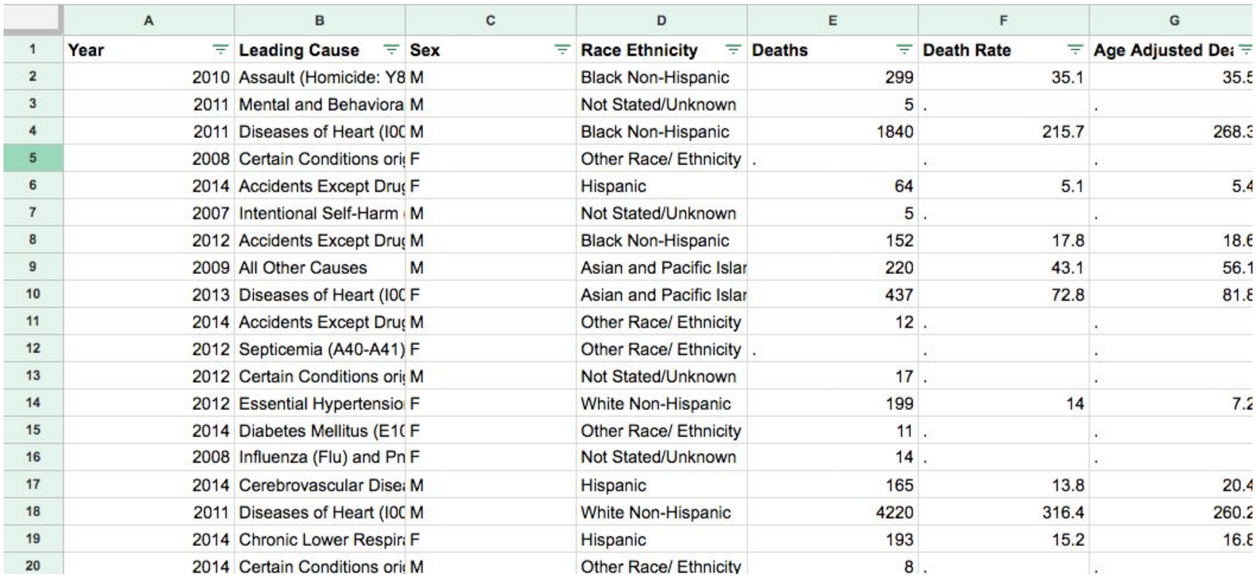
Assignment 2 Data Report - James Fan

The dataset used was taken from the NYC Open Data portal under the Health datasets. It describes the leading causes of death in New York City categorized by ethnicity and gender. The original file format was in CSV, and the scrubbed file format was also a CSV but with various data changes and formatting changes.

Data Link: https://data.cityofnewyork.us/Health/New-York-City-Leading-Causes-of-Death/jb7j-dtam

Raw Data Example



	A	B	C	D	E	F	G
1	Year	Leading Cause	Sex	Race Ethnicity	Deaths	Death Rate	Age Adjusted De
2	2010	Assault (Homicide: Y8	M	Black Non-Hispanic	299	35.1	35.5
3	2011	Mental and Behaviora	M	Not Stated/Unknown	5	.	.
4	2011	Diseases of Heart (I0C	M	Black Non-Hispanic	1840	215.7	268.3
5	2008	Certain Conditions ori	F	Other Race/ Ethnicity	.	.	.
6	2014	Accidents Except Drug	F	Hispanic	64	5.1	5.4
7	2007	Intentional Self-Harm	M	Not Stated/Unknown	5	.	.
8	2012	Accidents Except Drug	M	Black Non-Hispanic	152	17.8	18.6
9	2009	All Other Causes	M	Asian and Pacific Islar	220	43.1	56.1
10	2013	Diseases of Heart (I0C	F	Asian and Pacific Islar	437	72.8	81.8
11	2014	Accidents Except Drug	M	Other Race/ Ethnicity	12	.	.
12	2012	Septicemia (A40-A41)	F	Other Race/ Ethnicity	.	.	.
13	2012	Certain Conditions ori	M	Not Stated/Unknown	17	.	.
14	2012	Essential Hypertensio	F	White Non-Hispanic	199	14	7.2
15	2014	Diabetes Mellitus (E1C	F	Other Race/ Ethnicity	11	.	.
16	2008	Influenza (Flu) and Pn	F	Not Stated/Unknown	14	.	.
17	2014	Cerebrovascular Dise	M	Hispanic	165	13.8	20.4
18	2011	Diseases of Heart (I0C	M	White Non-Hispanic	4220	316.4	260.2
19	2014	Chronic Lower Respir	F	Hispanic	193	15.2	16.8
20	2014	Certain Conditions ori	M	Other Race/ Ethnicity	8	.	.

Data Problems & Audit Trail

While the data was formatted in a CSV document there were no formula for the death rates, or procedures for age-adjusted death rates. On top of this many numbers for deaths were blank due to unavailable information with placeholder periods. This meant a few things. I needed to delete the last two columns because I do not know exactly how they achieved those statistics and wanted to make my own, and secondly to delete any data that was not relevant or missing in order to make my own calculations simpler.

Data was downloaded in CSV format, put into the same directory as a python data-scrubbing file made in Spyder and then output to a output.csv written by the scrubbing program.

Aggregated Data

Average Deaths per Year	Total Deaths From 2007 - 2017	Max Deaths by One Cause	Min Deaths by One Cause
38636.18182	424998	7050	5

Above is the aggregated data statistics shown here. These are statistics that can be seen as overal death statistics that are not clear from just looking at the raw data. The first statistic shows the Average Deaths in NYC per year, while the next shows the Total Deaths in NYC from 2007 - 2017.

Thirdly, the Max Death Cause is a statistic that shows the most deaths from one cause (Heart Disease) and the 4th statistic shows the Minimum deaths of one particular cause.