CDC- Drug Mortality Rate in the US

By Hasnat Tahir

Data Description:

- Data has been taken from a shared Dropbox folder and contains the drug poisoning deaths in the states of the US, and includes variables like Age, Sex, Race, Population, etc for the years 1999-2015.

Data Source URL-
https://www.dropbox.com/sh/wybpws86l930sd8/AACNDmzNk9YGP6NplyJlPzuMa?dl=0

Data Overview:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	year	sex	age	race_and_	state	deaths	populatio	crude_dea	standard_	low_confi	upper_cor	age_adjus	standard_	lower_cor	upper_cor	state_cruc	us_crude_	us_age_ad
2	1999	Both Sexe	All Ages	All Races-/	Alabama	169	4,430,143	3.8148	0.29344	3.2396	4.3899	3.8521	0.29657	3.2708	4.4334	1.8-7	6.0382	6.057
3	2000	Both Sexe	All Ages	All Races-/	Alabama	197	4,447,100	4.4299	0.31561	3.8112	5.0485	4.4857	0.31985	3.8588	5.1126	1.8-7	6.1882	6.1749
4	2001	Both Sexe	All Ages	All Races-/	Alabama	216	4,467,634	4.8348	0.32896	4.19	5.4795	4.8915	0.33329	4.2382	5.5447	1.8-7	6.8057	6.7922
5	2002	Both Sexe	All Ages	All Races-/	Alabama	211	4,480,089	4.7097	0.32423	4.0742	5.3452	4.7619	0.32868	4.1177	5.4062	1.8-7	8.1766	8.1957
6	2003	Both Sexe	All Ages	All Races-/	Alabama	197	4,503,491	4.3744	0.31166	3.7635	4.9852	4.4333	0.31701	3.812	5.0547	1.8-7	8.8881	8.8765
7	2004	Both Sexe	All Ages	All Races-/	Alabama	283	4,530,729	6.2462	0.3713	5.5185	6.974	6.3542	0.37944	5.6105	7.0979	1.8-7	9.366	9.3831
8	2005	Both Sexe	All Ages	All Races-/	Alabama	283	4,569,805	6.1928	0.36813	5.4713	6.9143	6.333	0.37832	5.5915	7.0745	1.8-7	10.0884	10.0699
9	2006	Both Sexe	All Ages	All Races-/	Alabama	398	4,628,981	8.598	0.43098	7.7533	9.4427	8.7498	0.44162	7.8842	9.6154	7-9.8	1	3
10	2007	Both Sexe	All Ages	All Races-/	Alabama	511	4,672,840	10.9355	0.48376	9.9874	11.8837	11.0885	0.49516	10.118	12.059	9.9-12.3	1	5
11	2008	Both Sexe	All Ages	All Races-/	Alabama	607	4,718,206	12.8651	0.52218	11.8416	13.8885	12.9811	0.53292	11.9366	14.0256	12.3-15.2	1	7

Fig-1

	COLUMN_NAME	DATA_TYPE
1	year	float
2	sex	nvarchar
3	age	nvarchar
4	race_and_hispanic_origin	nvarchar
5	state	nvarchar
6	deaths	float
7	population	float
8	crude_death_rate	float
9	standard_error_for_crude_rate	float
10	low_confidence_limit_for_crude_rate	float
11	upper_confidence_limit_for_crude_rate	float
12	age_adjusted_rate	float
13	standard_error_age_adjusted_rate	float
14	lower_confidence_limit_for_age_adju...	float
15	upper_confidence_limit_for_age_adju...	float
16	state_crude_rate_in_range	nvarchar
17	us_crude_rate	float
18	us_age_adjusted_rate	float

Fig-2

```
In [25]: #descriptive statistics
         data.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 2703 entries, 0 to 2702
         Data columns (total 18 columns):
         year                                          2703 non-null int64
         sex                                           2703 non-null object
         age                                           2703 non-null object
         race_and_hispanic_origin                      2703 non-null object
         state                                         2703 non-null object
         deaths                                        2703 non-null object
         population                                    2703 non-null object
         crude_death_rate                              2703 non-null float64
         standard_error_for_crude_rate                 2703 non-null float64
         low_confidence_limit_for_crude_rate           2703 non-null float64
         upper_confidence_limit_for_crude_rate         2703 non-null float64
         age-adjusted_rate                             1071 non-null float64
         standard_error_age-adjusted_rate              1071 non-null float64
         lower_confidence_limit_for_age-adjusted_rate  1071 non-null float64
         upper_confidence_limit_for_age-adjusted_rate  1071 non-null float64
         state_crude_rate_in_range                     1071 non-null object
         us_crude_rate                                 2703 non-null float64
         us_age-adjusted_rate                          2703 non-null float64
```

Fig-3

- The dataset contains 18 columns and 2703 columns.
- Fig-1 above gives an overview of the data stored in the dataset. There are 5 qualitative and 13 quantitative variables in the dataset.
- Fig-2 shows the sql output describes the data type of all columns in the dataset.

Problems with the data:

- The data had bad column names that included spaces and some invalid characters that sql does not support. I used python to change the column names, code used is available in the last page of this report.
- The dataset has missing values in some columns as evident from the output shown in Fig-3. Python was used to generate the output.

- The ‘States’ column contains ‘United States’ as an observation. Every time for analysis where ‘State’ was used I had to filter out ‘United States’.
- The ‘Age’ column contains values in range, had it been continuous data it would have been easier to work with and more analysis would have been possible.
- In the dataset, state wise observations are available only for ‘Both sexes’ values. This make the ‘Sex’ column obscure for proper analysis.

General Statistics:

- Total number of records in the dataset: **2703**
- Number of distinct values in each categorical column:

	Year	Age	Race	State	Sex
1	17	9	4	52	3

- Maximum number of drug poisoning deaths between 1999-2015:

	HIGHEST_DEATH_COUNT	STATE	YEAR
1	4659	California	2015

- Minimum number of drug poisoning deaths between 1999-2015:

	LOWEST_DEATH_COUNT	STATE	YEAR
1	12	North Dakota	1999

- Average number of drug poisoning deaths in the US year on year(bottom left table):

	YEAR	AVG_DEATHS
1	1999	330
2	2000	341
3	2001	380
4	2002	461
5	2003	506
6	2004	538
7	2005	585
8	2006	675
9	2007	706
10	2008	715
11	2009	726
12	2010	752
13	2011	811
14	2012	814
15	2013	862
16	2014	923
17	2015	1028

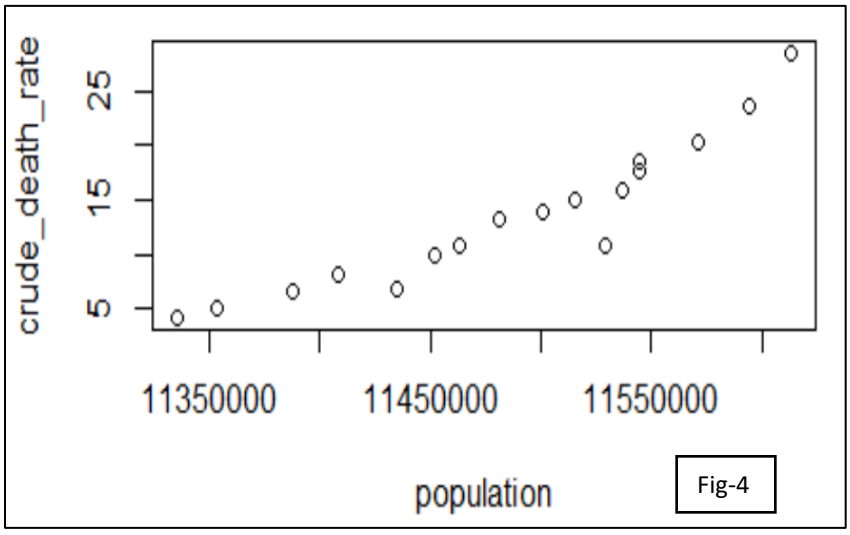


Fig-4

- Average statistics of **Ohio**:

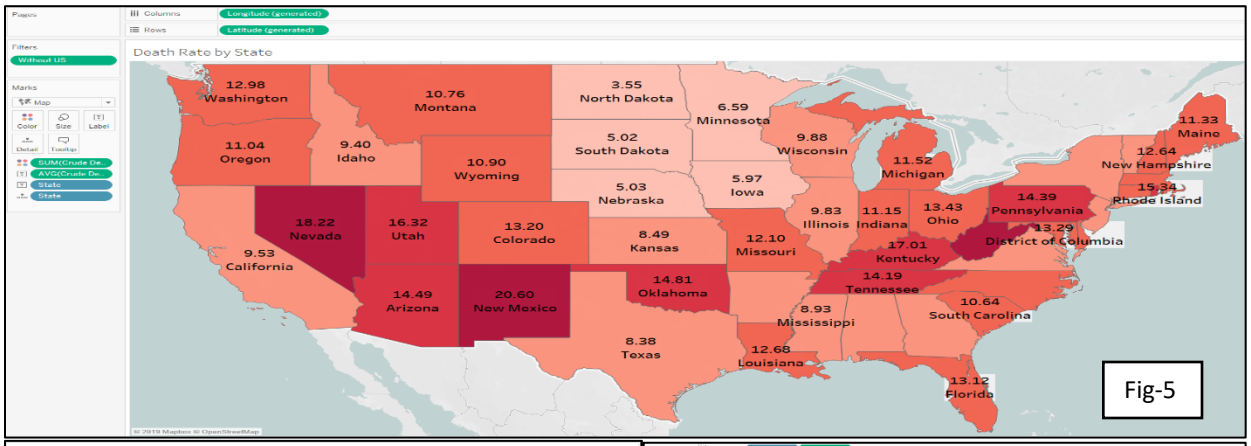
	state	Avg_Deaths	Avg_Population	Avg_DeathRate
1	Ohio	1547.88	11486134.76	13.43

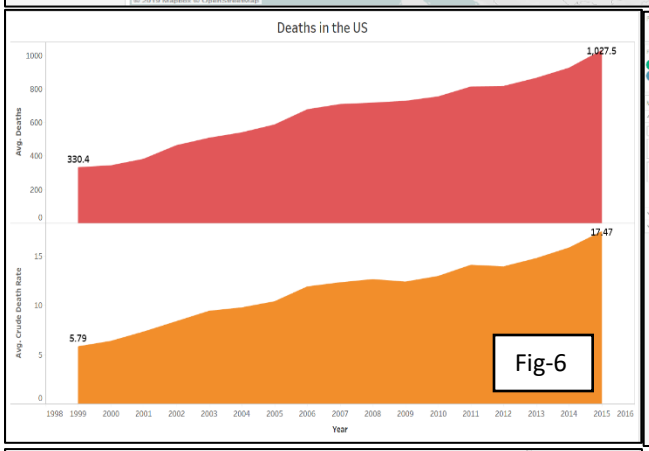
- Average Statistics of the **United States**:

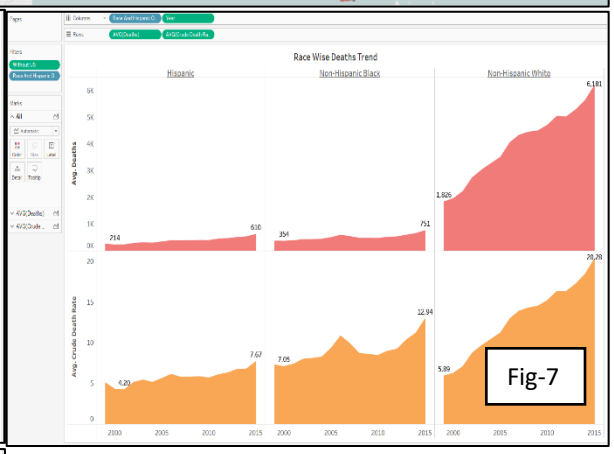
	Avg_Deaths	Avg_Population	Avg_DeathRate
1	655.94	5896907.67	11.49

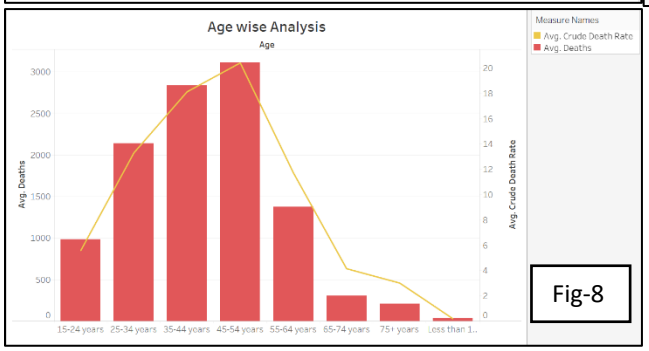
- Fig-4 represents the scatter plot plotted in **R** of Crude Death rate vs Population of Ohio. The Correlation between these two variables is 0.9374643, which shows a very strong correlation.

Additional Analysis:









The above charts were created using **Tableau,** had to get rid of 'United States' from the states column before creating the charts**.** I connected my SQL sever to the Tableau software for making these charts. Some inferences can be made from these charts and are mentioned below:

Findings:

- Fig-5 shows the State-wise crude death rate in the United States. We can see in the map that the highest death rates are shared by neighboring states of Nevada, Arizona, Utah and New Mexico. A pattern can be observed from the chart.

- Fig-6 shows the year on year trend of Average Deaths in the US and Average Crude Death Rate, we can tell there is a marginal increase in both the figures. Alarmingly, the average death rate increased by around 210 % and the average crude death rate increased by 200%.
- Fig-7 shows us the Race-wise segregated year on year charts of average deaths and average crude death rate. It is evident from the charts that "Non-Hispanic Whites" community has the highest number of deaths, also the highest death rate. That too increasing at a very high rate. Also, the gap in Non-Hispanic Whites and other communities is big.
- Fig-8 shows the age-wise drug poisoning deaths in the United States. It is clear from the chart that 35-54 years is the most affected age group.

Summary:

Based on all the analyses, following conclusion can be drawn on the CDC_mortality rate data.

- Overall, State-wise California had the highest number of drug poisoning deaths (Count = 4659) so far, in the year 2015. And, North Dakota had the lowest number of drug poisoning deaths (Count = 12) so far, in the year 1999.
- The overall Crude death rate and death counts in the United states are increasing year on year at an alarming rate. There is a 210% and 200% increase in the average deaths and average crude death rate respectively.
- Non-Hispanic White community is the most affected community in the United States.
- Age wise analysis shows that majority of the deaths people from the age group of 25 -54 years. Out of which 35-54 years age group had the most deaths by drug poisoning.
- Ohio leads the country in drug overdose deaths per capita.
- Ohio's overdose rate has grown almost nine-fold between 1999-2016. Also, the average death rate and average deaths for Ohio is above the country average.
- Using R, I tried to fit a regression model to predict the death rate, the p-values for the coefficients were promising until I did the residual analysis. As per, the obtained residual plots some transformation are required to get a perfect model.
- Software and languages used for this HW: **SQL, Python, R and Tableau**.

Challenges faced:

- The dataset had bad column names which I fixed using python.
- The state column had 'United States' in it, due to which I had to filter it out every time I had to do an analysis. Cannot remove it completely as it might degrade the dataset.
- In the dataset, state wise observations are available only for 'Both sexes' values. This make the 'Sex' column obscure for proper analysis.

Please find all the relevant codes related to this case in the subsequent pages.

Codes:

SQL

```sql
CREATE DATABASE H5
USE H5

-- Data: CDC Mortality in the US
SELECT * FROM DBO.CDC_MORTALITY
SELECT COUNT(*) as Total_Obsrv FROM DBO.cdc_mortality

-- checking the column types of the table

SELECT COLUMN_NAME, DATA_TYPE
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_NAME = 'CDC_MORTALITY'

-- duplicate values
select count(*)
from dbo.CDC_MORTALITY
GROUP BY
HAVING COUNT(*)>1
--

-- NULL VALUES
SELECT COUNT(AGE_ADJUSTED_RATE) FROM DBO.CDC_MORTALITY
WHERE AGE_ADJUSTED_RATE IS not NULL


-- MISSING VALUES FOUND ON PYTHON

--General statistics

-- Checking distinct number of values in each categorical column
select count(distinct year) as Year,
count(distinct age) as Age,
count(distinct race_and_hispanic_origin) as Race,
count(distinct state) as State,
count(distinct sex) as Sex
  from dbo.cdc_mortality

--State wise  average polpulation along the years
SELECT STATE, YEAR, AVG(POPULATION) AS AVG_POP
FROM DBO.CDC_MORTALITY
GROUP BY STATE, YEAR
ORDER BY STATE;

-- Number of distinct values in the columnS, TELLS YOU MORE ABOUT THE DATA

SELECT DISTINCT AGE FROM DBO.CDC_MORTALITY
ORDER BY AGE

SELECT DISTINCT STATE FROM DBO.CDC_MORTALITY
ORDER BY STATE
-- Contains United States and District of Columbia, so 52

-- SEX WISE NUMBER OF DATA
SELECT sex, COUNT(*) AS 'count' FROM DBO.CDC_MORTALITY
GROUP BY SEX

SELECT * FROM DBO.CDC_MORTALITY
```

```sql
-- TOTAL POPULATION OF THE US YEAR WISE
SELECT YEAR, SUM(POPULATION) AS TOTAL_POP_US FROM DBO.CDC_MORTALITY
WHERE STATE != 'UNITED STATES'
GROUP BY YEAR
ORDER BY YEAR
-- HAD TO REMOVE UNITED STATES FOR CORRECT FIGURES

-- MAXIMUM NUMBER OF DEATHS
SELECT TOP 1 DEATHS AS HIGHEST_DEATH_COUNT, STATE, YEAR
FROM DBO.CDC_MORTALITY
WHERE STATE != 'UNITED STATES'
ORDER BY DEATHS DESC

--MINIMUM NUMBER OF DEATHS
SELECT TOP 1 DEATHS AS LOWEST_DEATH_COUNT, STATE, YEAR
FROM DBO.CDC_MORTALITY
WHERE STATE != 'UNITED STATES'
ORDER BY DEATHS

-- AVERAGE DEATHS YEAR ON YEAR IN THE US
SELECT YEAR, ROUND(AVG(DEATHS),0) AS AVG_DEATHS
FROM DBO.CDC_MORTALITY
WHERE STATE != 'UNITED STATES'
GROUP BY YEAR
ORDER BY YEAR

PRINT @@SERVERNAME
-- data contains male and female values only in case of state = united states, so some
comparisons are be difficult

select state,
       Round(avg(Deaths),2) as Avg_Deaths,
       round(avg(population),2) as Avg_Population,
       round(avg(crude_death_rate),2) as Avg_DeathRate
from dbo.cdc_mortality
where state = 'ohio'
group by state

select Round(avg(Deaths),2) as Avg_Deaths,
       round(avg(population),2) as Avg_Population,
       round(avg(crude_death_rate),2) as Avg_DeathRate
from dbo.cdc_mortality
where state !=  'united States'
```

Python:

```python
#Import Data
data = pd.read_csv('E:\CDC_Mortality_State.csv')
data

#Cleaning bad column names that contained spaces and invalid characters
clean_names = data.columns.str.strip().str.lower().str.replace(' ', '_').str.replace('(', '').str.replace(')', '')
data.columns= clean_names
data
#Downloading the data
data.to_csv(r'E:\clean_cdc_data.csv')
```

R:

```
library(readxl)

data = read_excel('E:/clean_cdc_data_r.xls')
data
head(data)
Population = data$population
Death_rate = data$crude_death_rate

reg_model = lm(Death_rate~Population)
summary(reg_model)
plot(reg_model)

cor(x,y) # 0.0672922
plot(Population,Death_rate)

#data for multiple sates were present so took for just on state to test the relationship
newdata <- subset(data, data$state == 'Ohio', select=c(population, crude_death_rate))
plot(newdata)
cor(newdata)

reg_model01 = lm(newdata$crude_death_rate~newdata$population)
summary(reg_model01)
plot(reg_model01)
```