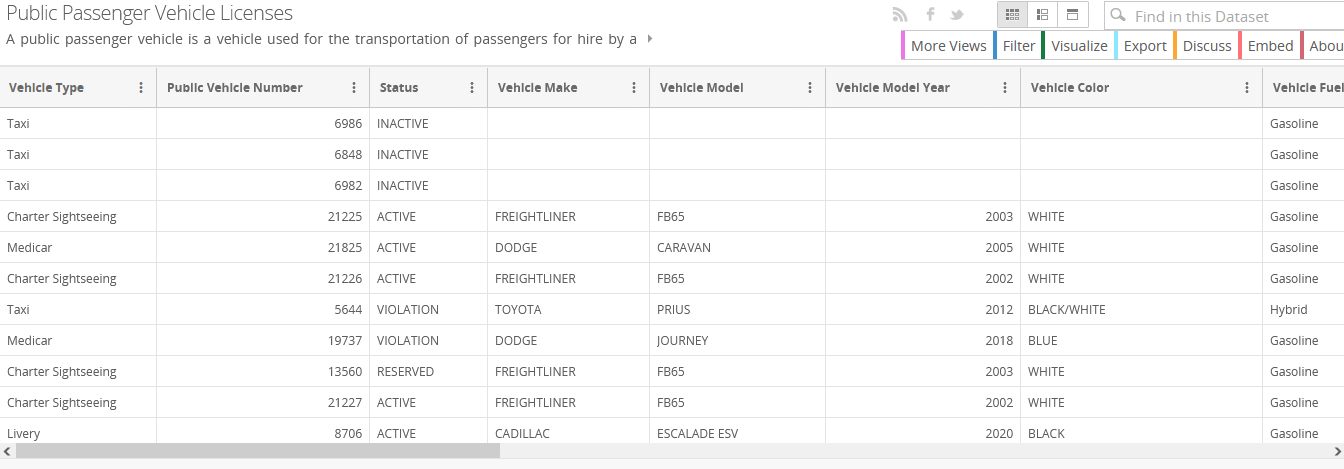
**NAME**

**REG NUMBER**

**TITLE: DATASET ANALYSIS**

**Introduction**

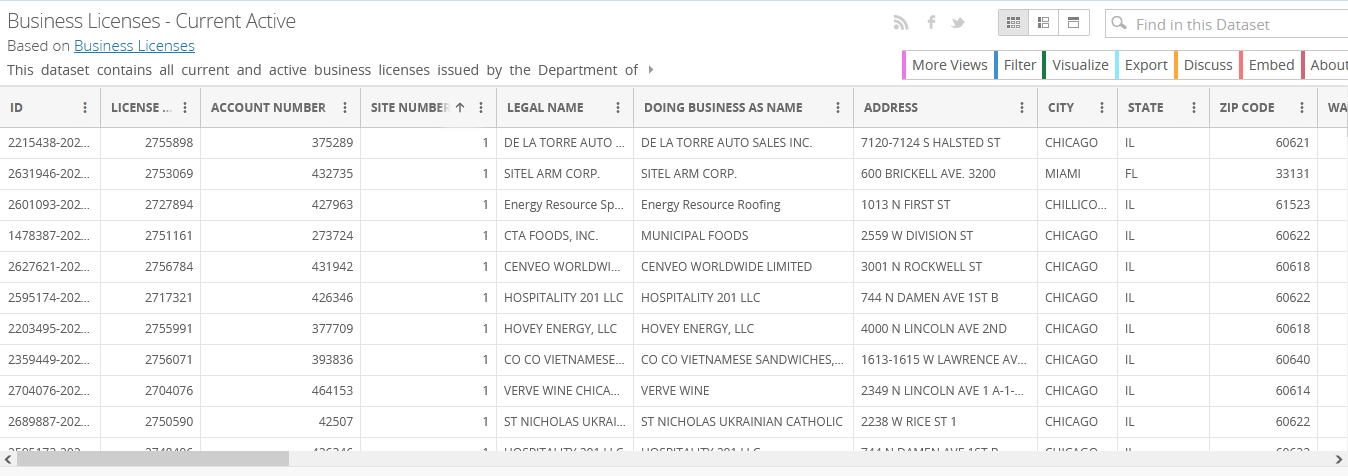
For this particular assignment, I picked the taxi trips dataset located at <https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew>, upon viewing the data, the following output was generated;



**Question 3**

**Data and variable description:**

There are a total of 17 variable columns for this particular dataset. A quick observation through this dataset revealed that there are items among this dataset that have empty column values. Another dataset that I picked was a publicly available data that is located at <https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses-Current-Active/uupf-x98q> . Upon previewing the dataset, the following is shown;



The dataset describes a series of variables as depicted by the business demographic details that include id, licence, account number, site number, legal name etc.

**Question 3**

**Analytical questions;**

Based on the two datasets proved above, we can form the following analytical questions from these sets:

1. What is the total number of taxi cabs that are currently in the database?
2. On a daily average, how many taxis ply the city?
3. Do all the taxi cars pay taxes? Which ones do not?
4. What is the relationship between the vehicle types and the city they come from?
5. What is the distribution of vehicle type per city?
6. What is the total number of business registered?
7. What is the distribution of these businesses per state?
8. What is the summary of the businesses categories in the cities?

**Question 4**

**In** this step I downloaded both the proposed dataset on the taxi cars, and the second dataset on business across Chicago. The fowling data wrangling techniques was perfumed on the datasets:

* Removal and replacement of spaces with null values
* Filtering columns to check for wrong attributes, for example, checking where there are strong literals in a column meant for integer values

**Data analysis**

For the passenger dataset, the independent variable, the independent variable here in the vehicle type, upon which other dependent variables scope. In the active business dataset, the in order answer the analytical questions arising from this study analyses was performed on the tow datasets using:

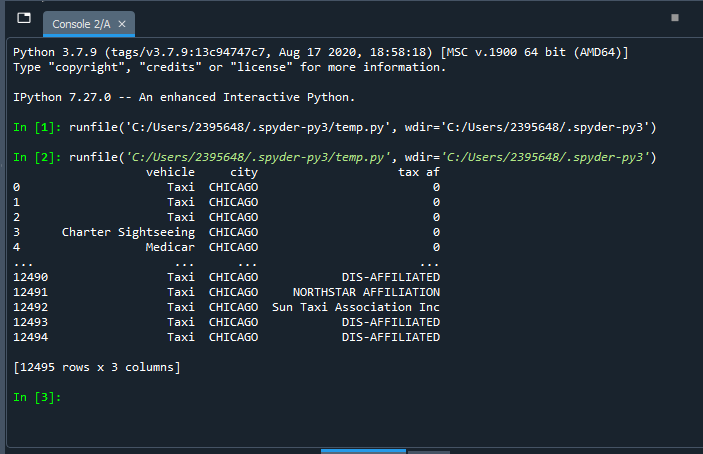
1. Spyder IDE
2. Python programming language and package such as Pandas and Matploob lib

Printing the headers of the vehicles passenger distribution file for the variables we want to use to answer the above questions:

import pandas as pd

df = pd.read\_csv (r'vpd\_one.csv')

print(df)



Answering the above question using the measures of central tendencies such as mean, median and mode, then getting the group by per city and tax affiliation from the above research questions

However, the above code threw an error and could not give a very accurate result of the measures of central tendencies since the datasets on passenger car distribution do not contained integer values but rather string charsets that python cannot use o calculate the values of central tendencies, a so a little modification was done by adding variable”speed” to the dataset and the result is smooth below. It answers our above questions:

import pandas as pd

df = pd.read\_csv (r'vpd\_one.csv')

print(df)

# block 1 - simple stats

mean1 = df['speed'].mean()

sum1 = df['speed'].sum()

max1 = df['speed'].max()

min1 = df['speed'].min()

count1 = df['speed'].count()

median1 = df['speed'].median()

std1 = df['speed'].std()

var1 = df['speed'].var()

# block 2 - group by

groupby\_sum1 = df.groupby(['city']).sum()

groupby\_count1 = df.groupby(['city']).count()

# Group by tax affiliation

groupby\_sum1 = df.groupby(['vehicle']).sum()

groupby\_count1 = df.groupby(['vehicle']).count()

# print block 1

print ('Mean speed: ' + str(mean1))

print ('Sum of speed ' + str(sum1))

print ('Max speed: ' + str(max1))

print ('Min speed: ' + str(min1))

print ('Count of speed: ' + str(count1))

print ('Median speed: ' + str(median1))

print ('Std of speed: ' + str(std1))

print ('Var of speed: ' + str(var1))

# print block 2

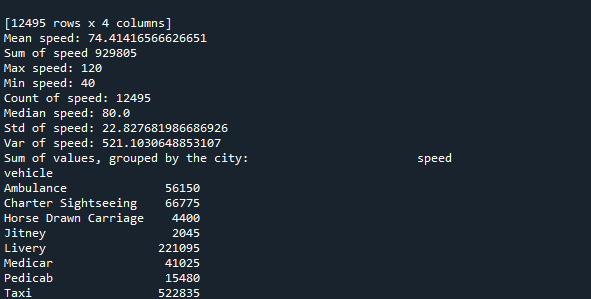
print ('Sum of values, grouped by the city: ' + str(groupby\_sum1))

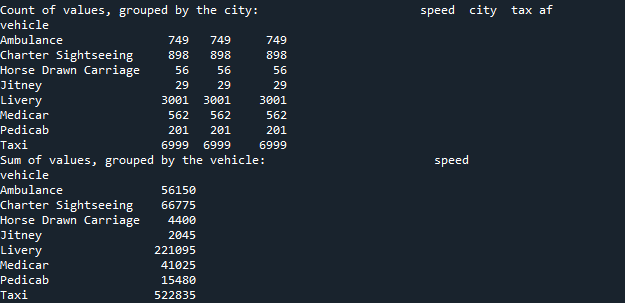
print ('Count of values, grouped by the city: ' + str(groupby\_count1))

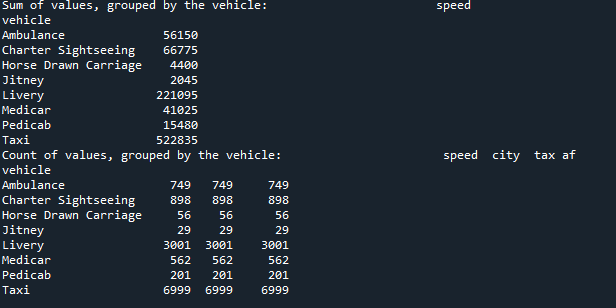
# print block 2

print ('Sum of values, grouped by the vehicle: ' + str(groupby\_sum1))

print ('Count of values, grouped by the vehicle: ' + str(groupby\_count1))





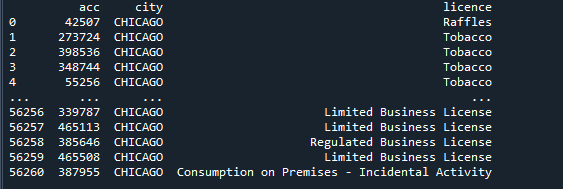


The same was also done for the second dataset on business in Chicago

import pandas as pd

df = pd.read\_csv (r'vpd\_one.csv')

print(df)



**Measures of central tendencies**

import pandas as pd

df = pd.read\_csv (r'bca2.csv')

print(df)

# block 1 - simple stats

mean1 = df['acc'].mean()

sum1 = df['acc'].sum()

max1 = df['acc'].max()

min1 = df['acc'].min()

count1 = df['acc'].count()

median1 = df['acc'].median()

std1 = df['acc'].std()

var1 = df['acc'].var()

# block 2 - group by

groupby\_sum1 = df.groupby(['city']).sum()

groupby\_count1 = df.groupby(['city']).count()

# Group by tax affiliation

groupby\_sum1 = df.groupby(['licence']).sum()

groupby\_count1 = df.groupby(['licence']).count()

# print block 1

print ('Mean acc: ' + str(mean1))

print ('Sum of acc ' + str(sum1))

print ('Max acc: ' + str(max1))

print ('Min acc: ' + str(min1))

print ('Count of acc: ' + str(count1))

print ('Median acc: ' + str(median1))

print ('Std of acc: ' + str(std1))

print ('Var of acc: ' + str(var1))

# print block 2

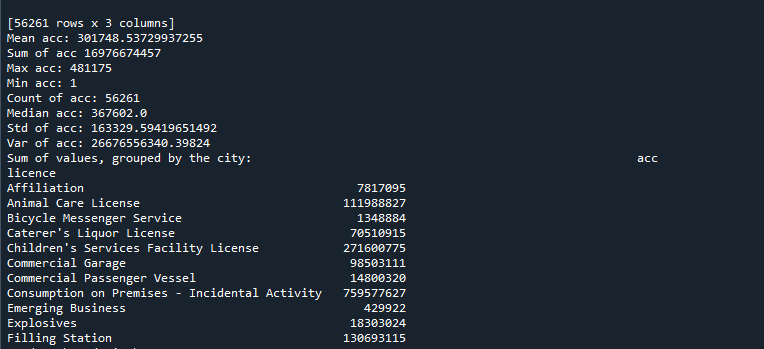
print ('Sum of values, grouped by the city: ' + str(groupby\_sum1))

print ('Count of values, grouped by the city: ' + str(groupby\_count1))

# print block 2

print ('Sum of values, grouped by the licence: ' + str(groupby\_sum1))

print ('Count of values, grouped by the licence: ' + str(groupby\_count1))

****