Capturing & Exploring Datasets

Charlie Quah 173271M
Nicolas Tan 172944L

Project Objectives

- To capture data sets and/or performing preliminary data explorations & analytics.

- Outcome will lead to the set up of a centralised data repository/source for teaching & learning

Task Distribution

Nicolas Tan

Web Scraping
Data Cleaning
Mapping of Address

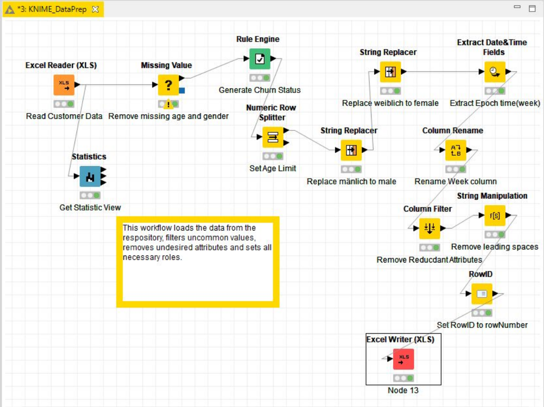
Charlie Quah

KNIME Exploration
Data Visualisation

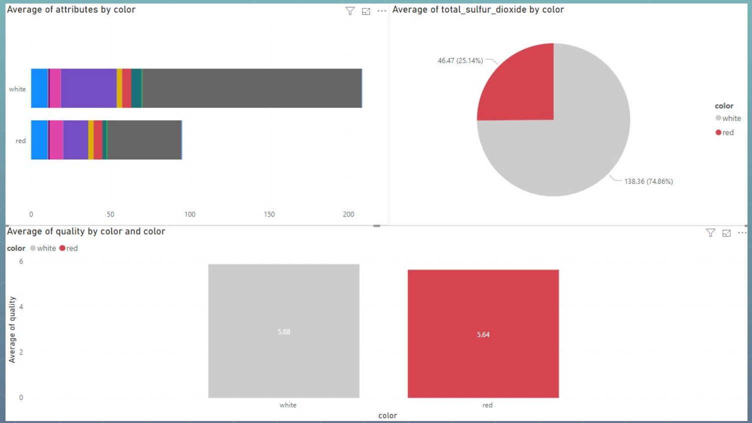
Recap

- Knime data preparation
- Knime analytics (decision tree & linear regression)
- Knime mapping
- HR, Wine, Crime and Telecom dataset visualisation using Tableau and Power BI.

Before

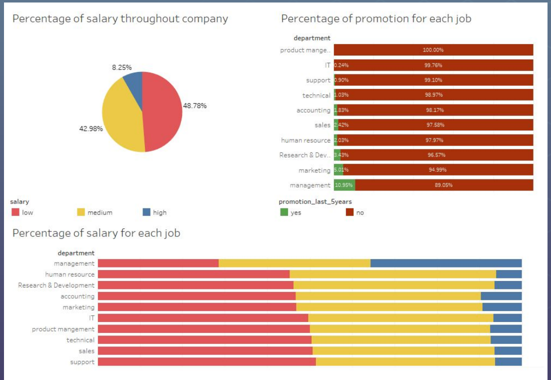


This workflow loads the data from the respository, filters uncommon values, removes undesired attributes and sets all necessary roles.


Average of attributes by color

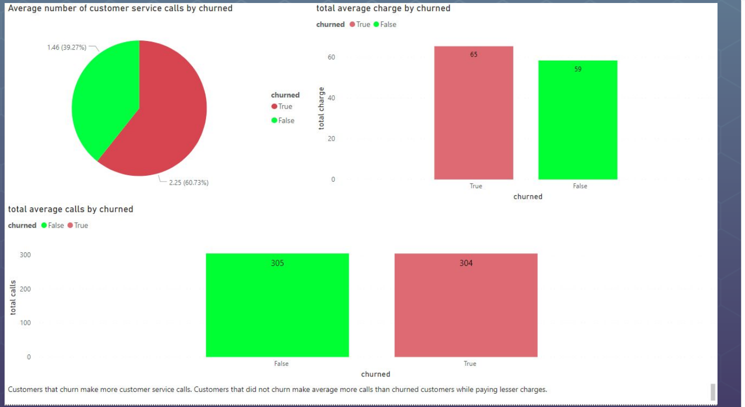
Average of total_sulfur_dioxide by color

Average of quality by color and color


Percentage of salary throughout company

Percentage of promotion for each job

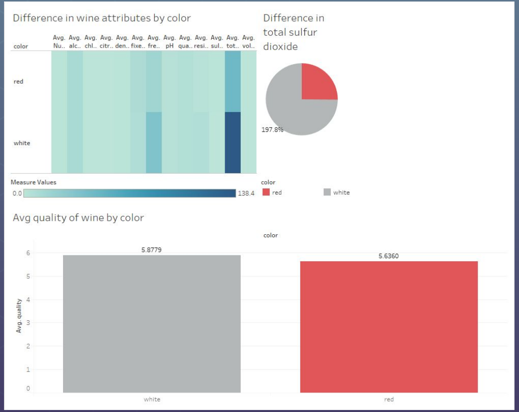
Percentage of salary for each job


Average number of customer service calls by churned

total average charge by churned

total average calls by churned

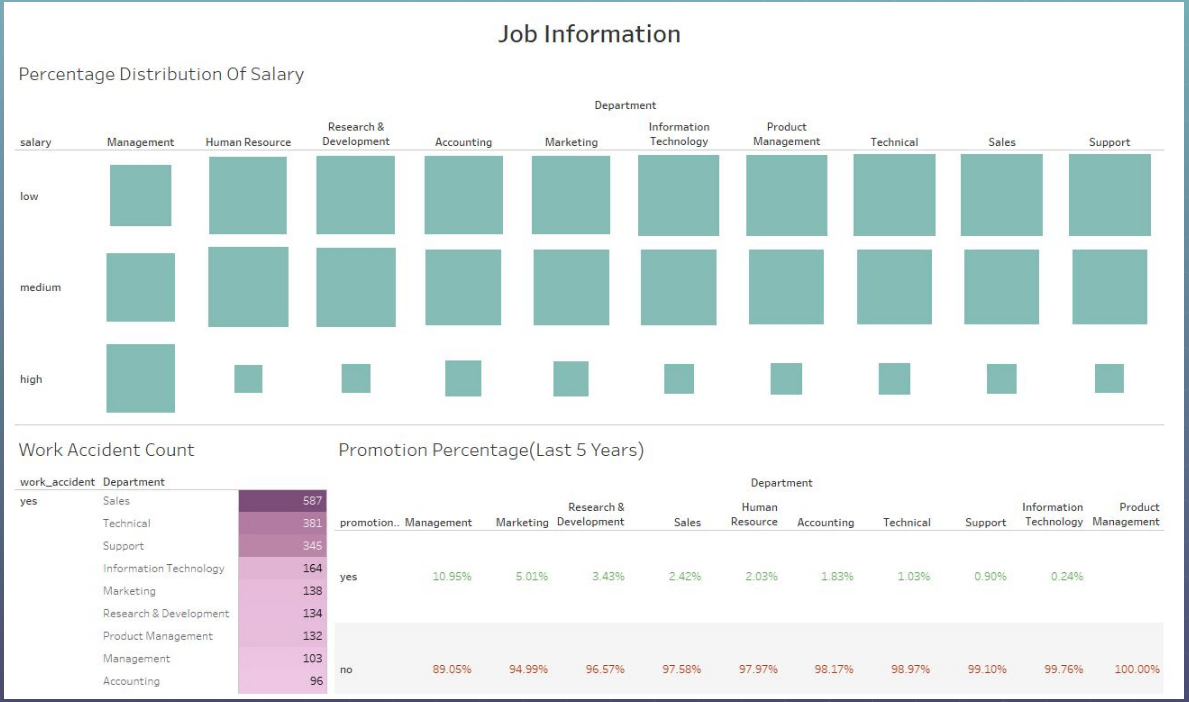
Customers that churn make more customer service calls. Customers that did not churn make average more calls than churned customers while paying lesser charges.


Difference in wine attributes by color

Difference in total sulfur dioxide

Avg quality of wine by color

HR dataset



Job Information

Percentage Distribution Of Salary

Work Accident Count

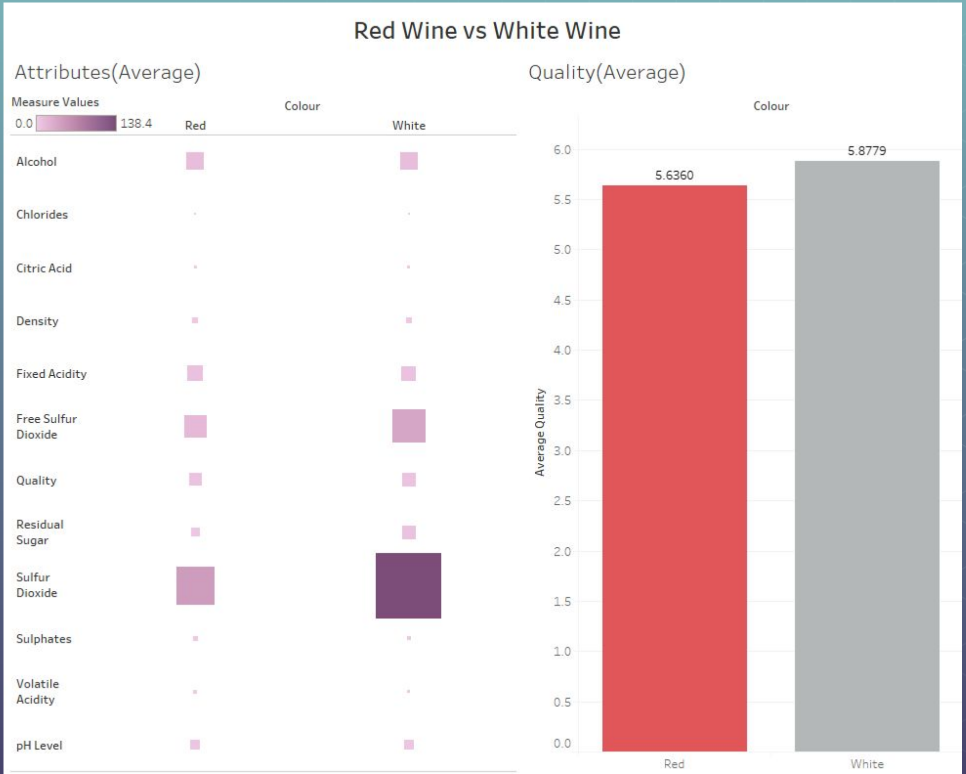
work_accident	Department	
yes	Sales	587
	Technical	381
	Support	345
	Information Technology	164
	Marketing	138
	Research & Development	134
	Product Management	132
	Management	103
	Accounting	96

Promotion Percentage(Last 5 Years)

	Department									
promotion..	Management	Marketing	Research & Development	Sales	Human Resource	Accounting	Technical	Support	Information Technology	Product Management
yes	10.95%	5.01%	3.43%	2.42%	2.03%	1.83%	1.03%	0.90%	0.24%	
no	89.05%	94.99%	96.57%	97.58%	97.97%	98.17%	98.97%	99.10%	99.76%	100.00%

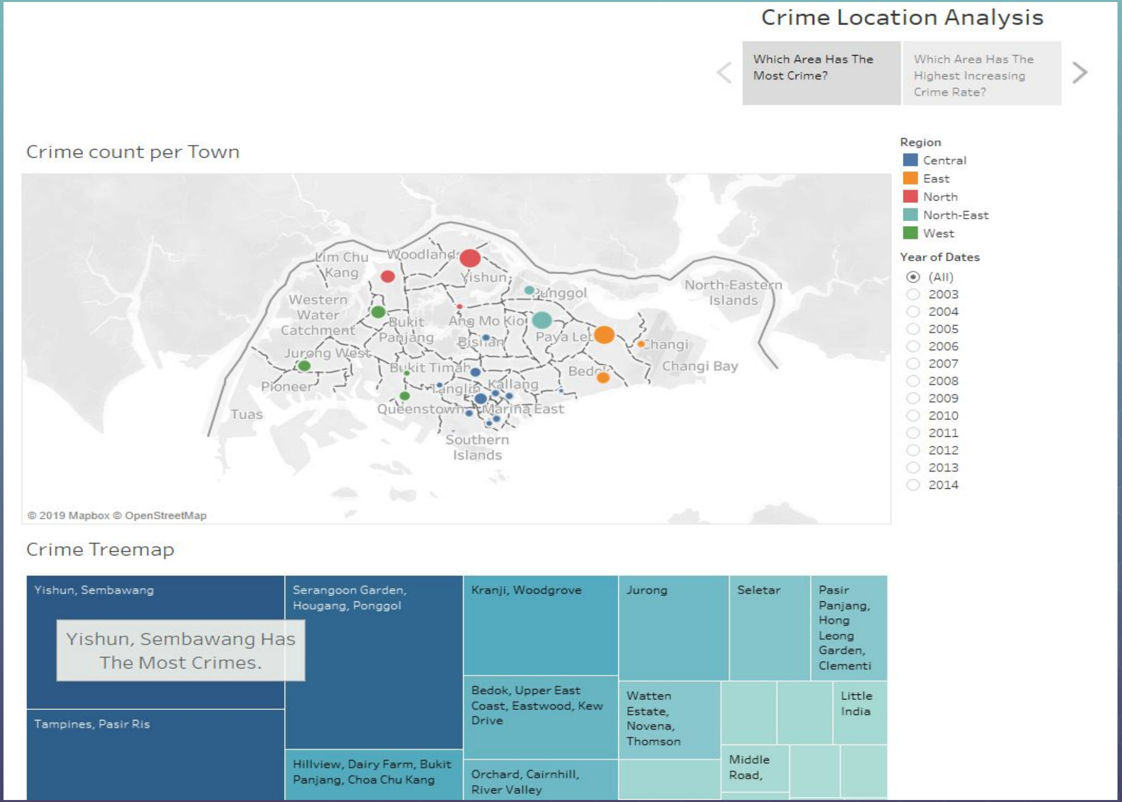
- View company insights.

Wine dataset



- Compare red and white wine.

Crime Dataset



Crime Location Analysis

Which Area Has The Most Crime?

Which Area Has The Highest Increasing Crime Rate?

Crime count per Town

Region

- 🟦 Central
- 🟧 East
- 🟥 North
- 🟩 North-East
- 🟩 West

Year of Dates

- ● (All)
- ○ 2003
- ○ 2004
- ○ 2005
- ○ 2006
- ○ 2007
- ○ 2008
- ○ 2009
- ○ 2010
- ○ 2011
- ○ 2012
- ○ 2013
- ○ 2014

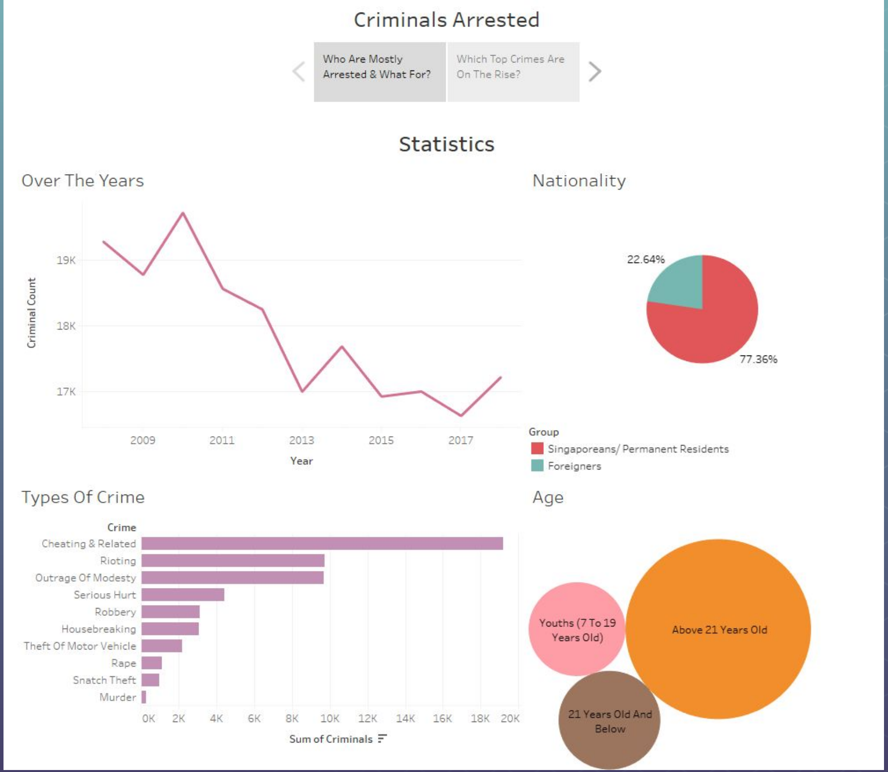
Crime Treemap

Yishun, Sembawang
Serangoon Garden, Hougang, Ponggol
Kranji, Woodgrove
Jurong
Seletar
Pasir Panjang, Hong Leong Garden, Clementi
Tampines, Pasir Ris
Bedok, Upper East Coast, Eastwood, Kew Drive
Watten Estate, Novena, Thomson
Little India
Middle Road,
Hillview, Dairy Farm, Bukit Panjang, Choa Chu Kang
Orchard, Cairnhill, River Valley

Yishun, Sembawang Has The Most Crimes.

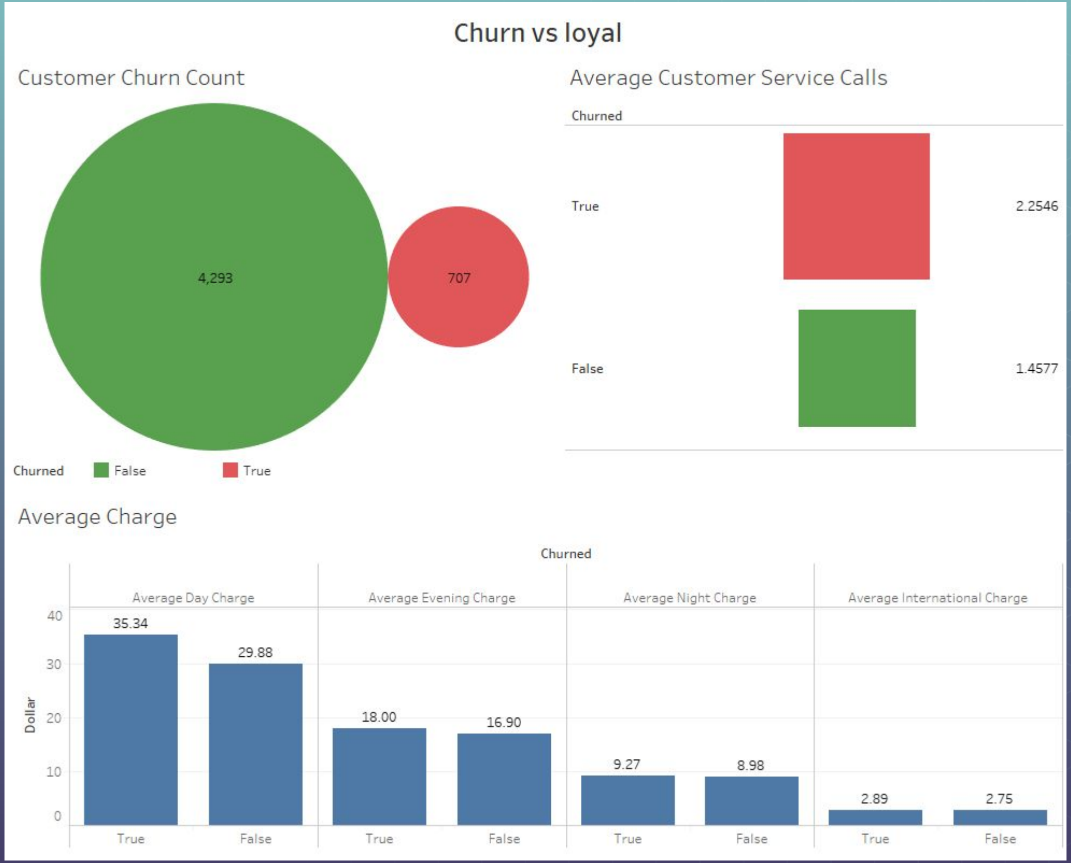
- View insights about location of crimes.

Criminal dataset



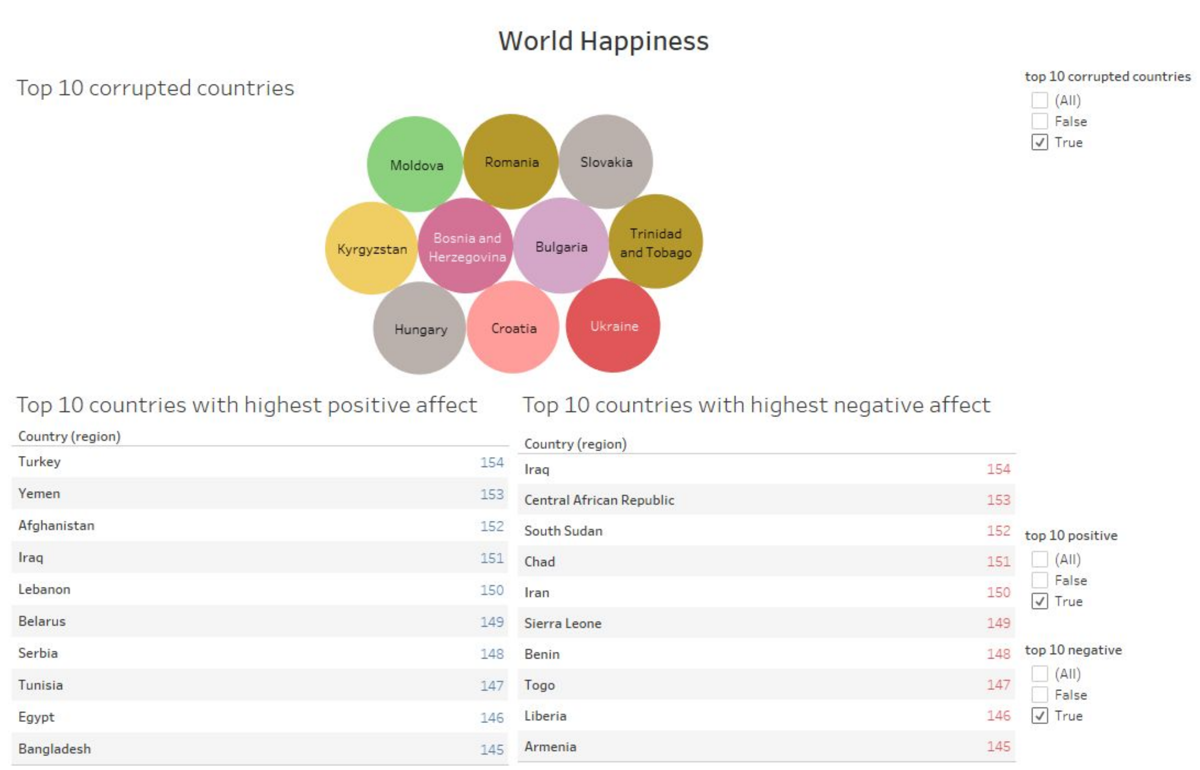
- Get criminals insights.

Telecom Customer Churn Dataset



- View insights about customer.

World happiness dataset



World Happiness

Top 10 corrupted countries

top 10 corrupted countries

- □ (All)
- □ False
- ☑ True

Top 10 countries with highest positive affect

Country (region)	
Turkey	154
Yemen	153
Afghanistan	152
Iraq	151
Lebanon	150
Belarus	149
Serbia	148
Tunisia	147
Egypt	146
Bangladesh	145

Top 10 countries with highest negative affect

Country (region)	
Iraq	154
Central African Republic	153
South Sudan	152
Chad	151
Iran	150
Sierra Leone	149
Benin	148
Togo	147
Liberia	146
Armenia	145

top 10 positive

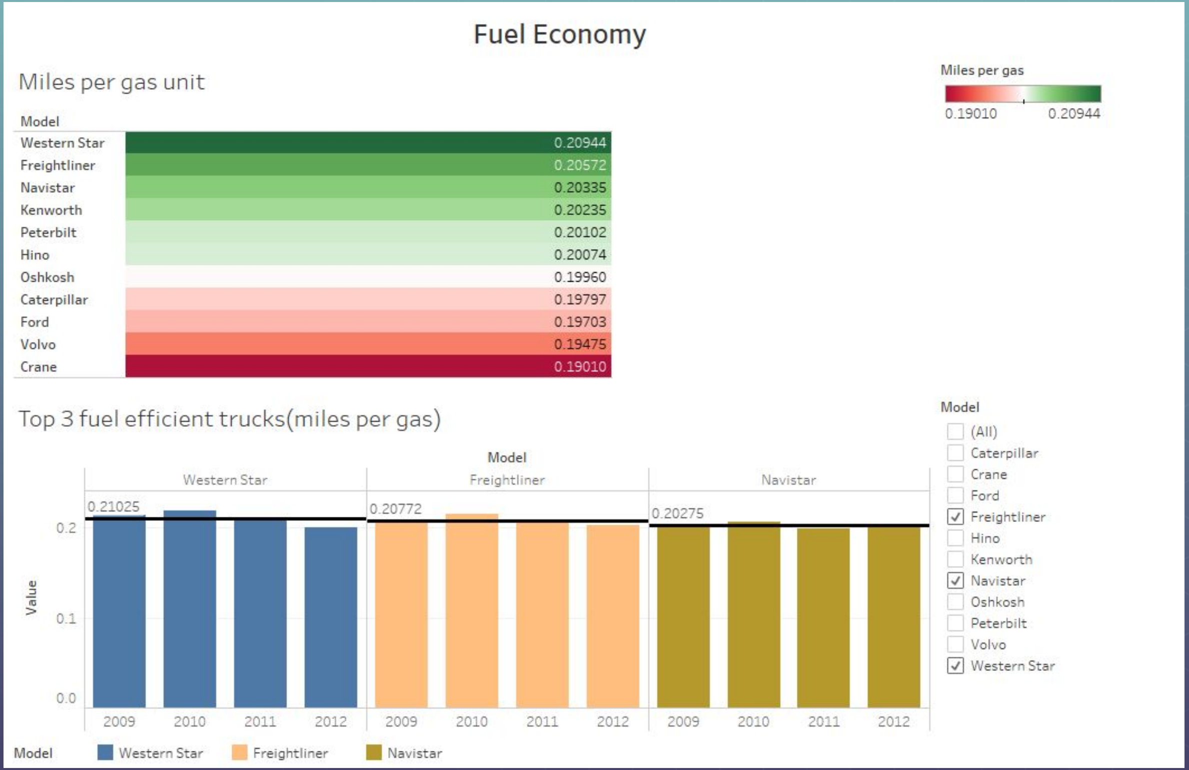
- □ (All)
- □ False
- ☑ True

top 10 negative

- □ (All)
- □ False
- ☑ True

- Understand countries' placing

Trucks dataset



Fuel Economy

Miles per gas unit

Model	
Western Star	0.20944
Freightliner	0.20572
Navistar	0.20335
Kenworth	0.20235
Peterbilt	0.20102
Hino	0.20074
Oshkosh	0.19960
Caterpillar	0.19797
Ford	0.19703
Volvo	0.19475
Crane	0.19010

Miles per gas

0.19010 0.20944

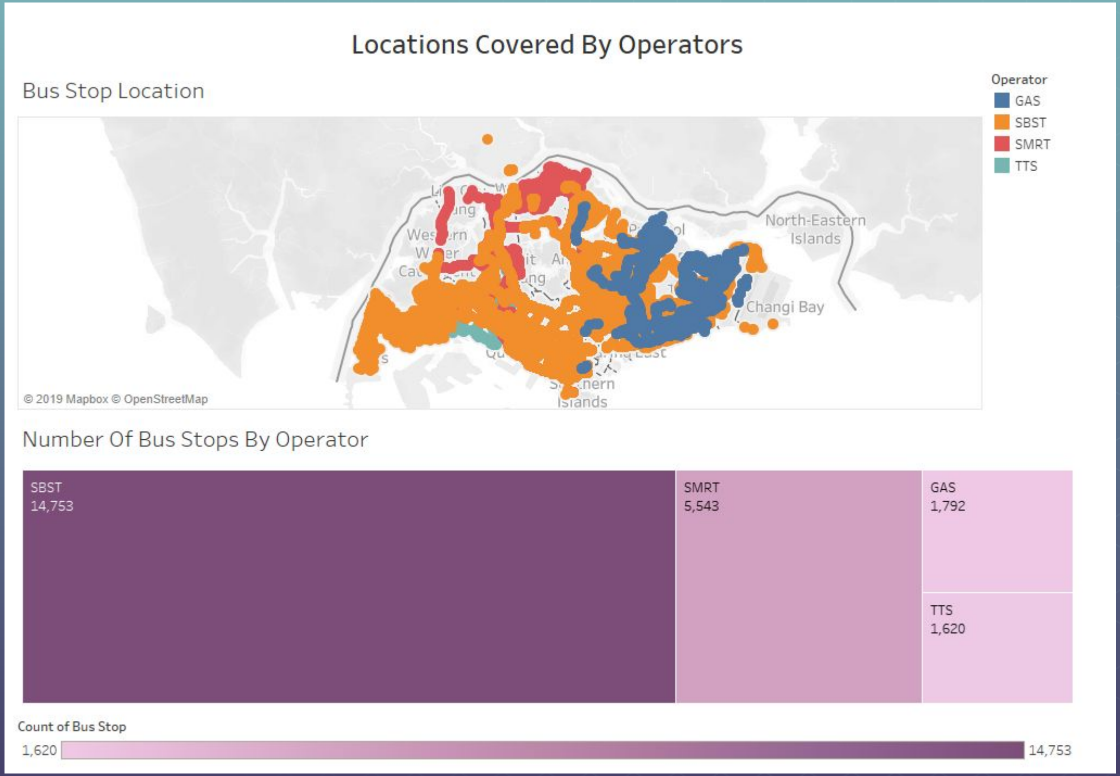
Top 3 fuel efficient trucks(miles per gas)

Model

- ☐ (All)
- ☐ Caterpillar
- ☐ Crane
- ☐ Ford
- ☑ Freightliner
- ☐ Hino
- ☐ Kenworth
- ☑ Navistar
- ☐ Oshkosh
- ☐ Peterbilt
- ☐ Volvo
- ☑ Western Star

- View information about different trucks.

Bus dataset



Locations Covered By Operators

Bus Stop Location

Operator
- 🟦 GAS
- 🟧 SBST
- 🟥 SMRT
- 🟩 TTS

Number Of Bus Stops By Operator

SBST 14,753

SMRT 5,543

GAS 1,792

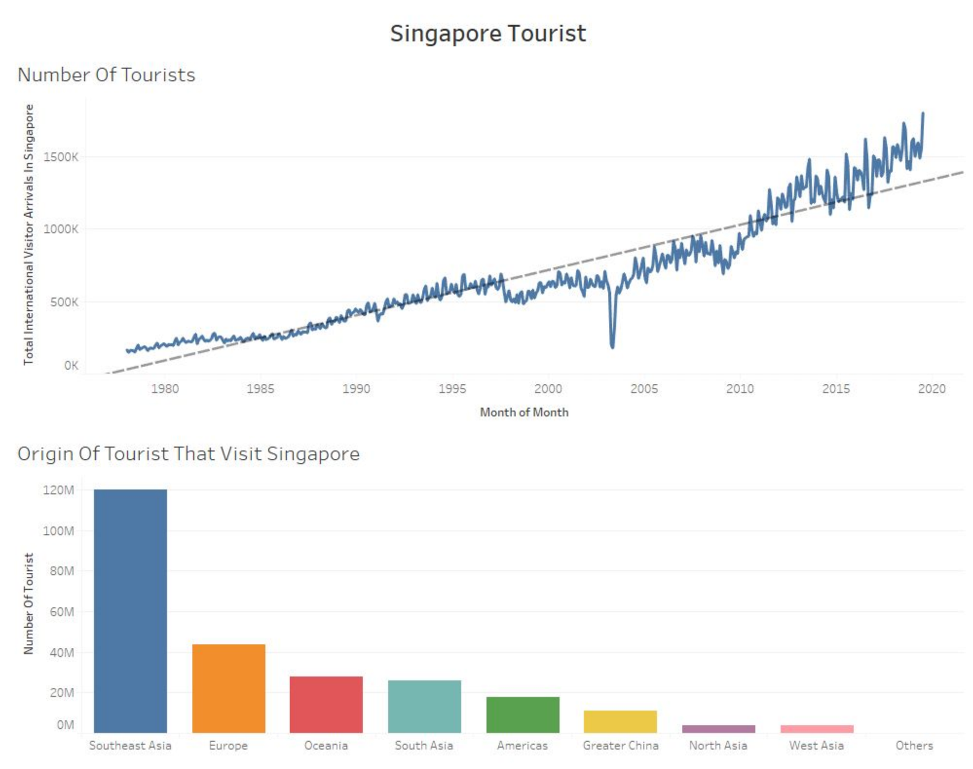
TTS 1,620

Count of Bus Stop

1,620 – 14,753

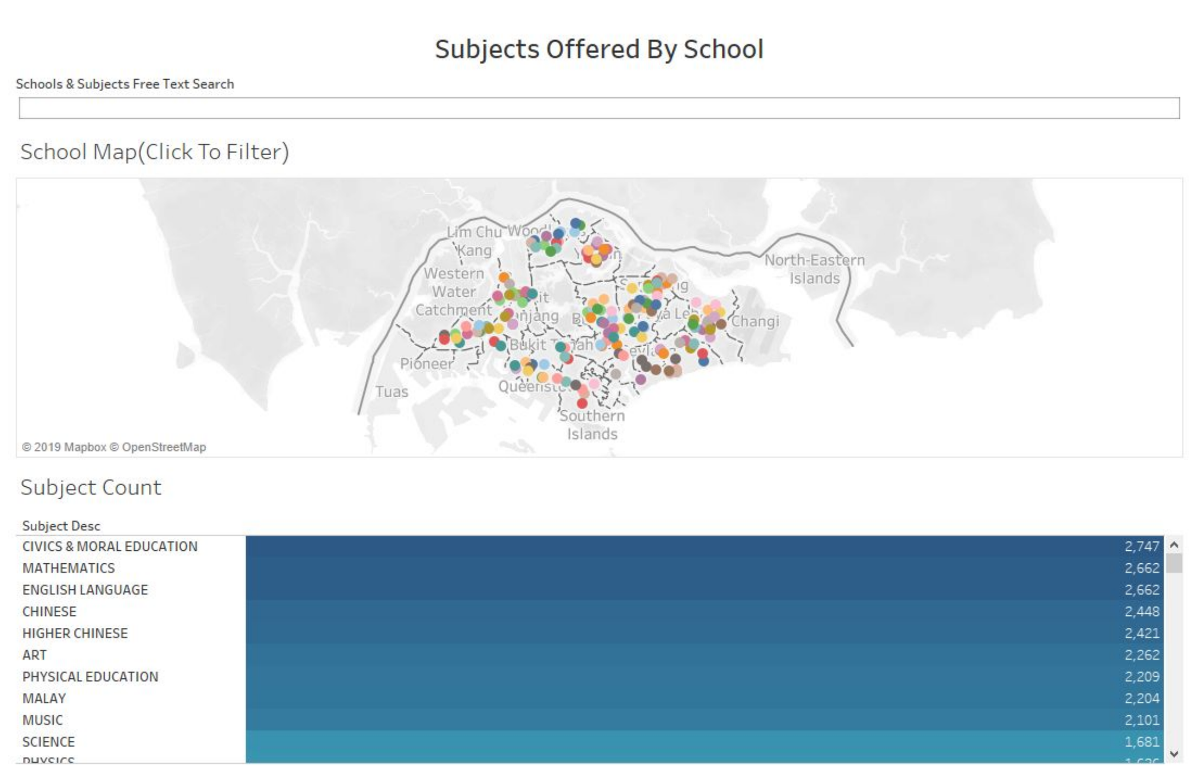
- Get bus services information.

Tourism dataset



- View tourist information.

School dataset



Subjects Offered By School

Schools & Subjects Free Text Search

School Map(Click To Filter)

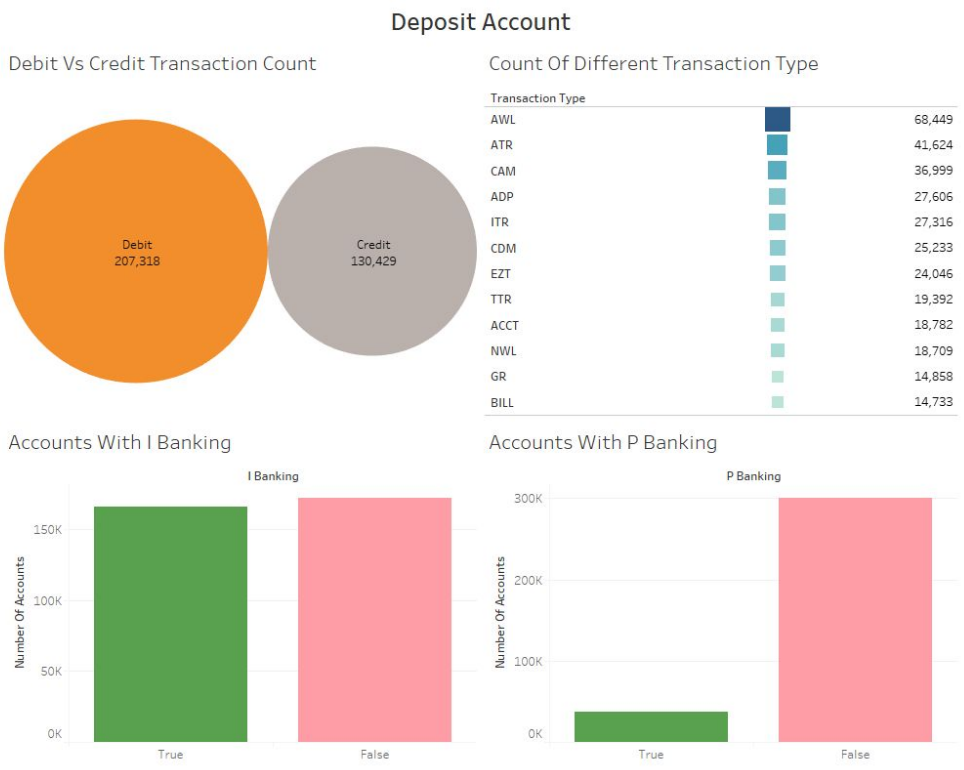
© 2019 Mapbox © OpenStreetMap

Subject Count

Subject Desc	
CIVICS & MORAL EDUCATION	2,747
MATHEMATICS	2,662
ENGLISH LANGUAGE	2,662
CHINESE	2,448
HIGHER CHINESE	2,421
ART	2,262
PHYSICAL EDUCATION	2,209
MALAY	2,204
MUSIC	2,101
SCIENCE	1,681
PHYSICS	[illegible]

- Find information about schools.

Banking dataset



Deposit Account

Debit Vs Credit Transaction Count

Count Of Different Transaction Type

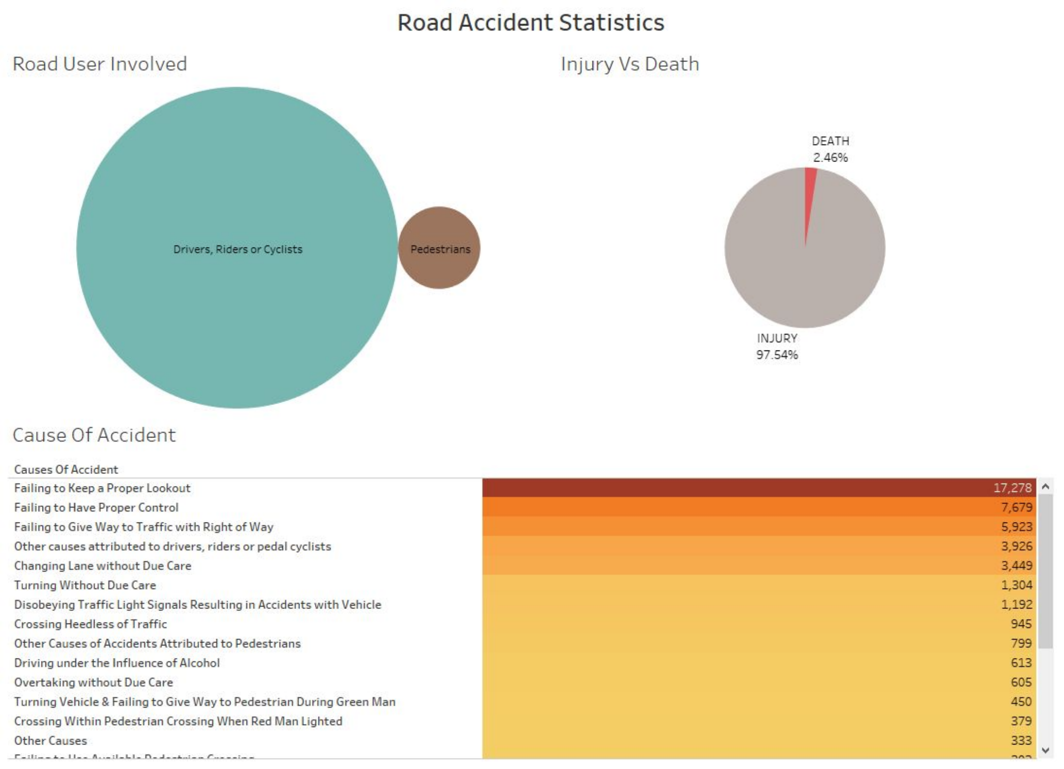
Transaction Type	
AWL	68,449
ATR	41,624
CAM	36,999
ADP	27,606
ITR	27,316
CDM	25,233
EZT	24,046
TTR	19,392
ACCT	18,782
NWL	18,709
GR	14,858
BILL	14,733

Accounts With I Banking

Accounts With P Banking

- View insights about bank accounts.

Road accident dataset



Road Accident Statistics

Road User Involved

Injury Vs Death

Cause Of Accident

Causes Of Accident	
Failing to Keep a Proper Lookout	17,278
Failing to Have Proper Control	7,679
Failing to Give Way to Traffic with Right of Way	5,923
Other causes attributed to drivers, riders or pedal cyclists	3,926
Changing Lane without Due Care	3,449
Turning Without Due Care	1,304
Disobeying Traffic Light Signals Resulting in Accidents with Vehicle	1,192
Crossing Heedless of Traffic	945
Other Causes of Accidents Attributed to Pedestrians	799
Driving under the Influence of Alcohol	613
Overtaking without Due Care	605
Turning Vehicle & Failing to Give Way to Pedestrian During Green Man	450
Crossing Within Pedestrian Crossing When Red Man Lighted	379
Other Causes	333
Failing to Use Available Pedestrian Crossing	[illegible]

- Get insights on road accidents.

Address Web Scrape

Why?

Required a set of local addresses which are tied to a region for my mapper

SRX

What is SRX

- SRX is a property website (Sale/Rent)

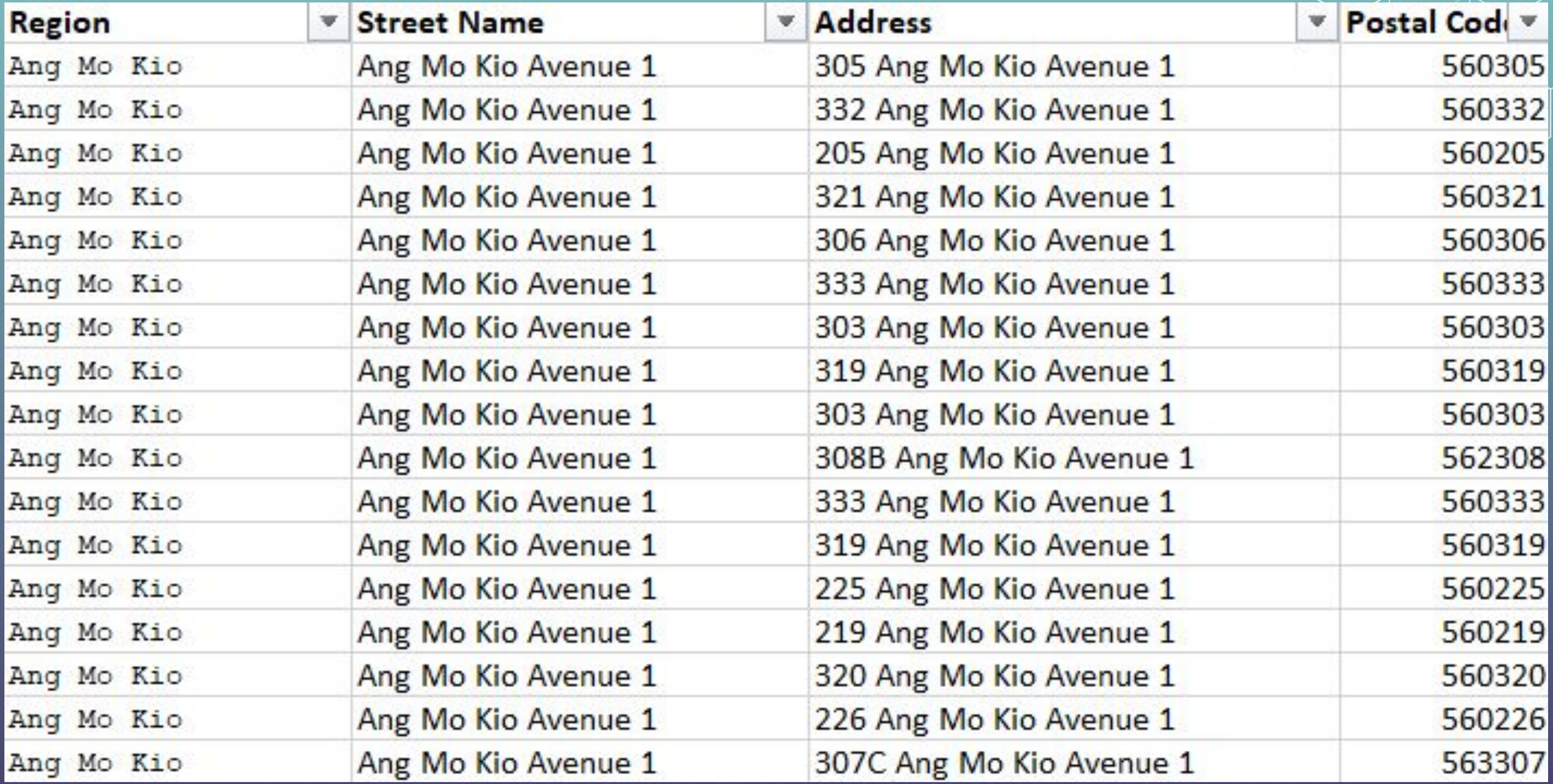
How?

- Used **python** (programming language) and 2 of libraries(**Pandas** & **BeautifulSoup**) on **Jupyter Notebook** (open source application for coding)

- Collected Addresses & their associated region

- Collected a total of approximately 10,000 Rows of Address Data.

Outcome of SRX Scrape



Region	Street Name	Address	Postal Cod
Ang Mo Kio	Ang Mo Kio Avenue 1	305 Ang Mo Kio Avenue 1	560305
Ang Mo Kio	Ang Mo Kio Avenue 1	332 Ang Mo Kio Avenue 1	560332
Ang Mo Kio	Ang Mo Kio Avenue 1	205 Ang Mo Kio Avenue 1	560205
Ang Mo Kio	Ang Mo Kio Avenue 1	321 Ang Mo Kio Avenue 1	560321
Ang Mo Kio	Ang Mo Kio Avenue 1	306 Ang Mo Kio Avenue 1	560306
Ang Mo Kio	Ang Mo Kio Avenue 1	333 Ang Mo Kio Avenue 1	560333
Ang Mo Kio	Ang Mo Kio Avenue 1	303 Ang Mo Kio Avenue 1	560303
Ang Mo Kio	Ang Mo Kio Avenue 1	319 Ang Mo Kio Avenue 1	560319
Ang Mo Kio	Ang Mo Kio Avenue 1	303 Ang Mo Kio Avenue 1	560303
Ang Mo Kio	Ang Mo Kio Avenue 1	308B Ang Mo Kio Avenue 1	562308
Ang Mo Kio	Ang Mo Kio Avenue 1	333 Ang Mo Kio Avenue 1	560333
Ang Mo Kio	Ang Mo Kio Avenue 1	319 Ang Mo Kio Avenue 1	560319
Ang Mo Kio	Ang Mo Kio Avenue 1	225 Ang Mo Kio Avenue 1	560225
Ang Mo Kio	Ang Mo Kio Avenue 1	219 Ang Mo Kio Avenue 1	560219
Ang Mo Kio	Ang Mo Kio Avenue 1	320 Ang Mo Kio Avenue 1	560320
Ang Mo Kio	Ang Mo Kio Avenue 1	226 Ang Mo Kio Avenue 1	560226
Ang Mo Kio	Ang Mo Kio Avenue 1	307C Ang Mo Kio Avenue 1	563307

Mapping Foreign Datasets to Local Context

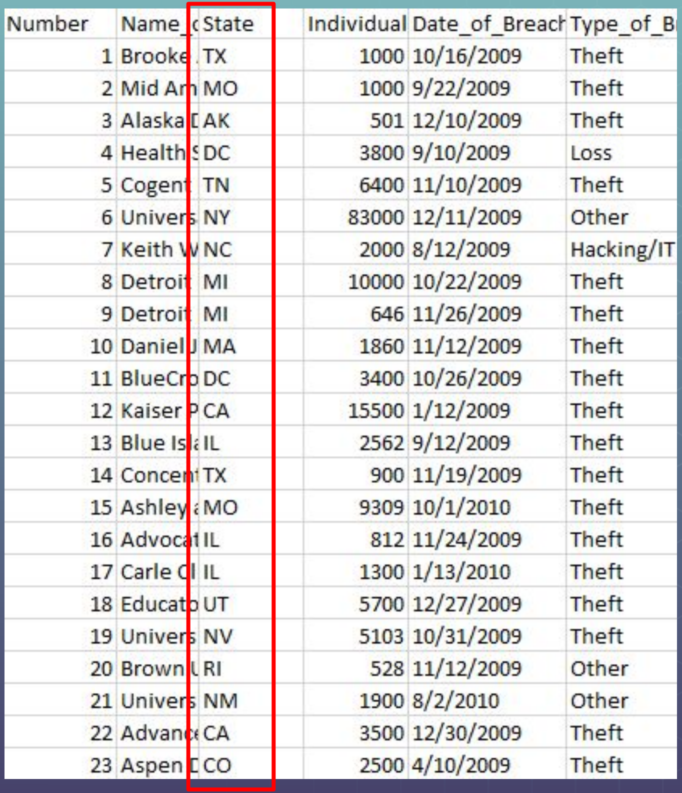
The datasets we are working on may not be local datasets. Hence the addresses in the datasets will be hard for students/adult learners to understand/visualize due to the unfamiliarity of the locations.

The purpose of mapping the datasets to singapore region is so that the user of the datasets can understand/visualize the datasets better
(User of the data will know where “Yishun” is instead of “TX” or “TEXAS”)

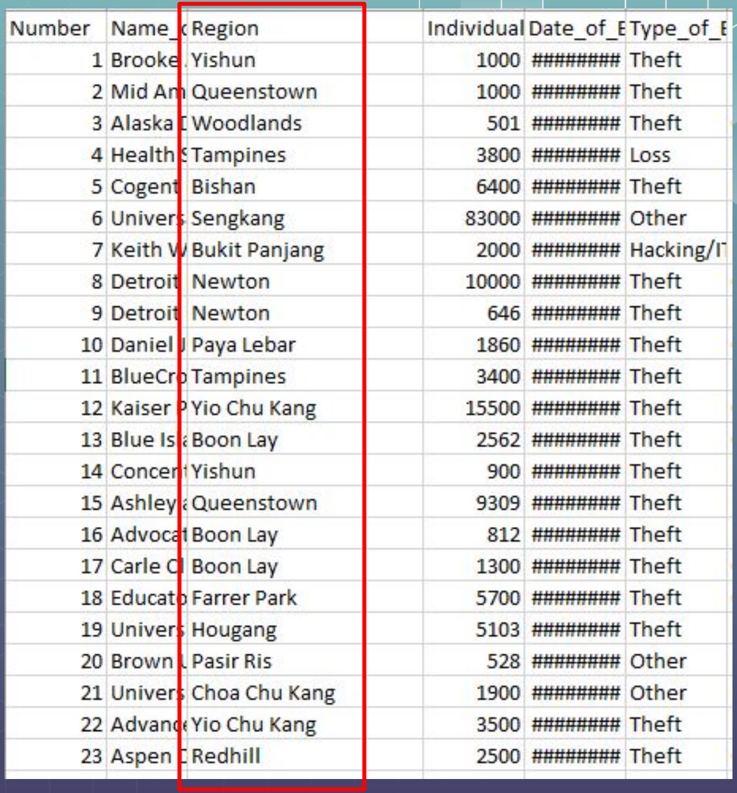
Recap of Previous Mapping Program

- Programmed using Python with the use of only “Pandas” Library
- Can only map Singapore Regions (Yishun, Yio Chu Kang, Sengkang etc.)
- Every dataset would require the modification of codes before being able to execute the program.

Results of old mapper

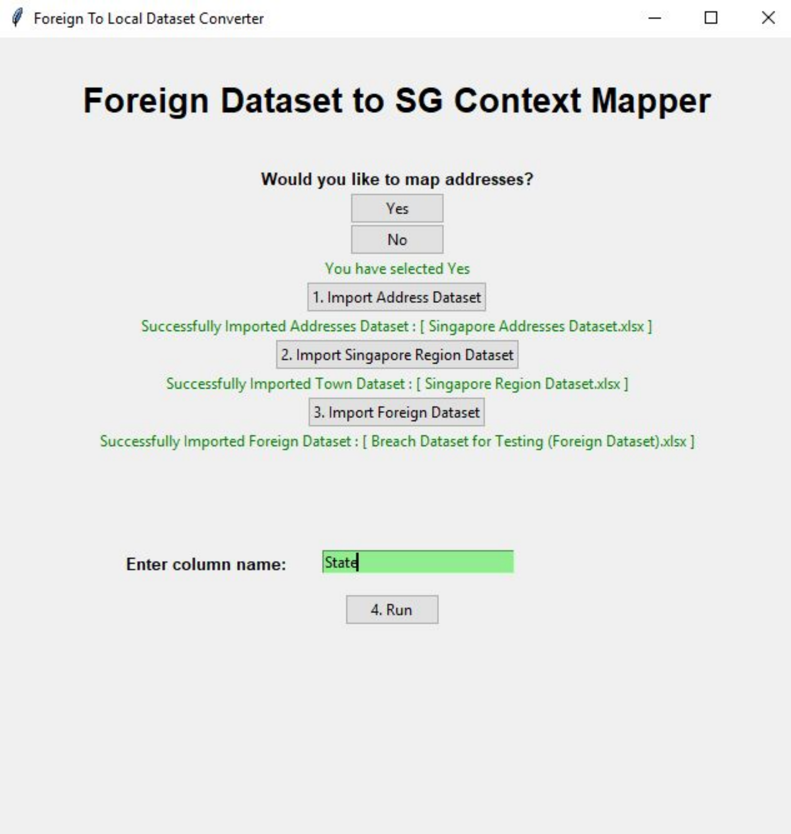


Number	Name_	State		Individual	Date_of_Breach	Type_of_B
1	Brooke	TX		1000	10/16/2009	Theft
2	Mid Am	MO		1000	9/22/2009	Theft
3	Alaska D	AK		501	12/10/2009	Theft
4	Health S	DC		3800	9/10/2009	Loss
5	Cogent	TN		6400	11/10/2009	Theft
6	Univers	NY		83000	12/11/2009	Other
7	Keith W	NC		2000	8/12/2009	Hacking/IT
8	Detroit	MI		10000	10/22/2009	Theft
9	Detroit	MI		646	11/26/2009	Theft
10	Daniel J	MA		1860	11/12/2009	Theft
11	BlueCro	DC		3400	10/26/2009	Theft
12	Kaiser P	CA		15500	1/12/2009	Theft
13	Blue Isl	IL		2562	9/12/2009	Theft
14	Concen	TX		900	11/19/2009	Theft
15	Ashley	MO		9309	10/1/2010	Theft
16	Advocat	IL		812	11/24/2009	Theft
17	Carle Cl	IL		1300	1/13/2010	Theft
18	Educato	UT		5700	12/27/2009	Theft
19	Univers	NV		5103	10/31/2009	Theft
20	Brown U	RI		528	11/12/2009	Other
21	Univers	NM		1900	8/2/2010	Other
22	Advanc	CA		3500	12/30/2009	Theft
23	Aspen D	CO		2500	4/10/2009	Theft



Number	Name_	Region		Individual	Date_of_E	Type_of_E
1	Brooke	Yishun		1000	#########	Theft
2	Mid Am	Queenstown		1000	#########	Theft
3	Alaska D	Woodlands		501	#########	Theft
4	Health S	Tampines		3800	#########	Loss
5	Cogent	Bishan		6400	#########	Theft
6	Univers	Sengkang		83000	#########	Other
7	Keith W	Bukit Panjang		2000	#########	Hacking/IT
8	Detroit	Newton		10000	#########	Theft
9	Detroit	Newton		646	#########	Theft
10	Daniel J	Paya Lebar		1860	#########	Theft
11	BlueCro	Tampines		3400	#########	Theft
12	Kaiser P	Yio Chu Kang		15500	#########	Theft
13	Blue Isl	Boon Lay		2562	#########	Theft
14	Concen	Yishun		900	#########	Theft
15	Ashley	Queenstown		9309	#########	Theft
16	Advocat	Boon Lay		812	#########	Theft
17	Carle Cl	Boon Lay		1300	#########	Theft
18	Educato	Farrer Park		5700	#########	Theft
19	Univers	Hougang		5103	#########	Theft
20	Brown U	Pasir Ris		528	#########	Other
21	Univers	Choa Chu Kang		1900	#########	Other
22	Advanc	Yio Chu Kang		3500	#########	Theft
23	Aspen D	Redhill		2500	#########	Theft

Updated Mapping Program


Foreign To Local Dataset Converter
Yes
No
1. Import Address Dataset
2. Import Singapore Region Dataset
3. Import Foreign Dataset
4. Run

Foreign Dataset to SG Context Mapper

Would you like to map addresses?

You have selected Yes

Successfully Imported Addresses Dataset : [Singapore Addresses Dataset.xlsx]

Successfully Imported Town Dataset : [Singapore Region Dataset.xlsx]

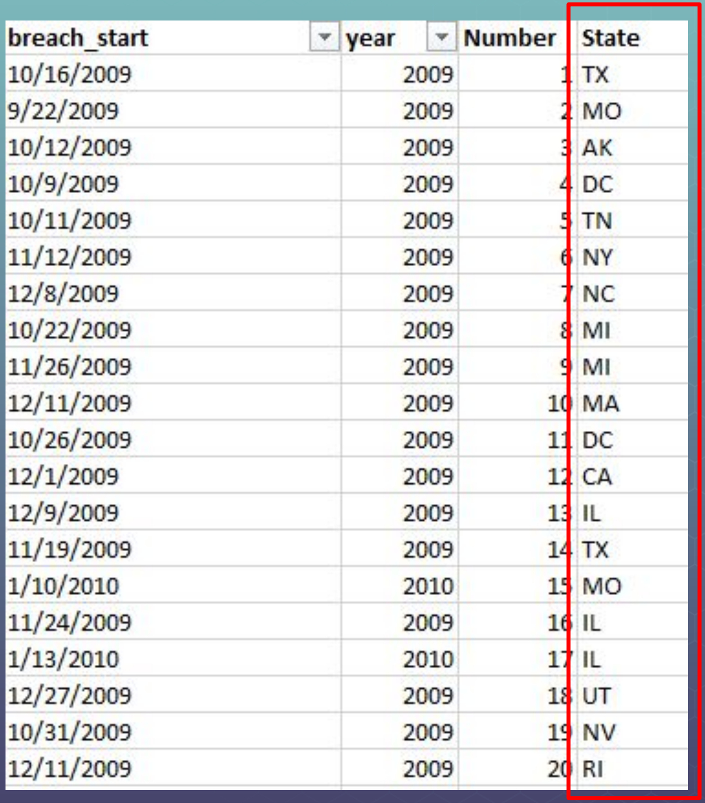
Successfully Imported Foreign Dataset : [Breach Dataset for Testing (Foreign Dataset).xlsx]

Enter column name: State

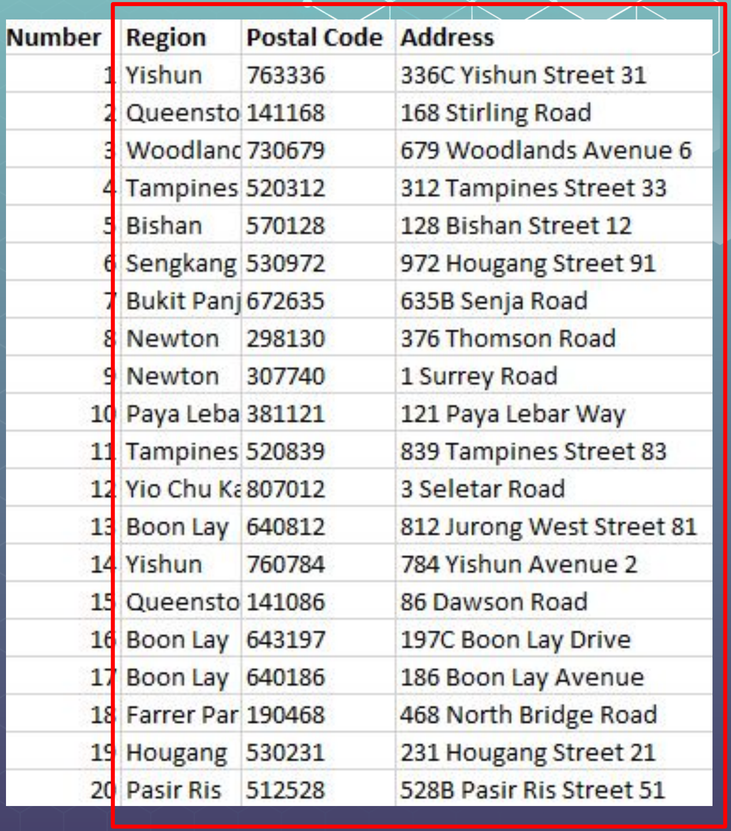
- Programmed using Python with the use of "Pandas" & "Tkinter" Library.
- Can map to local Regions & assign them new columns such as "Address" & "Postal Code" to make dataset richer for analysis purposes.
- Graphical User interface which allows mapping process to be executed without any modification of codes (easier to use)

Mapper Demo

Results of new mapper



breach_start	year	Number	State
10/16/2009	2009	1	TX
9/22/2009	2009	2	MO
10/12/2009	2009	3	AK
10/9/2009	2009	4	DC
10/11/2009	2009	5	TN
11/12/2009	2009	6	NY
12/8/2009	2009	7	NC
10/22/2009	2009	8	MI
11/26/2009	2009	9	MI
12/11/2009	2009	10	MA
10/26/2009	2009	11	DC
12/1/2009	2009	12	CA
12/9/2009	2009	13	IL
11/19/2009	2009	14	TX
1/10/2010	2010	15	MO
11/24/2009	2009	16	IL
1/13/2010	2010	17	IL
12/27/2009	2009	18	UT
10/31/2009	2009	19	NV
12/11/2009	2009	20	RI



Number	Region	Postal Code	Address
1	Yishun	763336	336C Yishun Street 31
2	Queensto	141168	168 Stirling Road
3	Woodlanc	730679	679 Woodlands Avenue 6
4	Tampines	520312	312 Tampines Street 33
5	Bishan	570128	128 Bishan Street 12
6	Sengkang	530972	972 Hougang Street 91
7	Bukit Panj	672635	635B Senja Road
8	Newton	298130	376 Thomson Road
9	Newton	307740	1 Surrey Road
10	Paya Leba	381121	121 Paya Lebar Way
11	Tampines	520839	839 Tampines Street 83
12	Yio Chu Ka	807012	3 Seletar Road
13	Boon Lay	640812	812 Jurong West Street 81
14	Yishun	760784	784 Yishun Avenue 2
15	Queensto	141086	86 Dawson Road
16	Boon Lay	643197	197C Boon Lay Drive
17	Boon Lay	640186	186 Boon Lay Avenue
18	Farrer Par	190468	468 North Bridge Road
19	Hougang	530231	231 Hougang Street 21
20	Pasir Ris	512528	528B Pasir Ris Street 51

How it works?

- 1. Application would gather all the required files
  - Singapore Address Dataset and/or Singapore Region Dataset - Web Scrapped Singapore Addressed from SRX property website.
  - Foreign Dataset

- 2. Read in all the dataset using Pandas
- 3. Assign an index to the unique values of foreign dataset (e.g. State) ⟶ [TX = 0, MO =1, AK = 2]
- 4. Assign an index to the unique values of Singapore Dataset (Region) ⟶ [Yio Chu Kang = 0, Hougang = 1, Sembawang = 2]
- 5. Match the index together and update the dataset. [TX becomes Yio Chu Kang, MO becomes Hougang, AK becomes Sembawang]
- 6. IF user selects that they would like address to be mapped,
  - 6a. The program loops through every row in the updated Singapore Context dataset.
  - 6b. Reads in the region of current row. [e.g. Ang Mo Kio]
  - 6c. Filters the Singapore Address Dataset to only "Ang Mo Kio"
  - 6d. Randomly select one of the address of region "Ang Mo Kio" and assign it.

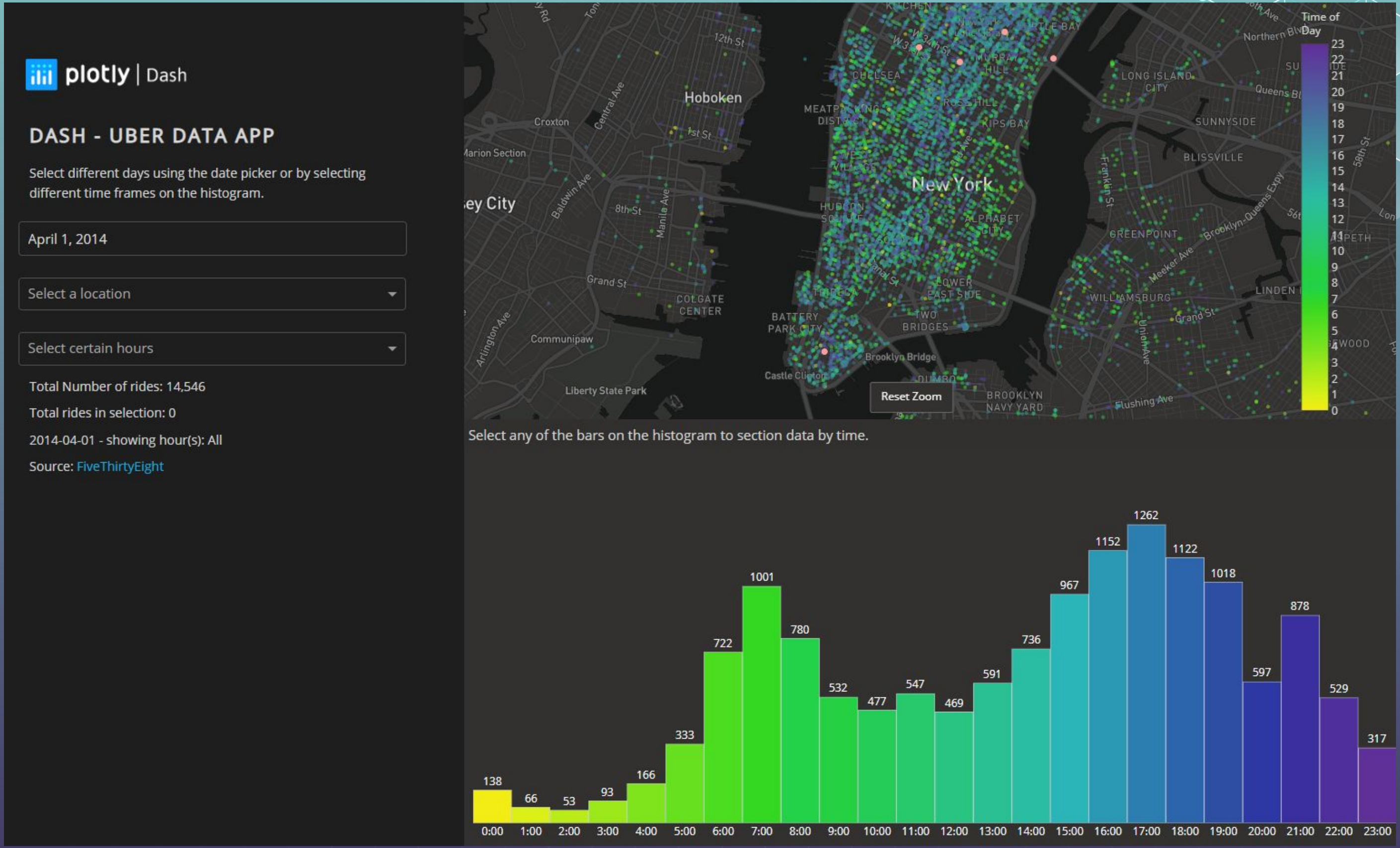
Research on Plotly Dash

What is Plotly Dash

- An open source Python framework for building responsive analytical web applications which do not require any JavaScript/HTML/CSS coding.

- Plotly Dash is built on top of :
  - **Flask** - A web framework which allows building of web applications without HTML/CSS
  - **Plotly.js** - A high-level, declarative charting library with over 40 chart types.
  - **React.js** - A JavaScript library for building user interfaces

Plotly Dashboard Example



DASH - UBER DATA APP

Select different days using the date picker or by selecting different time frames on the histogram.

April 1, 2014

Select a location ▾

Select certain hours ▾

Total Number of rides: 14,546

Total rides in selection: 0

2014-04-01 - showing hour(s): All

Source: FiveThirtyEight

Select any of the bars on the histogram to section data by time.

My attempt on Plotly Dash

Thoughts of Plotly Dash

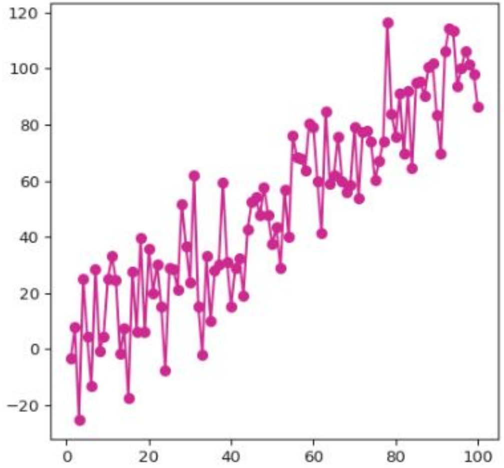
- Nice looking
- Very Flexible
- Not many resources available explaining how to use Plotly.
- Time consuming compared to visualization applications (Tableau/SAS etc)
- Datasets requires a lot of data preparation before being able to be used for charting. (Very Tedious)

Other Open Sourced Charting Frameworks?

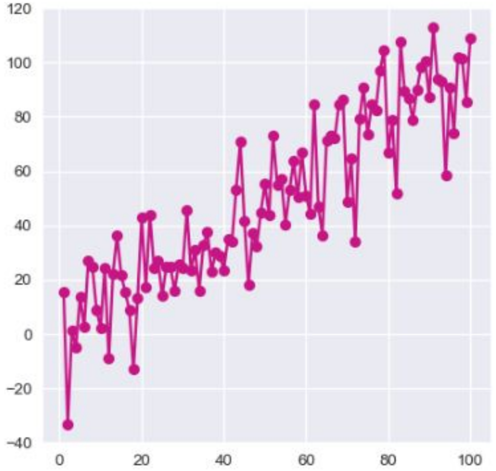
Matplotlib

Seaborn

- A n
  t
- V th to



#106 Matplotlib style



#106 .. with Seaborn style

Why use Plotly?

- Plotly has a wide variety of chart types which looks appealing.
- The charts are not rendered as an Image (Making it interactive)
- The charts can be published onto a web server making it visible to anyone with the link.
  - Includes viewing on mobile devices
- **FREE**, does not require any licensing.

THANK YOU