Health Status Indicators

Project Introduction

1. Goal of The Project

1. Analysing health status indicators and understand how behavioural factors such as obesity, tobacco use, diet, physical activity, drugs, alcohol usage and others contribute and relate to leading causes of death like obesity, heart disease, cancer etc and determine what affects them at the county level.
2.Using Predictive Modeling for predicting and understanding the leading causes of death like Lung Cancer, Breast Cancer, Colon Cancer, Poverty, Heart Diseases etc using various feature selection techniques.

2. Expected Outcome

1.After exploratory data analysis, a one-time report to be prepared providing insights and measures to be taken to improve healthcare across the country.

2.Predictive Analytics to be done using Machine Learning based models to provide and assist with evidence based decisions for better healthcare on Leading Causes of Death like Diabetes, Cancer, Heart Diseases etc.

3.Spread awareness about the issues affecting public health and a tool to local public health agencies for improving their community's health and provide them with insights to assist in the development of public policies, health programs and prioritise funding in the most effective pathway.

3. About the Data

1.HEALTHY_PEOPLE_2010.csv (Healthy People 2010 Targets and the U.S. Percentages or Rates)
2.DEMOGRAPHICS.csv (Demographics indicator domain) 3141 rows, 44 columns
3.LEADING_CAUSES_OF_DEATH.csv (Leading Causes of Death indicator domain) 3000+ rows, 235 columns
4.SUMMARY_MEASURES_OF_HEALTH.csv (Summary Measures of Health indicator domain) 3141 rows, 141 columns
5.MEASURES_OF_BIRTH_AND_DEATH.csv (Measures of Birth and Death indicator domain) 3000+ rows, 141 columns
6.RELATIVE_HEALTH_IMPORTANCE. (Relative Health Importance indicator domain) 3141 rows, 28 columns
7.VULNERABLE_POPS_AND_ENV_HEALTH.csv (Vulnerable Populations and Environmental Health) 3141 rows, 28 columns
8.PREVENTIVE_SERVICES_USE. (Preventive Services indicator domain) 3141 rows, 43 columns
9.RISK_FACTORS_AND_ACCESS_TO_CARE.csv (Risk Factors and Access to Care indicator domain) 3141 rows, 31 columns

4. Goals

- We strive to provide as much useful insights as possible from the dataset to spread awareness about general issues pertaining to health and give correlations of each of the above attribute with other attributes in the dataset and find actionable insights.
- We intend to determine the important factors which are responsible or have correlation for higher values for each of the attributes in leading causes of death.
- Predictive Modeling for leading causes of death like Average Life Expectancy, Breast Cancer, Lung Cancer, Diabetes etc. on a county level.

5. About Data Cleaning

A data dump was picked up from the link below in a zip file consisting of 10 csvs. The data dump consisted of the data dictionary as well as the description of the default values therein.

1. All missing values and default values were replaced to nan or handled appropriately for plotting purposes.

Default Values= [-9999,-2222,-2222.2,-2,-1111.1,-1111,-1,-9998.9]

2.During Modeling Phase missing values in all numerical attributes were replaced by the mean of the column

In [1]:
```
def DataCleaner(df):
    ListofNans=[-9999,-2222,-2222.2,-2,-1111.1,-1111,-1,-9998.9]
    df=df.replace([i for i in ListofNans], np.NAN)#replacing odd values with nan
    return(df)#except this no cleaning was required. For Modelling phase missing values were imputed with mean of the col
import os
from os import listdir
from os.path import isfile, join
codepath=r'C:\Users\Varun\Desktop\Studies\IDS Project\Codes'
datapath=r'C:\Users\Varun\Desktop\Studies\IDS Project\Dataset'
os.chdir(datapath)
onlyfiles = [f for f in listdir(datapath) if isfile(join(datapath, f))]
os.chdir(codepath)
%run LibrariesImport.py # loading libraries
%run DD.py # loading datadictionary
```

```
All Libraries loaded
```

6. Data Dictionary (Description About All CSVs and Columns)

In [2]:
```
DD.head(5) #manually change the integer to display more rows
```

Out[2]:

	PAGE_NAME	COLUMN_NAME	DESCRIPTION	IS_PERCENT_DATA
0	Demographics	State_FIPS_Code	Two-digit state identifier, developed by the N...	N
1	Demographics	County_FIPS_Code	Three-digit county identifier, developed by th...	N
2	Demographics	CHSI_County_Name	Name of county	N
3	Demographics	CHSI_State_Name	Name of State or District of Columbia	N
4	Demographics	CHSI_State_Abbr	Two-character postal abbreviation for state name	N

7. Exploratory Data Analysis

We have studied every attribute in all the files, merged all of them on the basis of the primary keys: 'State_FIPS_Code', 'County_FIPS_Code', 'CHSI_County_Name','CHSI_State_Name', 'CHSI_State_Abbr', 'Strata_ID_Number'.

1.The Data Granularity is of a county level.
2.Each row represents various values in percentages/or numeric of a particular county of a particular state.
3.The scope of the dataset is entire population of United States of America.
4. One Time Survey Data 1993-2003. There are 3141 records in total, One record(row) per county of United States.
5. Merged CSV of 9CSVs (Check About the Data Heading)

Independent Attributes: 'No_Exercise', 'Few_Fruit_Veg', 'Obesity', 'High_Blood_Pres', 'Smoker', 'Uninsured', 'Elderly_Medicare', 'Disabled_Medicare', 'Prim_Care_Phys_Rate', 'Dentist_Rate', 'FluB_Rpt', 'HepA_Rpt', 'HepB_Rpt', 'Meas_Rpt', 'Pert_Rpt', 'CRS_Rpt', 'Syphilis_Rpt', 'FluB_Rpt%', 'HepA_Rpt%', 'HepB_Rpt%', 'Meas_Rpt%', 'Pert_Rpt%', 'CRS_Rpt%', 'Syphilis_Rpt%', 'Pap_Smear', 'Mammogram', 'Proctoscopy', 'Pneumo_Vax', 'Flu_Vac', 'Pap_Smear%', 'Mammogram%', 'Proctoscopy%', 'Pneumo_Vax%', 'Flu_Vac%', 'Population_Size', 'Population_Density', 'Poverty', 'Age_19_Under', 'Age_19_64', 'Age_65_84', 'Age_85_and_Over', 'White', 'Black', 'Native_American', 'Asian', 'Hispanic', 'No_HS_Diploma', 'No_HS_Diploma%', 'Unemployed', 'Unemployed%', 'Sev_Work_Disabled', 'Sev_Work_Disabled%', 'Major_Depression', 'Major_Depression%', 'Recent_Drug_Use', 'Recent_Drug_Use%', 'Ecol_Rpt', 'Salm_Rpt', 'Shig_Rpt', 'Toxic_Chem', 'All_Death', 'Health_Status', 'Unhealthy_Days', 'LBW', 'VLBW', 'Premature', 'Under_18', 'Total_Births', 'Total_Deaths', 'Total_Births%', 'Total_Deaths%', 'Over_40', 'Unmarried', 'Late_Care', 'Infant_Mortality', 'IM_Neonatal', 'IM_Postneonatal' 'Homicide', 'Homicide%'

Dependent Attributes(Leading Causes of Death): 'ALE', 'Diabetes', 'Lung_Cancer', 'Brst_Cancer', 'Col_Cancer', 'Stroke', 'Suicide', 'Injury', 'CHD'

After Careful EDA of all attributes we found many Observations, some of the intuitive ones have been listed below:

In [17]:
```
PSU_Demo_VPEH_SMOH_RFAC_df.head()# merged, collated, cleaned ready for analysis dataset
```

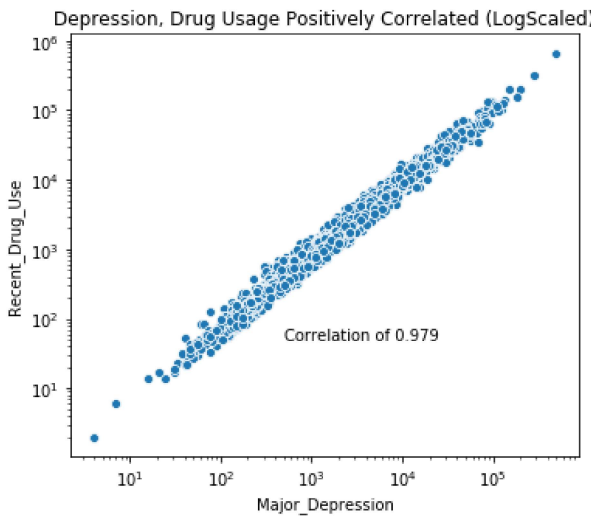
Out[17]:

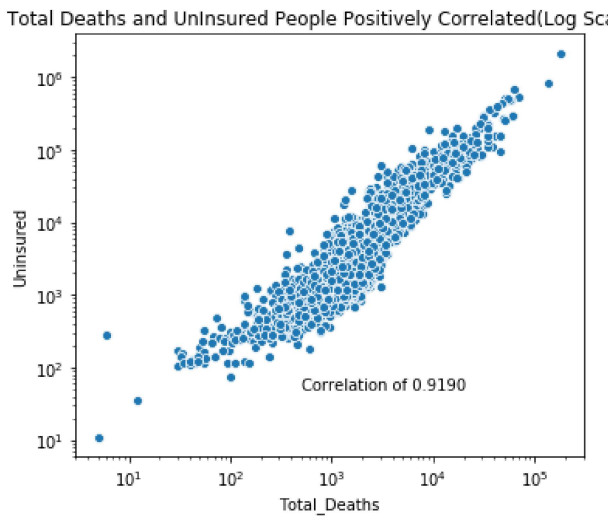
	State_FIPS_Code	County_FIPS_Code	CHSI_County_Name	CHSI_State_Name	CHSI_State_Abbr	Strata_ID_Number	No_Exercise	Few_Fruit_Veg	Ob
0	1	1	Autauga	Alabama	AL	29	27.8	78.6	
1	1	3	Baldwin	Alabama	AL	16	27.2	76.2	
2	1	5	Barbour	Alabama	AL	51	NaN	NaN	
3	1	7	Bibb	Alabama	AL	42	NaN	86.6	
4	1	9	Blount	Alabama	AL	28	33.5	74.6	

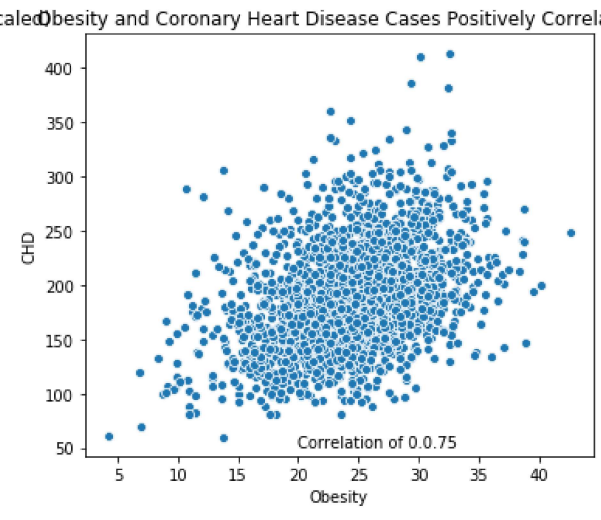
5 rows × 71 columns

In [4]:
```
os.chdir(codepath)
%run EDA.py
os.chdir(codepath)
%run plotter1.py
```

1.Higher cases of depression in a county are correlated with higher drug usage
2.Counties with more uninsured people, have higher death rates
3.Higher Coronary Heart Disease Cases in a county seen in counties with more obesity levels

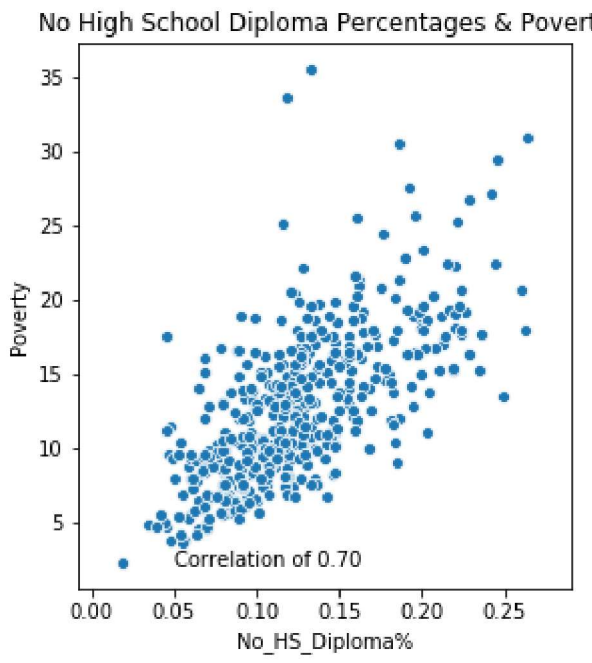


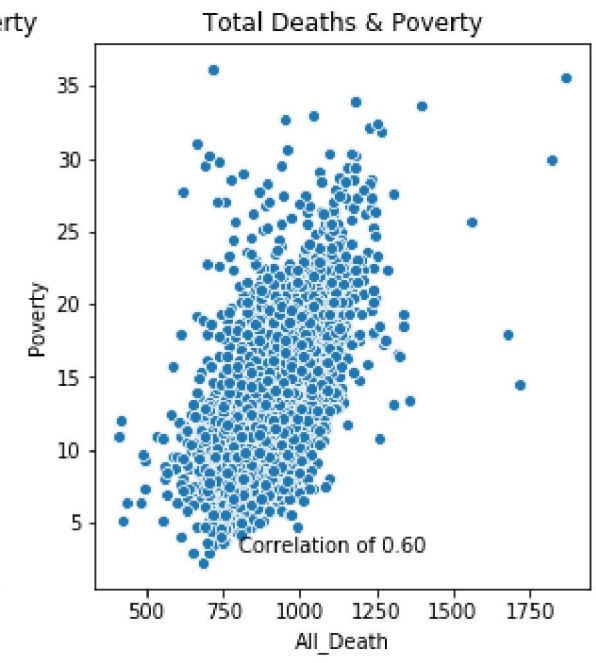


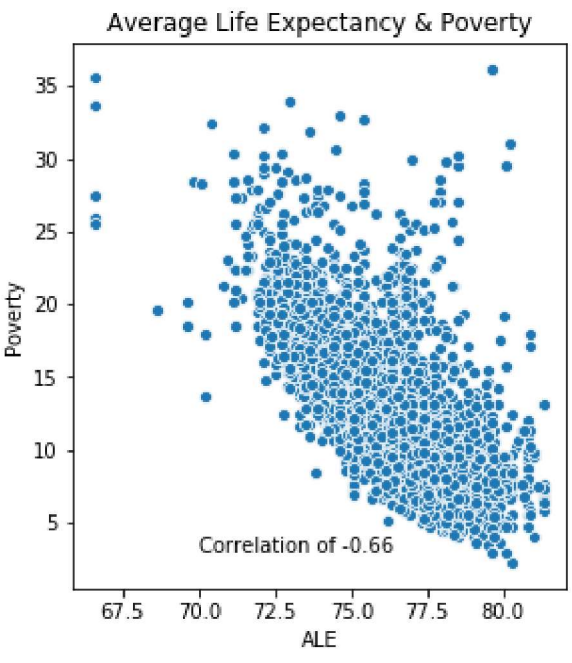


In [5]:
```
os.chdir(codepath)
%run plotter2.py
```

4 . Counties with Higher Povery Level have more people who are less educated
5 . Counties with Higher Povery Level have more death rates
6 . Counties with Higher Povery Level have lower average life expectancy levels

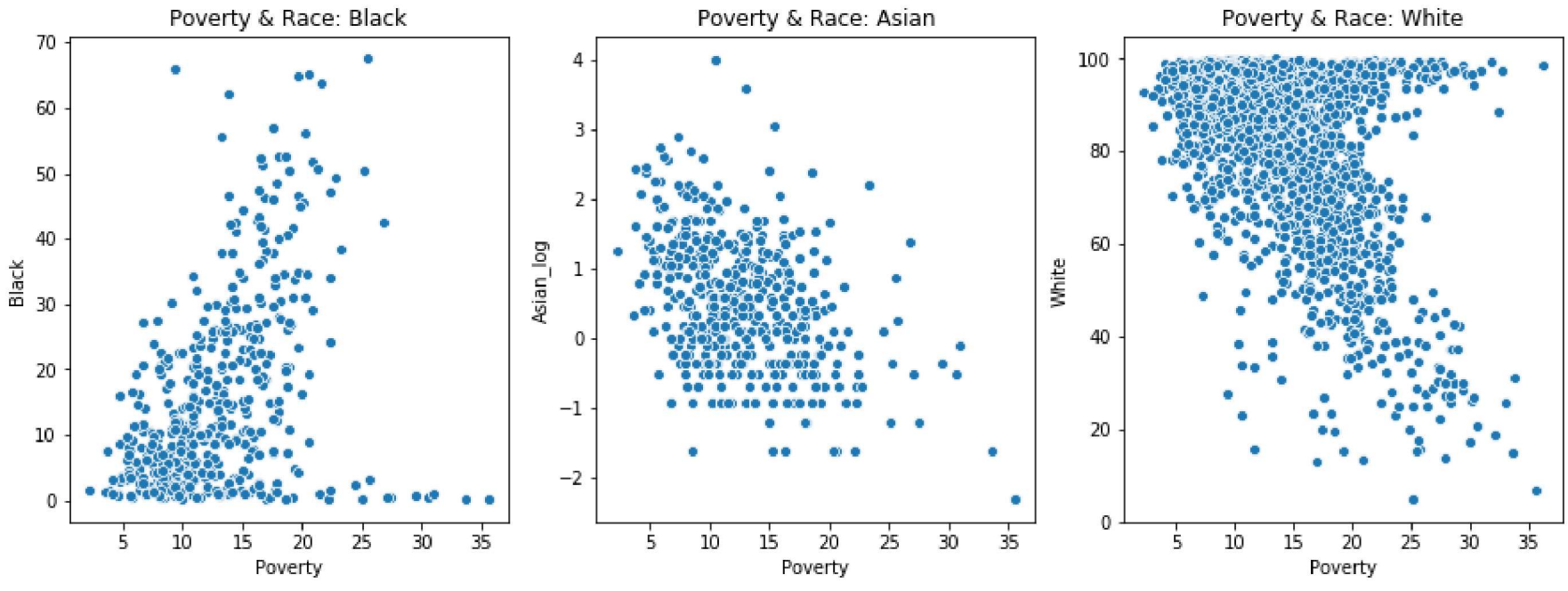






In [6]:
```
os.chdir(codepath)
%run plotter3.py
print("Poverty Ridden Counties are the ones which have higher population of Blacks")
```

```
Poverty Ridden Counties are the ones which have higher population of Blacks
```



In [82]:
```
os.chdir(codepath)
%run Plotter_7.py
```

1.No High School Percentages Positively correlated with drug use, depression and unemployment rate

2.Average life expectancy negatively correlated with obesity and high blood pressure

3.No Excercsie Positively Correlated with Obesity and High Blood Pressure

4.Lung Cancer and Smoker, Positively Correlated.

5. Poverty and homicide % of a county positively correlated.

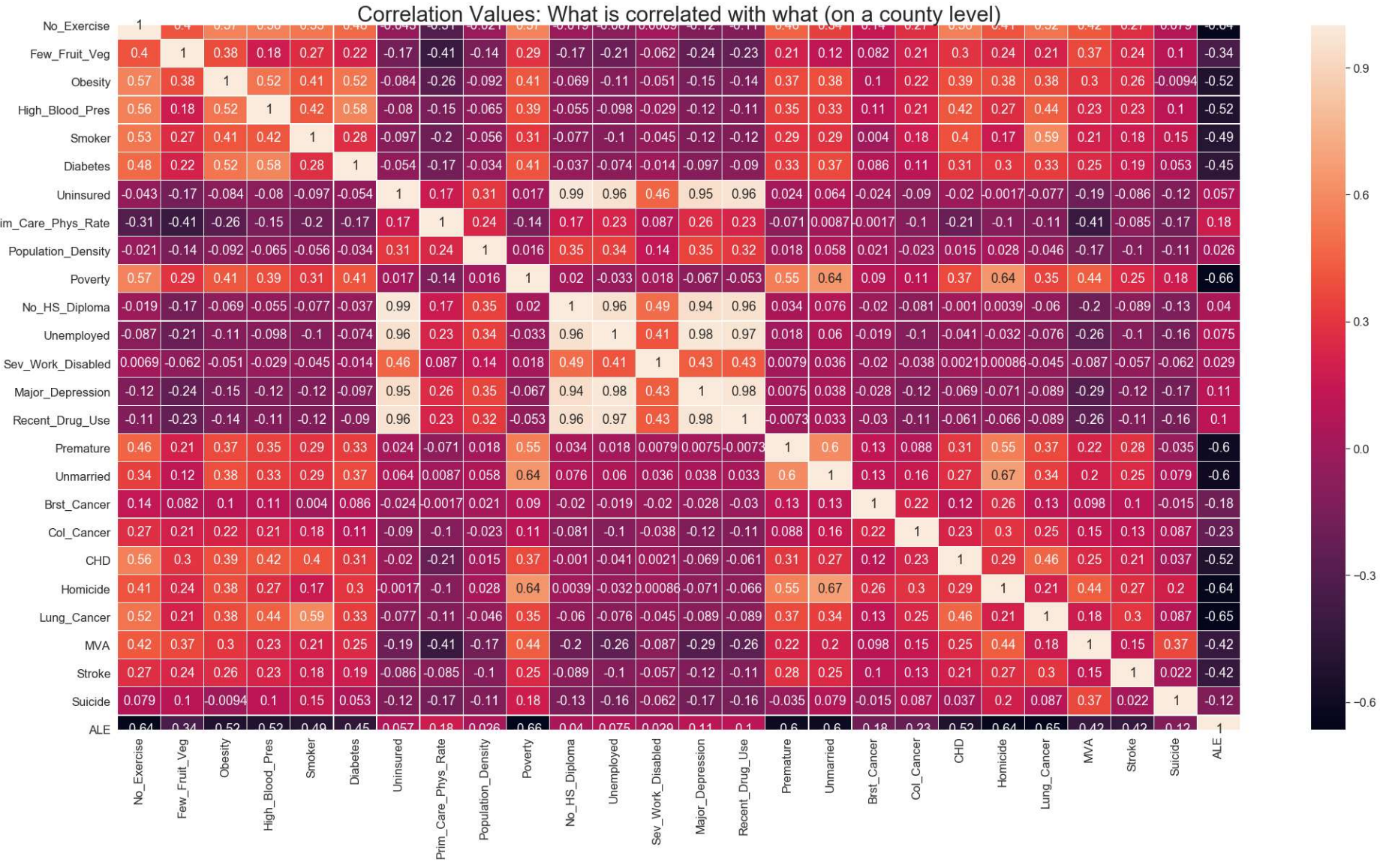
6. Poverty and premature% of a county positively correlated.

7. Poverty and unmarried(unmarried women% who give birth) positively correlated.

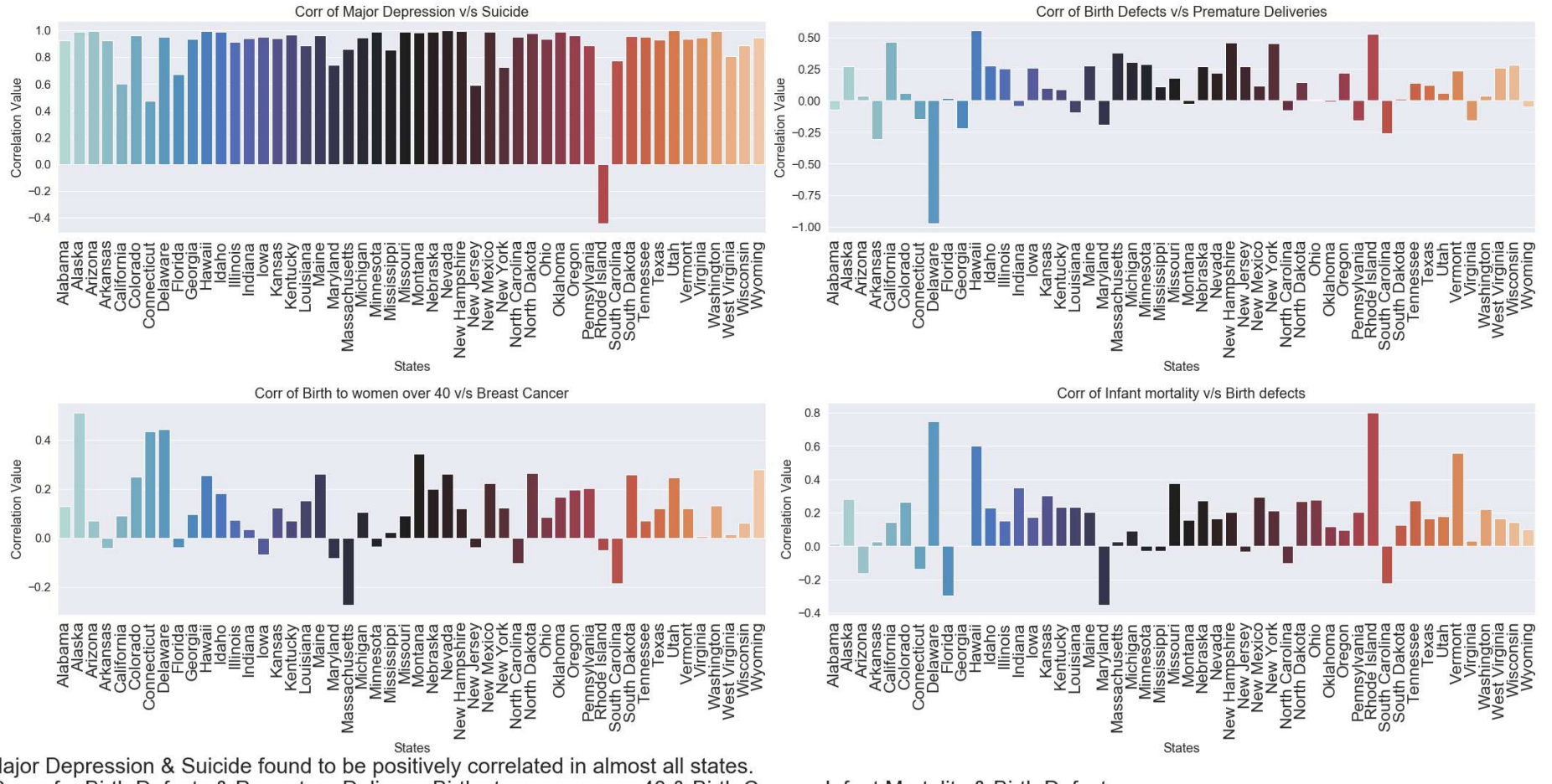
8.No High school diploma%, Drug Use, Unemployment rate, Major depression positvely correlated.

9.Obesity, High Blood Pressure and Diabetes Positively Correlated.

10.No Excercsie correlated with obesity, high blood pressure, diabetes and smoker%



In [83]:
```
os.chdir(codepath)
%run Vis3_Ap.py
```



Major Depression & Suicide found to be positively correlated in almost all states.
Same for Birth Defects & Premature Delivery, Births to women over 40 & Birth Cancer, Infant Mortality & Birth Defects

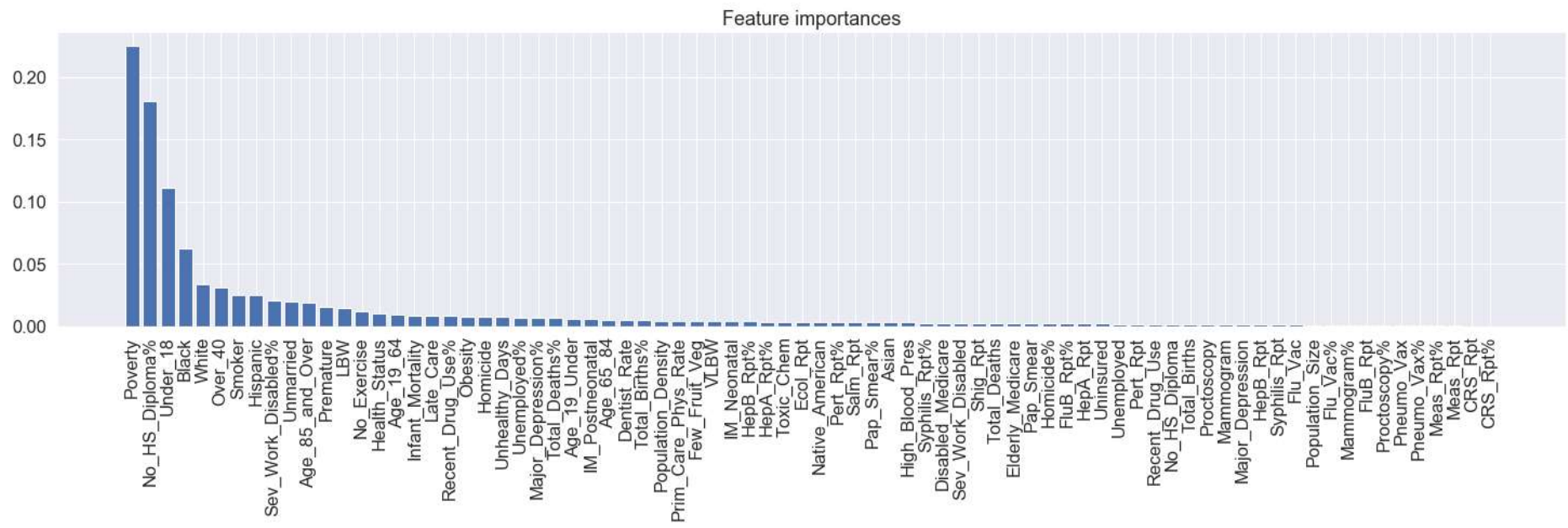
In [84]:
```
## We will try to predict each of these using the predictor attributes and find out which are the most important ones in d
os.chdir(codepath)
%run Modelling_Phase_1.py

ToBePredicted=['ALE','Diabetes','Lung_Cancer','Brst_Cancer','Col_Cancer','MVA','Stroke', 'Suicide','CHD']
strval='ALE' # choose one of the leading causes from the list above eg. ALE(Average Life Expectancy)
colname=ToBePredicted[ToBePredicted.index(strval)]
colname,top5,mlmodelx=modelrun(colname,mlmodel,X1)
```

```
------------------Data Specs-------------------------
Amount of Training Data 2197
Amount of Training Labels Data 2197
Amount of Testing Data 942
Amount of Testing Labels Data 942
-----------------------------------------------------
Baseline : Train Root Mean Squared Error: 1.9876988568598943
RandomForest: Train Root Mean Squared Error: 0.3574332590237172


Baseline : Test Root Mean Squared Error: 2.024027642619827
Random Forest: Test Root Mean Squared Error: 0.9963986756134902


Feature ranking:
Attribute Predicted     ALE
Predictor Columns
 36           Poverty
47    No_HS_Diploma%
65          Under_18
42             Black
41             White
Name: cols, dtype: object
```



8. Modelling & Inferencing

```
# choose one of the leading causes from the list above eg. ALE(Average Life Expectancy)
# the baseline model predicts the median always
```

In [10]:
```
###### Better Model Construction After Grid Search##############################
strval='ALE' # choose one of the leading causes from the list above eg. ALE(Average Life Expectancy)
colname=ToBePredicted[ToBePredicted.index(strval)]
mlmodel=mlmodel_bak
#grisearchdriver(colname,mlmodel,X1)
```

Prediction Tool: Leading Causes of Death. (Example/Mock County Sample)

In [11]:
```
pred_df=custom_df[X1].iloc[1:2]
pred_df['Under_18']=0 # % of women who give birth and are below 18 - Change this value or any of the values in the sample
pred_df
```

Out[11]:

	No_Exercise	Few_Fruit_Veg	Obesity	High_Blood_Pres	Smoker	Uninsured	Elderly_Medicare	Disabled_Medicare	Prim_Care_Phys_Rate	Dentist_R
1	27.2	76.2	23.6	30.5	24.6	19798.0	22635.0	3839.0	67.0	3

1 rows × 78 columns

In [12]:
```
print("Average Life Expectancy", mlmodelx.predict(pred_df))# average life expectancy
```

```
Average Life Expectancy [77.0624]
```

In [24]:
```
%%capture
os.chdir(codepath)
%run Modelling_Phase_2.py #Including All Models & Weights
```

In [25]:
```
ToBePredicted=['ALE','Diabetes','Lung_Cancer','Brst_Cancer','Col_Cancer','MVA','Stroke', 'Suicide','CHD']
pred_df=custom_df[X1].iloc[1:2]
pred_df['Under_18']=12
# % of women who give birth and are below 18, Change this value or any attribute's value to see change in leading causes c
pred_df#non-null values
```

Out[25]:

	No_Exercise	Few_Fruit_Veg	Obesity	High_Blood_Pres	Smoker	Uninsured	Elderly_Medicare	Disabled_Medicare	Prim_Care_Phys_Rate	Dentist_R
1	27.2	76.2	23.6	30.5	24.6	19798.0	22635.0	3839.0	67.0	3

1 rows × 78 columns

In [26]:
```
for i in range(0,len(ToBePredicted)):
    print(str(list1[i][1]), list1[i][0].predict(pred_df))# average life expectancy
```

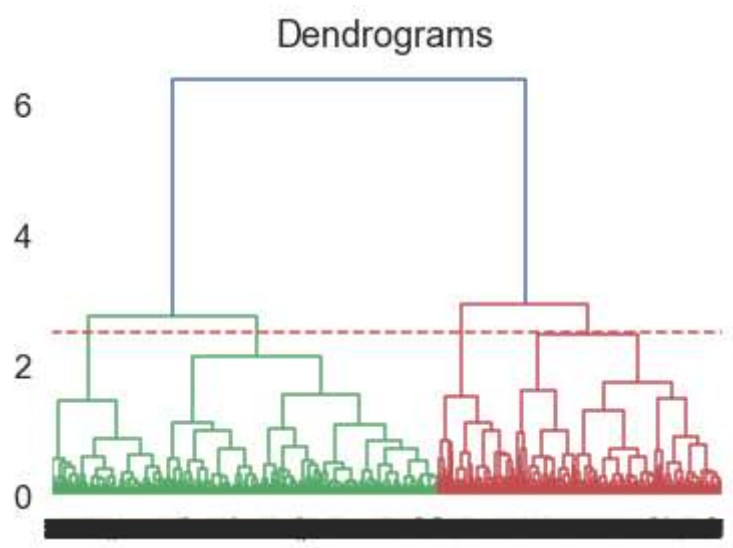
```
ALE [76.2391]
Diabetes [7.2715]
Lung_Cancer [62.1182]
Brst_Cancer [24.1394]
Col_Cancer [18.857]
MVA [22.6834]
Stroke [51.2348]
Suicide [14.3503]
CHD [165.9681]
```

Clustering of Counties based on Leading Causes of Death

(This will help local health agencies to know which counties are at a higher risk and to give more resources and attention to them)

In [29]:
```
os.chdir(codepath)
%run Modelling_Phase_3.py
plt.figure(figsize=(6, 4))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(data_scaled, method='ward'))
plt.axhline(y=2.5, color='r', linestyle='--')
```

Out[29]: <matplotlib.lines.Line2D at 0x2102b6bf2c8>



In [31]:
```
print("Counties in Cluster 3 are at the higher risk - Low ALE, Higher Cancer ")
result.groupby(['Cluster']).mean()
```

Counties in Cluster 3 are at the higher risk - Low ALE, Higher Cancer

Out[31]:

Cluster	ALE	Diabetes	Lung_Cancer	Lung_Cancer%	Brst_Cancer%	Brst_Cancer	Col_Cancer	Col_Cancer%	MVA	Stroke	Suicide
0	76.576532	7.523022	57.549842	0.004098	0.001348	26.415301	20.848420	0.001281	27.125486	64.237578	13.599494
1	76.304914	7.972392	56.970024	0.003503	0.001325	26.246221	21.574378	0.001289	24.650083	65.438769	12.986207
2	77.795238	6.301338	51.281707	0.003916	0.001049	25.638182	19.959091	0.001268	21.493770	58.599401	14.832014
3	74.932030	8.925163	69.031895	0.004598	0.001445	26.740486	22.544574	0.001338	29.049333	60.038606	13.961555

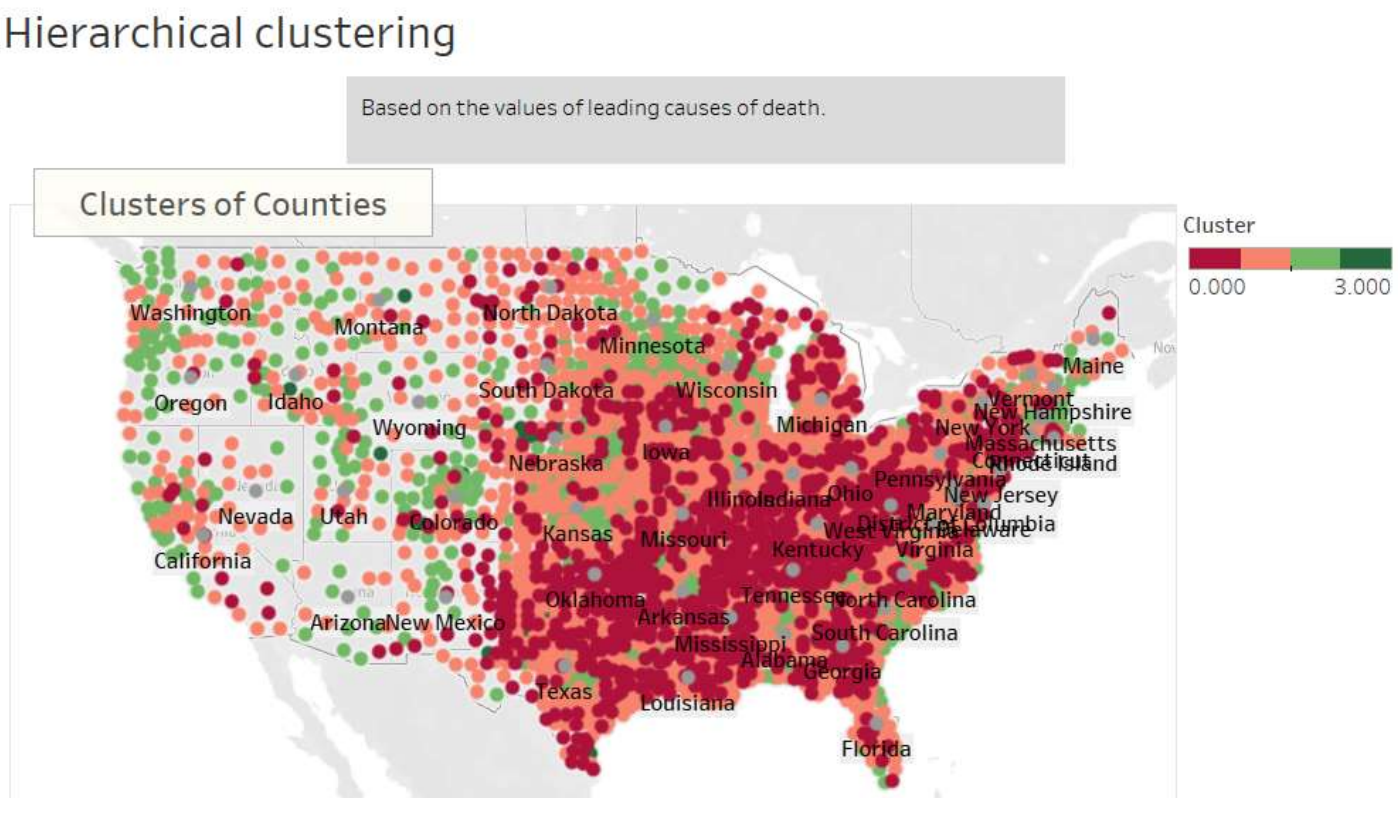
In [37]:
```
os.chdir(codepath)
%run Modelling_Phase_4.py
print("Few Counties belonging to cluster 3")
temp=tempdf[['CHSI_County_Name_x', 'CHSI_State_Name_x', 'ALE', 'Diabetes','Lung_Cancer', 'Brst_Cancer','Col_Cancer',  'Str
temp.head()
```

Few Counties belonging to cluster 3

Out[37]:

	CHSI_County_Name_x	CHSI_State_Name_x	ALE	Diabetes	Lung_Cancer	Brst_Cancer	Col_Cancer	Stroke	Suicide	CHD	Cluster
26	Escambia	Alabama	73.2	11.4	67.1	32.3	14.4	72.0	11.0	221.2	3
113	Boone	Arkansas	76.9	6.7	67.4	20.7	21.8	68.3	13.8	221.1	3
117	Chicot	Arkansas	73.9	15.6	66.0	29.4	26.7	76.1	9.8	235.1	3
118	Clark	Arkansas	75.5	6.3	61.5	27.8	21.6	74.9	10.0	230.8	3
119	Clay	Arkansas	75.0	10.2	84.8	20.2	26.7	68.5	13.1	282.9	3

In [86]:
```
%%HTML
<img src="outputcluster2.png" style="width:700px;height:400px;">
```



9. Reflection, Results & Progress

Observations : Exploratory Data Analysis:

1. Positive correlation of poverty and unemployment.
2. Negative correlation of poverty and population density.
3. No relationship between poverty and depression observed.
4. Poverty and homicide % of a county positively correlated.
5. Poverty and premature%(premature births%) of a county positively correlated.
6. Poverty and unmarried(unmarried women% who give birth) positively correlated.
7. Positive correlation between poverty and No High School Diploma Percentages.
8. No High school diploma%, Drug Use, Unemployment rate, Major depression positvely correlated.
9. Lung Cancer and Smoker% Positively correlated.
10. Obesity, High Blood Pressure and Diabetes Positively Correlated.
11. No Excercsie positively correlated with obesity, high blood pressure, diabetes.
12. Strong positive correlation between population density and depression.
13. Population size and E.Colli, Salmonella and Shigella Correlated (Hygiene Related Diseases)
14. Depression & Drug Use Positive Correlation
15. Povery & Number of Deaths Positive Correlation
16. Poverty & Average Life Expectancy Negative Correlation
17. UnInsured People Vs Number of Deaths Positive Correlation
18. Depression Vs Suicide Rate Positive Correlation
19. Heart Disease Vs Obesity Positive Correlation
20. Obesity and Smoker% Positively Correlated.

Results: Random Forest Regressor(9 Models) for each leading cause of death & Variable Importance (determining attribute for that cause):

1.Average Life Expectancy Train MSE=0.361304375017007 Test MSE=0.977328605007327
1.Poverty, 2.No_HS_Diploma, 3.Under_18, 4.Black, 5.Over_40
Eg. Main factors affecting ALE are poverty, no high school diploma, % of women under 18 getting pregnant etc.

---

2. Diabetes Train MSE=0.814102003689945 Test MSE=2.29140400291951
1.Obesity, 2.No_HS_Diploma, 3.Unmarried, 4.Poverty, 5.Recent_Drug_Use%

---

3. Lung_Cancer: Train MSE=3.35822064794277 Test MSE=8.40868361996625
1. Smoker, 2. NO_HS_Diploma, 3. Poverty, 4. Major_Depression, 5. Sev_Work_Disabled%

4.Breast_Cancer: Train MSE=1.96858600852315 Test MSE=5.09185013278892
1.Major_Depression, 2.Recent_Drug_Use%, 3.Black, 4.Unemployed%, 5.Population_Size

---

5.Colon_Cancer Train MSE=1.40279823772743 Test MSE=3.99720364350038
1.Unmarried, 2.Hispanic, 3.No_Exercise, 4.Smoker, 5.Major_Depression

---

6.Motor Vehicle Injuries: Train MSE=2.2402199328038 Test MSE=5.96104038324399
1.Under_18, 2.Population_Density, 3.Recent_Drug_Use%, 4.Asian, 5.No_HS_Diploma

---

7.Heart Stroke Train MSE=4.58634474894955 Test MSE=13.4173059174712
1.Black, 2.High_Blood_Pres, 3.No_HS_Diploma, 4.Premature, 5.Mammogram

---

8. Suicide: Train MSE=1.15818004036267 Test MSE=3.12497294467384
1.Population_Density, 2.Unemployment, 3.Recent_Drug_Use%, 4.Major_Depression, 5.Under_18

---

9. Coronary Heart Disease: Train MSE=12.4806249718433 Test MSE=34.9709746818494
1.No_Excercise, 2.No_HS_Diploma, 3.Smoker, 4.Unemployment, 5.Major_Depression

---

GITHUB: cs418-s20-cs418_spring20_datacrunchers_HSI/Project_Report_08_04/Project Report_08_04.ipynb