CMSC – 691
DATA ANALYSIS FOR STROKE MORTALITY RATES AND FAST FOOD RESTAURANTS IN VARIOUS U.S. STATES

ARSHITA JAIN
DEPARTMENT OF CSEE
UMBC
Email: a253@umbc.edu

TEJHAN BHARADWAJ RAMKUMAR
DEPARTMENT OF CSEE
UMBC
Email: kh50252@umbc.edu

INTRODUCTION

- Analyzing trends and patterns in stroke mortality rates among U.S. citizens by gender, ethnicity and state
- Studying for co-relation between stroke mortality rates and number of fast food restaurants
- Is Stroke Mortality Rate directly proportional to the number of fast food restaurants in a state? Meaning, the states having highest number of fast food joints, perhaps lead to a greater number of people consuming unhealthy food and thereby increasing the risk of heart related diseases.

MOTIVATION

- Stroke is the fifth leading cause of death for Americans, but the risk of having a stroke varies with race and ethnicity.
- Risk of having a first stroke is nearly twice as high for blacks as for whites, and blacks have the highest rate of death due to stroke.
- Though stroke death rates have declined for decades among all race/ethnicities, Hispanics have seen an increase in death rates since 2013.

ABSTRACT

- Our analysis showed that Stroke Mortality Rates in any state do depend upon the number of fast food restaurants in the area, but that’s certainly not an independent factor deciding it. The mortality rates also depend upon the Gender and ethnicity of the population.
- Also, there are various other factors affecting these rates, such as Alcohol consumption, or smoking habits, or lack of physical fitness, perhaps.

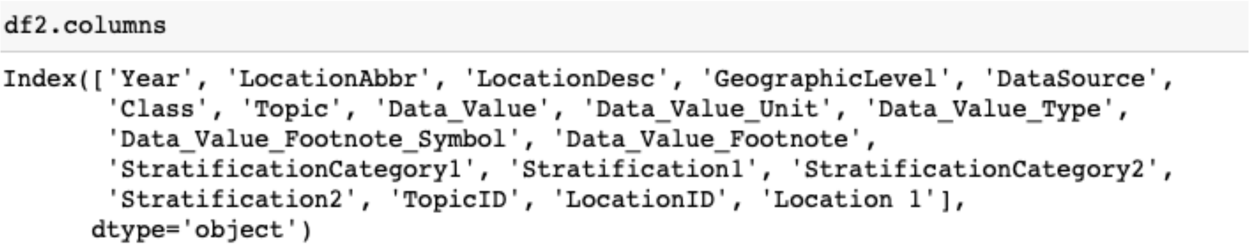
DATASETS USED

Worked with two datasets for our analysis.

- Stroke Mortality Data Among US Adults (35+) by State/Territory and County by CDC (Centers for Disease Control and Prevention)

This dataset had 18 columns that had information about various US states, their counties, then had Stroke Mortality Rate per 100,000 of the population, including classification based on Gender, Race/Ethnicity, etc. The attributes which did not seem to be important for the analysis were dropped at the beginning of the analysis.

The following picture depicts the original attributes:



```
df2.columns
```

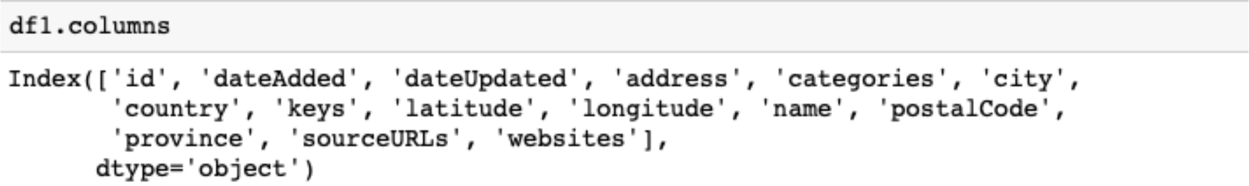
```
Index(['Year', 'LocationAbbr', 'LocationDesc', 'GeographicLevel', 'DataSource',
       'Class', 'Topic', 'Data_Value', 'Data_Value_Unit', 'Data_Value_Type',
       'Data_Value_Footnote_Symbol', 'Data_Value_Footnote',
       'StratificationCategory1', 'Stratification1', 'StratificationCategory2',
       'Stratification2', 'TopicID', 'LocationID', 'Location 1'],
      dtype='object')
```

Fig: Attributes

- Fast Food Restaurants across America: A list of 10,000 restaurants and their locations.

This dataset had various US states and their counties, and the names of various fast food restaurants present in those counties. It also had restaurant's address, locations etc. which were not needed for our analysis, and hence dropped.

The following picture depicts the original attributes:



```
df1.columns
```

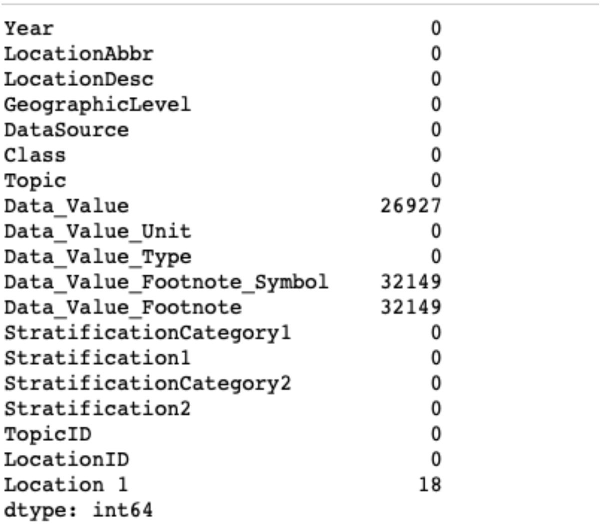
```
Index(['id', 'dateAdded', 'dateUpdated', 'address', 'categories', 'city',
       'country', 'keys', 'latitude', 'longitude', 'name', 'postalCode',
       'province', 'sourceURLs', 'websites'],
      dtype='object')
```

Fig: Attributes

DATA EXPLORATION

1. Stroke Mortality Rate Dataset

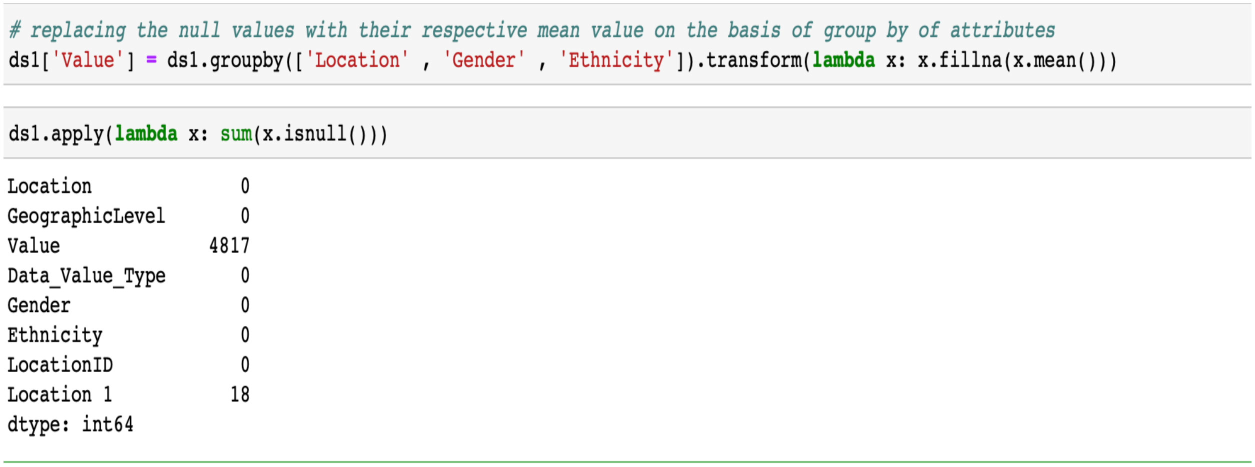
- First step was to study for the Null Values in the data and surprisingly we had more than half of the values as null. Thus, dealing with these was a major task.



```
Year                           0
LocationAbbr                   0
LocationDesc                   0
GeographicLevel                0
DataSource                     0
Class                          0
Topic                          0
Data_Value                 26927
Data_Value_Unit                0
Data_Value_Type                0
Data_Value_Footnote_Symbol 32149
Data_Value_Footnote        32149
StratificationCategory1        0
Stratification1                0
StratificationCategory2        0
Stratification2                0
TopicID                        0
LocationID                     0
Location 1                    18
dtype: int64
```

Fig: Check for Null Values

- Dealing with Null Values: Grouped the values on Location, Gender, Ethnicity; replaced the missing values with this grouped by Mean. This seemed to be the most sensible choice.



```
# replacing the null values with their respective mean value on the basis of group by of attributes
ds1['Value'] = ds1.groupby(['Location' , 'Gender' , 'Ethnicity']).transform(lambda x: x.fillna(x.mean()))
```

```
ds1.apply(lambda x: sum(x.isnull()))
```

```
Location            0
GeographicLevel     0
Value            4817
Data_Value_Type     0
Gender              0
Ethnicity           0
LocationID          0
Location 1         18
dtype: int64
```

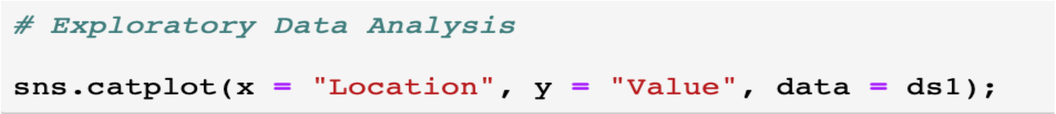
Fig: Dealing with missing values

- There were still 4817 Null values left. For there were states which had missing values for say, entire ethnicity or gender attributes and hence their mean also came out to be zero.

- Simply deleted these tuples.

- The further processing included removal of redundant attributes, renaming a few for convenience, checking for unique values and so.

- Exploratory Data Analysis: Plotted Catplot and Boxplot for visualizing potential outliers in the dataset. There was one that was way too deviated from the general range of values. Removed it using max() function.



```
# Exploratory Data Analysis

sns.catplot(x = "Location", y = "Value", data = ds1);
```

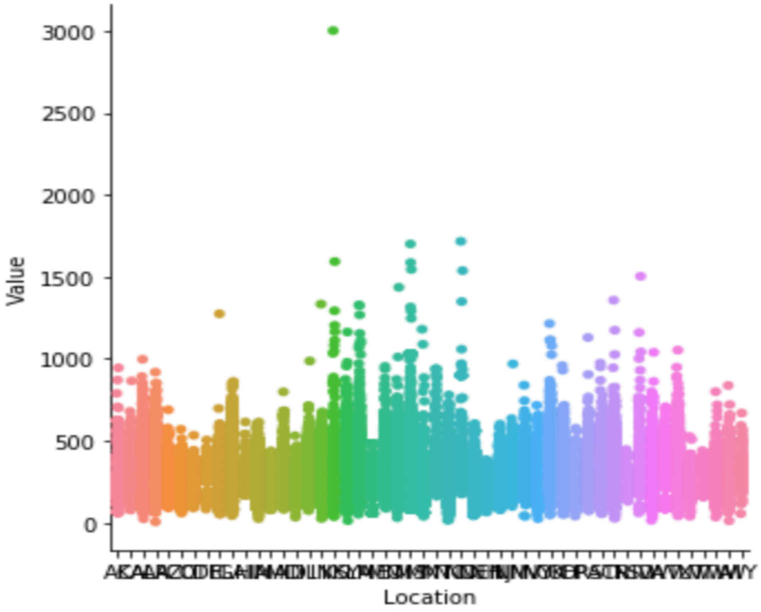
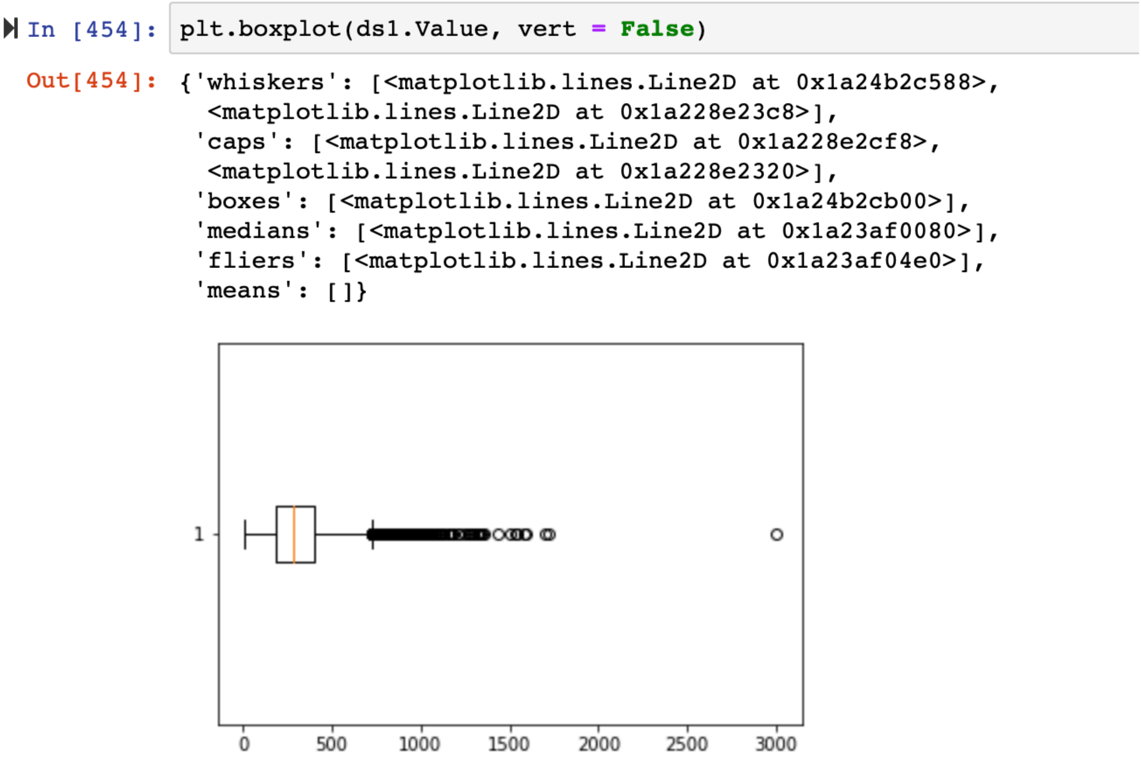


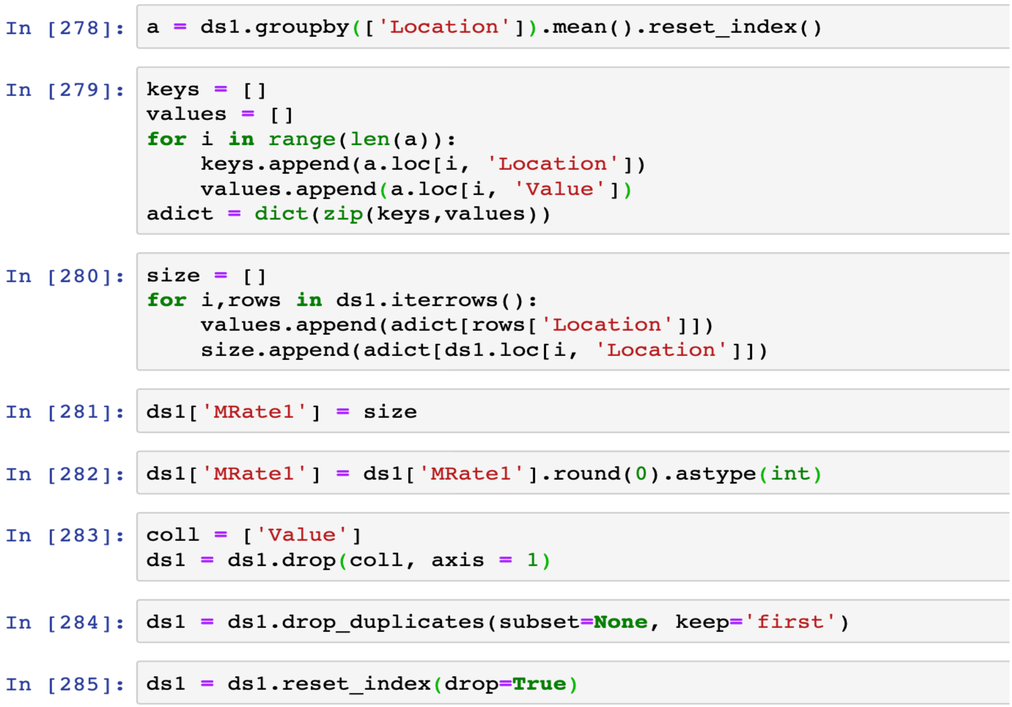
Fig: Catplot for Location VS Mortality Rates



```
In [454]: plt.boxplot(ds1.Value, vert = False)

Out[454]: {'whiskers': [<matplotlib.lines.Line2D at 0x1a24b2c588>,
            <matplotlib.lines.Line2D at 0x1a228e23c8>],
           'caps': [<matplotlib.lines.Line2D at 0x1a228e2cf8>,
            <matplotlib.lines.Line2D at 0x1a228e2320>],
           'boxes': [<matplotlib.lines.Line2D at 0x1a24b2cb00>],
           'medians': [<matplotlib.lines.Line2D at 0x1a23af0080>],
           'fliers': [<matplotlib.lines.Line2D at 0x1a23af04e0>],
           'means': []}
```

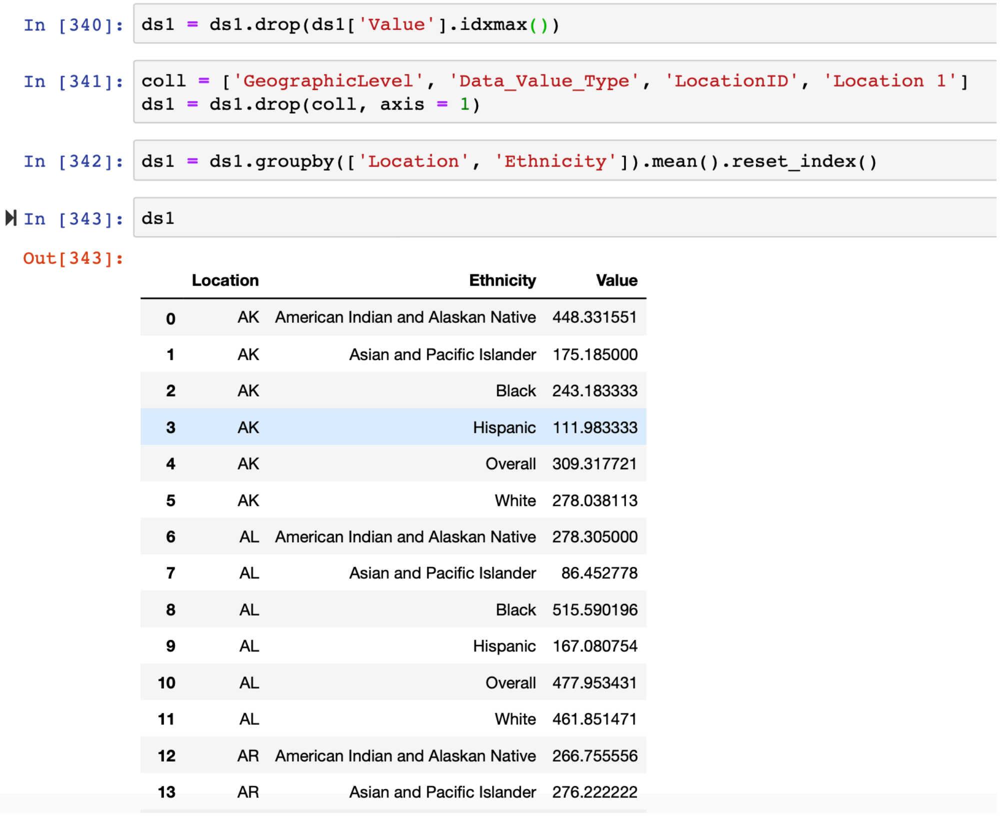
Fig: Boxplot for Location VS Mortality Rates

- This pre-processing of the data being performed, we first grouped by this dataset on Location, and then on Location and Ethnicity and Gender, for analyzing trends based on these factors.



```
In [278]: a = ds1.groupby(['Location']).mean().reset_index()

In [279]: keys = []
          values = []
          for i in range(len(a)):
              keys.append(a.loc[i, 'Location'])
              values.append(a.loc[i, 'Value'])
          adict = dict(zip(keys,values))

In [280]: size = []
          for i,rows in ds1.iterrows():
              values.append(adict[rows['Location']])
              size.append(adict[ds1.loc[i, 'Location']])

In [281]: ds1['MRate1'] = size

In [282]: ds1['MRate1'] = ds1['MRate1'].round(0).astype(int)

In [283]: coll = ['Value']
          ds1 = ds1.drop(coll, axis = 1)

In [284]: ds1 = ds1.drop_duplicates(subset=None, keep='first')

In [285]: ds1 = ds1.reset_index(drop=True)
```

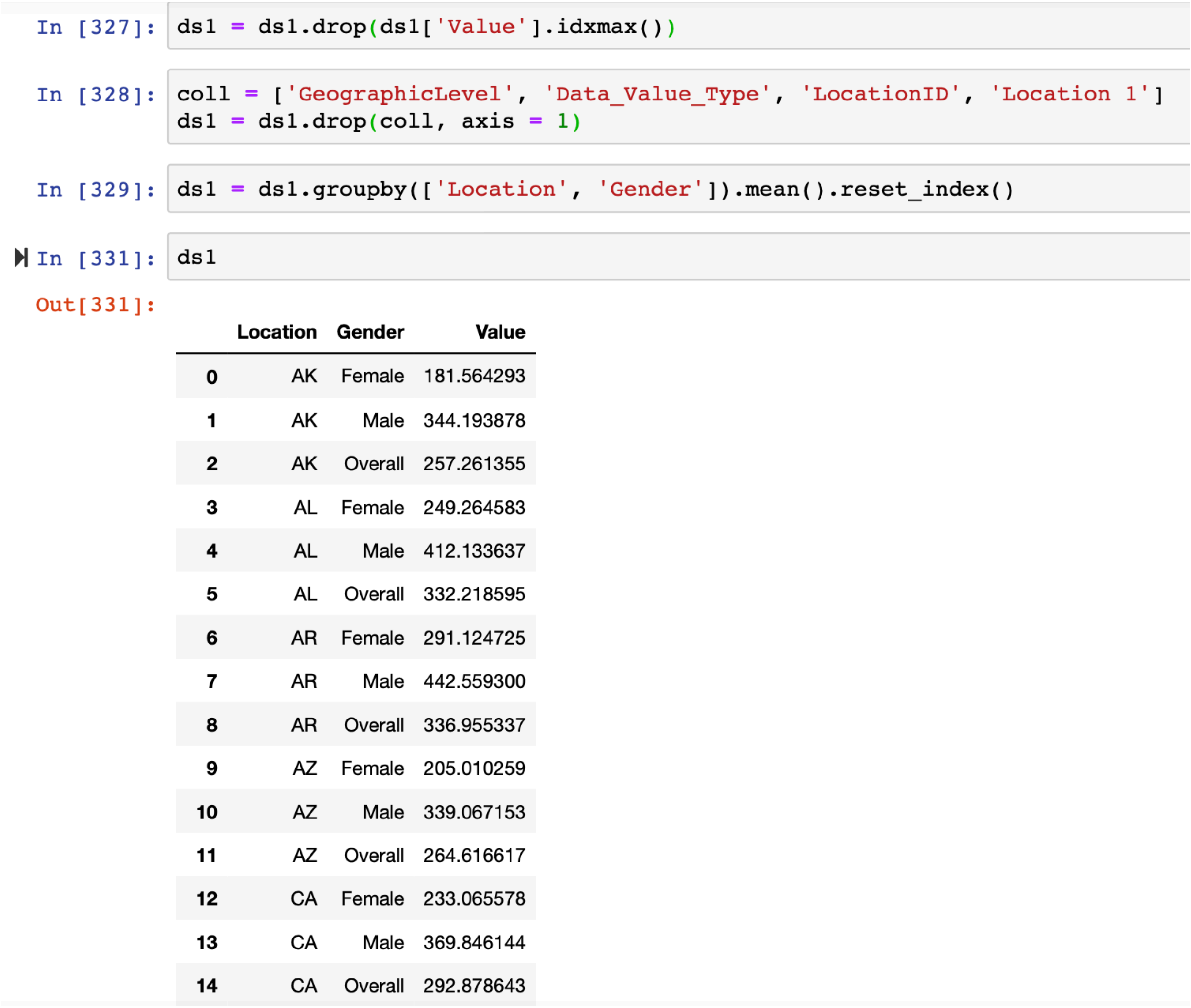
Fig: Groupby Location



```
In [340]: ds1 = ds1.drop(ds1['Value'].idxmax())

In [341]: coll = ['GeographicLevel', 'Data_Value_Type', 'LocationID', 'Location 1']
          ds1 = ds1.drop(coll, axis = 1)

In [342]: ds1 = ds1.groupby(['Location', 'Ethnicity']).mean().reset_index()

In [343]: ds1

Out[343]:
```

	Location	Ethnicity	Value
0	AK	American Indian and Alaskan Native	448.331551
1	AK	Asian and Pacific Islander	175.185000
2	AK	Black	243.183333
3	AK	Hispanic	111.983333
4	AK	Overall	309.317721
5	AK	White	278.038113
6	AL	American Indian and Alaskan Native	278.305000
7	AL	Asian and Pacific Islander	86.452778
8	AL	Black	515.590196
9	AL	Hispanic	167.080754
10	AL	Overall	477.953431
11	AL	White	461.851471
12	AR	American Indian and Alaskan Native	266.755556
13	AR	Asian and Pacific Islander	276.222222

Fig: Groupby Location and Ethnicity

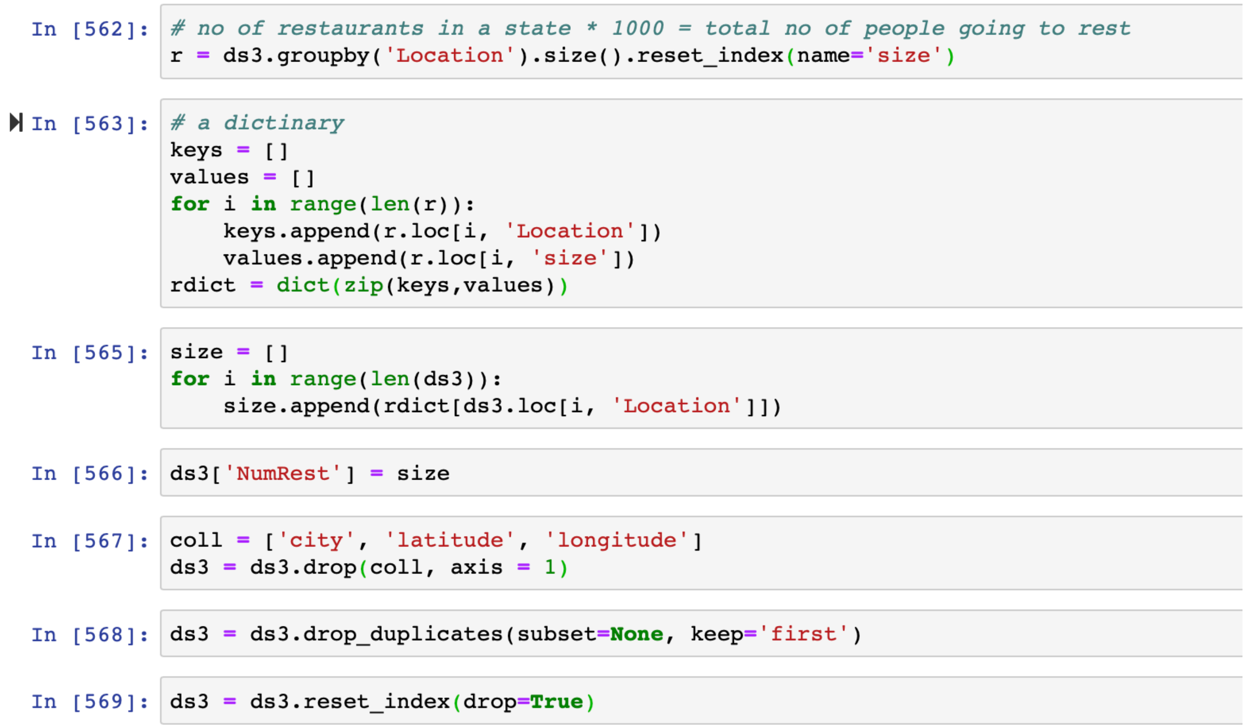


```
In [327]: ds1 = ds1.drop(ds1['Value'].idxmax())

In [328]: coll = ['GeographicLevel', 'Data_Value_Type', 'LocationID', 'Location 1']
          ds1 = ds1.drop(coll, axis = 1)

In [329]: ds1 = ds1.groupby(['Location', 'Gender']).mean().reset_index()

In [331]: ds1

Out[331]:
       Location  Gender       Value
    0        AK  Female  181.564293
    1        AK    Male  344.193878
    2        AK Overall  257.261355
    3        AL  Female  249.264583
    4        AL    Male  412.133637
    5        AL Overall  332.218595
    6        AR  Female  291.124725
    7        AR    Male  442.559300
    8        AR Overall  336.955337
    9        AZ  Female  205.010259
   10        AZ    Male  339.067153
   11        AZ Overall  264.616617
   12        CA  Female  233.065578
   13        CA    Male  369.846144
   14        CA Overall  292.878643
```

Fig: Groupby Location and Gender

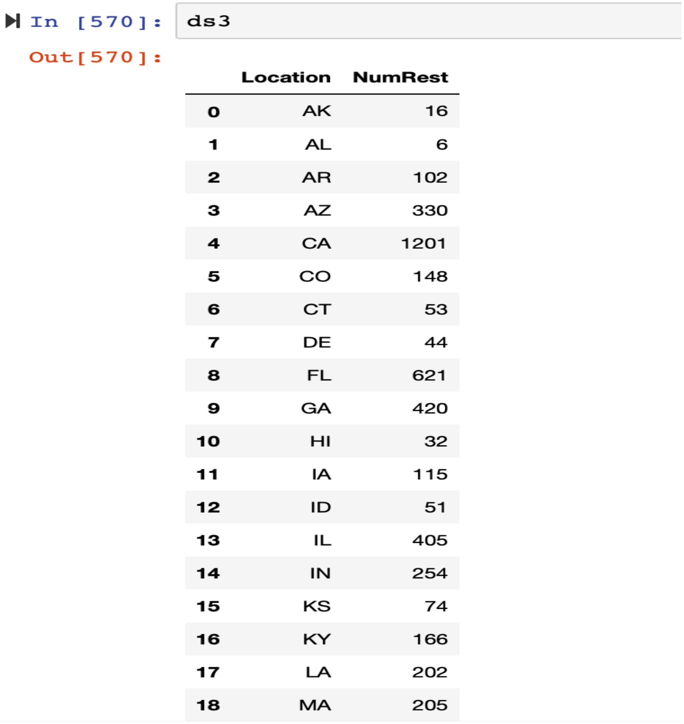
2. <u>Fast Food Restaurants Dataset</u>

- This dataset was comparatively easy to deal with. It had no Null values.
- Removed the redundant columns and performed renaming to a few for ease of reading.
- We were concerned with the total number of restaurants per state. Hence, grouped the data on ‘Location’ i.e., various states in the US. It returned total number of states per state.



```
In [562]: # no of restaurants in a state * 1000 = total no of people going to rest
          r = ds3.groupby('Location').size().reset_index(name='size')

In [563]: # a dictionary
          keys = []
          values = []
          for i in range(len(r)):
              keys.append(r.loc[i, 'Location'])
              values.append(r.loc[i, 'size'])
          rdict = dict(zip(keys,values))

In [565]: size = []
          for i in range(len(ds3)):
              size.append(rdict[ds3.loc[i, 'Location']])

In [566]: ds3['NumRest'] = size

In [567]: coll = ['city', 'latitude', 'longitude']
          ds3 = ds3.drop(coll, axis = 1)

In [568]: ds3 = ds3.drop_duplicates(subset=None, keep='first')

In [569]: ds3 = ds3.reset_index(drop=True)
```

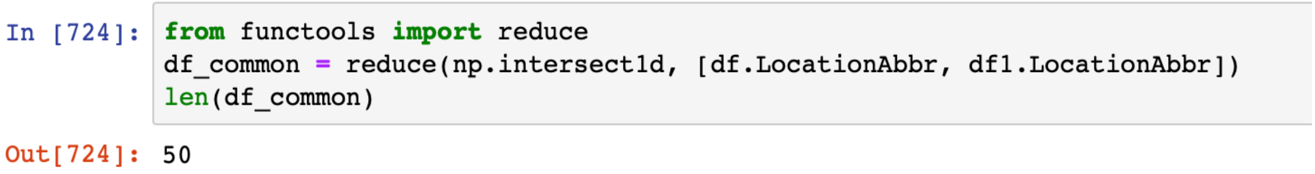
- A view of Final Dataset



```
In [570]: ds3
Out[570]:
```

	Location	NumRest
0	AK	16
1	AL	6
2	AR	102
3	AZ	330
4	CA	1201
5	CO	148
6	CT	53
7	DE	44
8	FL	621
9	GA	420
10	HI	32
11	IA	115
12	ID	51
13	IL	405
14	IN	254
15	KS	74
16	KY	166
17	LA	202
18	MA	205

- The two datasets had unequal number of states. One had states from both North and South America, while the other had only from North America. Hence, we discarded the extra states from one of the datasets, for a fair analysis.

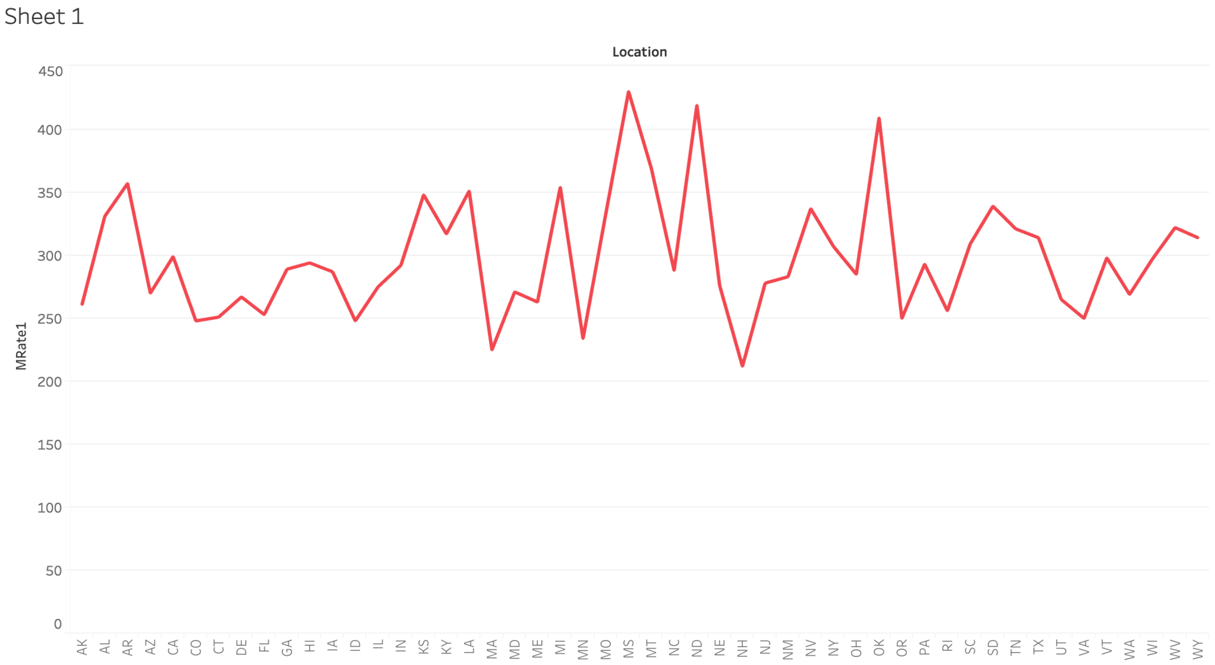


```
In [724]: from functools import reduce
          df_common = reduce(np.intersect1d, [df.LocationAbbr, df1.LocationAbbr])
          len(df_common)

Out[724]: 50
```

DATA VISUALIZATION

Used **Tableau Public** for almost all of the visualizations.

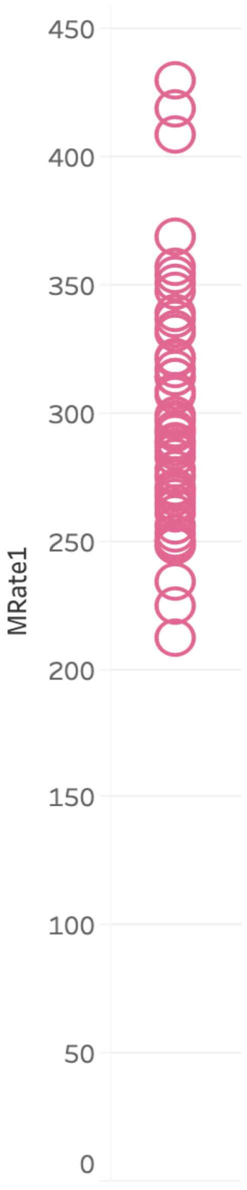
- Line Graph depicting relationship between mean Stroke Rates, Location wise.



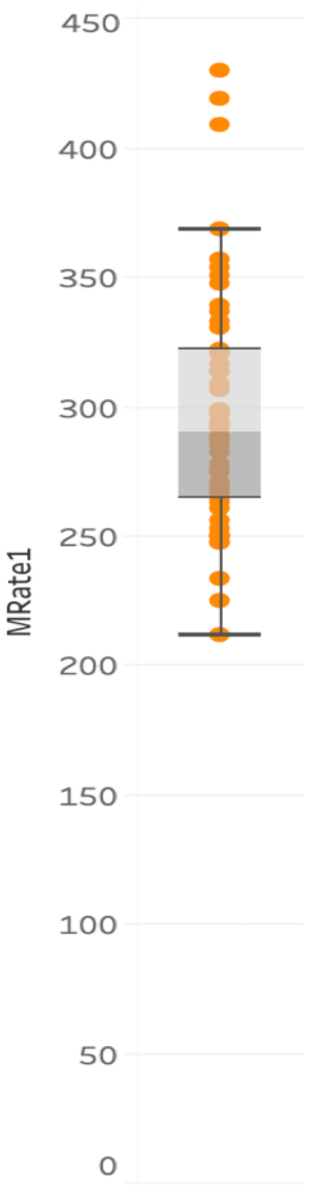
Insight gained: Mississippi (MS) has the highest mean rate and New Hampshire(NH) has the lowest. Maybe, most of the states have the mean values between 250-300, per 100000 population. Not sure. Made some more visualizations.

- Graphs depicting relationship between mean Stroke Rates, Location wise.

Sheet 2



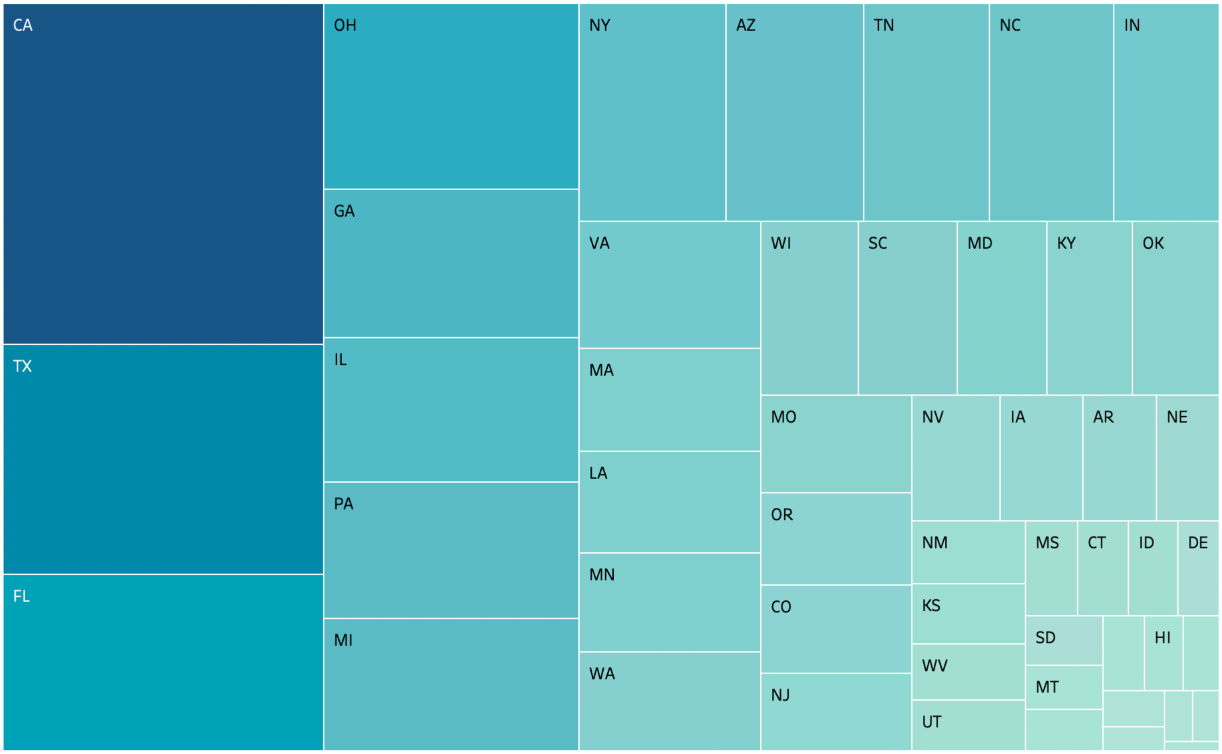
Sheet 3



Insight: Most of the states DO have mean mortality rates between 250-300 per 100,000 population. A few states have in a slightly higher range of mean of about 400 deaths per 100,000 population.

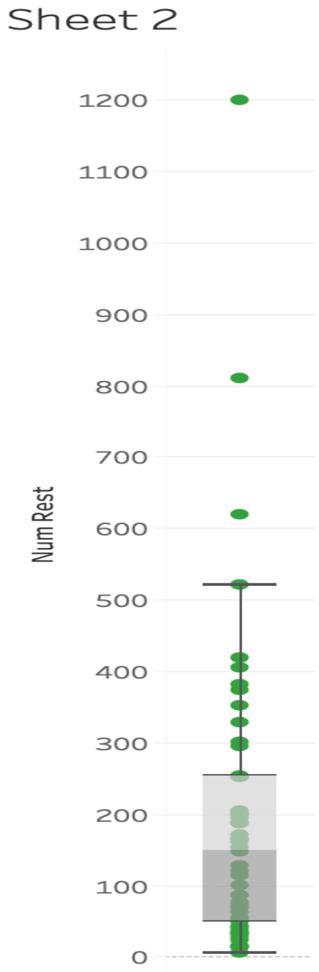
- Tree Map depicting total number of restaurants present in the United States, state-wise.

Sheet 1



Insight: Clearly, California has the highest number of fast-food restaurants with a total of 1201 restaurants in total followed by Texas and Florida.

- Whisker plot for range of total number of fast food joints



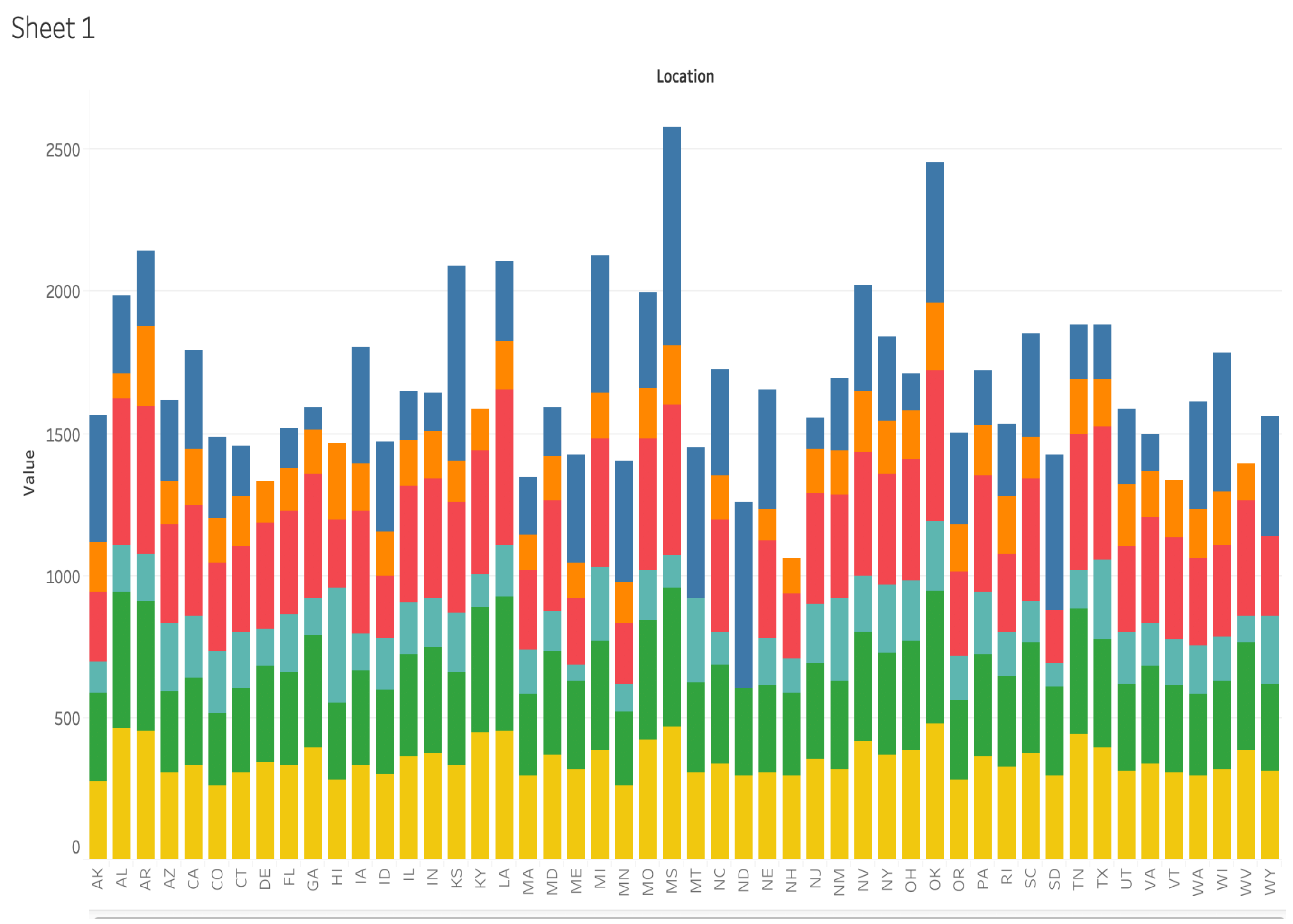
Insight: Mostly states have total fast food restaurants in the range of 0-120. Whereas, a few states have about 600, 700 and even an exceptionally high of about 1200 restaurants. As known from the previous graph, this is the state of California, this could be probably because California is a larger state by area.

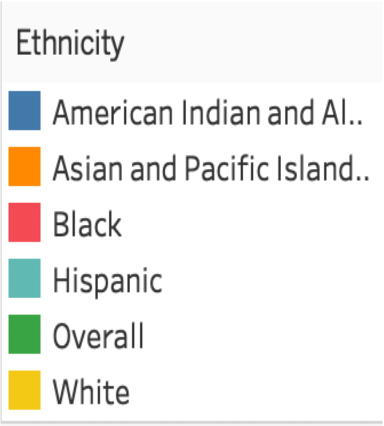
After making the visualizations for Stroke Mortality Rates and total number of restaurants in each state, we plotted the two datasets together. There were interesting results observed.

DESCRIPTION OF OUR ATTEMPT

Basically, what we are trying to do here is we have combined both the datasets to see if there is any direct correlation between stroke related deaths and number of fast-food restaurants in a state, since number of cardio-vascular diseases are known to be caused due to over consumption of fast-food. So here we are trying to find out, whether the state which has a greater number of fast-food restaurants have more deaths caused because of stroke. We are also trying to find is there any relation between Ethnicity and stroke deaths and between Gender and stroke deaths, where a particular ethnicity/gender is being affected more because of stroke compared to others.

- Bar Graph demonstrating Stroke Rates in various states, classified on the basis of Ethnicity, in the US



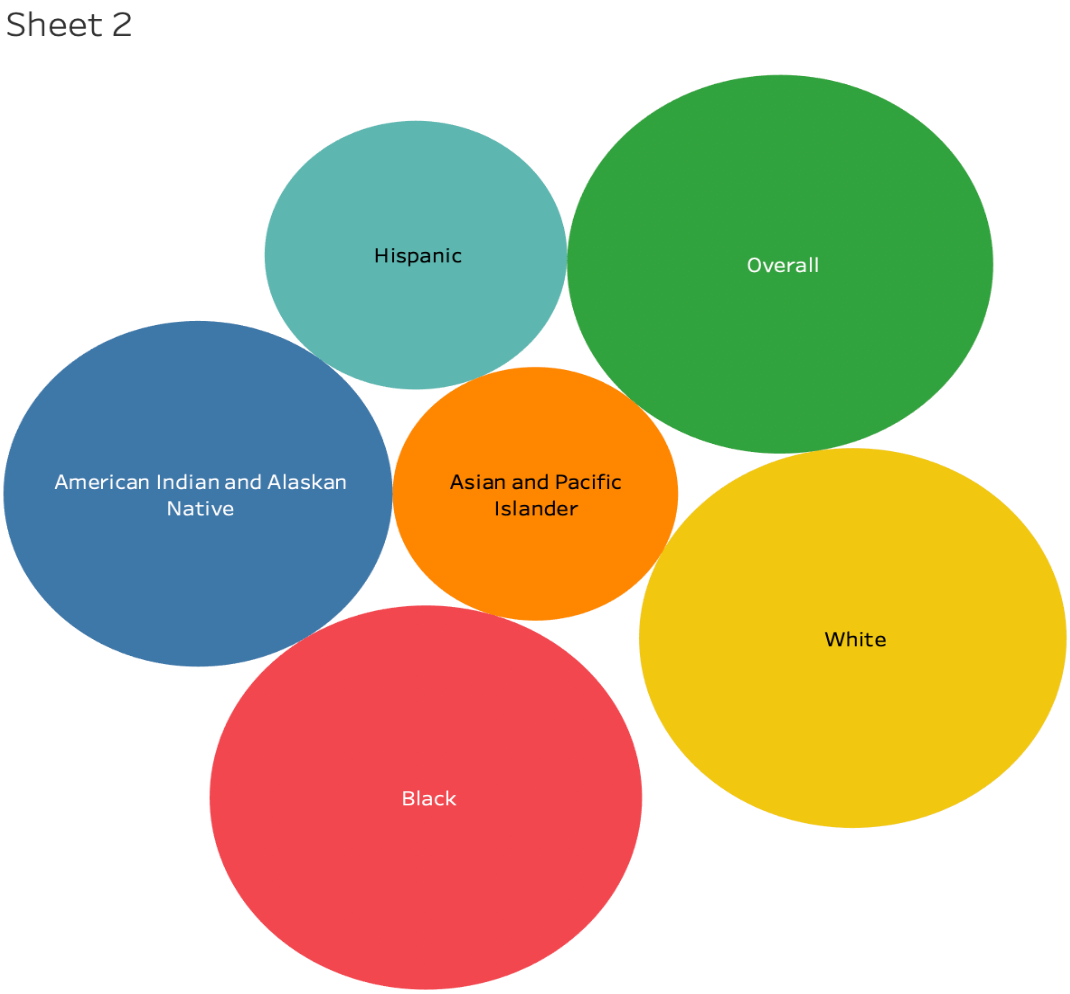


Ethnicity

- 🟦 American Indian and Al..
- 🟧 Asian and Pacific Island..
- 🟥 Black
- 🟦 Hispanic
- 🟩 Overall
- 🟨 White

Insight: American Indians suffer these deaths more than other ethnicities. More than half of the states depicts this. After American Indians there are a greater number of blacks who have died because of stroke, followed by other ethnicities. Mississippi which has the highest number of deaths compared to any other state has the greatest number of American Indians dead due to stroke. So, we can say that either there are a greater number of American Indians who live in Mississippi or American Indians in Mississippi consume more fast-food compared to other people.

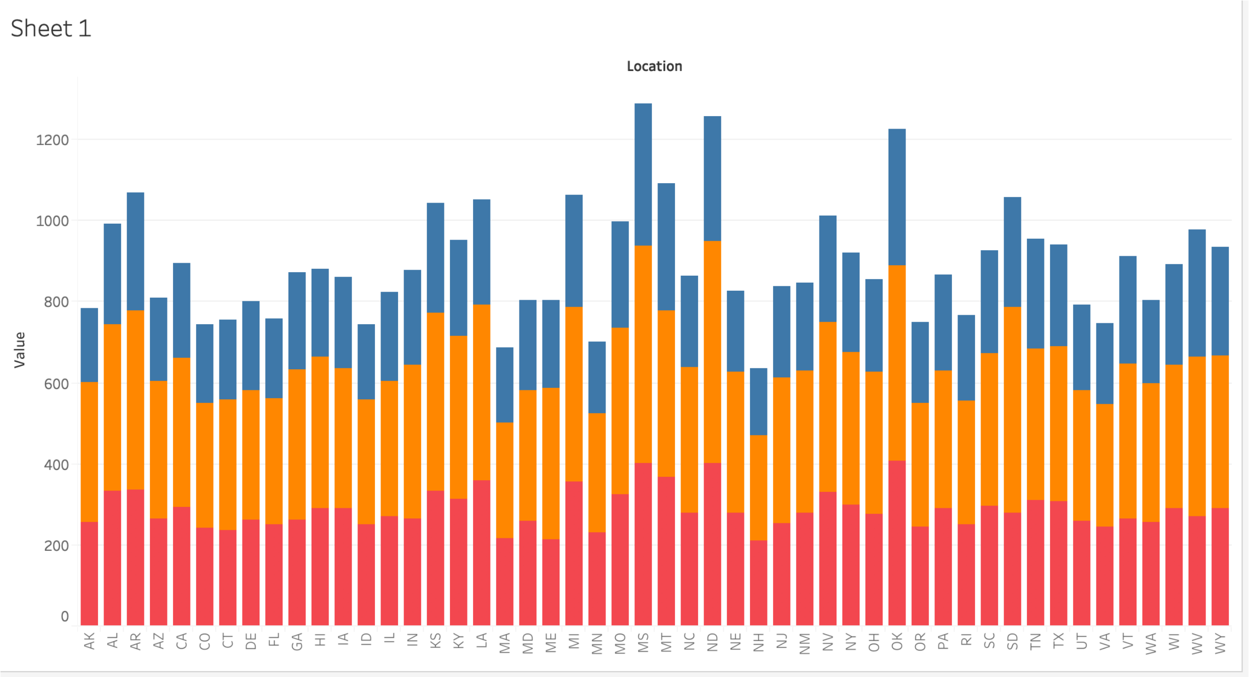
- Bubble Graph showing prevalence of total number of different Ethnicities, in the US

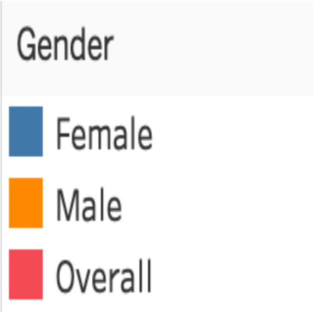


Insight: We know that the population of white’s and black’s are more in United States but compared to the population the death due to stroke is reasonable. Whereas if you consider American Indian and Alaska native’s the population is really less compared to other ethnicity, but the death rate is more.

Thus, Ethnicity could not really be accounted as a good factor contributing to stroke mortality rates.

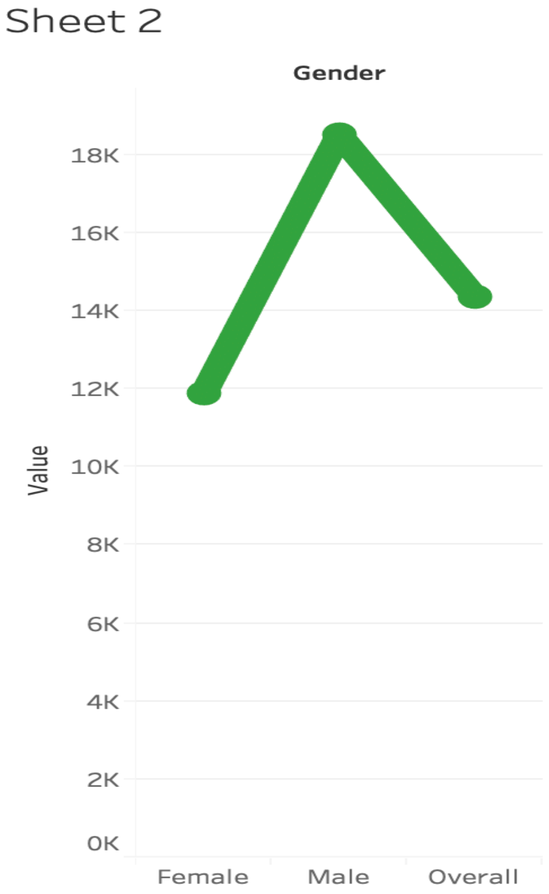
- Bar Graph demonstrating Stroke Rates in various states, classified on the basis of Gender, in the US





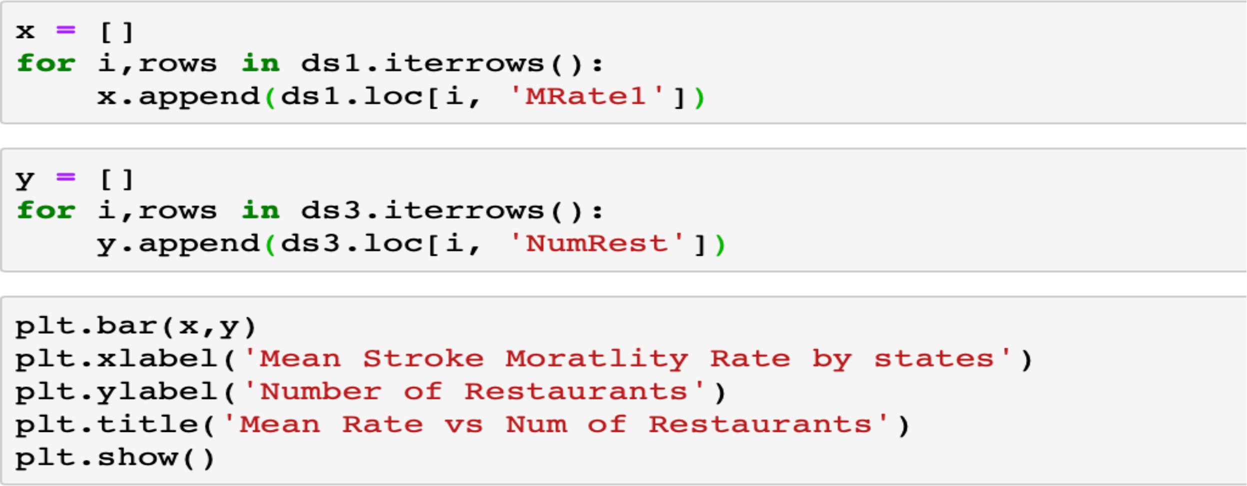
Insight: The trend shows that males are more affected than females. This shows that irrespective of the ethnicity men are getting more affected compared to woman. State of MS has a higher number of Male affected, than Female.

- Line Graph demonstrating Stroke Rates in various states, classified on the basis of Gender, in the US



Insight: Yes, males are affected more than females.

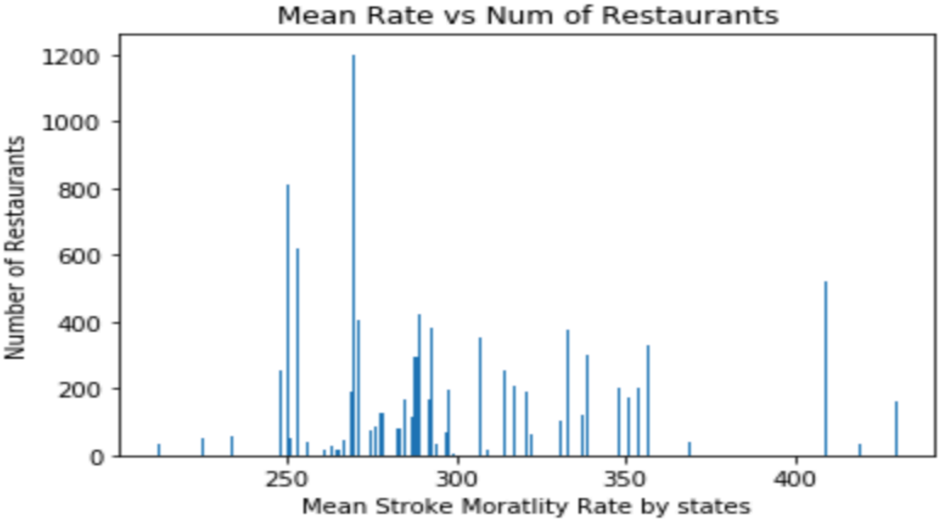
- Bar Graph between mean Stroke Rates grouped by various states in the US and total number of restaurants.



```
x = []
for i,rows in ds1.iterrows():
    x.append(ds1.loc[i, 'MRate1'])
```

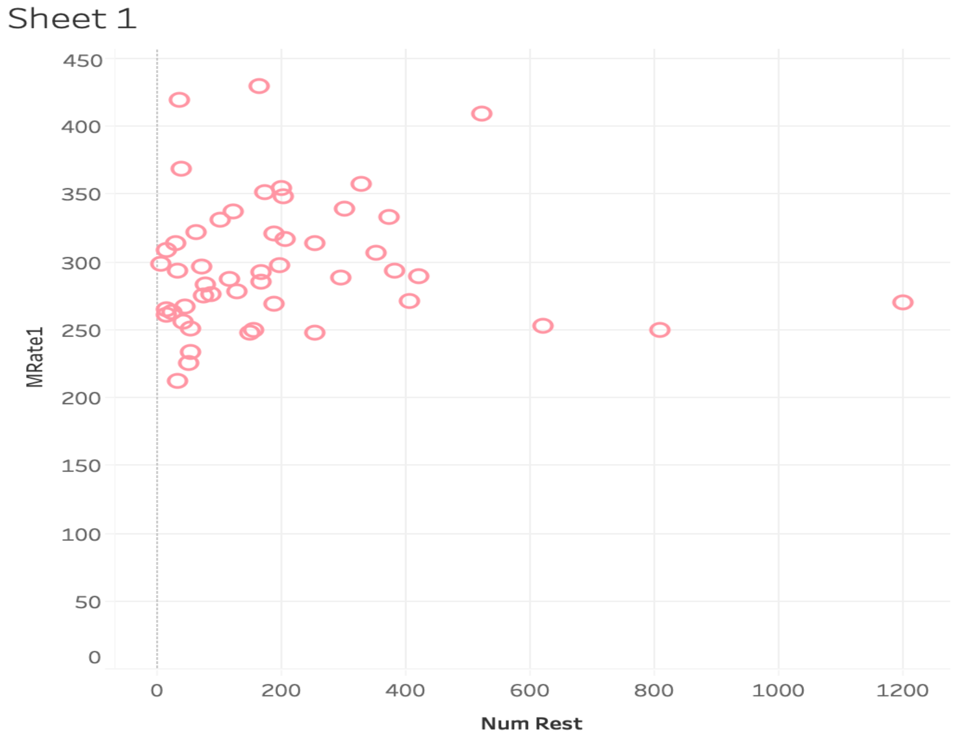
```
y = []
for i,rows in ds3.iterrows():
    y.append(ds3.loc[i, 'NumRest'])
```

```
plt.bar(x,y)
plt.xlabel('Mean Stroke Moratlity Rate by states')
plt.ylabel('Number of Restaurants')
plt.title('Mean Rate vs Num of Restaurants')
plt.show()
```



Insight: There is no absolute pattern observed. A few states have higher total number of fast food restaurants but low stroke mortality rates, whereas some has it other way round.

- Scatterplot between mean Stroke Rates grouped by various states in the US and total number of restaurants.



Insight: This graph shows that stroke mortality rates does not depend on the number of restaurants in a particular state. California which has 1201 restaurants only has a stroke rate close to 250 whereas Mississippi which has the highest stroke rate of around 450 has only 55 fast-food restaurants. Also, the average stroke rate to the number of fast-foods were around the range of 280-300. This shows that there are some other factors which results in death's due to stroke.

OBSERVATION

Average Stroke Mortality Rates in a state is not directly proportional to the total number of fast food restaurants in that state. Whereas, it does depend upon the Gender of the population. Ethnicity does not really play an important role. There are other factors that accounts for these rates, for sure.

FUTURE SCOPE

- We wish to work with other datasets like alcohol consumption or smoking habits of people in US states, or probably data relating to physical fitness of people, which again are few factors affecting these rates. Performing a detailed analysis on these factors would enable us to understand the main factors causing these deaths.

- The risk of Americans developing and dying from cardiovascular diseases would be substantially reduced if major improvements were made across the U.S. population in diet and physical activity, control of high blood pressure and cholesterol, smoking cessation etc.