AIT 580 – Final Project

500 Cities:
Local Data for Better Health, 2019 release

Introduction:

This is the complete dataset for the 500 Cities project 2019 release. This dataset includes 2017, 2016 model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (9). Data were provided by the Centers for Disease Control and Prevention (CDC), Division of Population Health, Epidemiology and Surveillance Branch. It represents a first-of-its kind effort to release information on a large scale for cities and for small areas within those cities. It includes estimates for the 500 largest US cities and approximately 28,000 census tracts within these cities. These estimates can be used to identify emerging health problems and to inform development and implementation of effective, targeted public health prevention activities. Because the small area model cannot detect effects due to local interventions, users are cautioned against using these estimates for program or policy evaluations.

Project and Its Scope:

The goal for this project is to clean the raw data and analyze the cleaned data to come up with some significant insights about the health of the population according to region and also portray the regions that have high health risk factor.

Nature of the Data Curation:

- Who (company, agency, organization) collected the data?

This data is collected by U.S Department of Health and Human Services and the dataset is given public access. The same organization also has another 1386 datasets in all the domains combined. The project of collecting the health data was funded by the Robert Wood Johnson Foundation in concurrence with Centers for Disease control and Prevention. (19ht10)

The United States Department of Health & Human Services is a cabinet-level department of the U.S. federal government with the goal of protecting the health of every single American and providing necessary human services. [hhs.gov]

- Why did they collect the data?

The data is collected so that the U.S government can evaluate and analyze the emerging health issues faced by the individuals and by analyzing this data the government can notify the disease control centers. so that, preventive measures can be taken to prevent huge scale spread of diseases and analyze major risk factors associated with particular deadly diseases, which aid in improving the health of people. (19ht8)

- Is the nature of the data given the purpose of the data collection?

The nature of the data provided best suites the purpose of data collection i.e. to bring out essential insights through the health data which contains both qualitative and quantitative values for each record.
The dataset needs to be preprocessed as there are Null values present and some insignificant columns for analysis.

- Leveraging the data

The data has insignificant columns which cannot be considered as parameters in our analysis, so the insignificant attributes have to be removed in the preprocessing and also the null values and outliers have to replaced. Coming to pros, the data has many useful attributes through which relationship between attributes can be developed.

Questions:

• Is there any privacy, quality, or other issues with this data?

Privacy:

The dataset is an open source public accessible data and can be accessed by any individual.

Quality:

The data set has few columns which are insignificant and have missing data.
There are missing data values in the dataset
After preprocessing, the data can be used to bring out actionable insights through the data attributes.
Other Issues:
The missing values in the dataset might cause inaccuracies and variability in the outcomes.

• Who can benefit from your data analysis?

The Population of united states can be benefited from this project as the health data consists of us records, the project is concerned to bring out conclusions that can improve the health of us.

Questions To Be Answered:

- What are the questions of your interests that can be answered through the data that you chose?

o the states with highest number of records?-the above questions will find the states with highest number of records which may lead to answer if the same states have highest unhealthy records and health risks.
o The cities with highest unhealthy category records?
-through the above question we can pull out the cities with highest health risks and cross check them with the states they are in and if there is any relation this result and the above question.
o Is there a correlation between crude prevalence data values and their corresponding population counts?
$H_{0:}$ Data values of Crude prevalence are correlated to their corresponding population count values.
$H_{1:}$ Crude prevalence and population count are not co-related.
o Linear Regression analysis between two variables which are crude prevalence high and crude prevalence low?
-To find if the variables are positively related or negatively related?
o Geospatial data visualization of cities with their frequency count?
-This question again links to the 1st two questions regarding link between highest number of record count and its link to the questions.
o Box plot analysis of Population count with respect to each category namely healthy records, unhealthy records and preventive records?
-It will let us know about the overall health scenario of the U.S , illustrating which category consists of how much population?

Requirements and Resources needed:

Hardware Resources:

- Processor – Intel Core i5
- Processor Speed – 2.7GHz
- RAM – 8GB
- System Type: 64-Bit MacOS

Software Resources:

- Rstudio – Data preprocessing, Transformation, and visualization.
- Tableau – Data visualization.

• What kinds of pre-processes were needed to make use of the data, and why?

The dataset contains NULL values which have to be replaced with '0' and there are columns which have no values and need to be removed and also some other insignificant column/attribute needs to be removed in order to avoid errors in the outcomes and to reduce unnecessary processing of data.
The target dataset is a cleaned dataset which can be used for data analysis and there will no or less occurrences of inaccuracies compared to raw dataset.

Descriptive Analysis:

• Briefly describe the dataset

The Dataset is massive with around 81,000 rows and 24 columns making it 1944000 tuples and also file size being 235.2 MB. The data is health data of 500 U.S cities. Its attributes being as follows:

StateAbbr: The attribute contains data of the state the record belongs to. This is a nominal data.
StateDesc: This attribute is containing the states which are mentioned in descending order. So, this data is nominal data. Example: Alabama, Alaska.
CityName: This attribute is containing city names from which the data has been collected. So, this data is a nominal data. Example: Birmingham, Abilene.
GeographicLevel: This attribute is containing the geographic place names from where the data has been collected. So, this data is a nominal data. Example: US, CITY.
Datasource: This attribute is containing the sources from where the data has been collected. So, this data is a nominal data. Example: BRFSS.
Category: This attribute is containing the types of categories on which the data has been collected. So, this data is a nominal data. Example: Prevention, Health Outcomes.
Measure: This attribute is containing the reasons for the health issues on which the data has been collected. So, this data is a nominal data. Example: Arthritis among adults aged >=18 Years, Binge drinking among adults aged >=18 Years.
DataValueTypeID: This attribute is containing the data value type Id of the data which has been collected. So, this data is a nominal data. Example: AgeAdjPrv, CrdPrv.
Data_Value_Type: This attribute is containing the data value types of the data which has been collected. So, this data is a nominal data. Example: Age-adjusted prevalence, Crude prevalence.
Data_Value: This attribute is containing the data values of the data which has been collected. So, this data is an Interval data. Example: 14.6, 11.6.
Low_Confidence_Limit: This attribute is containing the value of low confidence limit of the data which has been collected. So, this data is an Interval data. Example: 14.3, 11.3.

High_Confidence_Limit: This attribute is containing the value of High confidence of the data which has been collected. So, this data is an Interval data. Example: 14.9, 11.8.

Data_Value_Footnote_Symbol: This attribute is containing very few symbols which can be avoided during the process of visualization. Example: *, #.

Data_Value_Footnote: This attribute is containing information related to Data value footnote of the data which has been collected. So, this data is a Nominal data. Example: Data based on states available from the 2016 BRFSS, Estimates suppressed for a population less than 50.

Population: This attribute is containing the population count of the are from which the data has been collected. So, this data is an Interval data. Example: 3629,3992.

CategoryID: This attribute is containing information related to Category Id of the data which has been collected. So, this data is a Nominal data. Example: PREVENT, HLTHOUT.

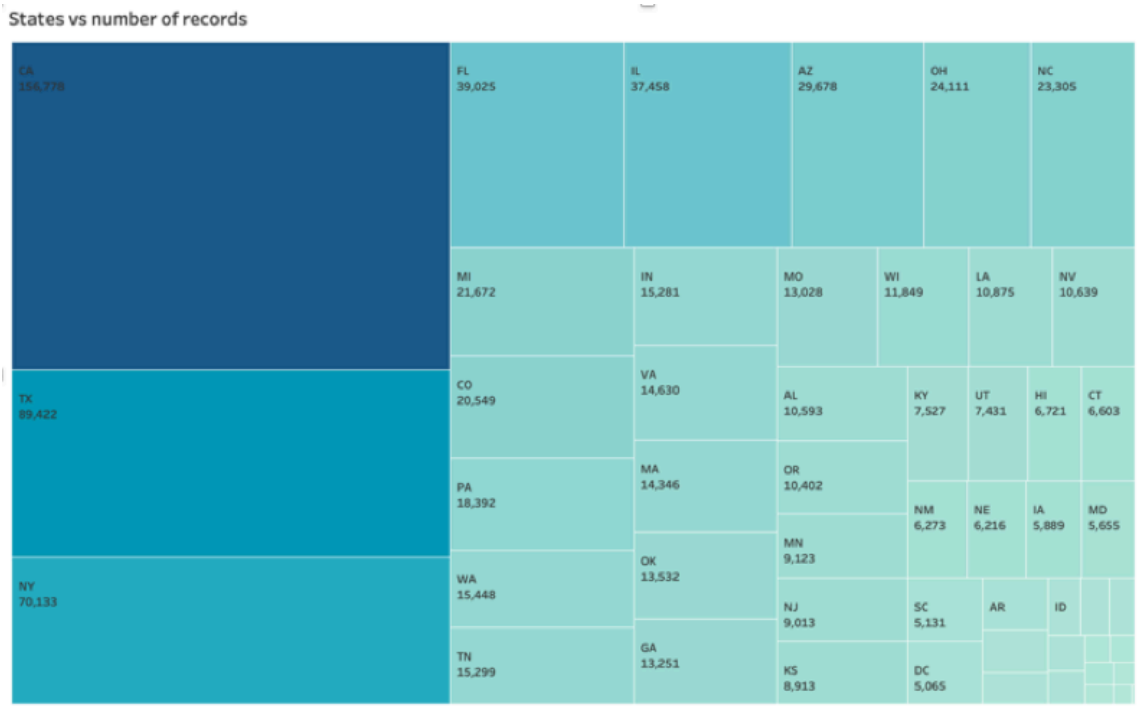
Descriptive Statistics:

```
> summary(pdata1)
      Year         StateAbbr           StateDesc              CityName
 Min.   :2016   CA     :146645   California:146645   New York     : 59264
 1st Qu.:2016   TX     : 82855   Texas     : 82855   Los Angeles  : 27829
 Median :2017   NY     : 68913   New York  : 68913   Chicago      : 22232
 Mean   :2017   FL     : 36231   Florida   : 36231   Houston      : 15444
 3rd Qu.:2017   IL     : 35973   Illinois  : 35973   Philadelphia: 10525
 Max.   :2017   AZ     : 26997   Arizona   : 26997   Phoenix      :  9612
                (Other):361735   (Other)   :361735   (Other)      :614443
     GeographicLevel   DataSource                    Category
 Census Tract:759349   BRFSS:759349   Health Outcomes    :353670
 City        :     0                  Prevention         :269629
 US          :     0                  Unhealthy Behaviors:136050



                UniqueID
 0107000-01073000100:    28
 0107000-01073000300:    28
 0107000-01073000400:    28
 0107000-01073000500:    28
 0107000-01073000700:    28
 0107000-01073000800:    28
 (Other)            :759181
                                                          Measure
 Arthritis among adults aged >=18 Years                       : 27210
 Binge drinking among adults aged >=18 Years                  : 27210
 Cancer (excluding skin cancer) among adults aged >=18 Years  : 27210
 Cholesterol screening among adults aged >=18 Years           : 27210
 Chronic kidney disease among adults aged >=18 Years          : 27210
 Chronic obstructive pulmonary disease among adults aged >=18 Years: 27210
 (Other)                                                      :596089
 Data_Value_Unit  DataValueTypeID                  Data_Value_Type     Data_Value
 %:759349        AgeAdjPrv:     0   Age-adjusted prevalence:     0   Min.   : 0.30
                 CrdPrv   :759349   Crude prevalence       :759349   1st Qu.:10.00
                                                                     Median :22.90
                                                                     Mean   :31.42
                                                                     3rd Qu.:46.00
                                                                     Max.   :95.70
```

```
 Low_Confidence_Limit High_Confidence_Limit Data_Value_Footnote_Symbol
 Min.    : 0.20       Min.    : 0.30          :759349
 1st Qu.: 8.80        1st Qu.:11.20         *:      0
 Median :20.70        Median :25.30         #:      0
 Mean   :29.67        Mean   :33.17         ~:      0
 3rd Qu.:43.20        3rd Qu.:49.20
 Max.   :94.60        Max.   :96.50

                                          Data_Value_Footnote PopulationCount
                                                      :759349   Min.    :   50
 Data based on states available from the 2016 BRFSS   :      0   1st Qu.: 2458
 Data not available for this state from the 2016 BRFSS:      0   Median : 3611
 Estimates suppressed for population less than 50     :      0   Mean   : 3786
                                                                  3rd Qu.: 4900
                                                                  Max.   :28960

                        GeoLocation       CategoryID         MeasureId
 (21.2620062174, -157.803375842):    28   HLTHOUT:353670   ARTHRITIS: 27210
 (21.2655150514, -157.817511576):    28   PREVENT:269629   BINGE    : 27210
 (21.2705966647, -157.781522276):    28   UNHBEH :136050   BPHIGH   : 27210
 (21.2710245012, -157.707137562):    28                    BPMED    : 27210
 (21.2713403723, -157.792698935):    28                    CANCER   : 27210
 (21.2719071793, -157.811002162):    28                    CASTHMA  : 27210
 (Other)                        :759181                    (Other)  :596089
    CityFIPS          TractFIPS                    Short_Question_Text
 Min.    :  15003   Min.   :1.073e+09   Annual Checkup          : 27210
 1st Qu.: 681666   1st Qu.:8.005e+09   Arthritis               : 27210
 Median :2622000   Median :2.608e+10   Binge Drinking          : 27210
 Mean   :2607002   Mean   :2.588e+10   Cancer (except skin)    : 27210
 3rd Qu.:4052500   3rd Qu.:4.011e+10   Cholesterol Screening   : 27210
 Max.   :5613900   Max.   :5.602e+10   Chronic Kidney Disease: 27210
                                       (Other)                :596089
>
```

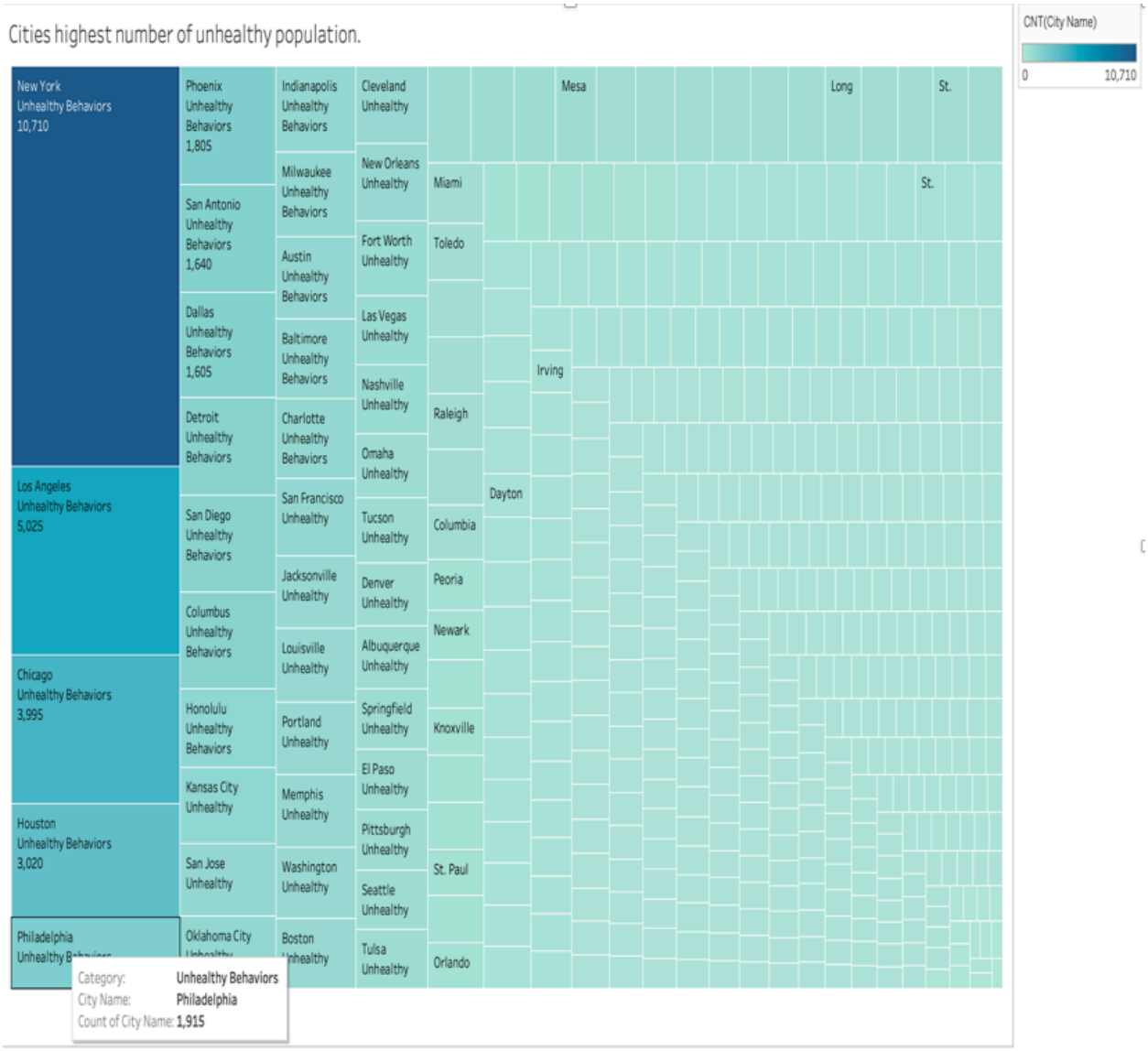
Results/Findings:

Tree map depicting the number of records each state contains:



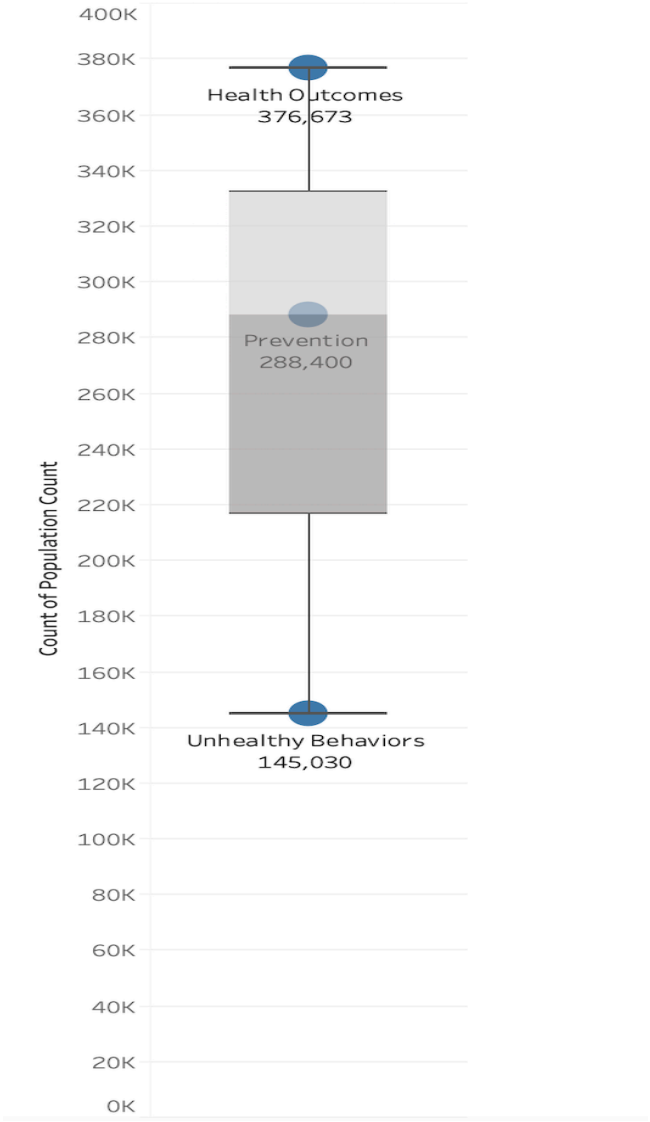
The above visualization depicts that the states California, Texas and New York has the highest number of records with count 156778, 89442, 70333 respectively. California being the state with highest number of health records, the health centers here have to be notified and necessary prevention steps have to be taken in order to make peoples life healthy.

Tree map Depicting cities with unhealthy records:



We find an answer to the question that the cities with highest number of unhealthy records are within the states which have highest total records. We can observe here that cities New York, Houston, Los Angeles have the highest unhealthy records which are in states New York, Texas, California respectively.

BOXPLOT of population frequency with respect to category:



Upper Whisker:	376,673
Upper Hinge:	332,536.5
Median:	288,400
Lower Hinge:	216,715
Lower Whisker:	145,030

The visualizations depicts the amount of population each category contains , through the above boxplot we can observe that majority population-376,673 belong to healthy category ,which is also the upper whisker, second highest comes the prevention category with 288,400 count and the last being unhealthy behavior category with 145,030 count , even though least it is a signification number as it is 18% of total population. So, we can say that the overall health of U.S is not in a highly risky situation as the number depicts.

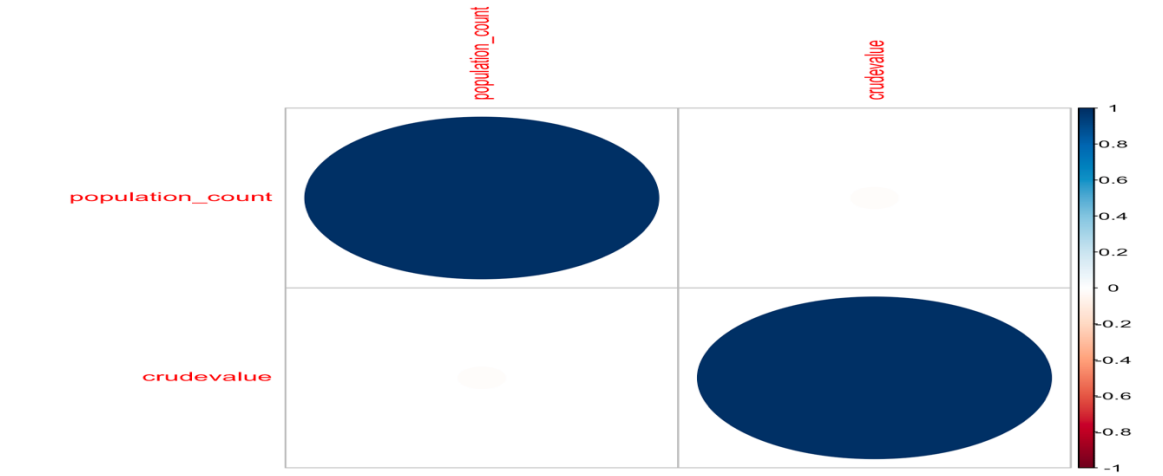
Hypothesis Testing:

H_0: Data values of Crude prevalence are correlated to their corresponding population count values.

H_1: Crude prevalence and population count are not co-related.

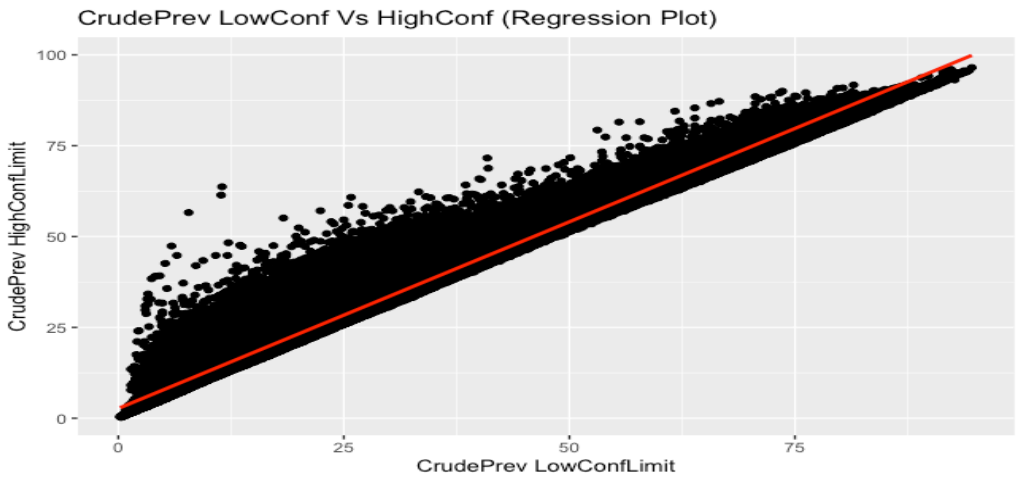
Correlation analysis between population count and crudevalue:

```
> cor.test(population_count,crudevalue)

        Pearson's product-moment correlation

data:  population_count and crudevalue
t = -14.905, df = 759347, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01935062 -0.01485354
sample estimates:
        cor
-0.01710217
```



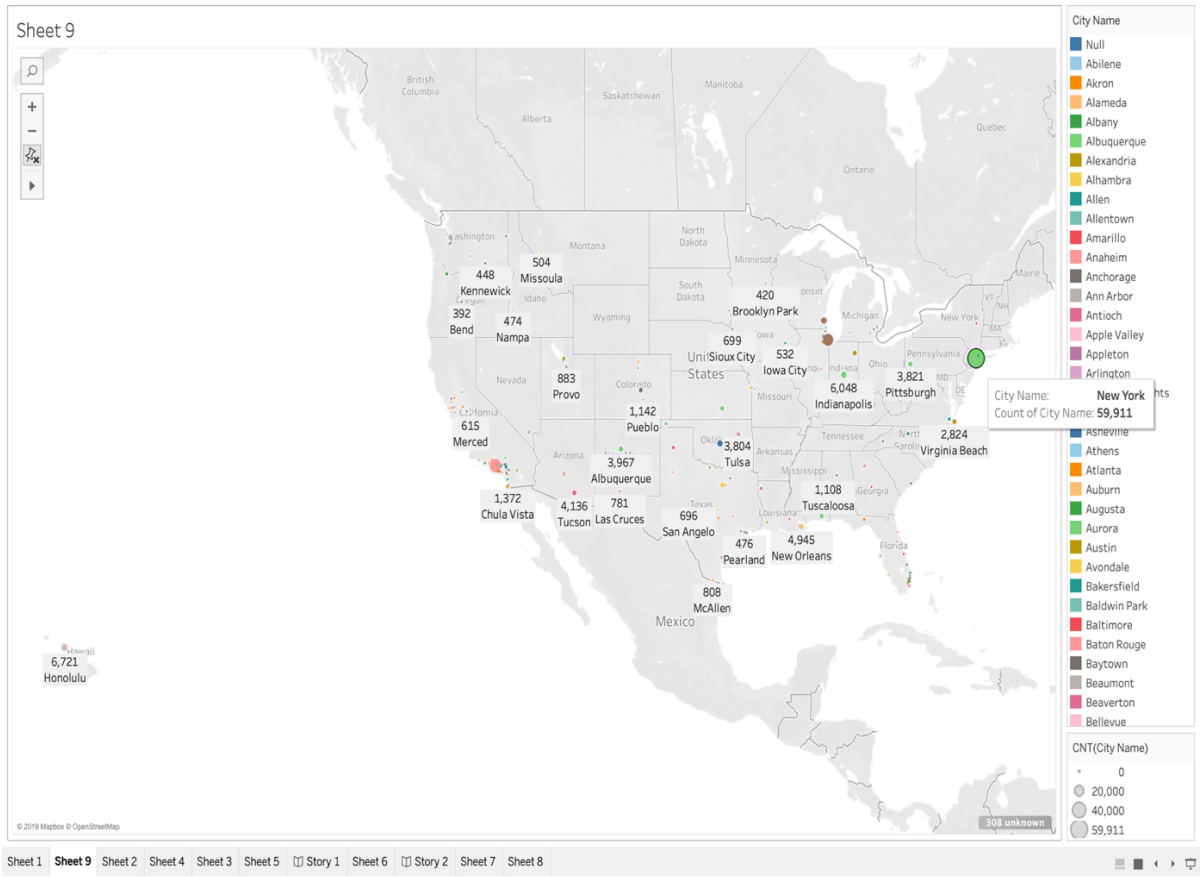
The above correlation hypothesis testing is done using pearson correlation method, the true correlation R-value obtained is -0.017 signifies alternate hypothesis is true i.e crude prevalence and population count are not correlated.

Regression analysis for CrudePrevalence LowConfLimit Vs HighConfLimit:



The visualization is done for Crude Prevalence where the Crude Prevalence has Low confidence limit and High Confidence Limit. In this Visualization, we can observe that most of the Crude Prevalence Low Confidence Limit and Crude Prevalence Low Confidence Limit values are in Linear form along the slope line. From this, it is also observed that the crude low and crude high are strongly and positively related with each other.

Geospatial data visualization of records frequency count with respect to cities:



The above visualization depicts the spread over U.S cities with Losangeles , Newyork, Houston having highest number of records.

Conclusion and Future Work:

The observations made through the analysis are the states with highest number of health records are California, New York, Texas etc. The cities with highest frequency of records are New York being 1st and Los Angeles coming 2nd shown in second visualization.

The 3rd visualization depicts tree map with unhealthy records count in each city , the observations are that new York, Los Angeles and Houston have highest number of unhealthy population concluding that larger the frequency of health records with respect to each state or city larger are the unhealthy records from above visualizations.

Comparing between population count and the crude value, you will find that there is no correlation. But then there is a linear relationship between CrudePrevalence LowConfLimit Vs HighConfLimit.

We can conclude that there should be more precautions to be taken in states like California, new York and Texas as they have higher risks when it comes to health. The is a similar occurence coming to cities like

New York , los Angeles , Houston , Chicago who needs attention from U.S health services and alert the health centers regarding the health risks in these regions which can lead to reduction in unhealthy percentage which is at 18% of total population , which is a significant amount of population.

Explain/define Terms

Linear Regression:

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The

Crude Prevalence:

A crude rate is the number of new cases (or deaths) occurring in a specified population per year, usually expressed as the number of cases per 100,000 population at risk. (19ht11)

Reference:

Retrieved Dec 2019, from https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-fc759
Retrieved Dec 2019, from https://www.hhs.gov/
Retrieved Dec 2019, from https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-fc759
(Retrieved Dec 2019, from https://surveillance.cancer.gov/joinpoint/crude.html
Retrieved Dec 2019, from https://en.wikipedia.org/wiki/Linear_regression
(2019, 12 15).
Retrieved from https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-fc759
(2019, 12 15).
Retrieved from https://www.hhs.gov/