIGHT		<input checked="" type="checkbox"/>	
C. RAINING			C. RAN OFF ROAD			
D. SNOWING			D. MAKING RIGHT TURN			
E. FOG/VISIBILITY			E. MAKING LEFT TURN			
F. OTHER			F. MAKING U TURN			B. VISION OBSCUREMENT <input type="checkbox"/>
G. WIND			G. BACKING		<input checked="" type="checkbox"/>	C. INATTENTION* <input checked="" type="checkbox"/>
LIGHTING			H. SLOWING/STOPPING			D. STOP & GO TRAFFIC <input type="checkbox"/>
A. DAYLIGHT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	I. PASSING OTHER VEHICLE			E. ENTERING/LEAVING RAMP <input type="checkbox"/>
B. DUSK - DAWN			J. CHANGING LANES			F. PREVIOUS COLLISION <input type="checkbox"/>
C. DARK - STREET LIGHTS			K. PARKING MANUEVER			G. UNFAMILIAR WITH ROAD <input type="checkbox"/>
D. DARK - NO STREET LIGHTS			L. ENTERING TRAFFIC			H. DEFECTIVE WEH EQUIP CITED <input type="checkbox"/> YES <input type="checkbox"/> NO
E. DARK - STREET LIGHTS NOT FUNCTIONING*			M. OTHER UNSAFE TURNING			
ROADWAY SURFACE			N. XING INTO OPPOSING LANE			
A. DRY	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	O. PARKED			I. UNINVOLVED VEHICLE <input type="checkbox"/>
B. WET			P. MERGING			J. OTHER* <input type="checkbox"/>
C. SNOWY - ICY			Q. TRAVELING WRONG WAY			K. NONE APPARENT <input type="checkbox"/>
D. SLIPPERY (MUDDY, OILY, ETC.)			R. OTHER*			L. RUNAWAY VEHICLE <input type="checkbox"/>

Hence, we will use the `pdfreader.getFields()` method to call the data and save it in the variable 'checkbox'.

```
19 checkbox = pdfreader.getFields()
20
21
22 # extract checkbox fields from PDF
23
24 # ---- #
25 time_of_day = ''
26
27 if checkbox.get('AM').value == '/ ':
28     time_of_day = 'AM'
29 else:
30     time_of_day = 'PM'
31
32 # ---- #
33 vehicle_damage = ''
34
35 if checkbox.get('Unknown').value == '/Yes':
36     vehicle_damage = 'unknown'
37 elif checkbox.get('None').value == '/Yes':
38     vehicle_damage = 'none'
39 elif checkbox.get('minor').value == '/Yes':
40     vehicle_damage = 'minor'
41 elif checkbox.get('Moderate').value == '/Yes':
42     vehicle_damage = 'moderate'
43 elif checkbox.get('major').value == '/Yes':
44     vehicle_damage = 'major'
45 else:
46     vehicle_damage = 'unknown'
47
48 # ---- #
49 moving_stopped = ''
50
51 if checkbox.get('Moving').value == '/ ':
52     moving_stopped = 'moving'
53 elif checkbox.get('Stopped in Traffic').value == '/ ':
54     moving_stopped = 'stationary'
55 else:
56     moving_stopped = 'unknown'
57
58 # ---- #
59 autonomous_conventional = ''
60
61 if checkbox.get('Autonomous Mode').value == '/ ':
62     autonomous_conventional = 'autonomous'
63 else:
64     autonomous_conventional = 'conventional'
65
66 # ---- #
67 weather = ''
68
69 if checkbox.get('WEATHER A 1').value == '/Yes':
70     weather = 'clear'
71 elif checkbox.get('WEATHER B 1').value == '/Yes':
72     weather = 'cloudy'
73 elif checkbox.get('WEATHER C 1').value == '/Yes':
74     weather = 'raining'
75 elif checkbox.get('WEATHER D 1').value == '/Yes':
76     weather = 'snowing'
77 elif checkbox.get('WEATHER E 1').value == '/Yes':
78     weather = 'fog/visibility'
79 elif checkbox.get('WEATHER F 1').value == '/Yes':
80     weather = 'other'
81 elif checkbox.get('WEATHER G 1').value == '/Yes':
82     weather = 'wind'
83 else:
84     weather = 'unknown'
85
```



```

86     # ---- #
87     lighting = ''
88
89     if checkbox.get('LIGHTING A 1').value == '/Yes':
90         lighting = 'daylight'
91     elif checkbox.get('LIGHTING B 1').value == '/Yes':
92         lighting = 'dusk-dawn'
93     elif checkbox.get('LIGHTING C 1').value == '/Yes':
94         lighting = 'dark-street_lights'
95     elif checkbox.get('LIGHTING D 1').value == '/Yes':
96         lighting = 'dark-no_street_lights'
97     elif checkbox.get('LIGHTING E 1').value == '/Yes':
98         lighting = 'dark-street_lights_malfunction'
99     else:
100         lighting = 'unknown'
101
102     # ---- #
103     roadway = ''
104
105     if checkbox.get('ROADWAY A 1').value == '/Yes':
106         roadway = 'dry'
107     elif checkbox.get('ROADWAY B 1').value == '/Yes':
108         roadway = 'wet'
109     elif checkbox.get('ROADWAY C 1').value == '/Yes':
110         roadway = 'snowy_ice'
111     elif checkbox.get('ROADWAY D 1').value == '/Yes':
112         roadway = 'slippery_mud_oil'
113     else:
114         roadway = 'unknown'
115
116     # ---- #
117     movement = []
118
119     movement_list = ['MOVEMENT A 1', 'MOVEMENT B 1', 'MOVEMENT C 1', 'MOVEMENT D 1',
120                     'MOVEMENT E 1', 'MOVEMENT F 1', 'MOVEMENT G 1', 'MOVEMENT H 1',
121                     'MOVEMENT I 1', 'MOVEMENT J 1', 'MOVEMENT K 1', 'MOVEMENT L 1',
122                     'MOVEMENT M 1', 'MOVEMENT N 1', 'MOVEMENT O 1', 'MOVEMENT P 1',
123                     'MOVEMENT Q 1', 'MOVEMENT R 1']
124
125     movement_name = ['stopped', 'proceeding_straight', 'ran_off_road', 'right_turn', 'left_turn', 'u_turn', 'backing',
126                     'slowing_stopping', 'passing_other_vehicle', 'changing_lane', 'parking', 'entering_traffic',
127                     'unsafe_turning', 'crossing_opposing_lane', 'parked', 'merging', 'wrong_way', 'others']
128
129     i=0
130
131     for key in movement_list:
132         if checkbox.get(key).value == '/Yes':
133             movement.append(movement_name[i])
134             i = i+1
135
136
137     # filter out the keys you do not want
138     filtered_dict = dict((k, dictionary[k]) for k in keys if k in dictionary)
139
140     # Insert the checkbox fields into dictionary
141     filtered_dict.update([('time_of_day', time_of_day), ('vehicle_damage', vehicle_damage),
142                         ('moving_stopped', moving_stopped), ('autonomous_conventional', autonomous_conventional),
143                         ('weather', weather), ('lighting', lighting), ('roadway', roadway), ('movement', movement)])
144

```

With the above codes, we can obtain the PDF data (be it form field or checkboxes) for each URL (i.e. AV Collision Report). We then append this data into the empty dataframe that we have created earlier.

```

145     # combined data
146     df = df.append(filtered_dict, ignore_index=True, sort=False)

```

Sample Table Showing the Extracted Data

1	df.head(193)									
	company_name	date	time	vehicle_year	make	model	num_of_vehicles_involved	autonomous_conventional	lighting	movement
0	Zoox	11/20/2022	1:24	2016	Toyota	Highlander	2	autonomous	daylight	[stopped]
1	MERCEDES-BENZ RESEARCH & DEVELOPMENT NORTH AMERICA	11/19/2022	10:45	2021	Mercedes-Benz	S 450	2	conventional	daylight	[left_turn_entering_traffic]
2	MERCEDES-BENZ RESEARCH & DEVELOPMENT NORTH AMERICA	11-17-2022	5:20	2021	Mercedes	S450	2	conventional	dark-street_lights	[proceeding_straight_changing_lane]
3	Cruise	11/16/2022	2:59	2023	Cruise	AV	2	conventional	daylight	[proceeding_straight_merging_other]
4	Zoox	11/14/2022	8:22	2016	Toyota	Highlander	2	autonomous	dark-street_lights	[slowing_stopping]
...
190	Waymo LLC	8/31/2021	7:30	2021	Jaguar	I-Pace	2	conventional	daylight	[entering_traffic]
191	Waymo LLC	8/31/2021	5:43	2021	Jaguar	I-Pace	2	conventional	daylight	[backing]
192	Cruise	08/30/2021	08:40	2020	Chevrolet	Bolt	2	conventional	dark-street_lights	[parked]

As the accident time is recorded in the format below, we can define a method and apply it to each individual row of the dataframe to convert the time to a 24-hour format (readable by PowerBI).

TIME OF ACCIDENT

06:54 ☐ AM ☒ PM

```

1 def convert24(row):
2
3     # is AM and first two elements are 12
4     if row['time_of_day'] == "AM" and row['time_split'][0] == "12":
5         row['time_split'][0] = "00"
6
7     # remove the AM
8     elif row['time_of_day'] == "AM":
9         row['time_split'][0] = row['time_split'][0]
10
11    # is PM and first two elements are 12
12    elif row['time_of_day'] == "PM" and row['time_split'][0] == "12":
13        row['time_split'][0] = row['time_split'][0]
14
15    else:
16        # add 12 to hours and remove PM
17        row['time_split'][0] = str(int(row['time_split'][0]) + 12)
18
19    return ':'.join(row['time_split'])

```

```

1 df2['time_24hr'] = df2.apply(lambda row : convert24(row), axis = 1)

```

Sample Table with 24-hour format

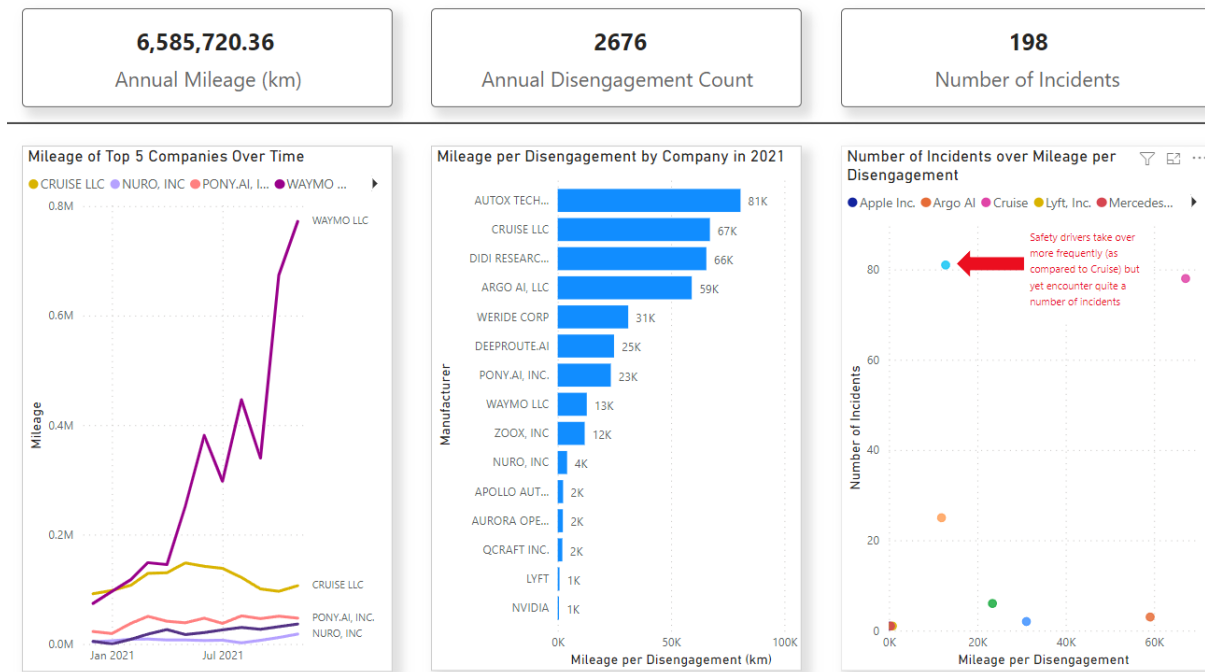
	company_name	date	time	time_of_day	time_24hr
0	Zoox	11/20/2022	1:24	PM	13:24
1	MERCEDES-BENZ RESEARCH & DEVELOPMENT NORTH AMERICA	11/19/2022	10:45	AM	10:45
2	MERCEDES-BENZ RESEARCH & DEVELOPMENT NORTH AMERICA	11-17-2022	5:20	PM	17:20

Now that we have our table, we can export as excel file and ingest the data into PowerBI for visualisation.

```
1 df2.to_excel("dmv_report.xlsx")
```

PowerBI Visualisation

PowerBI Report Page 1 (Mileage vs. Disengagement Count)



As we can see from the above visualisation, **Waymo has been increasing its on-road activities** over the past year. In November 2021, it clocked close to 800,000km of mileage. That said, its **frequency of disengagement is rather high** as compared to its competitors (i.e. 13,000km as compared to Cruise's 67,000km).

Even with such high disengagement rate, Waymo is involved in a relatively high number of AV incidents (autonomous). **Almost 50% of the total incidents belong to Waymo.**

From the above scatter plot of Number of Incidents over Mileage per Disengagement, we cannot really see any correlation between the two parameters. It will be good if DMV can provide the VIN number of the vehicle involved in the accident. We can then use this to plot the incident count vs. mileage per disengagement for each vehicle (instead of for each company).

If we look at the number of incidents over the years, we can see that **incidents have dropped during the onset of COVID-19 (in March 2020)**, possibly due to the reduced activities?

The number of **conventional vs. autonomous incidents is about 50%-50%**. From this, can we say that autonomous vehicles are not more unsafe as compared to conventional vehicles with drivers?

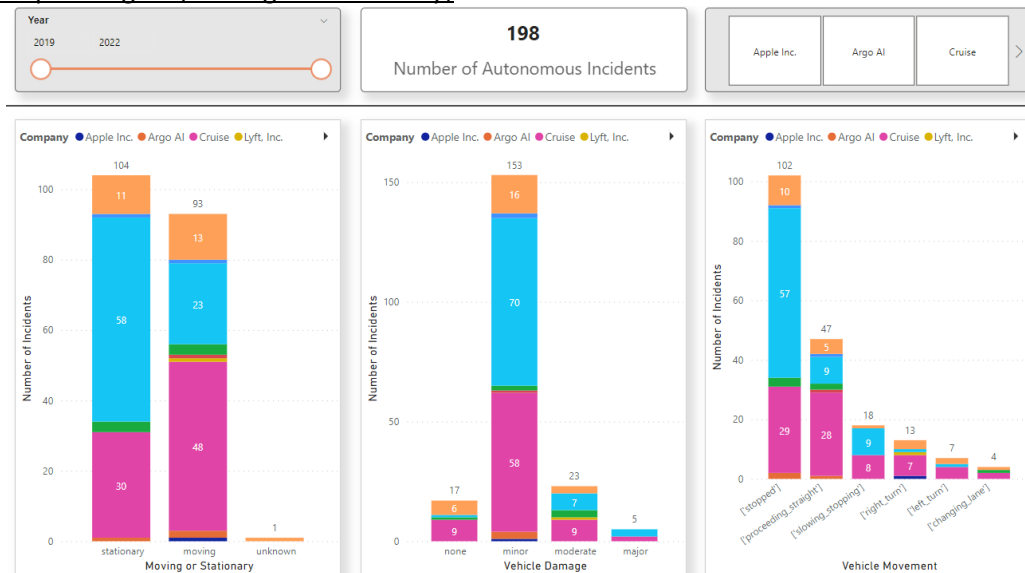
PowerBI Report Page 2 (Num of Incidents over Time)



Next, we look at whether there is a trend in the AVs' movement vs. them getting into accidents. A key observation from the visualisation below is that as compared to Waymo (light blue), a **huge proportion of Cruise's incidents (pink) happened while the vehicle is moving.**

Evident from the Vehicle Movement chart, we can see that the number of incidents while Cruise's AVs were **stopped vs. proceeding straight is similar 50%-50%**. Would be useful to dive deeper into what happened while the AVs were moving straight (was it due to a design issue of certain autonomous technology).

PowerBI Report Page 3 (Moving vs. Stationary)



Filter to show Waymo's and Cruise's incidents



Most of the incidents happen in daylight, and when the weather is clear and road is dry. We can observe from the Lighting chart below that quite a number of Cruise's incidents (pink) happened while the lighting is dark with street lights.

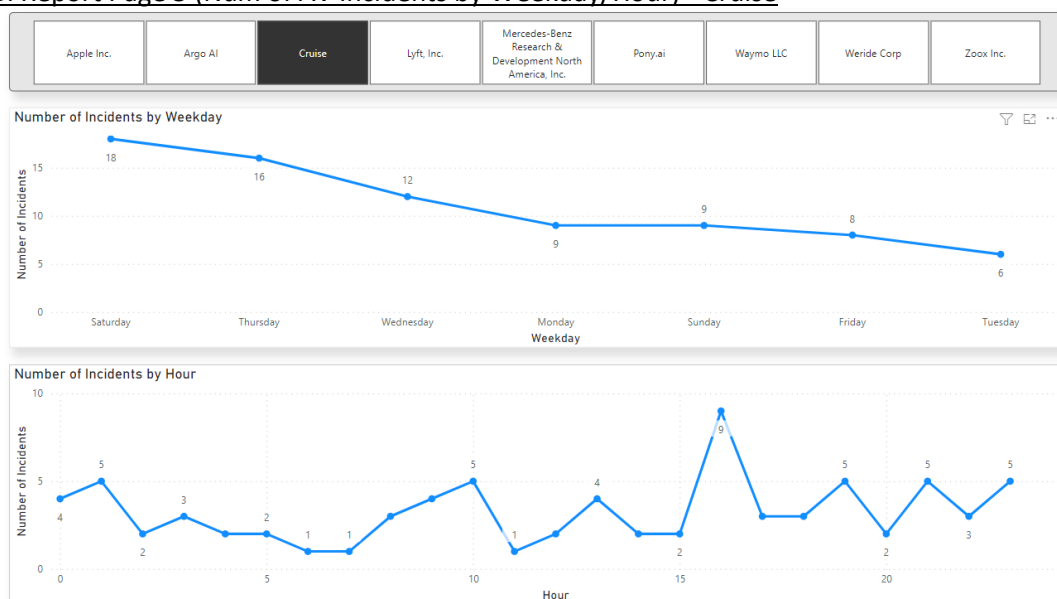
PowerBI Report Page 4 (Weather Lighting Roadway)



To see if this was due to increased activities during the night or an issue with Cruise's "night vision", we would require more data.

Indeed, by comparing the number of incidents by hour charts, we see that **Cruise's AVs get into incidents more frequently than others during the wee hours (i.e. from 9pm to 5am).**

PowerBI Report Page 5 (Num of AV Incidents by Weekday/Hour) - Cruise



All Companies except Cruise



An interesting observation: **The number of AV incidents is lowest on Sunday.**