project-021:I523:Project:Health Consequences of Smoking

Tanmoy Choudhury
F16-DG-4015
Indiana University
Bloomington, IN
tdchoudh@iu.edu

Madhavi Pilli
F16-DG-4055
Indiana University
Bloomington, IN
mpilli@iu.edu

Vipin Singh
F16-DG-4071
Indiana University
Bloomington, IN
vipsingh@iu.edu

ABSTRACT

Tobacco usage is considered to be a leading cause of health consequences in the United States. As part of this project, we have analyzed few datasets like, tobacco consumption, tobacco related mortality rates. This data is compared against US population to derive inferences between mortality rates and tobacco use. Data has been visualized using python plots like seaborn[1] and correlation as well as Tableau[2] to generate insightful dashboards. We have also worked on a software installation process to install hadoop[3] deployment on a multi-node Chameleon cloud cluster. This deployment is done using cloudmesh[4] client and it internally uses the ansible[5] script for deployment.

1. INTRODUCTION

"Cigarette Smoking is Injurious to Health" - is the warning that is printed on every cigarette packet. It has been an effort to bring public awareness on the harmful effects of smoking and use of tobacco products. We all know that immunity is our body's natural way of fighting against several diseases. Cigarette smoking compromises our immunity and hence makes humans more susceptible to cancer, diabetes, cardio-vascular diseases, vision imparity, blood and reproductive system. As per betobaccofree.hhs.gov[6], It is known that nearly 480,000 people die prematurely in the United States each year due to smoking cigarettes or exposure to smoking.

2. RELATED WORKS

One of the most authoritative work in this subject is the 2014 Surgeon General's report[7]. It marks the 50th anniversary of the initial report published in 1964. The report analyzes historical perspective of smoking and its consequences based on the recent evidence and link them back to 1964 report. It further investigates into patterns of smoking and the impact of the tobacco controlled environment since 1964. At the end it provides a vision to outline future strategy to curb tobacco consumption in future.

3. SOFTWARE AND TOOLS

The following software and tools were used for this project. Project was developed and tested on MAC OS X 10.1

- Virtual box with Ubuntu 16.04
- Python 2.7
- Cloudmesh client
- Hadoop
- Tableau 10.1.1

4. DATA ACQUISITION

Behavioral Risk Factor Surveillance System hosted on Centers for Disease Control(CDC) and Prevention is a state based system that collects data about modifiable risk factors for chronic disease and other leading causes for death. Data has cigarette smoking status as well as cigarette smoking prevalence by demographics.

In this report we have used below datasets

- Behavioral Risk Factor data by state for the years 1996-2010[8]
- Behavioral Risk Factor data by state for the years 2011-2015[9]
- Smoking attributable mortality (SAM) dataset,2005-2009[10]
- Average US population by state,2000-2010[11]

Behavioral Risk Factor data sets was provided by Centers for Disease Control and Prevention (CDC) using state Tobacco Activities Tracking and Evaluation (STATE) System. The Smoking attributable mortality (SAM) dataset from Centers for Disease Control(CDC) provides average annual death caused by cigarette smoking from 2005 to 2009. These dataset was accessed using python program and processed in memory and only the cleansed data was stored under "/data" folder in project directory for visualization. Average US population by state wise data was acquired and a average population column was populated by doing average between 2000 and 2010 numbers. These data files are stored into "/data" folder under project directory for easy access as the information was in html format. Cleansed dataset was used to develop a correlation graph between tobacco usages and number of deaths.

4.1 Challenges

We have encountered several challenges on data acquisition. There was no public dataset from 1996-2015 for tobacco causes death. We tried to contact Centers for Disease Control and Prevention (CDC) to acquire some of the dataset citing academic reason which was unanswered. We were able to gather data needed for the initial analysis, however when looking for additional data to find correlations we noticed that availability of data was not easy and not free. We had two datasets on Centers for Disease and Control (CDC) for

the years 1995 to 2010 and then for 2011 and beyond. The measures for this data was not similar to easily merge these datasets. We had to analyze them separately.

5. DATA ANALYSIS

Behavioral Risk Factor data is analyzed to come up with criteria to merge the two datasets. Data in the file 2010 and above has few additional attributes compared to the file 1995-2010.

Once the data is fetched using python, it is collected into data frames. Further analysis was done to see what makes the data more consumable. Data was available as Mean or Percentage, however few records did not have data for mean and hence this needs to be filtered out.

Using iterative process, further look at the data shows us different cuts based on Race, Gender, Education level and Age groups.

There were different measures being captured in the dataset. We limit our analysis to current smoking.

Plan is to make the following sets of data for analysis of current smoking percentages across different age groups.

- Male/Female, All Ages
- Overall, All Ages
- Overall, 18-24 Years
- Overall, 25-44 Years
- Overall, 45-64 Years
- Overall, 65 Years and Up
- Overall, Races, All Ages
- Overall, Education level

6. DATA PRE-PROCESSING

Data prepossessing can be divided into two subsection.One is pre-processing for visualization and other is for correlation between tobacco usages and number of deaths.

6.1 Pre-processing for Visualization

Following pre-processing task has been performed

- Extract only the first four digits of the year and change the type to integer.
- Filter datasets with data value type of percentage.
- Merge the two datasets to create one data frame with data points from both datasets for all years.
- Filter dataset for All Races where the measure is Current Smoking.
- To process the data further, the data is split again into multiple datasets based on criteria decided above.

Cleansed dataset has the following attributes used for further analytics:

- Year - 1995 to 2015
- LocationAbbr - State
- Data Value - Smoking Percentage
- Gender - Male, Female, Overall
- Race - Races(White, Hispanic, African American etc.,)
- Age - Age groups
- Education - Education levels
- GeoLocation - Latitude and Longitude

6.2 Preprocessing for Correlation

For correlation we have used three datasets. These are Behavioral Risk Factor data by state for the years 1996-2010[8], Smoking attributable mortality (SAM) dataset,2005-2009[10] and Average US population by state,2000-2010[11]. We plan on using these datasets to prove a correlation between tobacco usages and death. The following pre-processing tasks were performed:

Behavioral Risk Factor data by state (1996-2010)

- Behavioral Risk Factor data by state has span from 1996-2010, however the Smoking attributable mortality (SAM) dataset was only available for 2005-2009. So for correlation we need to consider Behavioral Risk Factor data from 2005 to 2009 only. The following filter was applied to the data set
  - Year having 2005 till 2009
  - Race having “All Races”
  - Age having “All Ages”
  - Measure Desc as “Current Smoking”
  - Data Value type as “Percentage”
  - Gender as “Male” and “Female”
- The Behavioral Risk Factor data value is averaged based on sample size. A percentage was derived based on sample size and the dataset was fitted to match with user population data.
- As next step, only statewide tobacco usage based on gender was considered.
- A new column “AvgSmoker” was introduced by taking mean on “UsageValue”. This pre-processing is required as the Smoking attributable mortality (SAM) dataset has annual average death number to ensure these two data points are comparable.
- These three attributes “LocationAbbr”, “AvgSmoker” and “Gender” were considered for further processing.

Smoking attributable mortality (SAM) dataset (2005-2009)

- The following filter was applied to the data set
  - Gender as “Male” and “Female”
  - Measure Desc as “Average Annual Deaths”
  - Location Abbreviation not in “US”
- These three attributes “LocationAbbr”, “Data_Value” and “Gender” were considered for further processing.

Average US population by state,2000-2010

- The data file does not have state abbreviation. Using a pre populated state abbreviation dictionary the two char state abbreviation was populated.

- Only state abbreviation ("StateAbbr") and "AvgPopulation" was considered for further processing.

Once all the above processing is done, a data mashup is performed between the three dataset to create a single data file for correlation and saved into "/data" folder under project directory in gitlab.

7. DATA VISUALIZATION

Python provides us great visualization options. Using the Seaborn plot the following plots have been generated. Seaborn[1] is a python visualization library based on matplotlib. It provides a great interface for generating insightful graphs.

The following graphs were generated using python to answer research questions. Data_Value represented on y-axis for all plots is the smoking percentage.

1. The plot below shows data points for all states for each year vertically with an average of those shown in the horizontal lines. Year on x-axis and smoking percentage for all ages by gender on y-axis.

Figure 1: Tobacco consumption across multiple years for genders.

Figure 1 illustrates that tobacco consumption is consistently higher in "Male" compared to "Female". It also indicates that overall consumption across both genders has decreased over the years.

2. Plots showing tobacco consumption percentage for different Age groups to answer our research question, "How is tobacco consumption across different age groups?" are provided below

Figure 2: Tobacco consumption in age group 18-24.

Figure 2 indicates a sharp decline in tobacco consumption in the age groups 18-24 which is good. It also shows that the tobacco consumption is highest in this age group.

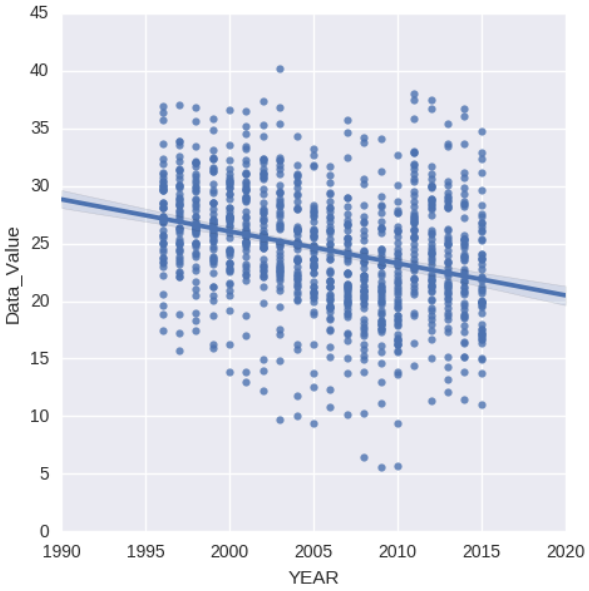

Figure 3: Tobacco consumption in age group 25-44

Figure 3 indicates that tobacco consumption in the age group 25-44 has steadily reduced over the years. The decrease percentage is little less compared to the age group 18-25.

Figure 4: Tobacco consumption in age group 45-64

Figure 4 similarly indicates a decline in tobacco consumption in the age group 45-64, however it is little less compared to the age groups 18-24 and 25-44.

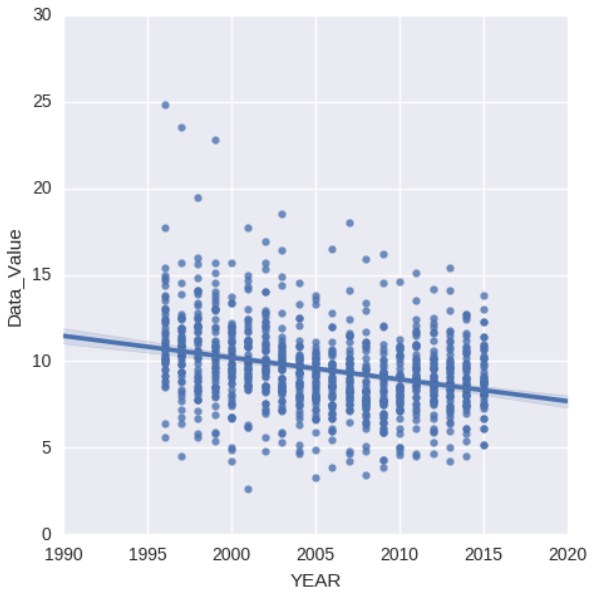

Figure 5: Tobacco consumption in age group 65 and up

Figure 5 shows that tobacco consumption is much less in the age group 65 and up

We wanted to put several of these visualizations together and hence researched for a tool that would provide us the capabilities to do so. This lead us to Tableau [2]. It was interesting to research the visualization capabilities within Tableau. It had great options to present the data we were trying to analyze.

Dashboard was a good way to present different visual representations of data in one page for easy comparison. We created two dashboards using Tableau.

3. Tobacco Consumption Dashboard showing different perspectives of data

Based on the visualizations we generated and assembled in the dashboard, we were able to make the following inferences. Our observations based on Figure 6 indicates,

- Smoking Trends across different age groups reaffirming our prior observations.
- Smoking Trends across genders
- Another interesting perspective is the smoking trends across different races. This shows that American Indian/Alaska Native smoke the highest and Asian/Pacific Islander smoke the lowest in the United States. Rest of the races are at the same level with not too much of difference in tobacco consumption.
- Smoking trends based on education levels is another interesting graph. This indicates that the higher the education the lower the tobacco consumption is. This probably explains why smoking trends have steadily gone down over the years.

4. An interactive Visualization showing tobacco consumption percentage across all years starting 1995 to 2010 with a slider for year. The visualizations are presented through a tableau reader, twbx[12] file. A few snapshots of the visualization across every 5 years is shown below.

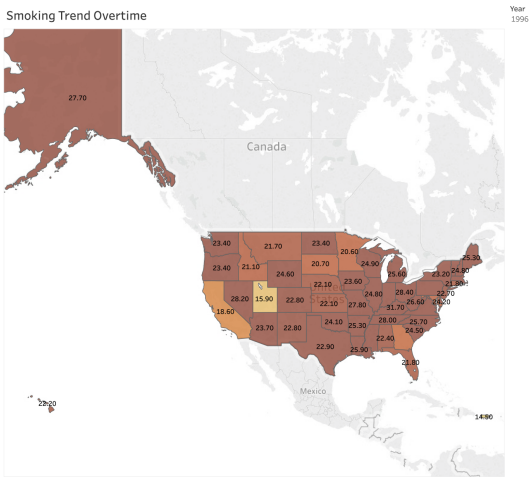

Figure 7: Tobacco consumption in 1996

Figure 7 is snapshot from the interactive visualization that shows the tobacco consumption for all the states across different years. This snapshot is for the year 1996. We see that tobacco consumption is lowest in Puerto Rico, Utah and California and highest in Kentucky, Indiana, Ohio, Nevada and Tennessee.

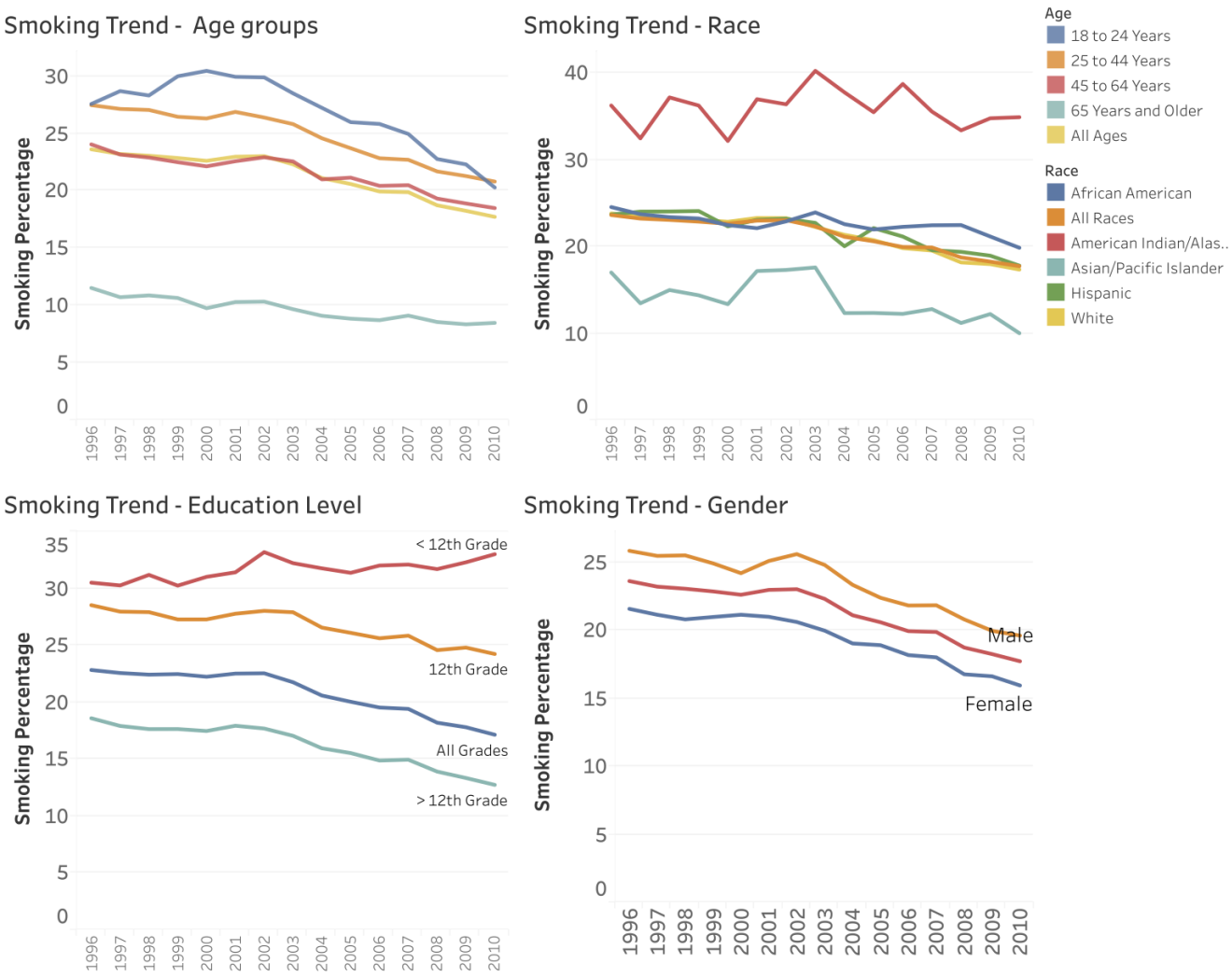

Figure 6: Tobacco consumption Dashboard

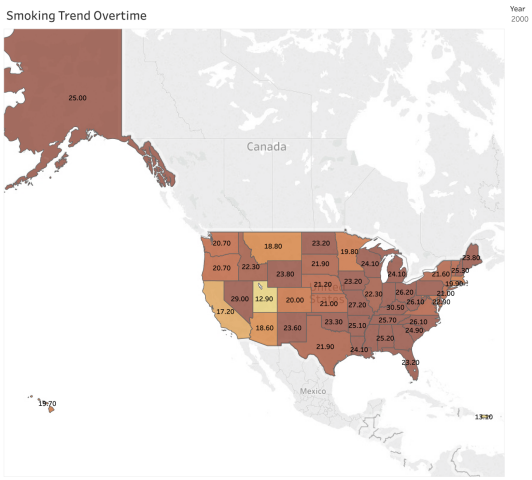

Figure 8: Tobacco consumption in 2000

Figure 8 is snapshot from the interactive visualization that shows the tobacco consumption for the year 2000. We see that tobacco consumption is lowest in Utah, Puerto Rico and California and highest in Kentucky, Nevada, Missouri, Ohio and Indiana.

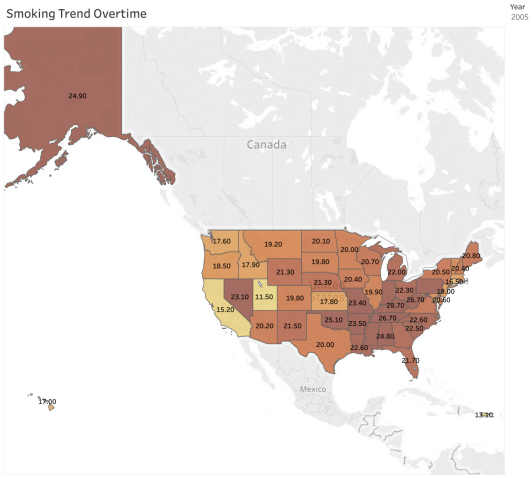

Figure 9: Tobacco consumption in 2005

Figure 9 is snapshot from the interactive visualization that shows the tobacco consumption for the year 2005. We see that tobacco consumption is lowest in Virgin Islands, Utah, Puerto Rico and California and highest in Kentucky, Indiana, West Virginia, Tennesse, Olkahoma.

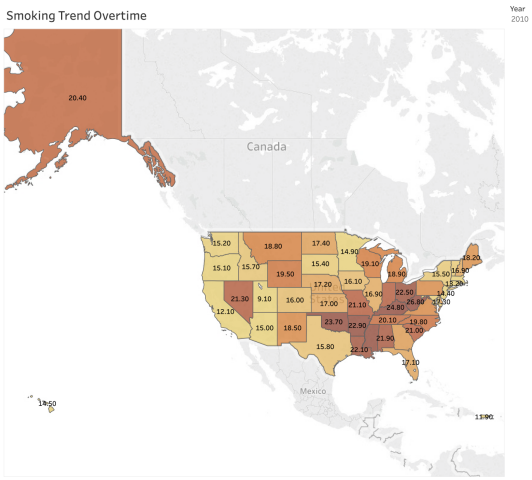

Figure 10: Tobacco consumption in 2010

Figure 10 is snapshot from the interactive visualization that shows the tobacco consumption for the year 2010. We see that tobacco consumption is lowest in Virgin Islands, Utah, Puerto Rico and California and highest in West Virginia, Guam, Kentucky and Olkahoma.

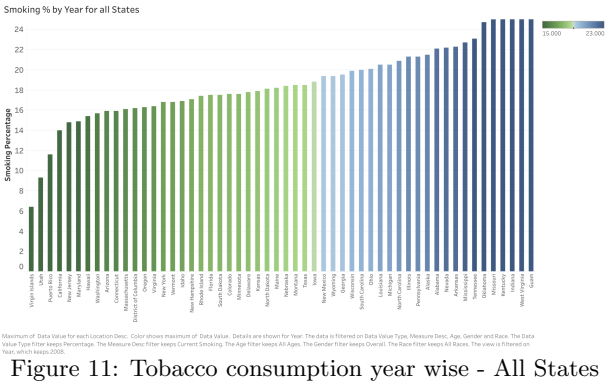

Figure 11: Tobacco consumption year wise - All States

Figure 11 is snapshot from the interactive visualization that shows the tobacco consumption for all states across different years. This graph can be viewed using Tableau reader. There is a slider that gives ability to slide across years to see tobacco consumption across all states in the United States. If we observe carefully we see that tobacco consumption continues to be low in states like Utah, Puerto Rico and California and high in states like Kentucky, West Virginia, Ohio, Indiana and Olkahoma.

8. CORRELATION

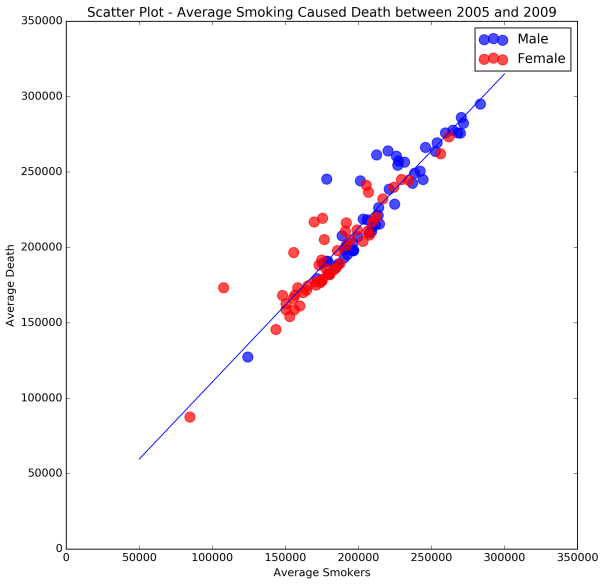

Figure 12: Correlation between tobacco consumption and death

Figure 12 shows the positive gradient between average smoker and deaths. It shows the linear association having a strong positive correlation. We do observe some outliers on the plot, however they are not too significant. The scatter plot was generated using *pyplot* from *matplotlib* library.

9. HADOOP

This section describes the Multinode hadoop cluster and its deployment on a chameleon cloud cluster.

9.1 Multinode Hadoop Cluster

Hadoop Cluster is a special type of computer cluster. It is designed for storing and analyzing huge amount of data - both structured and unstructured. Hadoop Cluster runs Hadoop's Open distributed processing software.

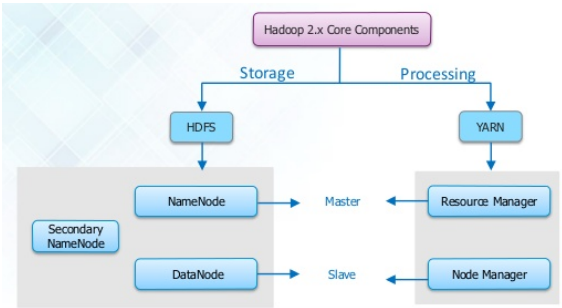

Figure 13: Infographic[3] created by Edureka

Figure 13 shows the Hadoop 2.x core components. Hadoop solves the problem of Storage and Processing the huge amount of data sets. The Storage is HDFS and Processing is Mapreduce or YARN in Hadoop Version 2. YARN does the resource negotiation.

Master component of HDFS is NameNode and Slave component is Data Node. Master component of YARN is Resource Manager and Slave component is Node Manager. NameNode and Resource Manager are also called Master Daemons, similarly the DataNode and NodeManager are called Slave Daemons.

These are the steps to create multi node hadoop cluster

- Download and Install Java on all the systems (Both Master and Slave)
- Specify the IP address of each system followed by their host names in hosts file inside etc folder of each system
- Configure Hadoop Configuration files(core-site,hdfs-site, mapred-site,yarn-site)
- Edit Slaves File on Master Node
- Format NameNode and start all hadoop services.
- Check Live nodes on Hadoop NameNode UI.

Our purpose is to learn Hadoop eco system and how to use data analytic software (ex. Python) on the Hadoop cluster.

9.2 Deployment of Hadoop

The deployment of hadoop on a multinode cluster is done by using the cloudmesh client, virtual env, chameleon cloud on a virtual box with ubuntu installed. The git repositories are downloaded from github. The SSH keys were generated and uploaded to github and chameleon. The ansible script is used to deploy the hadoop on the nodes in a cluster on chameleon cloud. A mapreduce example is executed to verify the results.

- Ubuntu Virtual machine - An Ubuntu virtual machine is used to deploy hadoop cluster.
- Cloudmesh Client - Cloudmesh client[4] allows to easily manage virtual machines, containers, HPC tasks, through a convenient client and API. Hence cloudmesh is not only a multi-cloud, but a multi-hpc environment that allows also to use container technologies.
- Chameleon Cloud - Chameleon Cloud[13] is a a large-scale platform to the open research community allowing them explore transformative concepts in deeply programmable cloud services, design, and core technologies. Chameleon will allow users to explore problems ranging from the creation of Software as a Service to kernel support for virtualization.
- Virtualenv - Virtual Environment on Python is private environment where the software can be installed and tested. It is a good environment to do lot of hit and trial. The benefit of virtual environment is that it does not affect the software installed on the base machine.
- Github - Github[14] is a publicly available, free service which requires all code (unless you have a paid account) be made open. Anyone can see code you push to GitHub and offer suggestions for improvement. GitHub currently hosts the source code for tens of thousands of open source projects.We used github to download big data open stack to be used in the hadoop deployment.

- SSH Keys - SSH keys[15] serve as a means of identifying yourself to an SSH server using public-key cryptography and challenge-response authentication. One immediate advantage this method has over traditional password authentication is that you can be authenticated by the server without ever having to send your password over the network. Anyone eavesdropping on your connection will not be able to intercept and crack your password because it is never actually transmitted. Additionally, using SSH keys for authentication virtually eliminates the risk posed by brute-force password attacks by drastically reducing the chances of the attacker correctly guessing the proper credential. We have used SSH keys to transmit, upload and download the files from github, gitlab and virtual nodes.

- Ansible - Ansible[5] is a free-software platform for configuring and managing computers. This platform combines multi-node software deployment, adhoc task execution, and configuration management.It manages nodes over SSH or over PowerShell.Modules work over JSON and standard output and can be written in any programming language. The system uses YAML to express reusable descriptions of systems. The cloudmesh uses the ansible script to install the Hadoop on all the nodes. The nodes in a cluster are assigned numbers in sequence. The node with the lowest number is the master node and will be used to run the map reduce examples.

- Mapreduce - MapReduce[16] is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. We downloaded few mapreduce programs from github and ran by doing an ssh to minimum number assigned node in a cluster, since it is a master node.

10. CONCLUSION

Our analysis indicates a steady decline in overall tobacco consumption. It also indicates that as more and more population get educated the trend is downward. It would be interesting to do a further analysis to see the education rates for these states and correlate them to the tobacco consumption patterns. This project provided us enough opportunities to learn python, Hadoop, cloud infrastructure as well as tableau for visualizations.

11. REFERENCES

[1] M. Waskom, "Seaborn: statistical data visualization," 2012-2015. [Online]. Available: "http://seaborn.pydata.org/"

[2] Tableau, "Analytics that work the way you think," 2003-2016. [Online]. Available: "http://www.tableau.com"

[3] Vardhan, "Setting up a multi node cluster in hadoop 2.x," 2015. [Online]. Available: "http://cdn.edureka.co/blog/wp-content/uploads/2015/11/1Hadoop-2-1.png"

[4] G. von Laszewski, "Installing cloudmesh client," 2016. [Online]. Available: "https://github.com/cloudmesh/client"

[5] Wikipedia, "Ansible (software)," Web Page, 2016. [Online]. Available: "https://en.wikipedia.org/wiki/Ansible_$software$"

[6] U. D. of Health & Human Services, "Betobaccofree.gov," Web Page, October 2016. [Online]. Available: http://betobaccofree.hhs.gov/

[7] U. S. P. H. S. O. of the Surgeon General, *The Health Consequences of Smoking–50 Years of Progress: A Report of the Surgeon General : Executive Summary*, 2014. [Online]. Available: "https://books.google.com/books?id=7liUDAEACAAJ"

[8] C. for Disease Control and P. (CDC)., "Behavioral risk factor data: Tobacco use (2010 and prior)," Web Page. [Online]. Available: https://chronicdata.cdc.gov/Survey-Data/Behavioral-Risk-Factor-Data-Tobacco-Use-2010-And-P/fpp2-pp25

[9] ——, "Behavioral risk factor data: Tobacco use (2011 to present)," Web Page. [Online]. Available: "https://data.cdc.gov/api/views/wsas-xwh5"

[10] ——, "Smoking-attributable mortality, morbidity, and economic costs (sammec) - smoking-attributable mortality (sam)," Web Page. [Online]. Available: https://chronicdata.cdc.gov/Health-Consequences-and-Costs/Smoking-Attributable-Mortality-Morbidity-and-Econo/4yyu-3s69

[11] ——, "U.s. population by state, 1790 to 2015," Web Page. [Online]. Available: http://www.infoplease.com/ipa/A0004986.html

[12] T. R. File, "Tableau packaged workbook file," 2015. [Online]. Available: "http://www.reviversoft.com/file-extensions/twbx"

[13] C. Cloud, "About," 2016. [Online]. Available: "https://www.chameleoncloud.org/about/chameleon/" abstract = "This is the information about chameleon cloud"

[14] Wikipedia, ""github"," 2016. [Online]. Available: "https://en.wikipedia.org/wiki/GitHub"

[15] Wiki, "Ssh keys," Web Page, 2016. [Online]. Available: "https://wiki.archlinux.org/index.php/SSH_keys"

[16] Wikipedia, "Mapreduce," Web Page, 2016. [Online]. Available: "https://en.wikipedia.org/wiki/MapReduce"

12. APPENDIX

12.1 Tableau

Tableau[2] provides the ability to develop insightful and interactive visualizations using data. To start with we downloaded trial version of Tableau Desktop Professional Edition 10.1.1.

Pre-processed data has been exported as an excel to be used in Tableau for few interactive, insightful visualizations and dashboards. This will be considered as the data source for our visualizations.

Here are some of the visualizations that provided good insights in to the tobacco consumption data.

Filter data based on measures required to produce the following visualizations:

- Tobacco consumption by age
- Tobacco consumption by gender
- Tobacco consumption by education level
- Tobacco consumption by race

Figure 14 illustrates how to make selections to choose data that is needed for visualizations.

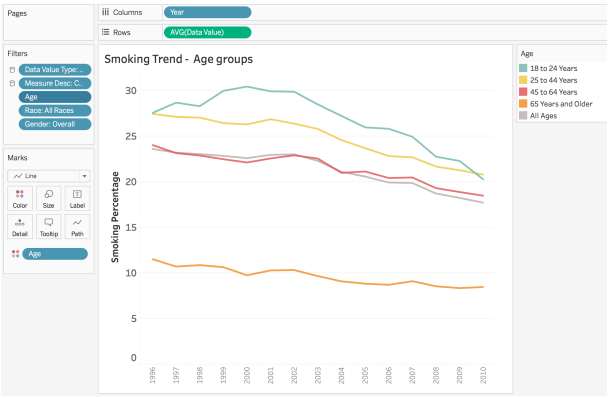

Figure 14: Illustrating Tableau worksheet selections for tobacco consumption by age

To be able to do a side by side comparison of these visualizations we could put together a dashboard. Dashboard provides ability to merge visualizations into a single board.

Here is a sample dashboard selection using Tableau[2] for the four different graphs generated.

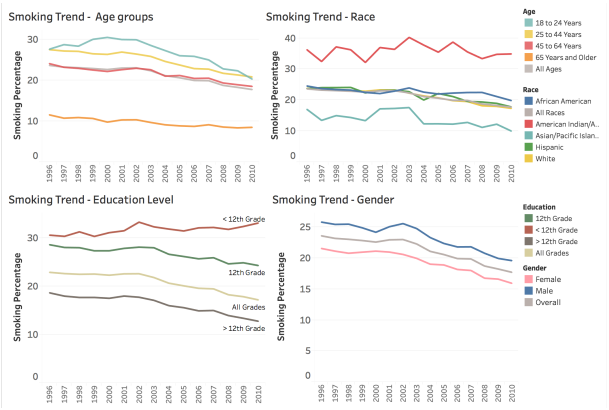

Figure 15: Illustrating Tableau dashboard merging 4 graphs indicating different perspectives to visualize data

Another interesting observation was to see how tobacco consumption was across different states through the years on one visualization. We used interactive visualization capability of a choropleath map using Tableau[2].

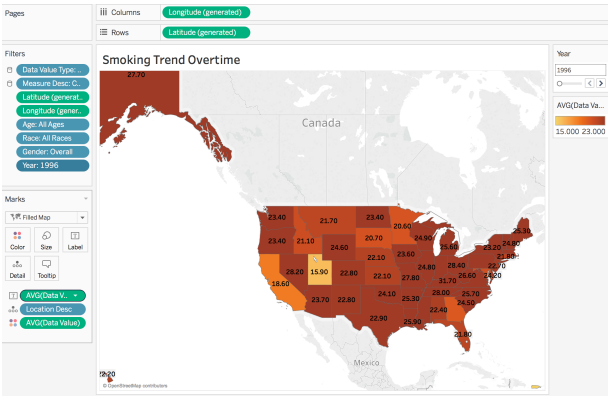

Figure 16: Illustrating Tableau interactive visualization indicating tobacco consumption across the states over the years 1995-2010 with a slider for years

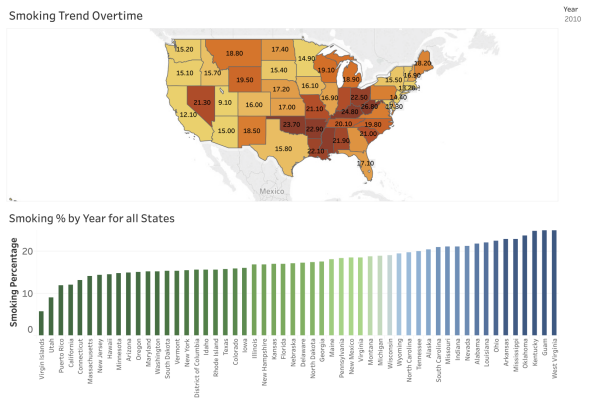

Figure 17: Illustrating Tableau interactive dashboard merging 2 visualizations indicating tobacco consumption across the states over the years 1995-2010 with one slider for years controlling both the visualizations

12.2 How to view Tableau visualizations

Tableau code is packaged as a twbx [12] and is checked into git under code. To be able to view data and make changes to the tableau visualizations a tableau desktop or online version 10.1.1 needs to be installed. However if the need is to be able to go through the visualizations and tell or understand the story, a tableau reader would be sufficient. For this project tableau reader 10.1.1 would be required as the visualizations are developed in this version.

One possible error you might encounter is with incompatible version of tableau reader if you have older version. This can be overcome by updating the tableau reader to the latest version.

12.3 Task Breakdown

Tanmoy Choudhury

- Worked on Data acquisition and correlation between tobacco usages and related death
- Worked on report using shareLaTex
- Worked on gitlab administration and maintenance
- Worked on python programming as below

  - fetchdata.py (Modularized earlier codebase contributed by Madhavi and added new methods and structure related to correlation data setup)
  - process.py (Modularized earlier codebase contributed by Madhavi and added new methods and structure related to correlation preprocessing)
  - visualize.py(Modularized earlier codebase contributed by Madhavi and added new methods and structure)
  - correlation.py
- Worked on requirement.txt and deployment setup
- participated in project planning, execution and collaboration events
- Contributed into README.rst file (Data source,Setup and Correlation visualization)
- Hadoop deployment testing

Madhavi Pilli

- Worked on data acquisition and visual analysis of data
- Worked on python programming as described below
  - Initial analysis of data and working on pre-processing of data
  - fetchdata.py (added python process to fetch data)
  - process.py (added python processing to pre-process data and prepared data for visualizations)
  - visualize.py (added python code to visualize data using seaborn[1] plots)
- Tableau Visualizations (Create trend graphs and dashboards for side by side comparison of charts and two interactive visualizations using a slider for year)
- Contributed to README.rst file
- Worked on report using shareLaTeX
- participated in project planning, execution and collaboration events
- Installation of cloudmesh client and setting up virtual environment.

Vipin Singh

- Worked on report section
- Worked on gitlab administration and maintenance
- Worked on deployment of hadoop
  - Worked on Data acquisition
  - Virtual Box creation and installation of Ubuntu
  - Setting up the SSH keys on the virtual machine, gitlab, github and chameleon.
  - Installed the bootstrap.sh to install the required software along with python.
  - Downloaded and installed the cloudmesh client
  - Setup the virtual machines using the cloudmesh client
  - Created a chameleon cloud cluster and installed hadoop on it.
  - Tested a mapreduce program on the chameleon cloud cluster
- Contributed to the README.rst File
- Created a README.rst in the report folder