

DR. ALVIN'S PUBLICATIONS

CRUNCHING BIG DATA WITH VAEX

DR. ALVIN ANG



1 | PAGE

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

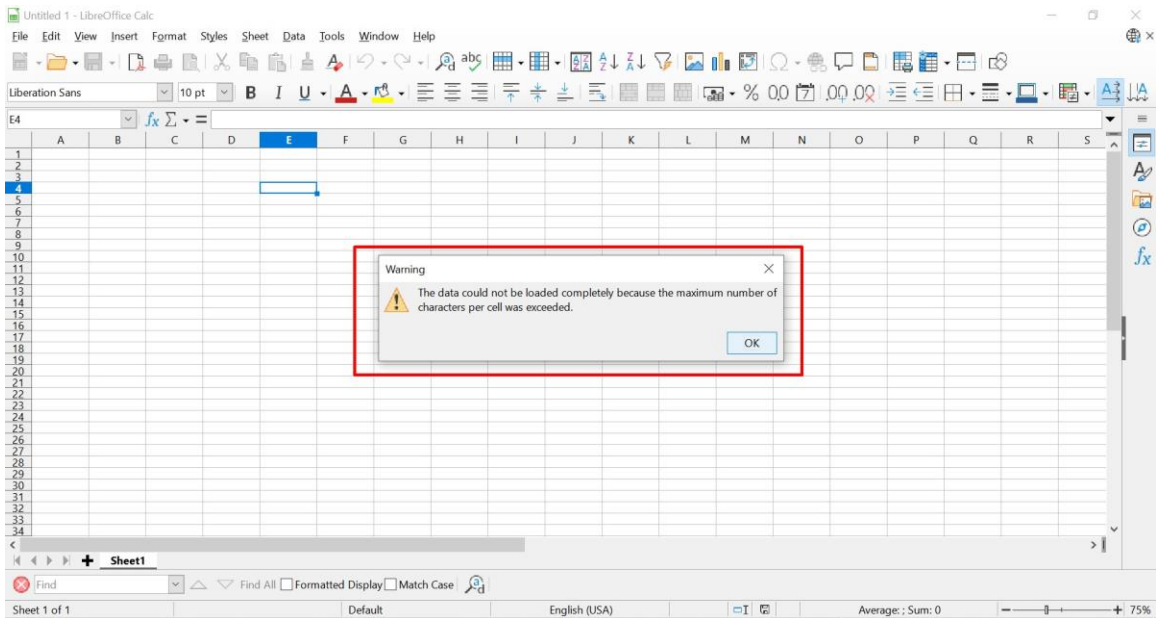
CONTENTS

<i>I. Why Do We Need Python To Do Big Data? Can't Excel Handle It?</i>	<i>3</i>
<i>II. Step 0: Download the Big Data (3GB).....</i>	<i>4</i>
<i>III. Step 1: Mount Google Drive</i>	<i>5</i>
<i>IV. Step 2: Try Using Pandas to REad in the CSV.....</i>	<i>6</i>
<i>V. Step 3: Try Using VAEX</i>	<i>7</i>
<i>About Dr. Alvin Ang</i>	<i>9</i>

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

References:

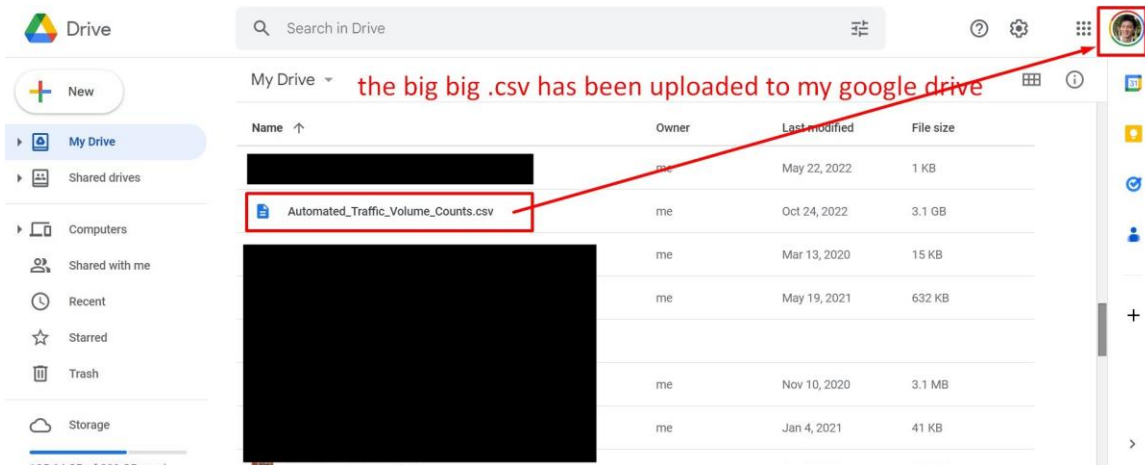
- https://www.alvinang.sg/s/VAEX_for_Crunching_Big_Data_by_Dr_Alvin_Ang.ipynb
- https://www.kaggle.com/datasets/aadimator/nyc-automated-traffic-volume-counts?select=Automated_Traffic_Volume_Counts.csv
- <https://medium.com/@jovan.veljanoski/8-powerful-vaex-dataframe-features-you-might-have-not-known-about-b2e2d3d3ee9>
- <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.



II. STEP 0: DOWNLOAD THE BIG DATA (3GB)

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

```
Step 1: Mount Google Drive

[1] import os

#Check the current Working Directory

os.getcwd()

'/content'

#Mounting Google Drive

from google.colab import drive
drive.mount('/content/drive')
```

The screenshot displays a Google Colab notebook interface. On the left, a file browser shows a directory structure with a file named 'Automated_Traffic_Volume_Counts.csv' highlighted. A red box is drawn around this file name, and a red arrow points to it from a text annotation: "the .csv in my google drive has been mounted to google colab...". A context menu is open over the file, with a red box around the 'Copy path' option and another red arrow pointing to it. The main code editor on the right shows the following code:

```
from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive

[ ] #Check the File Size...

!du -h /content/drive/MyDrive/Automated_Traffic_Volume_Counts.csv
print()
```

The output of the command shows the file size as 3.1G. Below the code editor, a title "Using Pandas to Read in the CSV" is visible, along with some partially obscured code: "In the data with Pandas", "ndas as pd", and "ne".

IV. STEP 2: TRY USING PANDAS TO READ IN THE CSV....

Step 2: Try Using Pandas to Read in the CSV

```
[ ] ### Read in the data with Pandas
import pandas as pd
import time

s = time.time()
df = pd.read_csv("/content/drive/MyDrive/Automated_Traffic_Volume_Counts.csv")
e = time.time()
print("Pandas Loading Time = {}".format(e-s))

Pandas Loading Time = 86.1985011100769

[ ] df.shape

(27190511, 14)
```

**pandas took over 1 min 26 secs
to read in the csv!!!!**

V. STEP 3: TRY USING VAEX ...

```
Step 3: Try Using VAEX

[ ] pip install vaex

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting vaex
  Downloading vaex-4.16.0-py3-none-any.whl (4.7 kB)
Collecting vaex-ml<0.19,>=0.18.1
  Downloading vaex_ml-0.18.1-py3-none-any.whl (58 kB)
    | 58 kB 3.2 MB/s
Collecting vaex-astro<0.10,>=0.9.3
  Downloading vaex_astro-0.9.3-py3-none-any.whl (20 kB)
Collecting vaex-core<4.17,>=4.16.0
  Downloading vaex_core-4.16.1-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.5 MB)
    | 4.5 MB 14.0 MB/s
Collecting vaex-viz<0.6,>=0.5.4
  Downloading vaex_viz-0.5.4-py3-none-any.whl (19 kB)
Collecting vaex-server<0.9,>=0.8.1
  Downloading vaex_server-0.8.1-py3-none-any.whl (23 kB)
Collecting vaex-jupyter<0.9,>=0.8.1
  Downloading vaex_jupyter-0.8.1-py3-none-any.whl (43 kB)
    | 43 kB 1.6 MB/s
Collecting vaex-hdfs<0.15,>=0.13.0
```

```
### Read in the data with Vaex
import vaex
import time

s1 = time.time()
df_lazy = vaex.open('/content/drive/MyDrive/Automated_Traffic_Volume_Counts.csv')
# Read lazily, not kept in RAM

e1 = time.time()
print("Vaex Loading Time = {}".format(e1-s1))
```

testID	Boro	Yr	M	D	HH	MM	Vol	SegmentID	WktGeom	street	fromSt	toSt	Direction
6	Queens	2015	6	23	23	30	9	171896	POINT (1052296 600156678 199785 26932711253)	94 AVENUE	207 Street	Francis Lewis Boulevard	WB
	Staten Island	2015	9	14	4	15	6	9896	POINT (942668 0589509147 171444 21206926)	RICHMOND TERRACE	Wright Avenue	Emeric Court	WB

Vaex Loading Time = 8.479580879211426 **only 8 seconds!!!**

- 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

```
[ ] 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.