CRUNCHING BIG DATA WITH VAEX

DR. ALVIN ANG

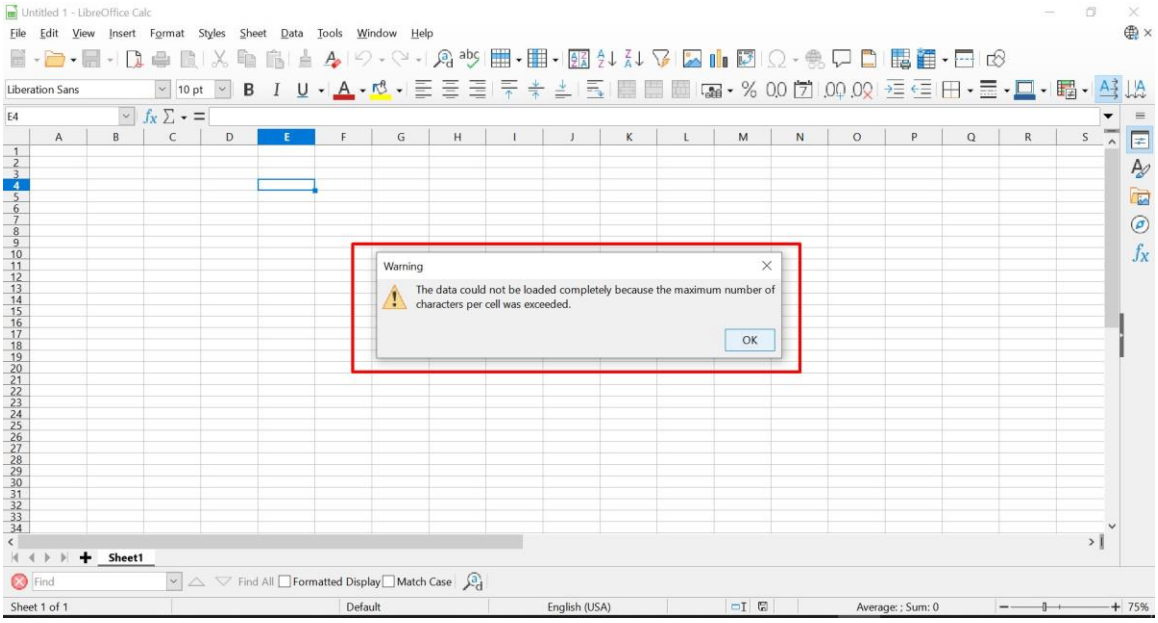
CONTENTS

I. WHY DO WE NEED PYTHON TO DO BIG DATA? CAN'T EXCEL HANDLE IT?

References:

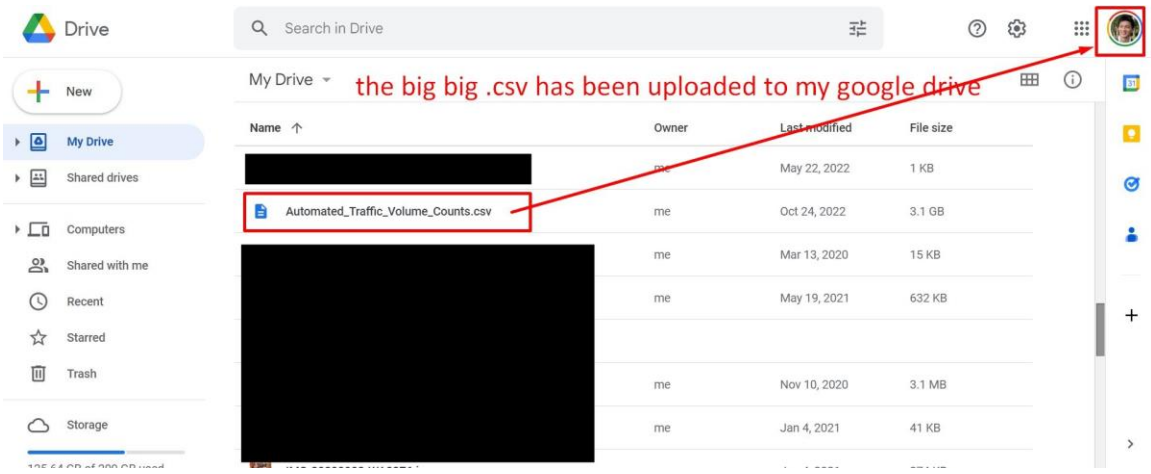
- https://www.alvinang.sg/s/VAEX_for_Crunching_Big_Data_by_Dr_Alvin_Ang.ipynb
- https://www.kaggle.com/datasets/aadimator/nyc-automated-traffic-volume-counts?select=Automated_Traffic_Volume_Counts.csv
- https://medium.com/@jovan.veljanoski/8-powerful-vaex-dataframe-features-you-might-have-not-known-about-b2e2d3d3ee9
- https://towardsdatascience.com/https-medium-com-jovan-veljanoski-flying-high-with-vaex-analysis-of-over-30-years-of-flight-data-in-python-b224825a6d56

- Excel crashed when I tried to open a CSV that is 3GB
- It has 27 million rows of data.
- But Excel maximum only has 1 million rows of data per sheet.

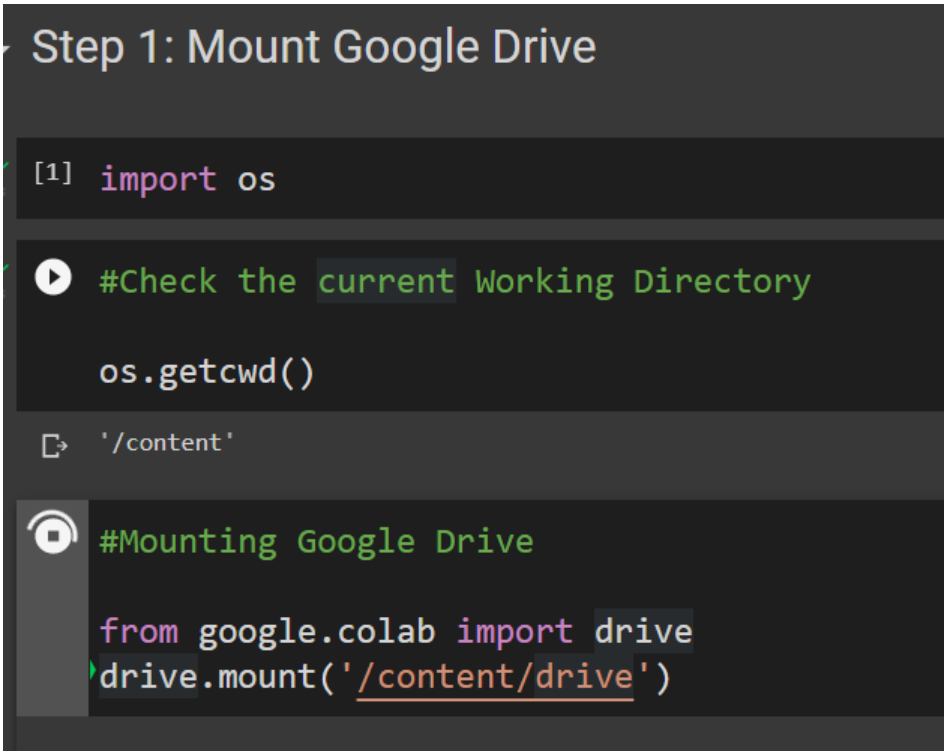


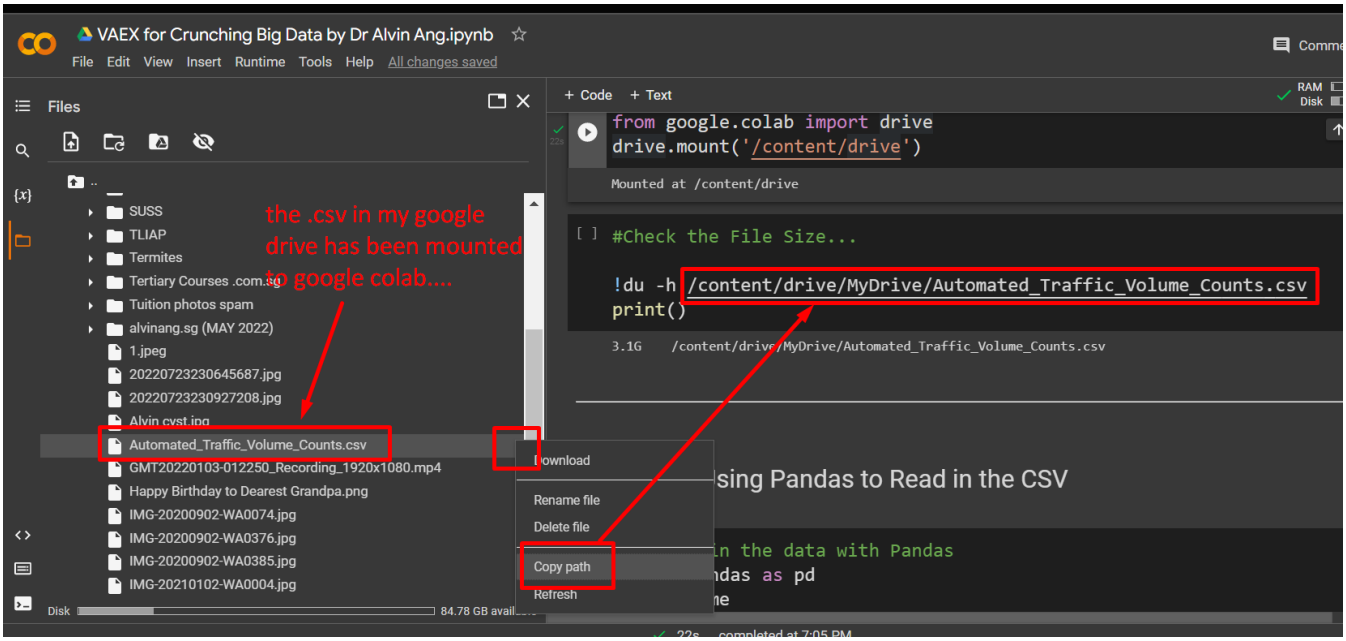
Step 0: In The Beginning....

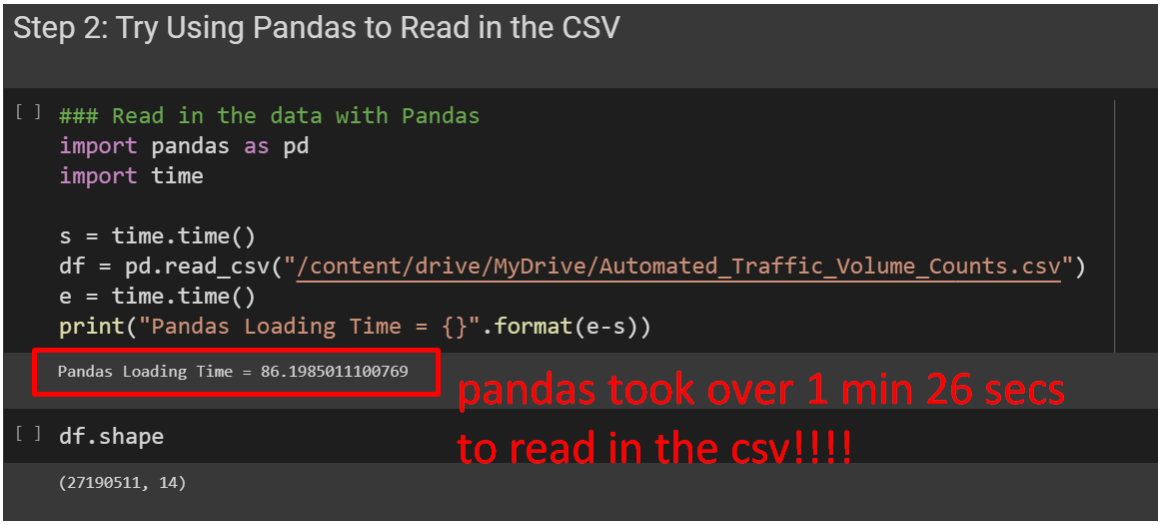
- Go to https://www.kaggle.com/datasets/aadimator/nyc-automated-traffic-volume-counts?select=Automated_Traffic_Volume_Counts.csv and download the .csv data (into your laptop)
- It is around 3 GB... so it will take some time....
- Once downloaded, upload it into your Personal Google Drive



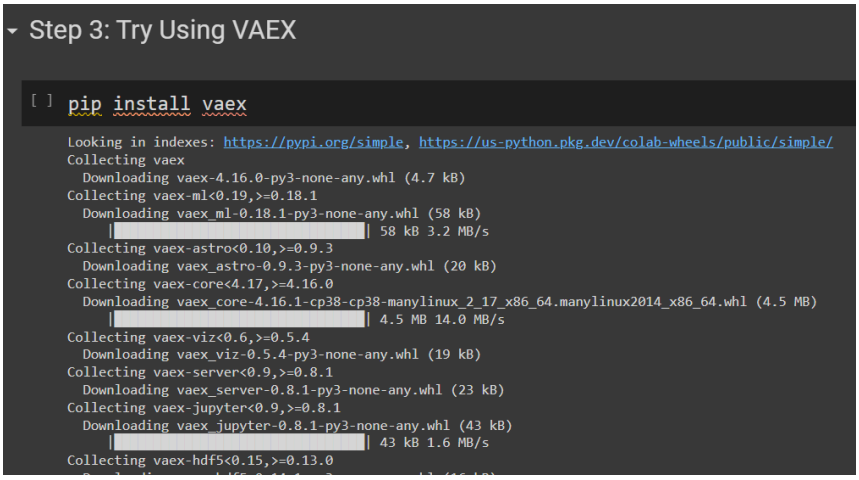
III. STEP 1: MOUNT GOOGLE DRIVE

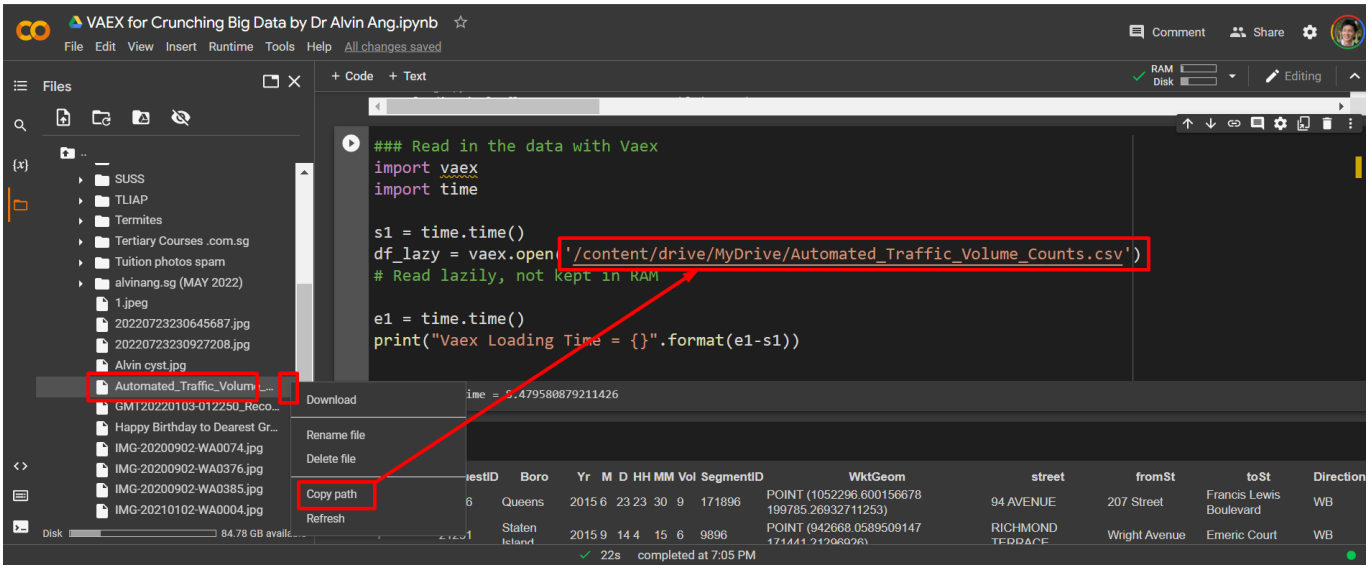






V. STEP 3: TRY USING VAEX …





- VAEX is able to read quickly because of "lazy" reading → it doesn't read in the entire dataset into the RAM
- It only scans the Metadata and displays it.

```
df_lazy
```

	#	RequestID	Boro	Yr	M	D	HH	MM	Vol	SegmentID	WktGeom	street	fromSt	toSt	Direction
	0	20856	Queens	2015	6	23	23	30	9	171896	POINT (1052296.600156678 199785.26932711253)	94 AVENUE	207 Street	Francis Lewis Boulevard	WB
	1	21231	Staten Island	2015	9	14	4	15	6	9896	POINT (942668.0589509147 171441.21296926)	RICHMOND TERRACE	Wright Avenue	Emeric Court	WB
	2	29279	Bronx	2017	10	19	4	30	85	77817	POINT (1016508.0034050211 235221.59092266942)	HUNTS POINT AVENUE	Whittier Street	Randall Avenue	NB
	3	27019	Brooklyn	2017	11	7	18	30	168	188023	POINT (992925.4316054962 184116.82855457635)	FLATBUSH AVENUE	Brighton Line	Brighton Line	NB
	4	26734	Manhattan	2017	11	3	22	0	355	137516	POINT (1004175.9505178436 247779.63624949602)	WASHINGTON BRIDGE	Harlem River Shoreline	Harlem River Shoreline	EB
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	27,190,506	28843	Queens	2019	1	11	3	30	13	48625	POINT (1010498.7141876142 198418.68389121862)	FOREST AVENUE	Greene Avenue	Bleecker Street	NB
	27,190,507	21109	Brooklyn	2015	9	28	18	45	65	45144	POINT (1007888.490398333 194036.05423851966)	IRVING AVENUE	Grove Street	Myrtle Avenue	SB
	27,190,508	23455	Queens	2016	5	23	13	30	55	149068	POINT (1024888.1767334475 211528.3190361287)	111 STREET	48 Avenue	47 Avenue	NB
	27,190,509	22517	Bronx	2016	3	19	9	0	82	89528	POINT (1026078.0476007946 269670.6741176347)	WHITE PLAINS ROAD	East 243 Street	Dead end	SB
	27,190,510	0450	Queens	2012	10	19	0	30	20	69002	POINT (1009571.4 220310.5)	23 AV	38 ST	STEINWAY ST	SB

27 million rows!!!

```
[ ] print(f'Number of Rows: {df_lazy.shape[0]:,}')

    print(f'Number of Columns: {df_lazy.shape[1]}')
```

```
Number of Rows: 27,190,511
Number of Columns: 14
```

THE END

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.