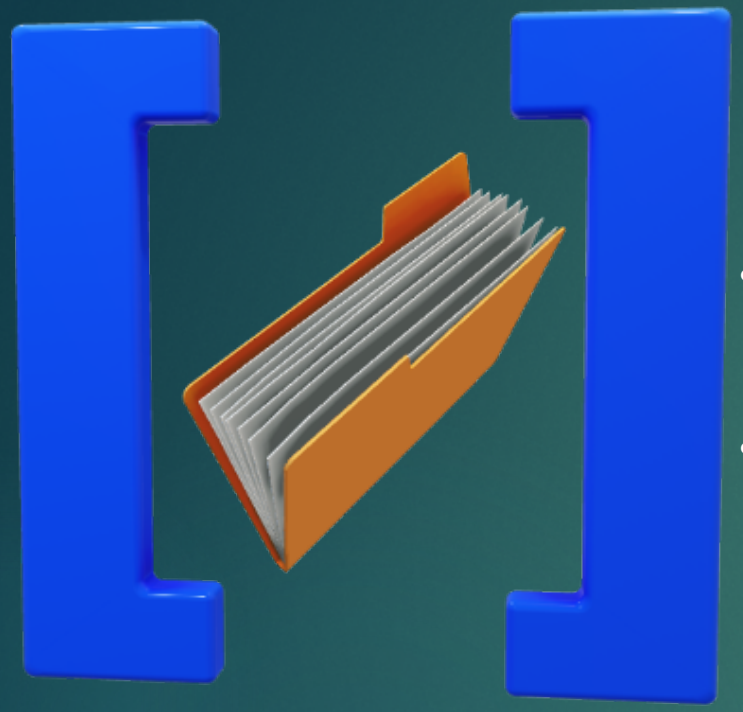
PROJECT PRESENTATION

ACRA DATA VISUALISATION



Dataset

- Choice of my Data set will be on data.gov.sg under Acra information entities.
- Base on my selection, I will be choosing this A company data set. I more interesting to find out why Singapore is called the well-known business hub in South East Asia.
- Since the economic is so bad now that I wanted to know whether in Singapore is the company really Sustainable

Steps by steps

Planning

Solve problem statement

List problem

Set Question to analysis

import all project csv file

combine all csv file into a dataframe

Create a important columns

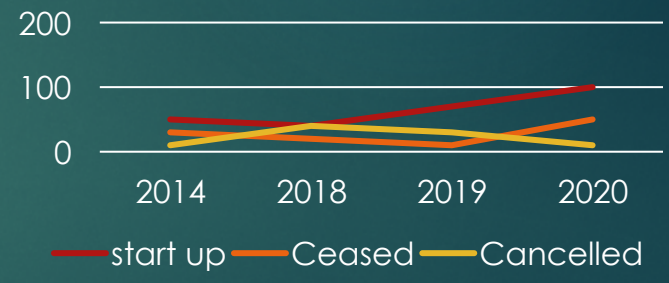
Clean the Dataset

Check the Dataset if its clean

insight

Analize data (Graph)

Illustration



Questions

what is the full address of the company?

Who is the oldest company in Singapore which is on 'live' status?

Is there any company which status is 'Ceased Registration' in Singapore ?

Is there any company which unable to sustain in Singapore? > 50years

Is there seasonality for Singapore company which unable to sustain over the years and month?

show the number Company's 'live' status over the years??

show the number Company status 'cancelled' over the years

Over the years, does Singapore company have monthly registration Seasonality?

the area for business location

the popular area for business location

Over the years, does singapore company able to survive?

Samples dataset

	business_constitution_description	primary_ssic_description	primary_user_described_activity	street_name	entity_status_description	annual_return_date	pos
0	Sole Proprietor	REPAIR OF DOMESTIC ELECTRICAL/ELECTRONIC APPLI...	na	HOUGANG AVENUE 8	Ceased Registration	na	
1	Sole Proprietor	CHARTERED BUS SERVICES (INCLUDING SCHOOL BUSES)	na	YISHUN STREET 41	Cancelled (Non-Renewal)	na	
2	Sole Proprietor	MANAGEMENT CONSULTANCY SERVICES FOR HEALTHCARE...	PROVIDE BUSINESS MANAGEMENT AND CONSULTANCY SE...	BUKIT BATOK STREET 31	Cancelled	na	
3	Sole Proprietor	LETTING AND OPERATING OF SELF-OWNED OR LEASED ...	SALES OF DRINKS & FOODS	TANJONG PAGAR PLAZA	Cancelled	na	
4	Sole Proprietor	MANAGEMENT CONSULTANCY SERVICES N.E.C.	na	TREVOSE CRESCENT	Cancelled	na	

_preference	entity_name	paid_up_capital2_others	...	paid_up_capital10_ordinary	paid_up_capital10_others	uen_of_audit_firm4	paid_up_capital7_others
na	A & E ENGINEERING SERVICES	na	...	na	na	na	na
na	A & Q TRANSIT	na	...	na	na	na	na
na	A AND Y CONSULTANCY SERVICES	na	...	na	na	na	na
na	A COFFEE SHOP	na	...	na	na	na	na
na	A DIXIT & ASSOCIATES	na	...	na	na	na	na

```
1  df1.info()
```

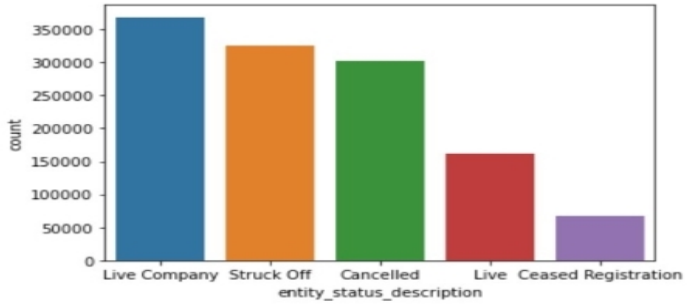
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 140209 entries, 0 to 140208
Data columns (total 95 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   business_constitution_description  70393 non-null   object
 1   primary_ssic_description        140209 non-null  object
 2   primary_user_described_activity  50108 non-null   object
 3   street_name                     139902 non-null  object
 4   entity_status_description       115391 non-null  object
 5   annual_return_date              48983 non-null   object
 6   postal_code                     139943 non-null  object
 7   paid_up_capital10_preference    0 non-null       float64
 8   entity_name                     140209 non-null  object
 9   paid_up_capital2_others         435 non-null     object
 10  name_of_audit_firm3             23 non-null      object
 11  address_type                    140209 non-null  object
 12  paid_up_capital9_currency       0 non-null       float64
 13  paid_up_capital8_ordinary       0 non-null       float64
 14  paid_up_capital6_others         1 non-null       object
 15  paid_up_capital6_ordinary       1 non-null       object
 16  level_no                        115294 non-null  object
 17  paid_up_capital8_others         0 non-null       float64
```

Using a dataset of a single csv file to examine, making sure that all my dataset columns are match before proceeding

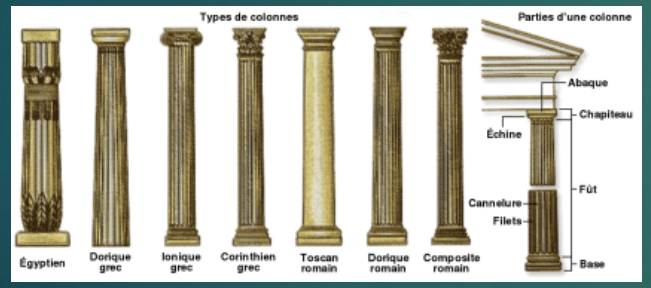
Insight

check the insights

```
In [22]:
1 sns.countplot(x="entity_status_description",data = create_df , order = pd.value_counts(create_df['entity_status_description'
2 plt.show()
```



This insight can be seen the highest data among all.



Using the importants and high data to select my columns and check through the data which I needed the most

1	primary_ssic_description	140209	non-null	object
3	street_name	139902	non-null	
4	entity_status_description	115391	non-null	
6	postal_code	139943	non-null	object
7	entity_name	140209	non-null	object
10	address_type	140209	non-null	object
13	level_no	115294	non-null	object
20	uen	140209	non-null	object
21	uen_issue_date	140209	non-null	object
22	no_of_officers	140209	non-null	int64
25	no_of_charges	140209	non-null	int64
32	block	139698	non-null	object
47	registration_incorporation_date	140209	non-null	object
52	primary_ssic_code	140209	non-null	int64
59	unit_no	115037	non-null	object
69	entity_type_description	140209	non-null	object

Challenge and difficulities

B) import all project csv file(🖌🖌🖌)

```
In [2]:
#i created a variable and loop all the csv files
list_of_csv_files = r"C:\Users\garyh\Downloads\project_multiple_excel_files\*.csv"
for fname in glob.glob(list_of_csv_files):
    df=pd.read_csv(fname)
    print(fname)
```

```
C:\Users\garyh\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3146: DtypeWarning: Columns (6) have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-a.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-b.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-c.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-d.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-e.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-f.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-g.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-h.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-i.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-j.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-k.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-l.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-m.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-n.csv
C:\Users\garyh\Downloads\project_multiple_excel_files\acra-information-on-corporate-entities-o.csv
```

Import all csv files into python

C) combine all csv file into a dataframe (🖌🖌🖌)

```
In [3]:
#after looping, i concat it

all_files = glob.glob(os.path.join(list_of_csv_files, r"C:\Users\garyh\Downloads\project_multiple_excel_files\*.csv"))
concatenated_df = pd.concat((pd.read_csv(f) for f in all_files), ignore_index=True)
```

```
<ipython-input-3-8c850802759b>:4: DtypeWarning: Columns (6) have mixed types.Specify dtype option on import or set low_memory=False.
  concatenated_df = pd.concat((pd.read_csv(f) for f in all_files), ignore_index=True)
```

```
In [4]:
#check the information of the dataframe
concatenated_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1713722 entries, 0 to 1713721
Data columns (total 94 columns):
 #   Column                          Dtype
```

Combine all the files

F8)Slice first 2 digit from existing column 'postal code', create it to a new column (🖌🖌)

```
In [20]:
#convert the column to pd Series in order to slice it
number = new_df.postal_code.to_dict()
s = pd.Series(number)
new_df['new_postal_code'] = s.str.slice(stop=2)
new_df.head()
```

Out[20]:

ode	no_of_officers	entity_type_description	new_uen_issue_date	new_registration_date	new_year	new_month	new_year_month	no_of_years	new_postal_code
446	1	Business	2008-09-12	2004-01-07 14:33:42	2004	1	2004-01	17.0	53
456	1	Business	2017-03-15	2017-03-15 12:31:24	2017	3	2017-03	4.0	76

Slice out the first 2 columns

```
In [51]:
###find district from URA website and create a csv file to merge it in

mapping_df = pd.read_csv(r"C:\Users\garyh\Downloads\project_postal_code\postal_code_general_location.csv")
mapping_df.head()
```

Out[51]:

	new_postal_code	general_location
0	1	Raffles Place, Cecil, Marina, People's Park
1	2	Raffles Place, Cecil, Marina, People's Park
2	3	Raffles Place, Cecil, Marina, People's Park
3	4	Raffles Place, Cecil, Marina, People's Park
4	5	Raffles Place, Cecil, Marina, People's Park

f9 a) format the column 'new_postal_code' from 1 digit to a 2 digit number (🖌)

```
In [52]:
mapping_df = pd.read_csv(r"C:\Users\garyh\Downloads\project_postal_code\postal_code_general_location.csv", converters = {'ne
mapping_df.head()
```

Out[52]:

	new_postal_code	general_location
0	01	Raffles Place, Cecil, Marina, People's Park
1	02	Raffles Place, Cecil, Marina, People's Park
2	03	Raffles Place, Cecil, Marina, People's Park

Mapping another created csv files and merge to find the region location

new_postal_code	general_location
53	Serangoon Garden, Hougang, Punggol
76	Yishun, Sembawang
65	Hillview, Dairy Farm, Bukit Panjang, Choa Chu ...
08	Anson, Tanjong Pagar
29	Watten Estate, Novena, Thomson

after combination of first 2 digit postal code to find location, results will be shown in the next column in dataframe

```
In [42]:
1 new_df['full_address'] = new_df['block'] + " " + new_df['street_name'] + " " + new_df['level_no'] + " " + new_df['unit_no']
2 new_df.head()
```

Out[42]:

_description	new_uen_issue_date	new_registration_date	new_year	new_month	new_year_month	no_of_years	new_postal_code	general_location	full_address
Business	2008-09-12	2004-01-07 14:33:42	2004	1	2004-01	17.0	53	Serangoon Garden, Hougang, Punggol	446 HOUGANG AVENUE 8 B1 1635 530446
Business	2017-03-15	2017-03-15 12:31:24	2017	3	2017-03	4.0	76	Yishun, Sembawang	456 YISHUN STREET 41 06 75 DEW SPRING @ YISHUN...
Business	2008-09-13	2005-04-29 01:38:25	2005	4	2005-04	16.0	65	Hillview, Dairy Farm, Bukit Panjang, Choa Chu ...	52 BUKIT BATOK STREET 31 09 08 MADEIRA, THE 65...
Business	2008-09-13	2006-04-10 10:36:33	2006	4	2006-04	15.0	08	Anson, Tanjong Pagar	1 TANJONG PAGAR PLAZA 01 53 082001

Add the details to a new column combination for full address

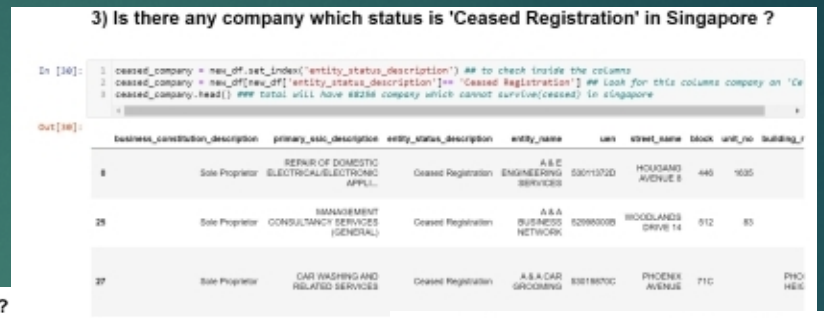
```
1 for col in new_df.columns:
2     pct_missing = np.mean(new_df[col].isnull())
3     print('{} - {}%'.format(col, round(pct_missing*100)))
```

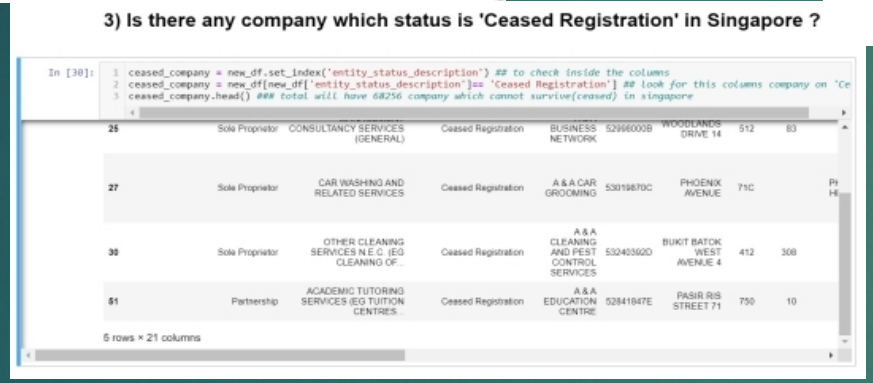
```
business_constitution_description - 0%
primary_ssic_description - 0%
entity_status_description - 0%
uen - 0%
entity_name - 0%
street_name - 0%
block - 0%
unit_no - 0%
level_no - 0%
building_name - 0%
postal_code - 0%
no_of_officers - 0%
entity_type_description - 0%
new_uen_issue_date - 0%
new_registration_date - 0%
new_year - 0%
new_month - 0%
new_year_month - 0%
no_of_years - 0%
new_postal_code - 0%
general_location - 0%
```

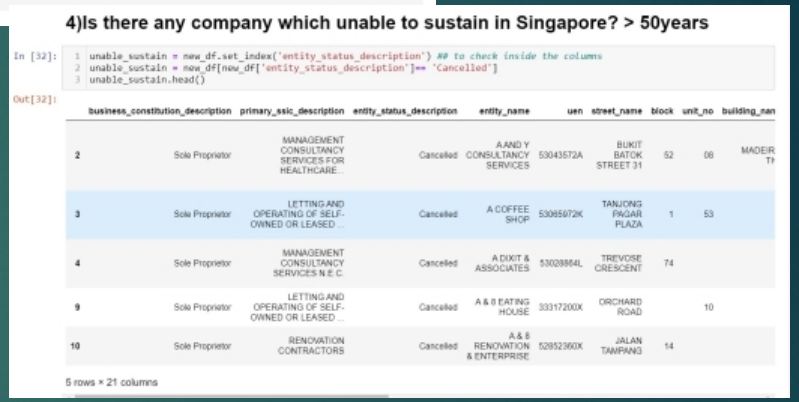
Check the data set to ensure that its cleaned

End of data cleaning

Use the insights







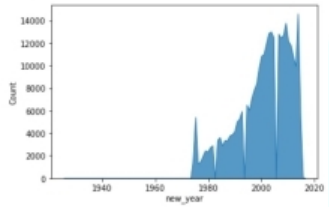
I have created the insights variable in order not to mess up with my data set

Analyse

9) Is there seasonality for Singapore company which unable to sustain over the years and month?

9a) Over the years

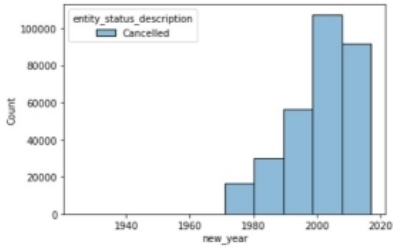
```
In [46]: 1 sns.histplot(data=unable_sustain, x="new_year", element='poly', bins=100)
         2 plt.show()
```



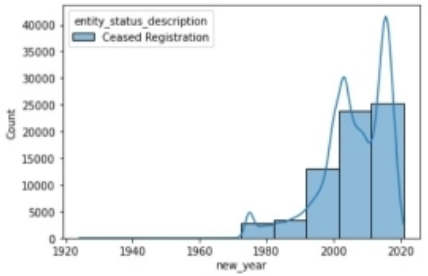
9b) monthly over the years

6) show the number Company status 'cancelled' over the years

```
In [36]: 1 sns.histplot(data=unable_sustain,x="new_year", hue="entity_status_description", bins = 10, )
         2 plt.show()
```

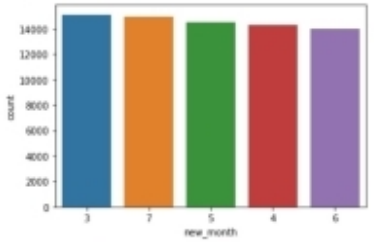



```
In [38]: 1
         2 sns.histplot(data=ceased_company,x="new_year", hue="entity_status_description", kde=True, bins=10)
         3 plt.show()
```



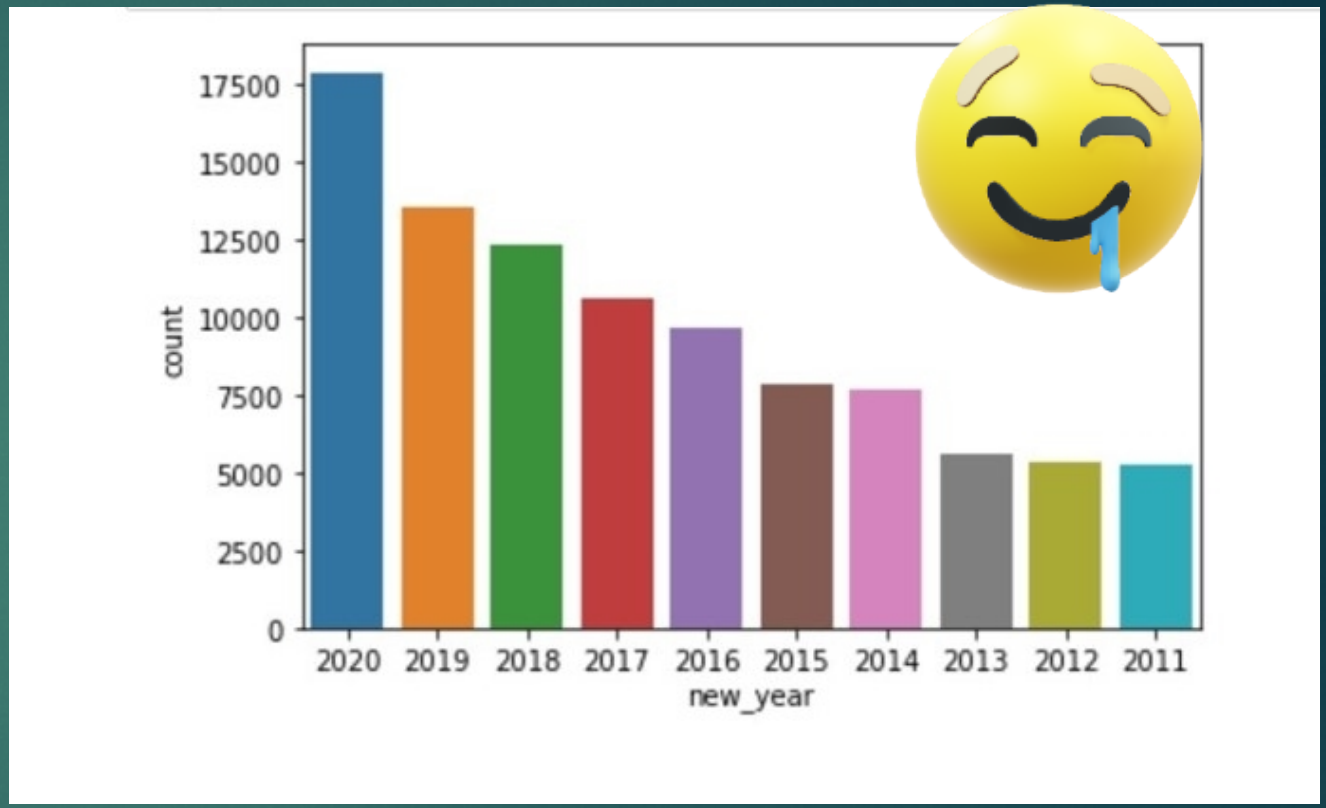
8)Over the years, does Singapore company have monthly registration Seasonality? ¶

```
n [40]: 1 # multiple="stack"
        2 sns.countplot(x="new_month",data = live_company,order = pd.value_counts(live_company['new_month']).iloc[:5].index)
        3 plt.show()
```



With the help of those dataframe and created variables

Problem Statement



THE END

GitHub

https://github.com/onggreat/G-Dreams