Developing a Framework for Establishing Inferences on U.S. Healthcare Data Using Data Analytics

An IT Project Proposal

Presented to the

Faculty of the School of Accountancy, Management,

Computing and Information Studies

Saint Louis University

In Partial Fulfillment

of the Requirements for the Course

IT 411

by

Erythrina Nicole C. Andres

Renphil Ian G. Balantin

Britanny M. Baldovino

Janine Q. Doria

Kathleen Kyle S. Faustino

Naider N. Mascardo

Melvin Dave H. Peralta

Christine B. Tagle

Mr. Dalos D. Miguel

Project Adviser

January 2019

Abstract

This research focuses on the application of data analytic models such as descriptive data analytics with regards to the data on the notifiable diseases of the United States. The purpose of this research is to develop and establish a process model for rendering the trends and patterns of the number of cases of notifiable diseases in the United States in relation to certain attributes. These attributes include the population density, topography, business industries, socioeconomic profile and the daily temperature of each state. With the help of descriptive data analytics, data visualization and machine learning, correlating these attributes with the number of cases of each notifiable disease will be possible to identify for trends and patterns. The resulting process model will in turn be used by healthcare stakeholders, researchers and information technologists as a template for analyzing healthcare data as well as developing tools that may help with it.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Context of the Study

The main focus of the research revolves around notifiable diseases, state attributes and data analytics. Notifiable diseases, otherwise known as reportable diseases, are any diseases that are to be reported to public health authorities as mandated by a governing body (Reportable diseases: MedlinePlus Medical Encyclopedia, n.d.). The process of storing data regarding notifiable diseases vary from one country to another. In the United States, each state is mandated to weekly report the number of cases for each notifiable disease as observed in that state. The government agency of the United States, National Notifiable Diseases Surveillance System (NNDSS) of Center for Disease Control (CDC), is assigned to summarize these reports. Given that the number of cases is continuously changing, the data regarding the number of cases for each notifiable disease every week vary from one another. Thus, a trend and pattern on when a certain disease occurs more frequently may be gathered from studying the increase or decrease in the number of cases of a notifiable disease.

The World Health Organization recognizes several factors that may influence the increase or the decrease in the spread of diseases. These factors include the climate and the living condition of the affected area (Environmental factors influencing the spread of communicable diseases, 2010). Other factors that also contribute to the spread of disease include the daily temperature, the size of the population, the topography, the socioeconomic profile and the business and industries of the affected area (Sarofim et al., 2016; Environmental factors influencing the spread of communicable diseases, 2010; Phung et al., 2011; Lambin, Tran, Vanwambeke, Linard & Soti, 2010). The researchers consider these other factors as state attributes. State attributes are defined by the researchers as any distinct characteristics of a given place. The researchers identified these five state attributes in order to study and investigate their relationship with the number of cases of a notifiable disease. The availability of the resources, the nature and probable origin of a disease such as vector-borne and water-borne causes were also considered in choosing the five state attributes.

The state attributes will be gathered based on each of the 50 states of United States through the use of their open databases and websites. The United States was chosen by the researchers due to the availability and completeness of the data to be gathered which will be the five state attributes and the weekly number of cases of notifiable diseases as reported and summarized by National Notifiable Diseases Surveillance System (NNDSS) of Center for Disease Control (CDC).

In order to find out the degree of correlation between the number of cases of the notifiable diseases and the five state attributes, the researchers will apply data analytics, such as descriptive data analytics. Descriptive data analytics is a sub-type of data analytics which deals

with summarizing historical data for the purpose of studying the changes and patterns of this data (Vesset, 2018). The researchers then propose an investigative study related to applying data analytic models, specifically descriptive data analytics, on the data of the notifiable diseases gathered from the open databases and websites of the United States. By studying the correlation of each state attribute and the number of cases of notifiable diseases, a process model for rendering visualizations and inferences on data regarding the notifiable diseases in the United States will be developed and established.

Background

For the research, it was decided that the cases of notifiable diseases recorded in the United States will be used as the basis. The United States has data that are up to date and are continuously studied, which is what the researchers need to achieve accurate results. The researchers also decided to use the United States as the basis since the data relevant to the notifiable diseases that are needed to be gathered are readily-available online. With that said, for this investigative and experimental research, the researchers have chosen five attributes to be related to the different notifiable diseases by reading and reviewing studies related to notifiable diseases. The researchers pointed out that correlating the different chosen attributes with notifiable diseases in the United States must be performed. The five chosen attributes will serve as fundamental attributes serving as independent variables to the dependent variable which is the cases of notifiable diseases.

The first attribute chosen for the research is temperature. Temperature refers to the degree of coldness or hotness that is measured on a definite scale. According to an article titled Temperature-Related Death and Illness found in the website globalchange.gov, the effects of mortality are observed even for the few differences from seasonal average temperatures. Temperature extremes have a direct impact to health by compromising the ability of the body to regulate its internal temperature. The loss of the internal temperature control may result in the occurrence of illnesses and may also worsen some conditions (Sarofim et al., 2016). Furthermore, on issues like airborne diseases as indicated in the review of Memarzadeh (2011), the infectivity of a virus passed by the airborne route in an indoor environment is the result of a lot of factors which include temperature, humidity and population density.

The second attribute is population density. According to one of the articles produced by Population Action International Health Families Healthy Planet, population matters to infectious diseases such as HIV/AIDS. As stated on the article, the unhealthy living conditions on urban slums and increased population may increase the transmission of infections. It also says that migration has the capability to enhance the vulnerability to diseases. It was explained that

humans go into the areas where people do not have the resistance to a specific disease. The article also emphasized that population growth has been adding challenges to the addressing of the spread of certain infections like AIDS/HIV (Why Population Matters to Infectious diseases and HIV/AID, n.d.). On another article published in Encyclopedia Britannica which talks about the effects of the surrounding environment on human disease, under the category of human activity specifically on the section of population density, it was explained how population density is connected to diseases. It stated that the problem caused by having a dense population, which is overcrowding, determines the ease with which infection spreads through a population (Christie, Feigin & Garg, 2018). The densely populated areas of cities can serve as breeding grounds for infectious agents, which may cause the birth of bacteria and viruses. This may result in the presence of new strains capable of causing serious diseases.

The third attribute used for the research is topography. Atieli et.al (2011), states in their article titled "Topography as a modifier of breeding habitats and concurrent vulnerability to malaria risk in the western Kenya highlands", that the topography of a certain place is one factor that could affect the spread of a certain disease. The article gives an idea on how the topography of the Western Kenya highlands affected the cases of Malaria. It shows that the topographic measures could be considered on the identification of high-risk malaria foci to have an important surveillance and do some control activities on the places where people are most needed. The researchers of the said article used data from the different kinds of topographic features to observe on how these features affect the existence of Malaria, which the research team of this paper wants to apply to the present cases of notifiable diseases that happen in the United States. Another article found in PMC was the study "Impact of Highland Topography Changes on Exposure to Malaria Vectors and Immunity in Western Kenya", where it was stated that the topography of the highlands in Western Kenya does have an impact on the exposure rates of the human to the parasites and malaria vectors (Wanjala & Kweka, 2016).

The fourth attribute chosen for the research is business and industries. There are various types of business and industries and each of them has an impact to the place where they are found. One impact of business and industries are its effect on the people involved. Some people may acquire diseases because of the work that they are into. According to Paul S. Peirce (1911) on his article called "Industrial Diseases", industrial diseases have been defined as morbid results of occupational activity traceable to specific causes or labor conditions, and followed by more or less extended incapacity for work." On his article, different kinds of diseases caused by the effects of business industries were presented along with how they affect the health of the workers. Another study titled "Industrial Development, Pollution and Disease: The Case of Swaziland" by K.D. Dlamini and P.N Joubert also figured out that each of this industry has its own health hazards that results on certain health problems which require consideration (Masuku, 2013).

The last and the fifth attribute chosen for this experimental research is the socioeconomic status. On an article titled "Effects Of Socioeconomic Factors On Obesity Rates in Four Southern States and Colorado", it was discovered that the level of increased obesity was connected to people with income below the level of poverty or lower income. Akil and Ahmad concluded that the lower income levels are equated to less consumption of healthy foods and instead food with poor quality (Akil & Ahmad, 2011). Another article reviewed by the researchers is an article titled "Socioeconomic Status and Coronary Heart Disease", where the result shows that the people belonging from the lower/middle social classes have the capacity of having a greater coronary heart disease risk than people who belong on higher social classes (Janati, Matlabi, Allahverdipour, Gholizadeh & Abdollahi, 2011).

Based on the different reviews and readings related to the research, the researchers have chosen temperature, topography, business and industries, socioeconomic profile and population density as the attributes that will be correlated to the different notifiable diseases present in the United States.

Statement of Objectives

The research aims to construct and establish a framework for identifying the trends and patterns of United States' healthcare data through visualization and inferences. The researchers identify the different objectives that must be followed and accomplished. The specific objectives of this research are stated below:

1) Gather data related to data analytics and the data for the notifiable diseases of the United States as well as each state's attributes specifically:

    a. Population Density

    b. Daily Temperature

    c. Business and Industries

    d. Capita/Socioeconomic Profile

    e. Topography

2) Make a pairwise or, if appropriate, a multi-variable correlation

3) Correlate these pairs and/or multi-variable pairs through simple linear and multiple linear regression as well as logistic regression for non-continuous data

4) Form inferences from the generated visuals through the use of k-means algorithm

Scope of the Project

The focus of the research is to apply data analytics in understanding the trends and patterns of notifiable diseases in the United States. Data ranging from notifiable diseases to particular factors that affect these diseases have been gathered to achieve the said objective. The data gathered specifically include the United States' weekly cases of notifiable diseases, population density, capita or socioeconomic profile, business and industries, topography, and daily temperature. Data will be gathered through manual collection and web scraping using Python. The latest available data is what will be gathered for the research. After collecting all the necessary data, data analytics is then applied also through using Python.

Significance of the Study

The findings of the research will be beneficial to the healthcare industry and its stakeholders considering that technology is important in health-related matters. By analyzing the results, healthcare agencies would have easier means of determining approaches to notifiable diseases. It will also serve as a future reference for researchers on the topic of trends and patterns of notifiable diseases and data analytic models. In the field of information technology, the developed process model could be used as a basis for creating tools and software used for healthcare. The research would also help develop data analytic models that are relevant in identifying and understanding widespread notifiable diseases not only in the United States but also in other countries such as the Philippines.

Chapter 2: Methodology

Formation of Research Team

The research will be conducted by a team of 8 Bachelor of Science in Information Technology (BSIT) students who will be involved with information about notifiable diseases. The team will be given guidance from start to finish by the SLU administration specifically the CIS Department of the School of Accountancy Management Computing and Information Studies.

Approach

The approach of the researchers for this research is an investigative and experimental study of the relationship of the number of cases of notifiable diseases and the state attributes through the use and application of data analytics models such as descriptive data analytics. Since the researchers are trying to find out how each state attribute can affect the given data on notifiable diseases, the state attributes are considered as independent variables while the number of cases is considered as the dependent variable. The number of cases then depends on the changes or the presence of each state attribute. The researchers expect a visualization of the trends and patterns to be realized by analyzing and correlating the said variables. The researchers provide the different steps and methods to be applied. Table 1 shows the mapping of objectives with their corresponding tools and methods to be applied and the expected output upon applying these tools and methods.

Table 1: Objectives with their the corresponding tools and methods to be applied and expected output

Objectives	Tools and Methods	Expected Output
Establish a framework for identifying the trends and patterns of United States' healthcare data through visualization and inferences	Research and the application of data analytic models such as descriptive data analytics	Process Model which may be used to develop healthcare tools and related software
Gather related data about data analytics, the data for notifiable diseases of the United States as well as each state's attributes such as:	Web scraper program through Python 3 using Selenium, Pandas and BeautifulSoup4	Scraped data, either in table format or in CSV format

1. Population Density 2. Daily Temperature (from December 2017 to June 2018) 3. Business and Industries 4. Capita/Socio-economic Profile 5. Topography (Glaciers, Locales, Beaches, Areas, Lakes, Streams, Swamps, Forests, Plains, Woods)	Manual searching and gathering of data through the open databases of the United States	A list of manually searched and gathered data which were not applicable to be web scraped
Make a pairwise or, if appropriate, a multi-variable correlation	Analysis of the gathered data	A list of pairs and/or multi-variable pairs
Correlation and Regression Analysis between the dependent variable, number of cases of notifiable diseases, and the other variables.	Scatter plots and applying Kendall's tau	• Verified Correlations • Visuals related to the correlation of state attributes and the number of cases • Visuals related to the trends and patterns of state attributes and the number of cases
	Simple Linear Regression	
	Multiple Linear Regression	
	Logistic Regression	
Form inferences from the generated visuals through the use of k-means algorithm	Analysis of the summarized data through the use of the k-means algorithm	• Inferences • Clustered data of each state with related and similar characteristic for better tracking and for future works

The method that will be applied for the collection of data is a mixture of the quantitative and qualitative method. The reason is that the researchers expect two types of data, quantitative, or numerical data, and qualitative, or categorical data. Most of the data to be gathered are considered quantitative data, except for business and industries as well as topography since there are different categories of topography such as plains and swamps.

Data Collection and Procedure

The collection of data needed for the research will be done in two different ways: web scraping, and manual searching and gathering. The scope of the collection and processing of data will only be limited to the 50 states of United States. The researchers will be collecting data regarding the number of cases of notifiable diseases and the five state attributes from the different open websites and databases of United States. The websites that will be accessed with the corresponding data to be gathered can be found in Table 2.

Table 2: Websites to be accessed and the corresponding data to be gathered

Data	Website Name	Website URL
Number of cases of each notifiable disease	Center for Disease Control and Prevention	data.cdc.gov
Population density	For the Area of Each State: Simple English Wikipedia	simple.wikipedia.org/wiki/List_of_U.S._states_by_area
	For Population by State: World Population Review	worldpopulationreview.com/states/
Daily Temperature	National Climatic Data Center	www.ncdc.noaa.gov
Business and Industries	careeronestop	www.careerinfonet.org
	Bureau of Economics Analysis	www.bea.gov
Capita/Socioeconomic Profile	Statista	statista.com
Topography	AnyPlaceAmerica	anyplaceamerica.com

The researchers will be gathering data through web scraping using Python 3 and its libraries Pandas, Selenium and BeautifulSoup4. Web scraping will be applicable to the number of cases of each notifiable disease, population density and topography. The other remaining data that will be gathered, business and industries, daily temperature and socioeconomic profile, must be manually gathered because of the several options and data that can only be configured and filtered through human intervention. The expected format for all the gathered data will be in table form and CSV file type for processing and analysis. The process of gathering data through web scraping and manual searching and gathering is shown below:

1. Each researcher will be assigned to gather one of the data needed for the research. Once assigned, the researchers will be gathering data concurrently.

2. The researchers will then be developing a web scraper for automating the gathering of the number of cases of each notifiable disease, population density and topography through the use of Python 3 and its libraries. The web scraper should follow the following objectives:

   a. The program must be able to scan the website for the needed data without human intervention after running it.

   b. The program must be able to perform data pre-processing to maintain understandable and consistent data. The web scraper program for population density must be able to automatically compute the population density of each state with its corresponding population and land area which will be scraped from the websites in Table 2. The web scraper program for topography must be able to automate the process of finding if a certain state has a certain topological characteristic. If there is a map regarding a certain topological characteristic for that given state, the value of 1 will be written for that specific topological column as a representation. The 0 value represents false which means that a certain topological characteristic does not exist in the state and 1 value stands for true which means that a certain or several landform or water bodies are existing in the state.

   c. The program must be able to store the gathered data into a CSV file

3. The researchers assigned for the manual collection of data will be manually searching and gathering the data for business and industries, daily temperature and socioeconomic profile through the websites shown in Table 2. The objectives of gathering each are as follows:

   a. The method that the researchers will be doing for gathering the socioeconomic profile is through manual searching of the GDP of each state. The researchers will also consider taking the TOP 5 states that has the highest GDP, TOP 5 states that has the lowest GDP. The result of the manual searching will be put into a CSV file.

b. The researchers must gather the minimum and maximum temperatures starting from December 2017 to June 2018 for each state in the USA and store it in a CSV file. The researchers will then create a Python program that will read the contents of each of the CSV file created to identify the average, mean, median and mode weekly temperature of each state and then store the processed data in another separate CSV file.

c. The researchers will also gather the largest business and industries of each state by determining the TOP 10 industries of each state. The data for the list will be gathered from a source published on 2015. The kind of industries with the highest number of employees will be identified, listed and stored in a CSV file by taking into consideration the description of the top largest companies.

4. After gathering all of the needed data, the researchers will integrate them into a single folder.

Data Processing and Analysis

The researchers will create Python programs which will use the data to be gathered as input to generate statistical information applicable to the research. This includes the mean, minimum value, maximum value, standard deviation, graphs, diagrams and other significant statistical information. Besides generating statistical information, correlation and linear regression will also be applied to the data to test their relationship between each other.

A. Correlation

The initial process to be done to the data is correlation analysis. With correlation analysis, the degree of the relationship between the variables is determined. Below are the steps in performing correlation analysis to the data:

1. The data for the notifiable diseases will be paired with the state attributes in order to identify how each of the state attributes affect the notifiable diseases. The number of cases of the notifiable diseases will be paired or grouped with the following:
    a. Population Density
    b. Temperature
    c. Types of Business and Industries
    d. Capita/Socioeconomic Profile
    e. Topography Categories
    f. Population Density and Temperature

2. Scatter plots will be created using the values of the data of each pair. Each scatter plot will show the correlation between the number of cases of notifiable diseases and the attribute tested with it. What will be seen in the scatter plot are points which ideally have a linear formation. The linear formation of the points will be used to identify the level of correlation between the variables.
3. Kendall's Tau or Kendall's Rank Correlation Coefficient will be applied to the quantitative data which will be gathered. Kendall's Tau measures the dependence of two variables in a nonlinear manner. Since scatter plots depend on the linear formation of plotted points, the researchers will use Kendall's Tau to check the quantified degree of correlation of the number of cases of notifiable diseases and each attribute tested with them in an approach where the results could be different.

B. Regression

Other than correlating the data to be gathered, regression analysis will also be performed to check the numerical relationship between the dependent variable and the independent variable. Simple linear regression is to be used between the number of cases of notifiable diseases and other quantitative variables. While logistic regression will be used for the qualitative data. Multiple linear regression may also be used. To apply regression analysis to the data, the following steps will be done:

1. Scatter plots will also be generated using the pairs or groups made during correlation.
2. Linear regression will be checked by creating a regression line in the generated scatter plots. The regression line is the best-fitting line formed in between the points based on a position where it is closest to all the points. Additionally, in order for a line to be considered as the regression line, the value of the distance between each of the points and the line formed between the points must be at its minimum.

C. K-means Algorithm

Using the k-means algorithm, the data will be grouped according to a calculated number of groups. The data with the most similar features will be grouped together. Each group will have a centroid or a value representing most of the data in the group. This will be used to organize and analyze the pairings and groupings.

Timeline

Table 3 shows the timeline of activities and their corresponding projected start and end date.

Table 3: Project Timeline

Start Date	End Date	Task
August 31, 2018	September 3, 2018	Discussion of the research problem regarding notifiable diseases and data analytics
September 5, 2018	September 26, 2018	Identifying and understanding the identified data model relevant to the research
September 26, 2018	October 3, 2018	Review of related literature
September 26, 2018	October 26, 2018	Establishing the specific problem to be solved, the objectives to be followed and the proposal to be made
September 21, 2018	September 28, 2018	Identifying the data to be gathered such as the state attributes and the notifiable diseases as well as their sources
October 1, 2018	December 17, 2018	Creating the collective table of the literature reviewed Gathering of data, either manual or automated using Python Creation and use of Python and its libraries for web scrapers and web automation to automate the gathering of data
October 3, 2018	November 10, 2018	Writing down the Chapter 1 of the research proposal
October 10, 2018	November 10, 2018	Writing down the Chapter 2 of the research proposal
October 5, 2019	October 9, 2018	Establishing the list of pairs and/or multi-variable pairs to be correlated
October 12, 2018	November 5, 2018	Processing and summarizing the data gathered

November 9, 2018	December 17, 2018	Correlating each state attribute with their designated pairings using Python and its libraries such as Matplotlib
October 26, 2018	November 10, 2018	Writing down the Chapter 3 and the remaining portions of the research proposal
November 9, 2018	January 10, 2019	Verifying and revising the research proposal
November 9, 2018	January 10, 2019	Constructing the framework through a process model
January 16, 2019	February 16, 2019	Application of correlation coefficient, linear and logistic regressions and statistical methods through descriptive data analytics, and k-means clustering algorithm
February 18, 2019	February 23, 2019	Generating visualizations from the processed and organized data
February 26, 2019	March 8, 2019	Interpretation of the trends and patterns generated
March 11, 2019	March 16, 2018	Refining the process model
March 18, 2019	March 28, 2019	Validating the process model established earlier with the results generated
March 29, 2019	April 25, 2019	Write up of the research
April 25, 2019	May 3, 2019	Revising and finalizing the research

Chapter 3: Outcomes and Results

Data Gathering

Number of Cases of Notifiable Diseases of Each State

Table 4 shows a sample of the data gathered through web scraping for number of cases of each notifiable disease per state. The complete data tabulates 46 notifiable diseases with their corresponding number of cases each week per state.

Table 4: Sample of the data gathered for the number of cases of each notifiable disease per state

Reporting Area	MMWR Week	MMWR Year	West Nile virus disease, Neuroinvasive Current week	West Nile virus disease, Nonneuroinvasive Current week	Zika virus disease, non-congenital Current week
ALABAMA	1	2018	0	0	0
ALASKA	1	2018	0	0	0
ARIZONA	1	2018	0	0	0
ARKANSAS	1	2018	0	0	0
CALIFORNIA	1	2018	0	0	0
COLORADO	1	2018	0	0	0
CONNECTICUT	1	2018	0	0	0
DELAWARE	1	2018	0	0	0
FLORIDA	1	2018	0	0	0
GEORGIA	1	2018	0	0	0
HAWAII	1	2018	0	0	0

Population Density

Table 5 shows the sample of the data gathered through web scraping for the population density of each state. The complete data contains the population density of the 50 states of the United States.

Table 5: Sample of the data gathered for the population density of each state

State	2018 Population	sq.miles	kmÂ²	Population_Density(kmÂ²)	Population_Density(sq.miles)
Alabama	4888949	50744	131426	37.19925281	96.34536103
Alaska	738068	567400	1481347	0.498241128	1.300789566
Arizona	7123898	113635	294312	24.20525837	62.69105469
Arkansas	3020327	52068	134856	22.39668239	58.00735577
California	39776830	155959	403933	98.47383106	255.046711
Colorado	5684203	103718	268627	21.16020728	54.80440232
Connecticut	3588683	4845	12548	285.9964138	740.6982456
Delaware	971180	1954	6030	161.0580431	497.0214944
Florida	21312211	53927	139670	152.5897544	395.2048325
Georgia	10545138	57906	149976	70.31216995	182.1078645
Hawaii	1426393	6423	16635	85.74649835	222.0758213

Temperature

Table 6 shows a sample of data gathered through manual searching and gathering for the daily temperature from December 2017 to June 2018 of Alaska from one of the weather stations situated there. This sample is the original data gathered and have not undergone processing and summarizing. The complete table of the daily temperature of each state contains thousands of rows of data from several hundreds or thousands of weather stations.

Table 6: Sample of the data gathered for temperature

STATION	NAME	DATE	TAVG	TAVG_ATTRIBUT	TMAX	TMAX_ATTRIBUT	TMIN	TMIN_ATTRIBUTE

				ES		**ES**		**S**
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/1/2017	-12.2	,,U	-10.2	H,,U	-13.1	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/2/2017	-7.5	,,U	-6.4	H,,U	-9.7	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/3/2017	-8.2	,,U	-7.6	H,,U	-9.3	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/4/2017	-10.4	,,U	-7.3	H,,U	-12.3	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/5/2017	-7.1	,,U	-4.6	H,,U	-12	H,,U

USR0000ASLR	SALMON RIVER ALASKA, AK US	12/6/2017	-6	,,U	-4.1	H,,U	-7.2	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/7/2017	-7	,,U	-6.3	H,,U	-8.4	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/8/2017	-10.2	,,U	-6.4	H,,U	-12.4	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/9/2017	-13.3	,,U	-12.6	H,,U	-14.3	H,,U
USR0000ASLR	SALMON RIVER ALASKA, AK US	12/10/2017	-13.1	,,U	-11.4	H,,U	-14.7	H,,U

After processing and summarizing through the use of Python 3 and its library, Pandas, Table 7 shows the average daily temperature of the same table shown in Table 6.

Table 7: Summarized table after processing

DATE	TMIN	TMAX
12/1/2017	-13.10100671	-6.519127517
12/2/2017	-10.43163265	-4.131292517
12/3/2017	-7.432094595	-1.858724832
12/4/2017	-5.915771812	0.05819398
12/5/2017	-6.062711864	0.564189189
12/6/2017	-6.694612795	0.063973064
12/7/2017	-7.393559322	0.553898305
12/8/2017	-5.277241379	0.545172414
12/9/2017	-6.807612457	-1.487889273
12/10/2017	-5.969097222	0.567820069

Business and Industries

Table 8 shows the data gathered for the business and industries through manual searching and gathering. Other business and industries not shown in the sample are as follows:

1. Wholesale trade
2. Retail trade
3. Transportation and warehousing
4. Information
5. Finance and insurance

Table 8: Sample of the data for business and industries of each state

State	Agriculture, forestry, fishing, and hunting	Mining, quarrying, and oil and gas extraction	Utilities	Construction	Durable goods manufacturing	Nondurable goods manufacturing
Alabama	1	1	1	1	1	1
Alaska	1	1	1	1	0	1
Arizona	1	1	1	1	1	1
Arkansas	1	1	1	1	1	1
California	1	1	1	1	1	1
Colorado	1	1	1	1	1	1
Connecticut	1	0	1	1	1	1
Delaware	1	1	1	1	1	1

Socioeconomic Profile Through Capita

Table 9 shows the sample of the data gathered for socioeconomic profile of each state through manual searching and gathering.

Table 9: Sample of socioeconomic profile of each state through capita

Alabama	37508
Alaska	63610
Arizona	39583
Arkansas	36714

California	60359
Colorado	54026
Connecticut	62633
Delaware	63955
Florida	39842
Georgia	45925

Topography

Table 10 shows the data gathered for topography through web scraping. Other categories of topography not shown in the sample is as follows:

1. Streams
2. Swamps
3. Forests
4. Plains
5. Woods

Table 10: Sample of the data gathered for topography

State	Glaciers	Locales	Beaches	Areas	Lakes
Alaska	1	1	1	1	1
Alabama	0	1	1	1	1
Arkansas	0	1	1	1	1
Arizona	0	1	1	1	1
California	1	1	1	1	1
Colorado	1	1	1	1	1
Connecticut	0	1	1	1	1

Accomplishing the collection and processing of data, the following results were achieved.

Figure 1 shows that Texas has the largest number of cases (with 110176 cases) among all states of US while Vermont has the least (having 418 cases).

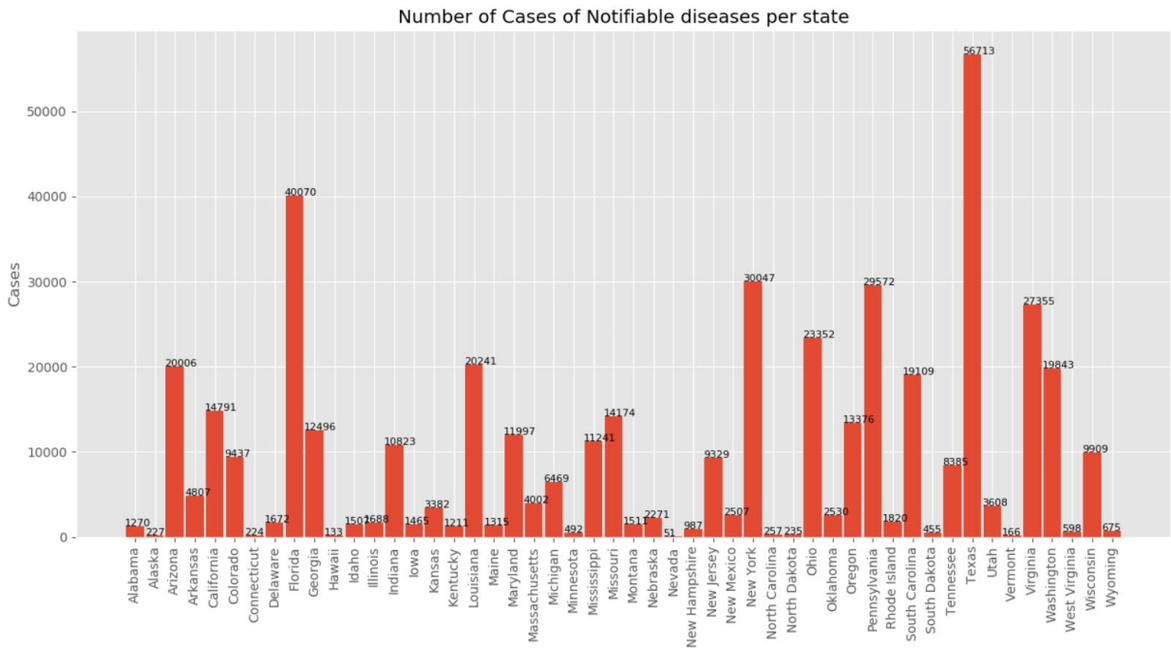

Figure 1: Total number of cases of notifiable diseases per state

Figure 2 compares the total cases of notifiable diseases in Texas to all the other States. Even if Texas had the largest number of cases of notifiable diseases, the summation of the number of cases of all other states is still a larger number.

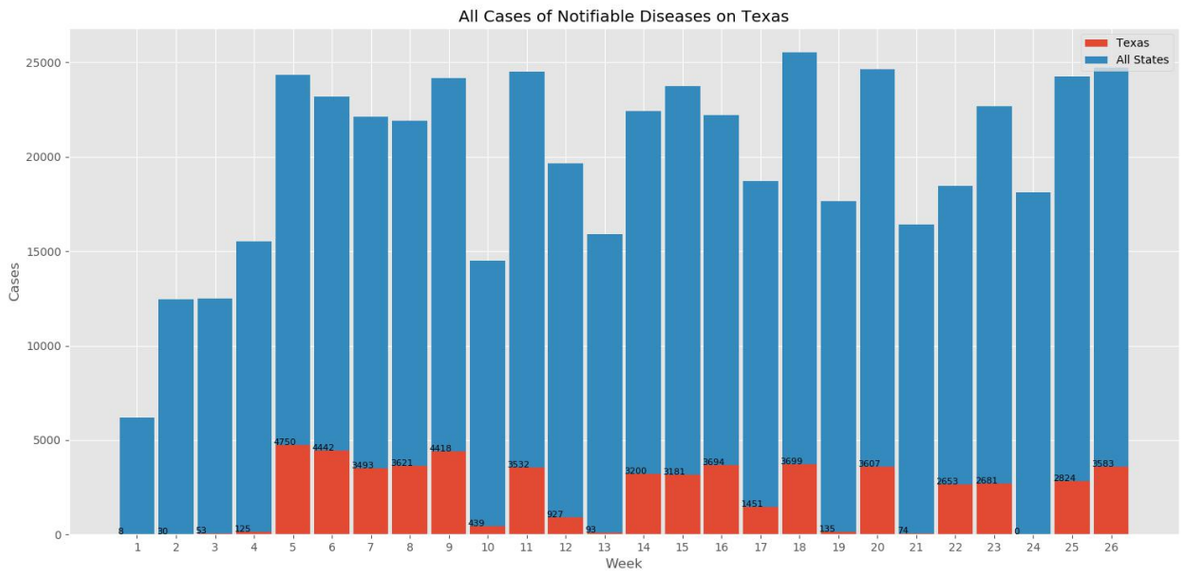

Figure 2: Cases of notifiable diseases in Texas compared to all states

Correlation Analysis

Figure 3: Correlation between cases of notifiable diseases and population density

Correlation coefficient of all cases of notifiable diseases and population density is 0.2767 with Chlamydia trachomatis infection with 0.2851 correlation coefficient with the population density which is the highest correlation among the notifiable diseases. This indicates an overall low positive correlation between the cases of notifiable diseases and population density which is noticeable on figure 3 while figure 4 shows the data points of the number of cases of notifiable diseases and population density on every state of the United States.

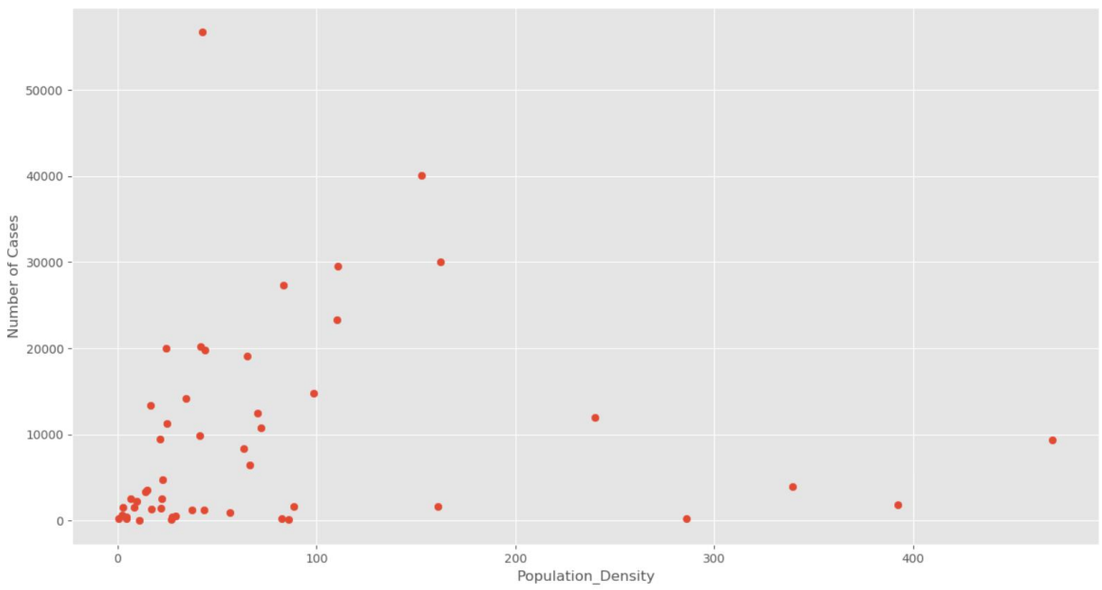

Figure 4: Number of cases of notifiable diseases and population density

The graphs placed in this part of the documentation are only samples of the graphs to be generated for the correlation of data. A specific type of graph is appropriate for each correlation to check if there is a relationship between the data to be correlated.

As a result of the research as a whole, a framework for rendering visualizations and analytics on data related to notifiable diseases in the United States will be developed. The currently formulated process includes the sequential steps shown in figure 5.

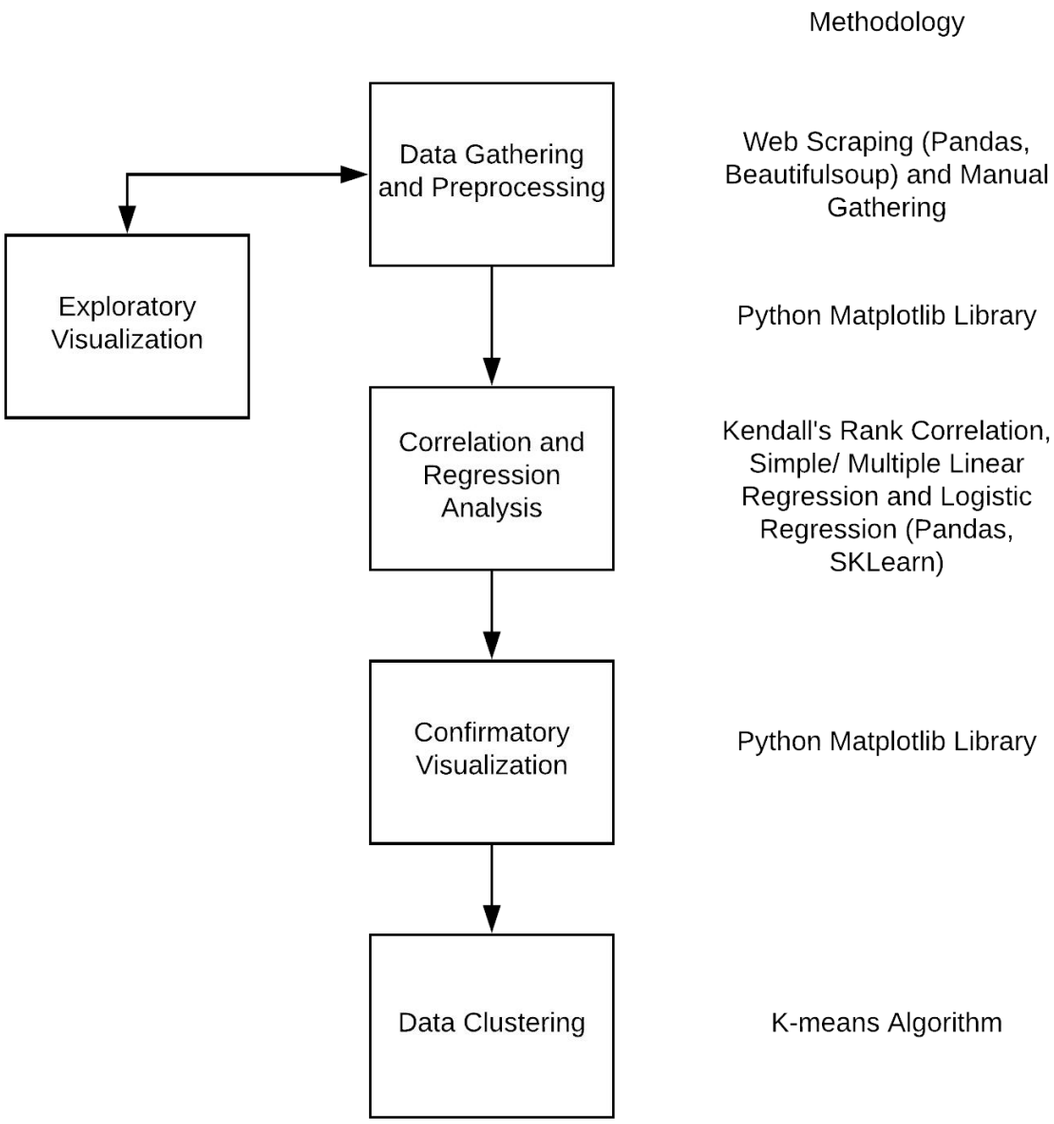

Figure 5: Framework/process model

- During Data Gathering and Preprocessing, the number of cases of notifiable diseases and the factors that might affect the diseases were either manually downloaded or web scraped and stored on different csv files. These factors serve as an identifier for certain characteristics of each state in the United States. Included in these factors are the temperature, population density and land mass, topography, business industries, and the socioeconomic profile of a state.
- During Exploratory Visualization, graphs were created to make sure that the factors may be correlated to the number of cases of notifiable diseases and to other other factors.
- The Data Correlation and Regression Analysis phase is still ongoing but multiple pairings of factors with the cases of the notifiable diseases have been tested. It is in this process where simple and linear regression will be applied.
- Confirmatory Visualization is where diagrams and conclusions will be rechecked.
- The last process which is Data Clustering aims to make use of the k-means clustering algorithm for data clustering.

Appendix

Table 11: Collective table summarizing the review of related literature conducted by the researchers

Title	Topic Area	Methodology	Findings
Literature Review of the Effect of Temperature and Humidity on Viruses	The paper is a review which explains that the infectious viruses are known to be transferred by the air as well as indirect and direct contact.	It is an extensive literature review of the researchers using more than 120 papers. Those papers were conducted on the effect of temperature and humidity on the transmission of the infectious viruses.	Both of the indirect and direct study shows the results of examining environmental conditions that could have an impact infectious disease aerosol transmission inside the enclosed environments.
Why Population Matters to Infectious diseases and HIV/AID	This paper presents topic about HIV/AID, Links and the State of Infectious Diseases between the Population and Infectious Diseases, Links between HIV/AID and Population, and also the Policy Considerations	The deaths from Infectious and Parasitic Diseases are High in Low-Income Countries graph and Swaziland's Age Structure Shaped by HIV/AIDS graph were analyzed.	Fertility, migration and urbanization has an impact on the spread of diseases including malaria,tuberculosis and HIV/AIDS. The The unhealthy living conditions in urban slums and increased population densities can let the transmission of the infections more easy. The migration might also increase the possibility of having a disease. HIV/AIDS which is a sample of

			infectious diseases have had a large effect on the trends of demography connected to the altering on the age structures of heavily countries that are affected. The paper emphasize that having an access to the family planning services has the power or ability to lessen the spread of disease, especially when it is integrated with the existing HIV prevention programs.
Topography as a modifier of breeding habitats and concurrent vulnerability to malaria risk in the western Kenya highlands	This paper shows the investigation of identifying whether the risk of infection with malaria parasites and the distribution of the local spatial malaria vectors located in the highlands is related into topography.	There are four villages and each of them has the measure of 9 Km2 lying between 1400-1700 m above sea level in the western Kenya highlands and they were categorized or divided into a pair of narrow and broad valley shaped terrain sites. The infection and indoor resting adult malaria vectors surveys were	The investigation shows that the broad flat bottomed valleys had an important higher number of anopheles dip or larvae in their location or habitats than in the narrow valleys during both the dry (1.89 versus 0.89 dip/larvae) and the rainy season (1.66 versus 0.89 dip/larvae). The same result,vector adult house or density in

		gathered or collected which originates from the bottom of the valley and ending at the hilltop on the both sides of the valley during the dry and rainy seasons. The data gathered at the distance of ≤500 m from the main stream or river were categorized as above as uphill and those as valley bottom The surveys on the Larval were categorized by habitat location while the infections and vectors by house location	broad valley villages shows higher density than those within narrow valley houses during both the rainy season (0.96 versus 0.09) and the dry (0.64 versus 0.40). The Asymptomatic malaria prevalence was importantly higher in participants residing in broad than those in the narrow valley villages during rainy (17.15% vs. 1.20%) season and the dry (14.55% vs. 7.48%) season. The malaria infections were widespread in most of the places in valley villages during both the rainy and dry season, where over 65 percent of the infections were flocked at the valley bottom in the narrow valley villages during both of the seasons.
Impact of Highland Topography Changes on Exposure to Malaria Vectors and Immunity in Western	This study was done to determine on how the major environmental terrain characteristics which	The study was done in the five different ecosystems located in the western Kenya highlands. The five	The results shows that the changes in the topography had an impact or implication on the transmission in

Kenya	control the breeding of the malaria vectors located in the western Kenya highlands could influence the exposure to transmission and the creation of an immune response	different ecosystems are Marani, Iguhu, Emutete, Fort Ternan, and Shikondi. It was done for 16 months, that ranges from age of 6- to 15-year-old children. The exposure to malaria was tested using circumsporozoite protein and merozoite surface protein immunochromatographic (CSP) antibody tests. The malaria parasite was investigated and examined using different kinds of tools, this includes microscopy that is based on blood smears, rapid diagnostic test based on HRP 2 proteins, and serology which based on the human immune response to the parasite and vector antigens was also examined in the highlands in comparison with different topographical systems of western Kenya.	highlands of western Kenya and right treatment, diagnosis and control tool are needed to be considered accordingly. Both (Shikondi) plateau and (Iguhu, Emutete)U-shaped valley found to have higher parasite density than (Marani, Fort Ternan)V-shaped valley. People in V-valley were less immune than in plateau and U-valley residents.

Industrial Development, Pollution and Disease: The Case of Swaziland	The aim is to specify problems caused by industrial development in Swaziland and specify the effect of industrialization on the environment; discover the level of corporate responsibility towards the environment by companies and make recommendations on the management of industrial pollution.	The data that was used were collected by primary and secondary sources, offering a representative selection of industries in Swaziland.	The study shows that each industry has its own risk on our health which can result in certain health problems that require special consideration. The cause of health problem depends on the awareness of the workforce, design of the factory,the nature of the final product and type of raw material used.
Effects of socioeconomic factors on obesity rates in four southern States and colorado	To investigate the association or connection between the increase in the body mass index and the socioeconomic factors like % below poverty line, income level, persons receiving food stamps and unemployment rates in Alabama, Mississippi, Tennessee,Louisiana, and Colorado.	The demographic data like ethnicity, sex, geographic location and BMI were gathered from the CDC's Behavioral Risk Factor Surveillance System 15 for the year ranging from 1995 to 2008. The researchers focused on the national data such as the data from Miss., Tenn.,Ala., La., and Colo. The data from Supplement Nutrition Assistance Program which is the percentage of people	The results from the study interprets a very strong association between the tested variables and obesity . The factors more closely that are founded related with obesity were the receipt of food stamps, income below poverty level, general income level and unemployment. These variables have coefficient of determination of 0.103,0.438, 0.427. and 0.018 respectively. The

		receiving food stamps was calculated from US Department of Agriculture data in their yearly national- and state-level reports for the year ranging from 1995 to 2008. The population rates of the target states of the researchers were gathered from the US Census Bureau for the year ranging from 1995 to 2008. The unemployment rates for the year ranging from 1995 to 2008 were gathered from the US Department of Labor. Lastly, the national and state median household income data and the percentage of people below the poverty level were gathered from the United States Census Bureau.	highest rate of obesity was determined in the place of Mississippi having 26.5% 6 4.13% and then followed by place of Alabama having 25.18% 6 4.41%. The Colorado had the lowest rate having 15.4% 6 2.63%. About ethnicity, African Americans had the highest rate having 32.64 6 5.99%. The researchers have found a important impact of consumption of food with low quality, caused by economic factors, on increased BMI. Other than physical activity, the quantity and the quality of the food are significant factors which contribute to obesity rates.
Socioeconomic Status and Coronary Heart Disease	The objective of the study was to define the socioeconomic and demographic characteristics, their Association or	The researcher conducted a cross-sectional descriptive study to analyze and find out the current Socioeconomic status	The results of the study shows that the less educated participants were more capable to coronary heart

	relation to the diseases, and to find the predictive risk of coronary heart disease in the place of Tabriz. Tabriz is the capital of East Azerbaijan Province and it is the fourth largest city in Iran.	of the patients having coronary heart disease. It was done in Tabriz and all patients having the number of 189 that were referred to the Shahid Madani Hospital which is the Central Referral Hospital for cardiac patients and the have decided to consider the patients on the range will start from 2009 to 2010. A researcher created questionnaire having 15 questions and it was used to gather data. The researcher used Descriptive statistics to define the basic features of the Socioeconomic status of the patients having coronary heart disease. The data analysis was delivered using the ver.16 of SPSS.	disease. Relating into the occupational status, the housewives and the retired men were in the level of having high risk of coronary heart disease than the other the people in the places. Patients participants from the study was also reported to be mostly coming from the urban areas and they were living in apartment.
Temperature-Related Deaths and Illness, focusing on Section 6 of Chapter 2 Author/s: Marcus C.	Investigative study on how temperature can increase the chances of mortality and illness in a specific	Analyzing a nationally representative database from the Healthcare Utilization	Several illnesses such as respiratory, cardiovascular and renal illnesses were found to be affected

Sarofim, Shubhayu Saha, Michelle D. Hawkins, David M. Mills, Jeremy Hess, Radley M. Horton, Patrick L. Kinney, Joel D. Schwartz and Alexis St. Juliana	area wherein the temperature is measured at.	Project (HCUP) which contains the data for heat-related illnesses not limited to hyperthermia and hypothermia.	by extreme heat.
The effects of socioecological factors on variation of communicable diseases: A multiple-disease study at the national scale of Vietnam Author/s: Dung Phung, Huong Xuan Nguyen, Huong Lien Thi Nguyen, Anh Mai Luong, Luong Manh Do, Quang Dai Tran and Cordia Chu	The research examines the effects of socioecological factors on multiple communicable diseases across Vietnam.The factors enumerated and gathered are: 1. Climatic data (temperature, humidity and cumulative rainfalls) 2. Population density 3. Monthly average income 4. % Illiteracy 5. % of households with supplied safe water 6. Number of the passengers by road	- Global Moran's I was applied to initially establish the spatial correlation of each disease. A visualization of the spatial representation is an expected output. This generated visualization will show how connected or disconnected the instance/s of a disease in a certain area. - Bayesian framework was applied in order to analyze the relationship between socio-ecological factors and variation of each communicable disease	The study revealed that most of the diseases were sensitive to climatic data while socioeconomic factors have varied influences wherein population density has the biggest influence. Additionally, the distribution of the spatial clustering of each disease revealed how certain communicable disease are endemic only on certain parts of Vietnam.
Pathogenic landscapes: Interactions between	The authors reviewed the eight case studies they conducted in	Analysis through visual spatial models were applied.	The researchers conclude that all of the case studies

land, people, disease, vectors, and their animal hosts Author/s: Eric F. Lambin, Annelise Tran, Sophie O. Vanwambeke, Catherine Linard, and Valérie Soti	Europe and West Africa regarding the relationship between land attributes and the emergence of vector-borne diseases and zoonoses.		they've conducted reveal that spatial variations regarding the risk for infection are affected by three sets of factors: 1. The pathogenic cycle and the biology of vectors, hosts and pathogens 2. Ecosystem processes at the landscape scale, as influenced by ecosystem structure and composition, landscape connectivity and configuration, climate, species interactions 3. Land use, human behaviour and mobility, knowledge and perception of disease risk, and socio-economic conditions

Tools and Applications Used

Table 12: The tools and applications used by the researcher

Tool/Application	Used As
Github	Repository for files and version control management
Google Docs	Platform for creating the project document and collaboration
Microsoft Excel	Platform for opening and viewing of the data gathered
Slack	Platform for communicating and collaborating with the group
Visual Studio Code	Code editor and debugger for the Python programs

Python Libraries and APIs Used

Table 13: The Python libraries and APIs used by the researcher

Library/API	Used For
BeautifulSoup4	Gathering data from websites
IO	Stream handling
Matlpotlib	Generating visualizations regarding the processed data
NumPy	Processing of the gathered data
OS	Getting the current directory of the code and for error-checking in case a certain file already exists or not

Pandas	Gathering data, data manipulation and analysis
Requests	Requesting an HTML copy for the gathering of the data which will be read to gather the specific data needed
Selenium	Gathering the data regarding the topography of each state which cannot be scraped by BeautifulSoup4
SKLearn	Machine Learning library that features regression and clustering algorithms.
Webdriver	The webdriver, chromedriver, was used as a requirement by Selenium. It allows the automation feature of Selenium to be run successfully. This is used when scraping the topography of each state.

List of Notifiable Diseases from data.cdc.gov

1. West Nile virus disease, Neuroinvasive

2. West Nile virus disease, Nonneuroinvasive

3. Zika virus disease, non-congenital

4. Vibriosis (Any species of the family Vibrionaceae, other than toxigenic Vibrio cholerae O1 or O139), Confirmed

5. Vibriosis (Any species of the family Vibrionaceae, other than toxigenic Vibrio cholerae O1 or O139), Probable

6. Tetanus

7. Varicella morbidity

8. Spotted Fever Rickettsiosis, Confirmed

9. Spotted Fever Rickettsiosis, Probable

10. Syphilis, primary and secondary

11. Salmonellosis (excluding Paratyphoid fever andTyphoid fever)

12. Shiga toxin-producing Escherichia coli

13. Shigellosis

14. Rabies, animal

15. Rubella

16. Rubella, congenital syndrome

17. Meningococcal disease, all serogroups

18. Mumps

19. Pertussis

20. Legionellosis

21. Malaria

22. Invasive Pneumococcal Disease, Age LT, Confirmed

23. Invasive Pneumococcal Disease, Age LT 5, Probable

24. Invasive Pneumococcal Disease, all ages, Confirmed

25. Invasive Pneumococcal Disease, all ages, Probable

26. Hepatitis (viral, acute, by type) , C, Confirmed

27. Hepatitis (viral, acute, by type), C, Probable

28. Hepatitis (viral, acute, by type), A

29. Hepatitis (viral, acute, by type), B

30. Ehrlichiosis and Anaplasmosis, Ehrlichia ewingii infection

31. Ehrlichiosis and Anaplasmosis, Undetermined Ehrlichiosis/Anaplasmosis

32. Giardiasis

33. Gonorrhea

34. Haemophilus influenzae, invasive disease (all ages, all serotypes)

35. Ehrlichiosis and Anaplasmosis, Anaplasma phagocytophilum infection

36. Ehrlichiosis and Anaplasmosis, Ehrlichia chaffeensis infection

37. Cryptosporidiosis

38. Dengue Virus Infections, Dengue

39. Dengue Virus Infections, Severe Dengue

40. Chlamydia trachomatis infection

41. Coccidioidomycosis

42. Carbapenemase-producing carbapenem-resistant Enterobacteriaceae, Klebsiella spp.

43. Carbapenemase-producing carbapenem-resistant Enterobacteriaceae, Escherichia coli

44. Carbapenemase-producing carbapenem-resistant Enterobacteriaceae, Enterobacter spp.

45. Babesiosis

46. Campylobacteriosis

References

Akil, L., PhD-C, & Ahmad, A. H., PhD, MBA. (2011). EFFECTS OF SOCIOECONOMIC FACTORS ON OBESITY RATES IN FOUR SOUTHERN STATES AND COLORADO. Retrieved January 5, 2019, from https://www.ethndis.org/priorarchives/ethn-21-01-58.pdf

Atieli, H. E. et al.. (2011). Topography as a modifier of breeding habitats and concurrent vulnerability to malaria risk in the western Kenya highlands. Retrieved December 29, 2018, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3269397/pdf/1756-3305-4-241.pdf

Christie, A., Feigin, R. D., & Garg, R. (2018, December 13). Infectious disease. Retrieved December 28, 2018, from https://www.britannica.com/science/infectious-disease/Population-density

Environmental factors influencing the spread of communicable diseases. (2010, December 12). Retrieved from https://www.who.int/environmental_health_emergencies/disease_outbreaks/communicable_diseases/en/

Janati, A., Matlabi, H., Allahverdipour, H., Gholizadeh, M., & Abdollahi, L. (2011). Socioeconomic Status and Coronary Heart Disease [Abstract]. *Health Promotion Perspectives,1*, 105-110.

Lambin, E. F., Tran, A., Vanwambeke, S. O., Linard, C., & Soti, V. (2010). Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. International journal of health geographics, 9, 54. doi:10.1186/1476-072X-9-54

Masuku, B. (2013). Socioeconomic analysis of beekeeping in Swaziland: A case study of the Manzini Region, Swaziland. Journal of Development and Agricultural Economics, 5(6), 236-241. doi:10.5897/jdae2013.002

Memarzadeh, F., PhD. (2011). Literature Review of the Effect of Temperature and Humidity on Viruses. Retrieved December 28, 2018, from https://www.orf.od.nih.gov/PoliciesAndGuidelines/Bioenvironmental/Documents/FINALPUBLISHEDPaperonHUMIDITYandViruses509.pdf

Phung, D. et al.. (2011). The effects of socioecological factors on variation of communicable diseases: A multiple-disease study at the national scale of Vietnam. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193246

Peirce, P. S. (1911). Industrial Diseases. The North American Review, 194, 529-540. Retrieved from https://www.jstor.org/stable/25107041.

Reportable diseases: MedlinePlus Medical Encyclopedia. (n.d.). Retrieved from https://medlineplus.gov/ency/article/001929.htm

Sarofim, Saha, S. et al.. (2016, April 04). Ch. 2: Temperature-Related Death and Illness. Retrieved from https://health2016.globalchange.gov/temperature-related-death-and-illness

Vesset, D. (2018, May 11). Descriptive analytics 101: What happened? Retrieved from https://www.ibm.com/blogs/business-analytics/descriptive-analytics-101-what-happened/

Wanjala, C., & Kweka, E. J. (2016, October 14). Impact of Highland Topography Changes on Exposure to Malaria Vectors and Immunity in Western Kenya. doi:https://doi.org/10.3389/fpubh.2016.00227

Why Population Matters to Infectious diseases and HIV/AID. (n.d.). Retrieved December 28, 2018, from https://pai.org/wp-content/uploads/2012/02/PAI-1293-DISEASE_compressed.pdf