**Describe the data (its source and preprocessing):**

* Title: New York City Air Quality data from 2005 to 2013.

We got the dataset from data.gov website in the .csv format. This dataset covers different areas (cities) across New York City. Data shows air quality values with different measures under consideration.

The dataset includes measure type, GeoName (Neighberhood Name), GeoID, GeoEntityName(cities), MeasureValue (concentration level).

We particularly looked which neighborhoods within NYC have what types of air quality measures

* For 2nd source with Amy used,

Title: USA Air Quality data from any years

Data provider looked at the entire United States -New York State was singled out through coding. The dataset includes healthy days vs. unhealthy days vs. hazardous days, and more.

**Data clean up:**

* We deleted few extra columns.
* We used fillna(0) to have a good data frame for analysis.

**Analysis 2 (Jinseo Bae): Age Groups affected/ Area affected**

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?

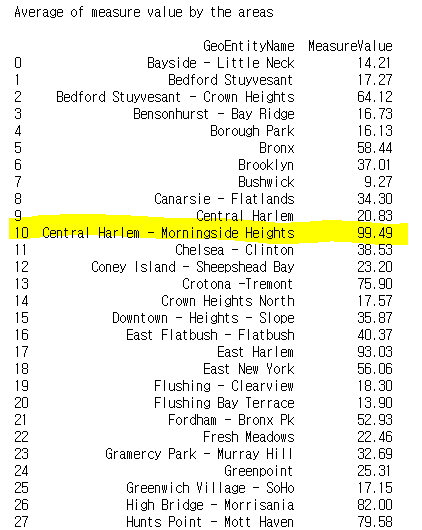
A:



(Analysis has made of not only in 2005-2007 but also of all years in the dataset 🡪 output part)

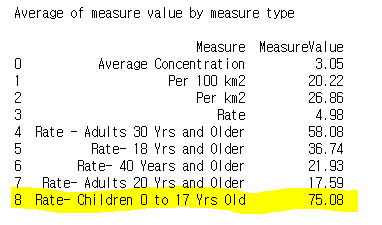
Q: Which age group has highest average of measure values?

A: Highest average for area was Central Harlem - Morningside Heights, which was 99.49.



Q: Which area has highest average of measure values?

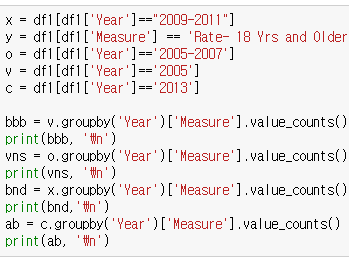
A: Highest average for measure type was age between children age 0 to 17 year old, which was 75.08.



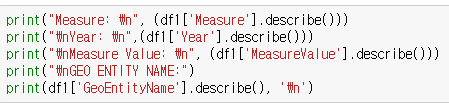
My part of analysis was focused on age groups and area affected.

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)



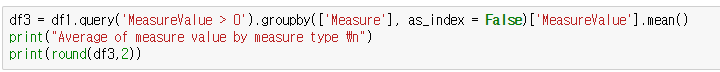
1. I used described() function to find each column’s most common usage, average, and other basic summaries.



1. I used query and groupby function to find the averages by areas and measure types.

(Which was the most affected age group)





**Overall description of the program:**

My part of program was purposed to find the total number of measure values of age groups, most affected age group, most affected city, and basic information of our dataset.

**Tasks and role of me:**

Three of us did different analysis on same dataset. As you can see in the code and output section, our parts can be seen clearly. As I presented in the class, I did analysis based on it.

**Code:**

import pandas as pd

#open file into csv

df = pd.read\_csv('Final\_Project.csv', encoding ='utf-8')

df1 = df.fillna(0) #fill all empty into zero

#print out list of columns

print('Column Labels', df1.columns.values.tolist())

#print out number of rows and columns

print('Number of rows and columns', df1.shape)

df1.dtypes #types of each columns

#print measure column summary including count, freq, top, and name

print("Measure: \n", (df1['Measure'].describe()))

#print year column summary including top, freq

print("\nYear: \n",(df1['Year'].describe()))

#print measure value summary including mean, std, min, and max

print("\nMeasure Value: \n", (df1['MeasureValue'].describe()))

#print Geo entity name summary including count,top,freq

print("\nGEO ENTITY NAME:")

print(df1['GeoEntityName'].describe(), '\n')

#Find total number of participated people in each group in dataset by measure type

e = df1.Measure.value\_counts()

print(e)

#Find average measure value of air quality in total of NYC

print ('Average Measure Value of air quality')

m = df1.MeasureValue.value\_counts()

avg = m.mean()

print(round(avg,2)) #round it into 2 decimal place

#set an alphabet specifically year between 2009 and 2011

x = df1[df1['Year']=="2009-2011"]

#set an alphabet specifically 18 years and older

y = df1[df1['Measure'] == 'Rate- 18 Yrs and Older']

#set an alphabet specifically year between 2005 and 2007

o = df1[df1['Year']=='2005-2007']

#set an alphabet specifically year 2005

v = df1[df1['Year']=='2005']

#set an alphabet specifically year 2013

c = df1[df1['Year']=='2013']

bbb = v.groupby('Year')['Measure'].value\_counts()

print(bbb, '\n') #get counts by measure type in 2005

vns = o.groupby('Year')['Measure'].value\_counts()

print(vns, '\n') #get counts by measure type in 2005-2007

bnd = x.groupby('Year')['Measure'].value\_counts()

print(bnd,'\n') #get counts by measure type in 2009-2011

ab = c.groupby('Year')['Measure'].value\_counts()

print(ab, '\n') #get counts by measure type in 2013

#To find the averages of measure value by each Geo entity Name

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))

#To find the averages of measure value by each Measure type

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))

**Output:**

Column Labels ['Year', 'Measure', 'GeoName', 'GeoID', 'GeoEntityName', 'MeasureValue']

In [6]:

Number of rows and columns (2328, 6)

In [8]:

Year object

Measure object

GeoName object

GeoID int64

GeoEntityName object

MeasureValue float64

dtype: object

Measure:

count 2328

unique 9

top Rate- 18 Yrs and Older

freq 576

Name: Measure, dtype: object

Year:

count 2328

unique 4

top 2009-2011

freq 960

Name: Year, dtype: object

Measure Value:

count 2328.000000

mean 39.200859

std 49.194294

min 0.000000

25% 8.075000

50% 21.250000

75% 49.900000

max 424.700000

Name: MeasureValue, dtype: float64

GEO ENTITY NAME:

count 2328

unique 54

top Upper East Side

freq 51

Name: GeoEntityName, dtype: object

Rate- 18 Yrs and Older 576

Rate- Children 0 to 17 Yrs Old 576

Rate- 40 Years and Older 192

Rate 192

Rate- Adults 20 Yrs and Older 192

Rate - Adults 30 Yrs and Older 192

Per 100 km2 168

Per km2 144

Average Concentration 96

Name: Measure, dtype: int64

Average Measure Value of air quality

3.52

Year Measure

2005 Per 100 km2 168

Average Concentration 96

Name: Measure, dtype: int64

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

Year Measure

2013 Per km2 144

Name: Measure, dtype: int64

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

28 Jamaica 39.58

29 Kingsbridge - Riverdale 34.32

30 Long Island City - Astoria 25.05

31 Lower Manhattan 27.06

32 Manhattan 47.94

33 Midtown Business District 34.67

34 New York City 38.82

35 Northeast Bronx 40.66

36 Pelham - Throgs Neck 47.03

37 Port Richmond 41.59

38 Queens 27.20

39 Ridgewood - Forest Hills 22.92

40 Rockaways 49.68

41 South Beach - Tottenville 16.59

42 Southeast Queens 28.06

43 Southwest Queens 28.01

44 Stapleton - St. George 34.50

45 Staten Island 25.88

46 Sunset Park 25.42

47 Union Square - Lower East Side 44.45

48 Upper East Side 25.95

49 Upper West Side 36.61

50 Washington Heights 46.83

51 West Queens 24.64

52 Williamsburg - Bushwick 63.48

53 Willowbrook 20.95

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