CS 5010 Project:
Chronic Disease Health Indicators And Impact of Social factors: Analysis on Counties in Virginia

Developed by:-
Akshat Verma, av2zf
Paul M. Cherian, pmc6dn
Prateek Agrawal, pa7sb

INDEX:

Introduction:

Project Scenario:

Chronic Diseases are the leading cause of mortality in the world accounting for over 60% of all deaths worldwide and around 70% in the United States.For our project we have chosen to compare death rates due to 4 chronic diseases at county level in the state of Virginia.

- *Cancer*
- *Respiratory Diseases*
- *Heart Diseases*
- *Diabetes*

We also seek to identify social factors which may affect health indicators. The various social indicators which we have used in our analysis are:

- Median Household Income
- Long term Care Hospital Admission
- Binge Drinking
- Smoking
- Primary Care
- No Insurance
- Obesity
- Unemployed persons

Objective:

To collect health data from various sources, perform a comparative analysis of health indicators pertaining to chronic diseases for various counties in the state of Virginia and to find interesting insights on the impact of social indicators on the mortality rates due to different chronic conditions

2.1 Queries Answered

Comparative analysis of the following indicators for counties in the state of Virginia:

1. Deaths due to Cancer.
2. Deaths due to Health Diseases.
3. Deaths due to Respiratory Diseases.
4. Deaths due to Diabetes.
5. Identifying the Best and the Worst performing Counties for each health Indicator

Predictive Analysis of all counties in Virginia state

6. Identifying the social factors which affect each health indicator.

Data Source:

We scraped data from Health Indicators Warehouse: http://www.healthindicators.gov/.

Health Indicator warehouse which is maintained by Centre of Disease Control, houses exhaustive data on various health and social indicators at county level for USA. The site also provides a REST API for retrieving data.

Methodology:

The whole applications was built in 5 Phases:

Phase 1: Data Collection
Phase 2: Data Cleaning
Phase 3: Descriptive Analysis
Phase 4: Predictive Analysis
Phase 5: Displaying Information in Chloropleth Maps

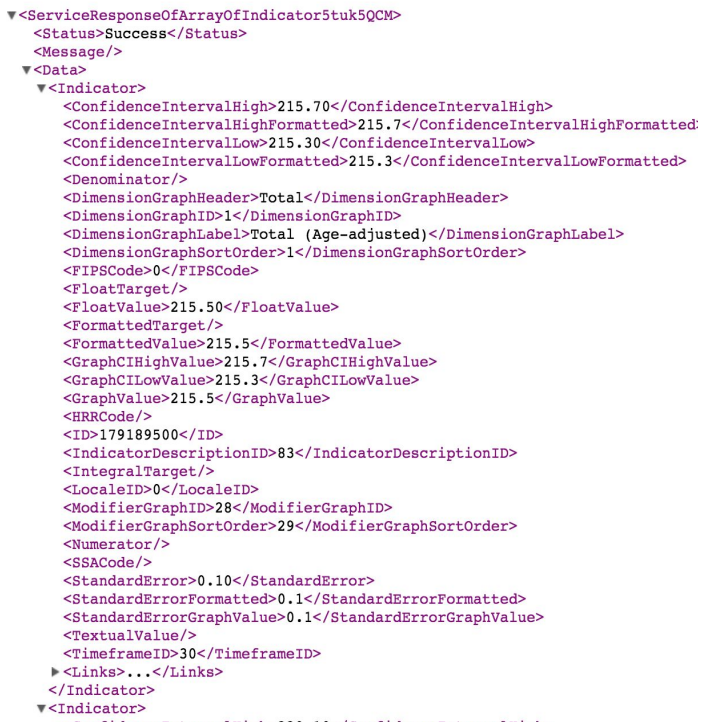
---

*Phase 1: **Data Collection:***

We have used the RESTful API provided by the Health Indicator Warehouse website to extract the required data. We used the 'requests' package in Python to interact with the API . To get the indicators of our interest, we invoked the 'Indicator Description' service of the API and specified the Indicator Description ID's in the request.

Response:

The response from the service is in the form of XML. A sample response is shown below:



```
▼<ServiceResponseOfArrayOfIndicator5tuk5QCM>
   <Status>Success</Status>
   <Message/>
  ▼<Data>
   ▼<Indicator>
      <ConfidenceIntervalHigh>215.70</ConfidenceIntervalHigh>
      <ConfidenceIntervalHighFormatted>215.7</ConfidenceIntervalHighFormatted
      <ConfidenceIntervalLow>215.30</ConfidenceIntervalLow>
      <ConfidenceIntervalLowFormatted>215.3</ConfidenceIntervalLowFormatted>
      <Denominator/>
      <DimensionGraphHeader>Total</DimensionGraphHeader>
      <DimensionGraphID>1</DimensionGraphID>
      <DimensionGraphLabel>Total (Age-adjusted)</DimensionGraphLabel>
      <DimensionGraphSortOrder>1</DimensionGraphSortOrder>
      <FIPSCode>0</FIPSCode>
      <FloatTarget/>
      <FloatValue>215.50</FloatValue>
      <FormattedTarget/>
      <FormattedValue>215.5</FormattedValue>
      <GraphCIHighValue>215.7</GraphCIHighValue>
      <GraphCILowValue>215.3</GraphCILowValue>
      <GraphValue>215.5</GraphValue>
      <HRRCode/>
      <ID>179189500</ID>
      <IndicatorDescriptionID>83</IndicatorDescriptionID>
      <IntegralTarget/>
      <LocaleID>0</LocaleID>
      <ModifierGraphID>28</ModifierGraphID>
      <ModifierGraphSortOrder>29</ModifierGraphSortOrder>
      <Numerator/>
      <SSACode/>
      <StandardError>0.10</StandardError>
      <StandardErrorFormatted>0.1</StandardErrorFormatted>
      <StandardErrorGraphValue>0.1</StandardErrorGraphValue>
      <TextualValue/>
      <TimeframeID>30</TimeframeID>
     ▶<Links>...</Links>
    </Indicator>
   ▼<Indicator>
```

---

*Phase 2: **Data Cleaning:***

To extract the required data from the above file, we have used the FIPS code as key, which is a 5 digit code for each county. As we are creating this application for the State of Virginia whose code is '51' therefore, all the FIPS code for all the counties in Virginia will start with 51.

example of FIPS code are shown below:

ACCOMACK 51001

ALBEMARLE 51003

After obtaining the XML file, we used XML 'ElementTree' package of python to parse the XML response and convert it in CSV file as shown below:

```
Indicator_Type,County,Value
Heart_Disease_Deaths,51010,3345
Median_Household_Income,51010,42221
No_Insurance,51010,22.6
Primary_Care,51059,122.7
Smoking,51095,11.1
Obesity,51041,27.7
College_Degrees,51003,52.2
Heart_Disease_Deaths,51010,3345
Median_Household_Income,51010
No_Insurance,51010,22.6
Primary_Care,51059,122.7
Smoking,51095,11.1
```

The above CSV data is imported into pandas dataframe and transformed, so that we get the value of all the indicators per county in one row.

Phase 3: ***Descriptive Analysis:***

To categorise the counties, we have divided the total mortality rates for each health indicator into 3 quantiles(0-33 percentile, 33-66 percentile and 66+ percentile).

- **The Top Performing Counties will be the one that lie the range of 1st Quantile.**
- **The moderate Performing Counties will be present in the 2nd Quantile**
- **The Worst Performing Counties will have the deaths in the range of 3rd Quantile.**

After Sorting the values of the counties we choose the best 3 and worst 3 performing counties for each health indicator.

Phase 4: ***Predictive Analysis:***

To predict the impact of Social Indicators on various Health Indicators we have used Statistical Models in Python. Since version 0.5.0 of statsmodels, we can use R-style formulas together with pandas data frames to fit our models Thus importing

1. statsmodels.formula.api
2. import numpy as np

After normalising all the values of social indicator from 0 to 1, we used the "glm" stats model package to carry out the regression and select the variables which satisfy p value threshold of 0.1. We used the coefficients in the normalized linear regression model to determine the relative importance of the variables.

Result:

The result is a list of list of tuples with the name and relative importance of the most influential social indicators for every health indicator.

Phase 5: ***Displaying Information in Chloropleth Maps:***

"Bokeh" Visualization package is used which consist of various graphical options for representation. Bokeh offer the capability to plot streaming data,through bokeh server. It also has functionality to add user interactions.

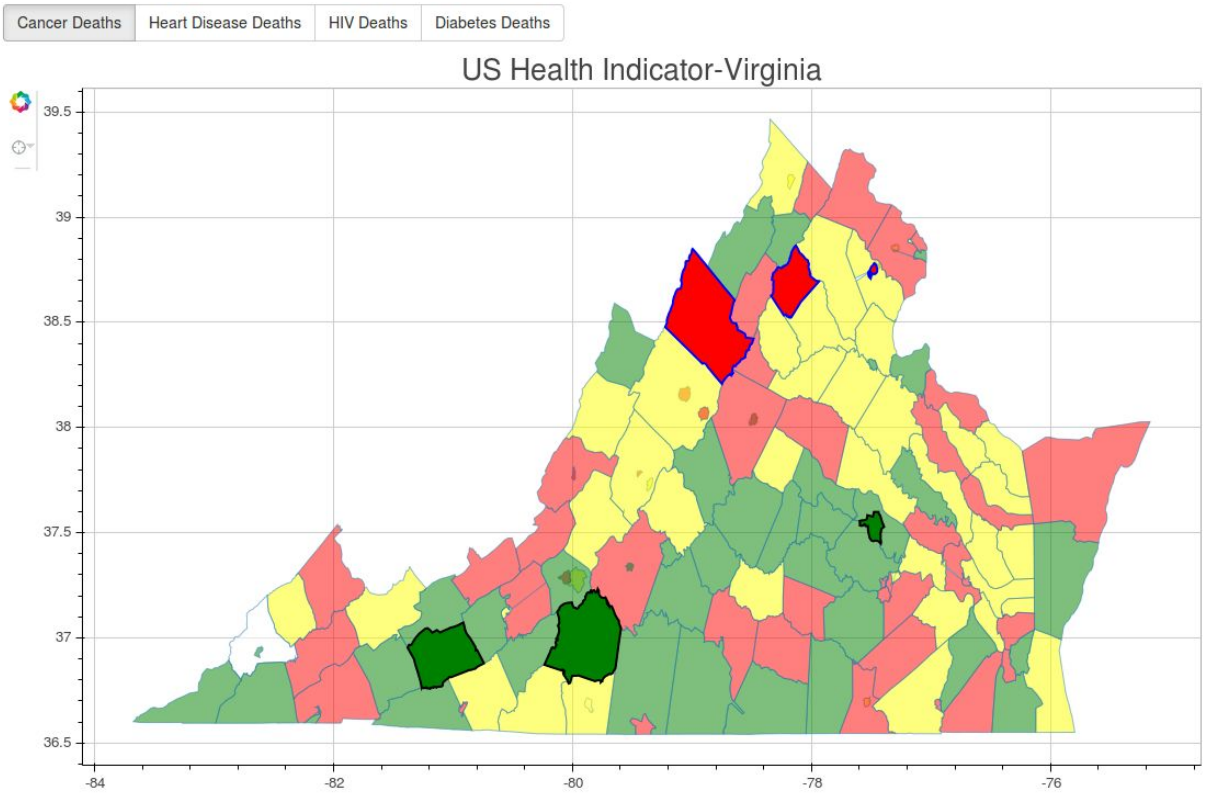
Chloropleth Maps:

Chloropleth maps are an excellent representation to show data that varies across geographical boundaries such as population density.

The whole method of developing a map is divided in steps:

1. **Creating a list of list for all FIPS values of each county**
2. **Taking the X-Y map of Virginia State where all the longitudes and latitudes values of counties are used.**
3. **Removing all missing values of Longitudes and Latitudes as in BOKEH, the missing values converts the list into a string.**
4. **As done in the Descriptive Analysis we have already divided our values in 3 quantiles**

5. Counties are then coloured according to the quantile they lie in.



The above chloropleth Map highlights the best and worst performing counties for a particular health indicator.

Conclusions and Trends found

After analysis we found that the following results for each health indicator:

S.No	***Health Indicator***	***Top Performing Counties***	***Worst Performing Counties***
1.	Deaths due to Cancer	Withe, Franklin, Richmond	Rockingham, Rappa Hannock, Manassas

2.	Deaths due to Respiratory	Withe, Franklin, Fredericksburg	Buchanan, Mathews, Hampton
3.	Deaths due to Heart Diseases	Withe, Franklin, Richmond	Charlotte, Giles, Galax
4.	Deaths due to Diabetes	Wythe, Culpeper, Bristol	Clarke, Greene, Waynesboro

Also, the key influential social Indicators affecting the health Indicators

S.No	***Health Indicator***	***Influential Social Indicator***
1.	Deaths due to Cancer	Median Household Income, Unemployed Persons
2.	Deaths due to Heart Diseases	Median Household Income, Unemployed Persons
3.	Death due to Respiratory Diseases	Median Household Income, No Insurance
4.	Deaths due to Diabetes	Unemployed Persons

Extra Credit Work in the Project

As mentioned in the Project Requirements, various extra requirement have been fulfilled in this project. i.e.

- *Web-scraping to obtain your data set instead of downloading a ready-made data set from a source.*

--The data collection phase is done by using REST API's rather than downloading ready made data, and then cleaning and formatting it is done by parsing the XML in 'eTree' package to usable format.

- *Have some user-interaction where the user may choose the kinds of queries to perform on the data. Retrieve/display only the appropriate result.*

-- Bokeh Visualization Package is used, which allows the user to choose the health indicator he/she wants to view and displays the graph in a unique chloropleth map format.

- *Use advanced queries or manipulate the data in another way (other data manipulation methods, etc..) and display the results. If you choose to do this, mention in your report how this goes beyond the basic/general queries you initially used.*

-- Data is Manipulated using statistical models. We have used 'statsmodel' package in python to fit generalized linear regression models to the data points contained in pandas data frames . We normalised the variable values from 0 to 1 and then employed the regression models to calculate relative importance of the variables.

By using such a methodology, we were able to analyse the impact of existing social factors on mortality due to chronic diseases. We calculated the quantitative impact of the social factors on different health indicators.

References

1. Health Indicators Warehouse: http://www.healthindicators.gov/
2. Bokeh package: http://bokeh.pydata.org/en/latest/