Group Members: Jin Bae, Amandine Lemmonaire, Dheeraj Menon IST 652

Air Quality in New York City

Following questions are answered in the report

- describe the data and its source(s), including any preprocessing

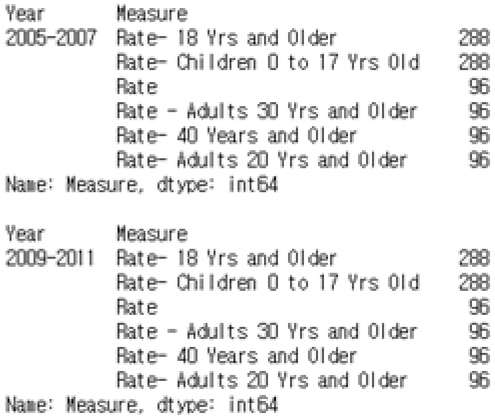
Our dataset is ‘airquality.csv’ file obtained from data.gov. This dataset includes data from NYC neighborhood areas and data values taken for years 2005 to 2013. Dataset explains different kind of chemical effects on people residing in these areas.

Another dataset includes the data for NY state as a whole and particularly focus on year 2009.

- describe your methods of analysis, including the questions that will be answered, what fields in the data will be used, and what the resulting output will be

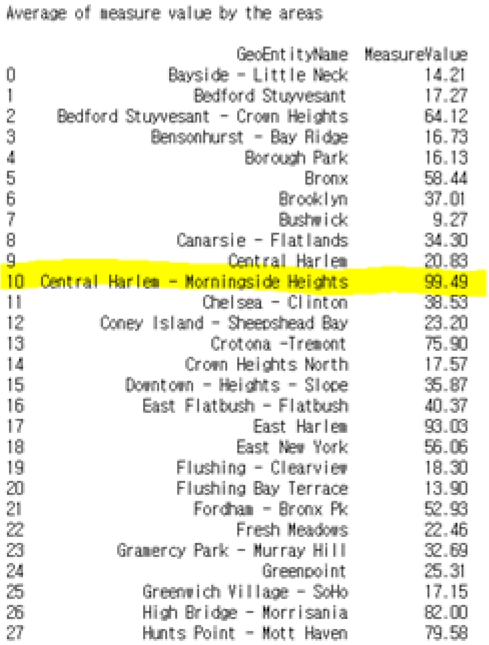
Analysis- Age Groups affected/ Area affected : Jinseo Bae

Q: What age group are considered in the 2005-2007 and 2009 and 2011 data?
A: 18 Yrs and Older, Children 0 to 17 years old, Adults 30 years and Older, 40 years and Older, and Adults 20 years and Older.
Q: How many people in particular age group in particular years?
A:



```
Year       Measure
2005-2007  Rate- 18 Yrs and Older           288
           Rate- Children 0 to 17 Yrs Old   288
           Rate                              96
           Rate - Adults 30 Yrs and Older    96
           Rate- 40 Years and Older          96
           Rate- Adults 20 Yrs and Older     96
Name: Measure, dtype: int64

Year       Measure
2009-2011  Rate- 18 Yrs and Older           288
           Rate- Children 0 to 17 Yrs Old   288
           Rate                              96
           Rate - Adults 30 Yrs and Older    96
           Rate- 40 Years and Older          96
           Rate- Adults 20 Yrs and Older     96
Name: Measure, dtype: int64
```

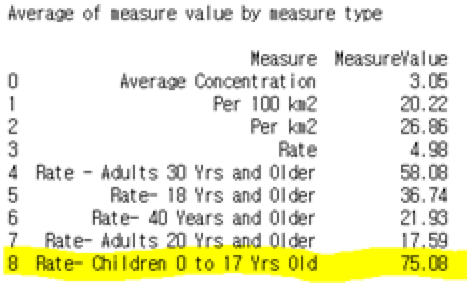
(Analysis has made of not only in 2005-2007 but also of all years in the dataset à output part)
Q: Which age group has highest average of measure values?
A: Highest average for area was Central Harlem - Morningside Heights, which was 99.49.



```
Average of measure value by the areas

                       GeoEntityName  MeasureValue
0                Bayside - Little Neck         14.21
1                   Bedford Stuyvesant         17.27
2    Bedford Stuyvesant - Crown Heights         64.12
3             Bensonhurst - Bay Ridge          16.73
4                        Borough Park          16.13
5                               Bronx          58.44
6                            Brooklyn          37.01
7                            Bushwick           9.27
8                Canarsie - Flatlands          34.30
9                        Central Harlem          20.83
10  Central Harlem - Morningside Heights          99.49
11                  Chelsea - Clinton          38.53
12         Coney Island - Sheepshead Bay          23.20
13                    Crotona -Tremont          75.90
14                 Crown Heights North          17.57
15           Downtown - Heights - Slope          35.87
16           East Flatbush - Flatbush          40.37
17                        East Harlem          93.03
18                      East New York          56.06
19               Flushing - Clearview          18.30
20               Flushing Bay Terrace          13.90
21                 Fordham - Bronx Pk          52.93
22                      Fresh Meadows          22.46
23            Gramercy Park - Murray Hill          32.69
24                         Greenpoint          25.31
25           Greenwich Village - SoHo          17.15
26           High Bridge - Morrisania          82.00
27              Hunts Point - Mott Haven          79.58
```

Q: Which area has highest average of measure values?

A: Highest average for measure type was age between children age 0 to 17 year old, which was 75.08.

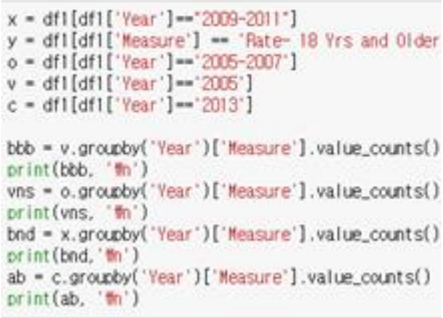


```
Average of measure value by measure type

                          Measure  MeasureValue
0           Average Concentration          3.05
1                      Per 100 km2         20.22
2                          Per km2         26.86
3                             Rate          4.98
4   Rate - Adults 30 Yrs and Older         58.08
5           Rate- 18 Yrs and Older         36.74
6         Rate- 40 Years and Older         21.93
7    Rate- Adults 20 Yrs and Older         17.59
8   Rate- Children 0 to 17 Yrs Old         75.08
```

My part of analysis was focused on age groups and area affected.

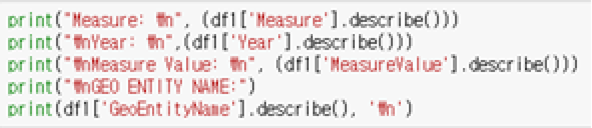
1) I determined number of people by different age ranges in different years. I selected measure column and year column and used groupby().value_counts function to find the results.

(Which was the total number of measure values of age groups)



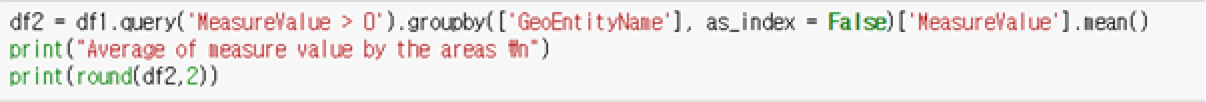
```
x = df1[df1['Year']=="2009-2011"]
y = df1[df1['Measure'] == 'Rate- 18 Yrs and Older
o = df1[df1['Year']=='2005-2007']
v = df1[df1['Year']=='2005']
c = df1[df1['Year']=='2013']

bbb = v.groupby('Year')['Measure'].value_counts()
print(bbb, '₩n')
vns = o.groupby('Year')['Measure'].value_counts()
print(vns, '₩n')
bnd = x.groupby('Year')['Measure'].value_counts()
print(bnd,'₩n')
ab = c.groupby('Year')['Measure'].value_counts()
print(ab, '₩n')
```

2) I used described() function to find each column's most common usage, average, and other basic summaries.

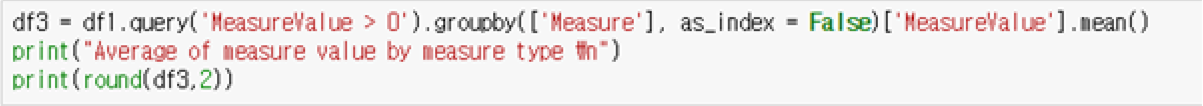


```
print("Measure: ₩n", (df1['Measure'].describe()))
print("₩nYear: ₩n",(df1['Year'].describe()))
print("₩nMeasure Value: ₩n", (df1['MeasureValue'].describe()))
print("₩nGEO ENTITY NAME:")
print(df1['GeoEntityName'].describe(), '₩n')
```

3) I used query and groupby function to find the averages by areas and measure types. (Which was the most affected age group)



```
df2 = df1.query('MeasureValue > 0').groupby(['GeoEntityName'], as_index = False)['MeasureValue'].mean()
print("Average of measure value by the areas ₩n")
print(round(df2,2))
```



```
df3 = df1.query('MeasureValue > 0').groupby(['Measure'], as_index = False)['MeasureValue'].mean()
print("Average of measure value by measure type ₩n")
print(round(df3,2))
```

Another set of analysis covers following (Dheeraj Menon)

what are the frequency of each column values with Measure and GeoType, total count of different types of Measure for each GeoName for the all years, total samples values performed for each type of chemical within each year for different neighbourhoods

- an overall description of the program

Firstly, the program shows the analysis of data for particular years and particular age and calculates the mean, total of people count. It also shows mean and total of each columns.

Secondly, the program shows the frequency of each column values with Measure and GeoType, total count of different types of Measure for each GeoName for the all years, total samples values performed for each type of chemical within each year for different neighbourhoods. This type of analysis gives the total idea about how many people are affected in different years with different effects, under what type of ages.

- if your project is a group project, describe the tasks and roles of each member of the group

Jin calculated the basic structure of dataset which includes calculating total column values and total count of different ages within different years.

Dheeraj analyzed that dataset and checked total count for different neighborhoods, different types of chemical effects on the neighbourhood for different ages

- (grad students) draw conclusions from your results about your data

- The dataset was very good to work with. However, there were certain limitations of the dataset which was a new learning. A new package was used to perform visualizations of dataset which gave us the general idea of healthy, unhealthy, moderate days of NY State. There was a lot of redundant data in our dataset since the values were accounted twice. Hence data cleanup was done.
- Data from other specific cities were not found, a comparison would have allowed for more depth of analysis

- The main dataset regarding New York city has gaps in years
- Population count was missing
- Factors such as the effect on the population are not recorded
- Graphs have been created for all of the years necessary looking at the moderate days and unhealthy days.
- Plans are to also look at hazardous days and unhealthy days for sensitive groups
- A graph may be made looking at the overall unhealthy or moderate days compared to healthy days in the state
- At the moment, I have not been able to properly format the data to sum the days.
- If time permits, a graph looking at the overall data over the years will be made
- The graph will look at the progression of the healthy and unhealthy days in New York State over the years used during the analysis

*****