PATTERN RECOGINITION IN SPREAD OF TUBERCULOSIS

AIT 580 (Final Project)

ANIRUDH MYAKALA
VAIBHAV TRIVEDI
RAKESH GANGAVARAPU
VENKATA JAYANDRA KUMAR LADE

1 TABLE OF CONTENTS

2 INTRODUCTION

TB is one of the world's deadliest diseases. About one-third of the people in the world are infected with TB. Not everyone who is infected with TB becomes sick with the disease, but in 2011, nearly 9 million people around the world became sick with TB. This means that we are still prone to the deadly disease TB. The number of current and future TB cases is an important measure in setting priorities among TB prevention and control efforts. As foreign-born persons make up an increasing proportion of TB cases in the United States, predicting TB trends will help of this consideration will help making more meaningful and better predictions

Here we are keen to know which states have large TB case and what are the groups of different categories which are prone for an attack on tuberculosis in future. Other than that other aspect which are potential variables for the prediction of TB are also considered. This projects aims at providing some reasonable solutions to the minimize the chance of new Tb infections in Usa. Also, this project focuses more on providing one target group for whom the appropriate precautionary measures can be suggested to prevent the spread of the disease.

Various visualizations have been done using tableau for the exploratory data analysis purpose. And many interesting facts were observed which helped in deciding which model would be suitable for the future predictions.

3 DATA

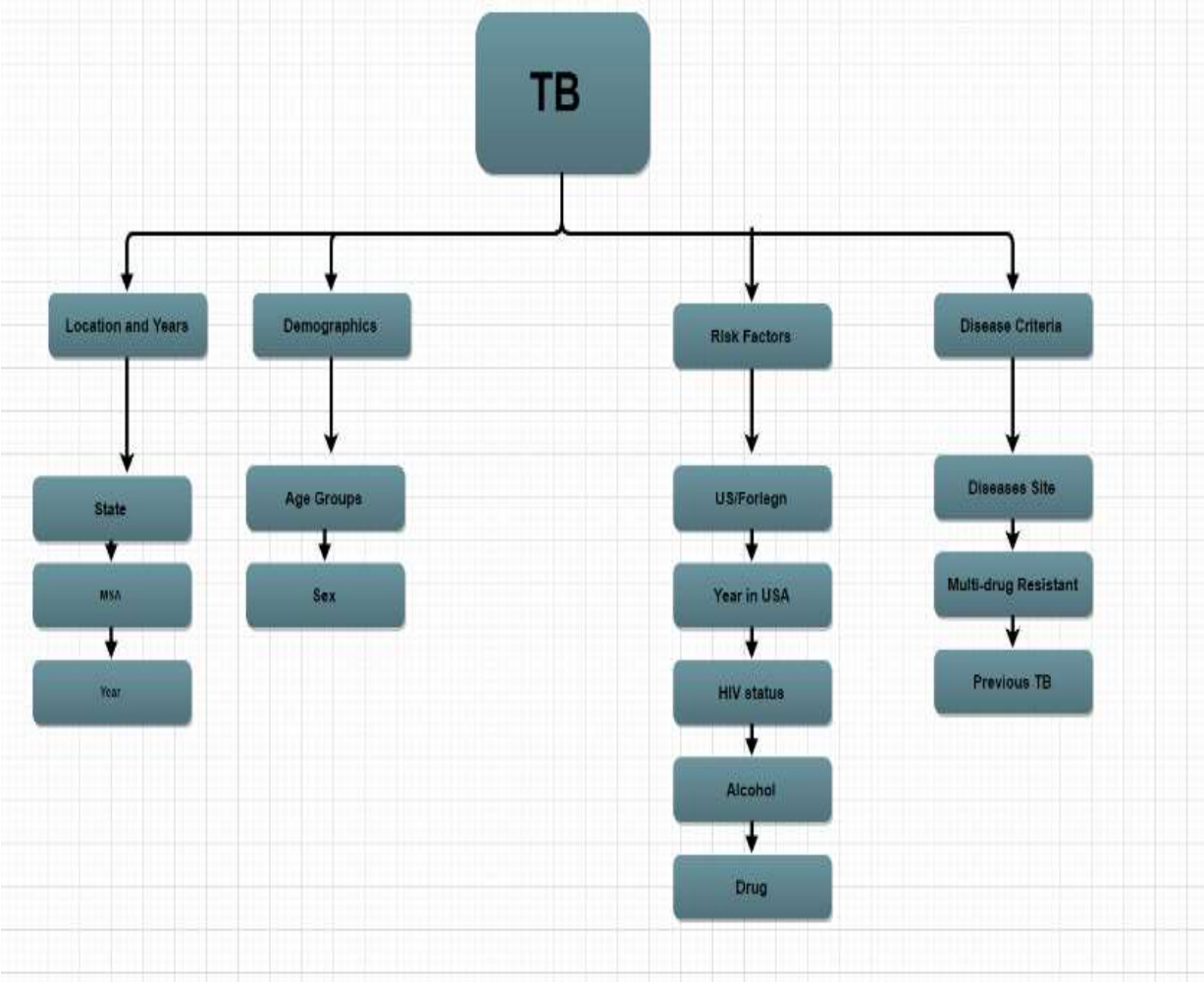
Data set contains 13 predictor variables. The target variable is total number of cases. The Online Tuberculosis Information System (OTIS) contains information on verified tuberculosis (TB) cases reported to the Centers for Disease Control and Prevention (CDC) by state health departments, the District of Columbia and Puerto Rico from 1993 through 2014. These data were extracted from the CDC national TB surveillance system.

Individual TB case information is collected at the local and state levels and transmitted electronically to CDC. The reporting areas are funded by DTBE through cooperative agreements to collect individual case data for surveillance purposes. Individual case data are collected using the RVCT form, which contains demographic and diagnostic information, the results of TB drug susceptibility testing, risk factors for TB disease, and treatment outcomes.

In accordance with CDC guidelines, confidentiality procedures were determined through careful examination of data by DTBE staff and state TB data providers. Aggregation, the grouping of continuous variables into specific intervals, is the main technique used by DTBE to protect the confidentiality of the national TB surveillance data.

3.1 Attribute Information

The data has a total of 4 categories i.e. Location and years, Demographics, Risk Factors and Disease.
Overall in all the categories there are 13 variables(Predictors).



3.2 Data Sources

The data was collected from these three sources

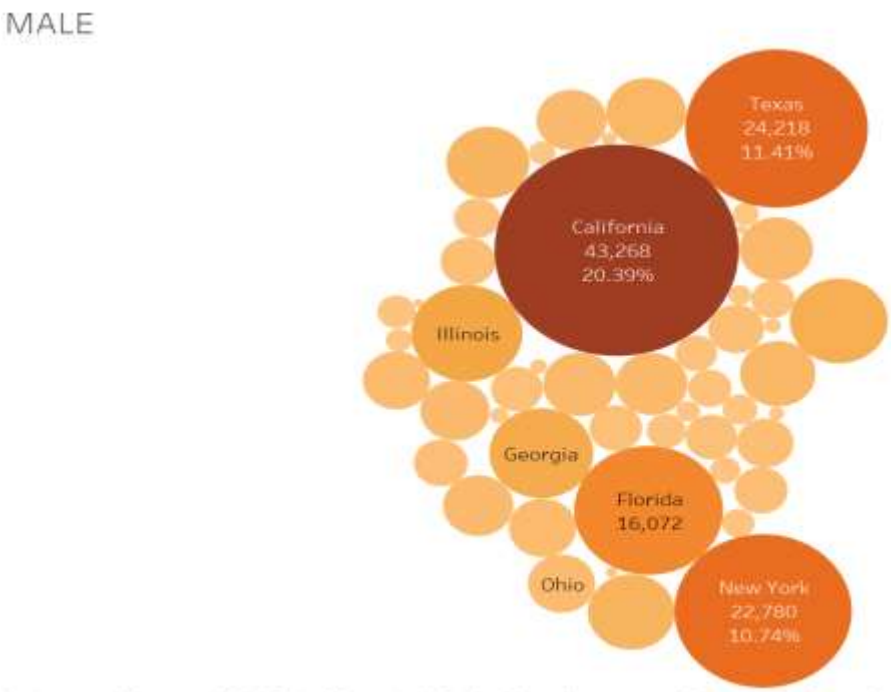
- ✓ Online Tuberculosis Information System (OTIS)
- ✓ National Tuberculosis Surveillance System
- ✓ United States and Centres for Disease Control and Prevention (CDC), Division of TB Elimination

4 DATA CLEANSING

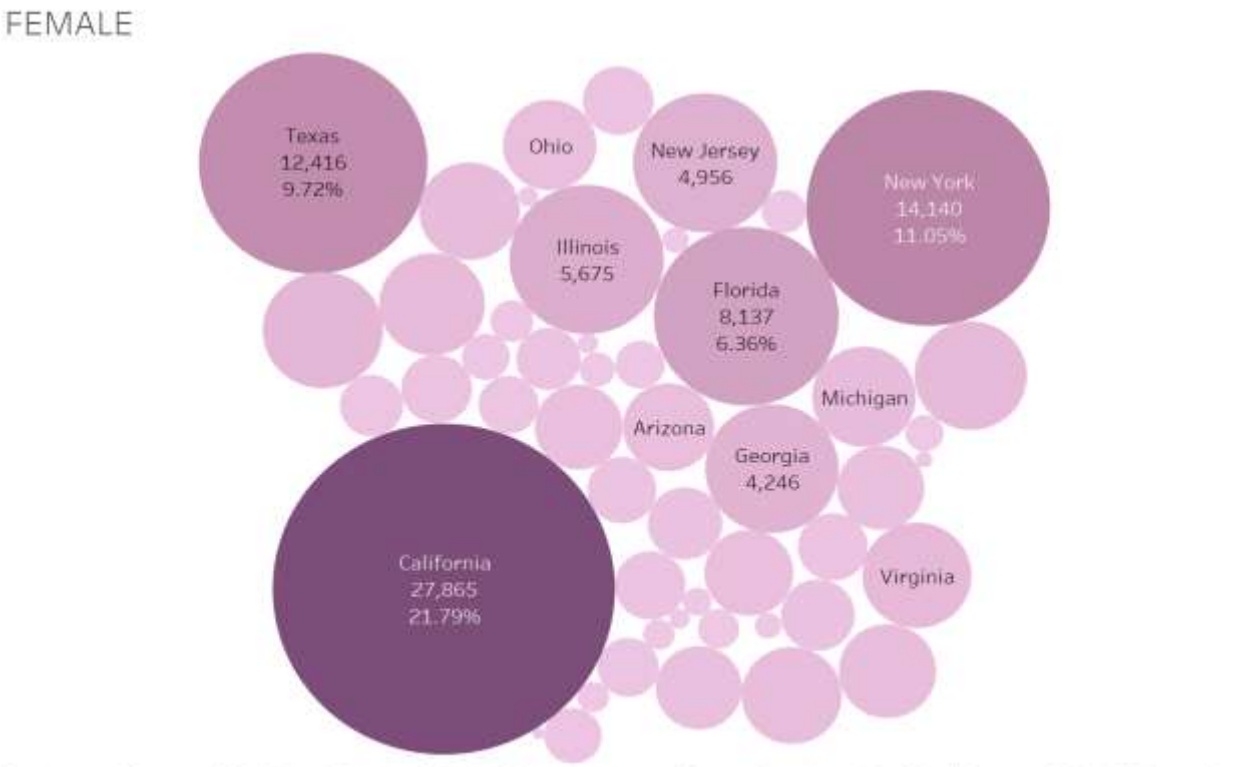
- ✓ The next data preprocessing included formatting the data using excel and making it into a readable csv format for loading in R.
- ✓ The data has lot of NA's and missing values and data integration was very critical
- ✓ Excel has been used for most of the cleaning.
- ✓ To normalized the data appropriate changes were done.
- ✓ Some of the unimportant predictors were removed.

5 VISUALIZATIONS

G1: Bubble chart drawn for no. of male cases in each state. Which shows This graph gives information which states have more male affected by TB



G2: Bubble chart drawn for no. of Female cases in each state. This graph gives information which states have more Female affected by TB.



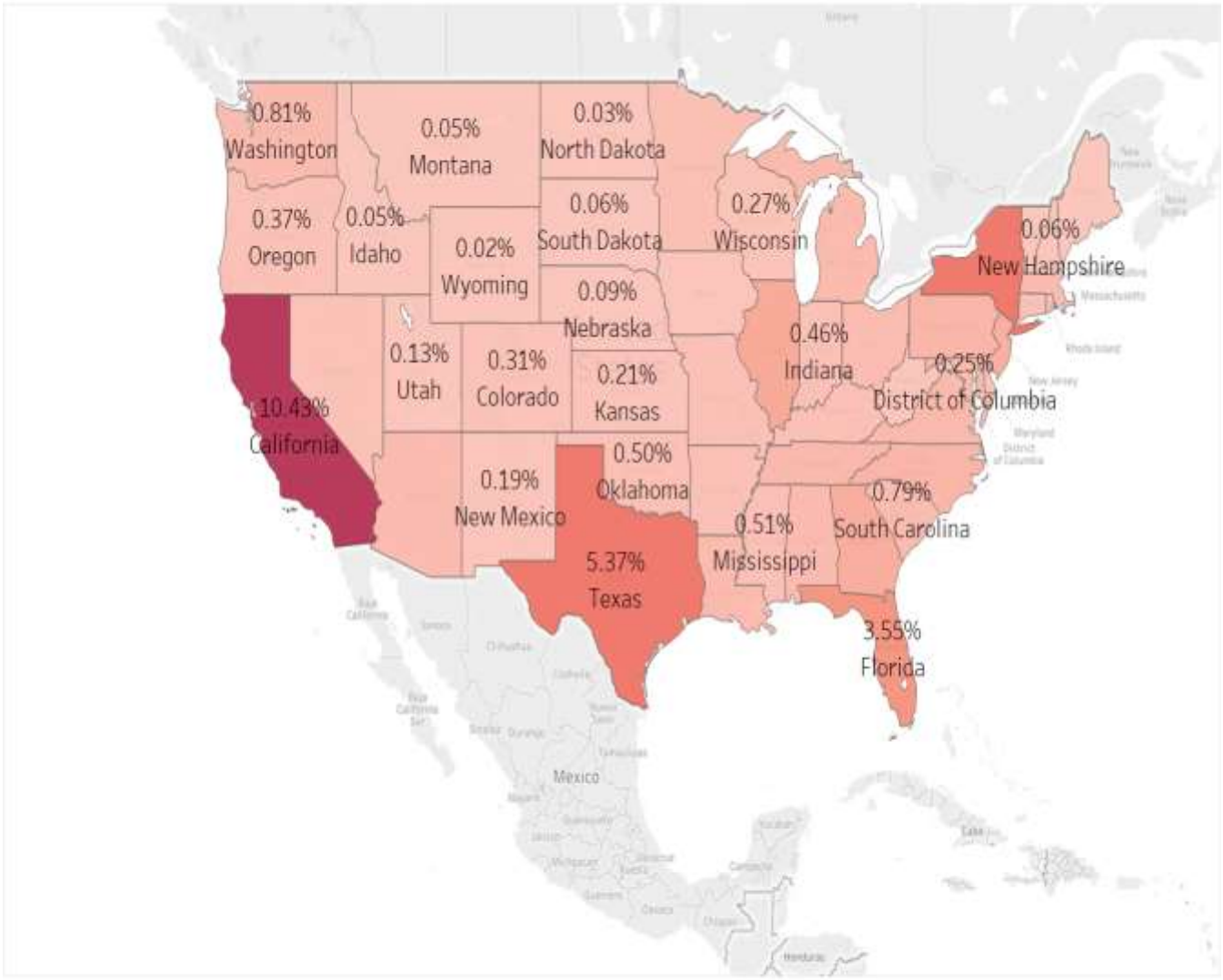
From the two above plots we can see that either for males or females California stand in the first place for total number of Tb cases. Also, we can see that we have more no. of case in females rather than male for rather same set of states. But a slight variation can be observed in states for the remaining positions i.e. for males New York has the next highest value while for males Texas holds the next highest value.

This is justified by the fact that these areas are the most urbanized cities and are the hub for globalization. So obviously, there will huge number of immigrants i.e. foreign born criteria will come into picture. Which is the most important predictor and we have already included that in the model.

G3: World-map color diverging with respective no. of cases in each state

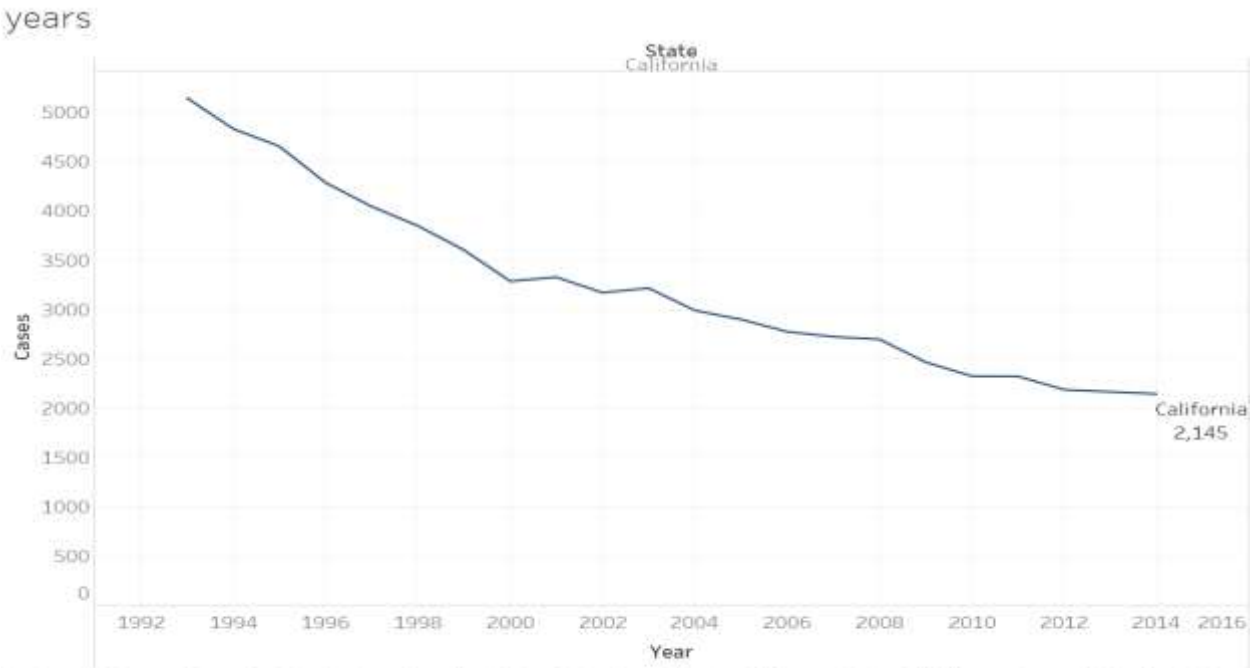
- ✓ By seeing these three graphs, we can say that California, Texas, Florida and New York are four major states where effected cases are more.
- ✓ So, now we are observing the trend how the no. of cases in the particular states are increasing/decreasing.

states



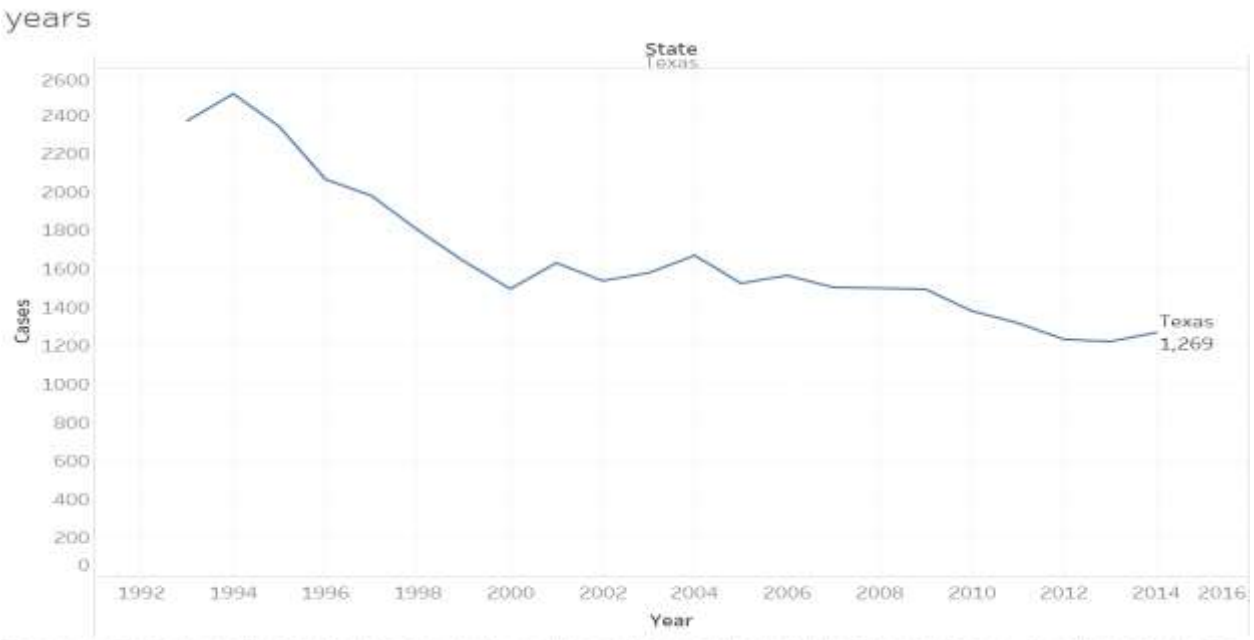
Map based on Longitude (generated) and Latitude (generated). Color shows % of Total Percent of Total. The marks are labeled by % of Total Percent of Total and State. Details are shown for State. The view is filtered on State, Latitude (generated) and Longitude (generated). The State filter has multiple members selected. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

G4: Line chart of California for total number of cases



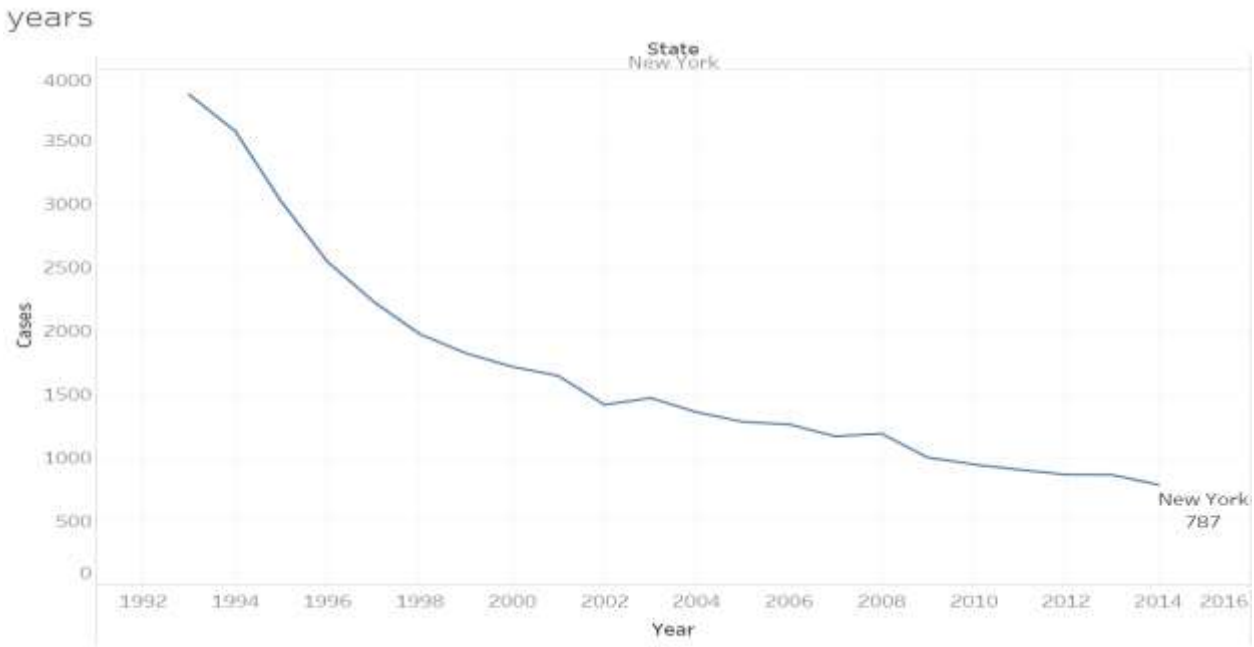
The trend of sum of Cases for Year broken down by State. Color shows % of Total Percent of Total. The marks are labeled by State and sum of Cases. Details are shown for State. The data is filtered on Year, which keeps 22 of 22 members. The view is filtered on State, which keeps California.

G5: Line chart of Texas for total number of cases



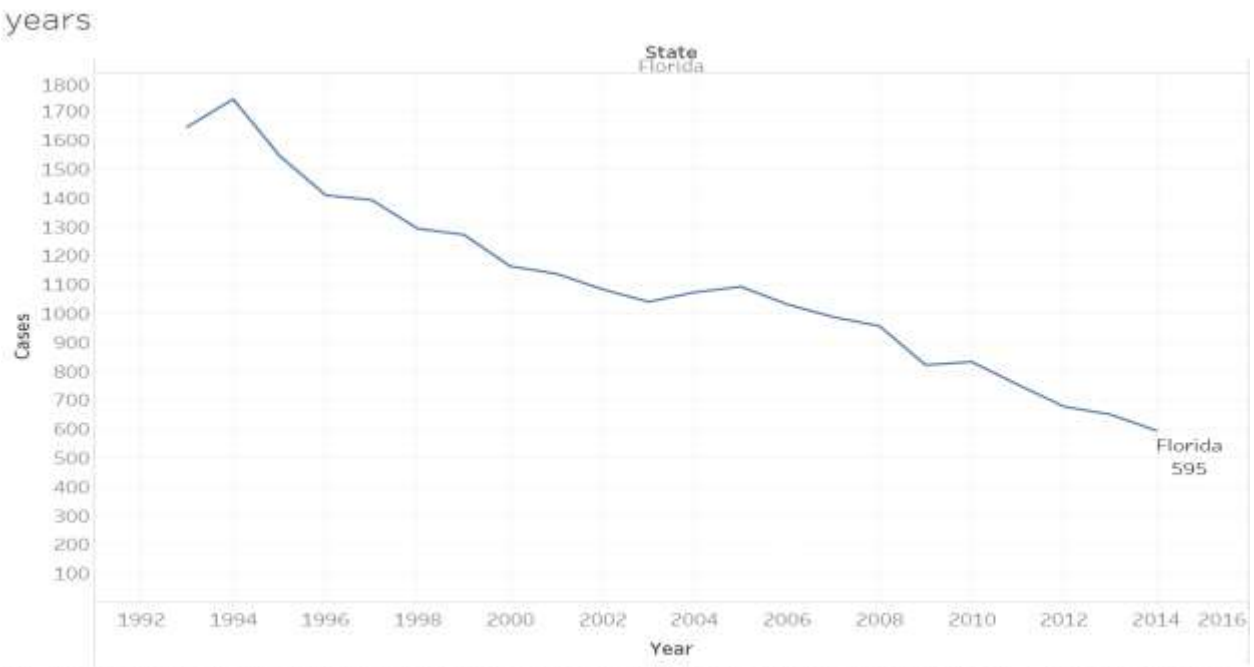
The trend of sum of Cases for Year broken down by State. The marks are labeled by State and sum of Cases. Details are shown for State. The data is filtered on Year, which keeps 22 of 22 members. The view is filtered on State, which keeps Texas.

G6: Line chart of Florida for total number of cases



The trend of sum of Cases for Year broken down by State. The marks are labeled by State and sum of Cases. Details are shown for State. The data is filtered on Year, which keeps 22 of 22 members. The view is filtered on State, which keeps New York.

G7: Line chart of New-York for total number of cases



The trend of sum of Cases for Year broken down by State. The marks are labeled by State and sum of Cases. Details are shown for State. The data is filtered on Year, which keeps 22 of 22 members. The view is filtered on State, which keeps Florida.

6 ANALYSIS

Our primary objective of this case study is to derive the most affected/ prone group of people to TB so that preventive measures can be taken on targeted groups. As mentioned earlier, our data is gathered from three different sources and integrated together using excel queries. We couldn't be sure about the correlation between our variables/ predictors as they are *categorical* while target/ response is *numerical* (Number of TB cases).

As a solution, we decided to use both Linear Regression and Time series Analysis for this scope. However, these methods are used to achieve different objectives as explained below.

6.1 Linear Regression

This statistical method is used to get following results.

- Correlation between variables
- Coefficient Estimates
- Significant Predictors

Now, we tried to get all the variables in single model but the data quality is not suitable for that. So, we decided to build three separate models based on the classification mentioned earlier. This makes sense as these categories are very different and hence independent from each other. The models with their predictors are shown below.

Model 1 – Demographics Factors

It contains following variables.

- Age Group
- Sex
- Race/ Ethnicity

The resulting coefficient estimates are in the below table.

Coefficeints	Estimates
Intercept	51.12
Year	-0.025
5-14 Years	-0.32
15-24 Years	0.86
25-44 Years	2.04
45-64 Years	1.94
65+ Years	1.61
Asian	2.71
Black/African	3.12
Hispanic/ Latino	3.15
Multiple Race	-1.21
Native Hawaian	-0.95
White	2.37
Male	0.25

We can see that the 25-44 has the highest estimates among age groups. It means that these people have more number of cases than others. Similarly, we can conclude that Black and Hispanic people are more prone to the disease.

Interesting observation is for Multiple Race group. It has negative coefficient showing that hybrid race has less chances or in other words safer than other. Possible reason is DNA mutation that increases biological immunity.

Model 2 – Risk Factors

Risk predictors are:

- US/Foreign Born
- Years in US
- HIV Status
- Alcohol Use
- Drug Use

Resulting Coefficient Estimates table:

Coefficients	Estimates
Intercept	-3.09E+03
Years	1.507
US Born	8.90E+01
1-4 Years in US	2.66E+01
5-14 Years in US	5.917
15+ Years in US	-5.90E-02
HIV Negative	2.83E+01
HIV Positive	-1.30E+02
Alcohol Use	-1.14E+02
Drug Use Injecting	1.14E+02

The above table shows that foreign born people living in US for less than 4 years are more prone to TB than others. It infers that immigrants may carry TB from outside. Here noticeable variable is HIV Positive which has negative estimates. However, in real world this is not the case. Probably it's because bad data (missing values) that we tried to eliminate.

Comparison

Model	R- Squared	p-value	Anderson Darling
Sex	0.9859	0.2005	0.49757
Age	0.9906	0.2187	0.48876
Race	0.9037	8.42E-14	5.5731
Demographics	0.9052	6.09E-14	5.6585
Disease Site	0.9932	0.1402	0.56209
Multidrug Resistant	0.994	0.0006142	1.5009
Previous TB	0.9986	0.1643	0.53169
Disease Criteria	0.9491	0.0004554	1.5793
US/Foreign Born	0.5953	0.2824	0.43747
Years in US	0.9023	0.1431	0.56046
HIV	0.8347	0.9754	0.13623
Alcohol	0.9597	0.04082	0.77409
Drug-Injecting	0.9863	0.4127	0.36869
Drug- Non-Injecting	0.9893	2.30E-09	3.6412
Risk Factors	0.5223	2.20E-16	28.978

The above table explains caparison between our model. As we can see that first two models i.e. Demographics Factors and Disease Criteria Factors have very unrealistic R-squared value. That shows overfitting model with actual data. However, they passed normality test with very low p-values.

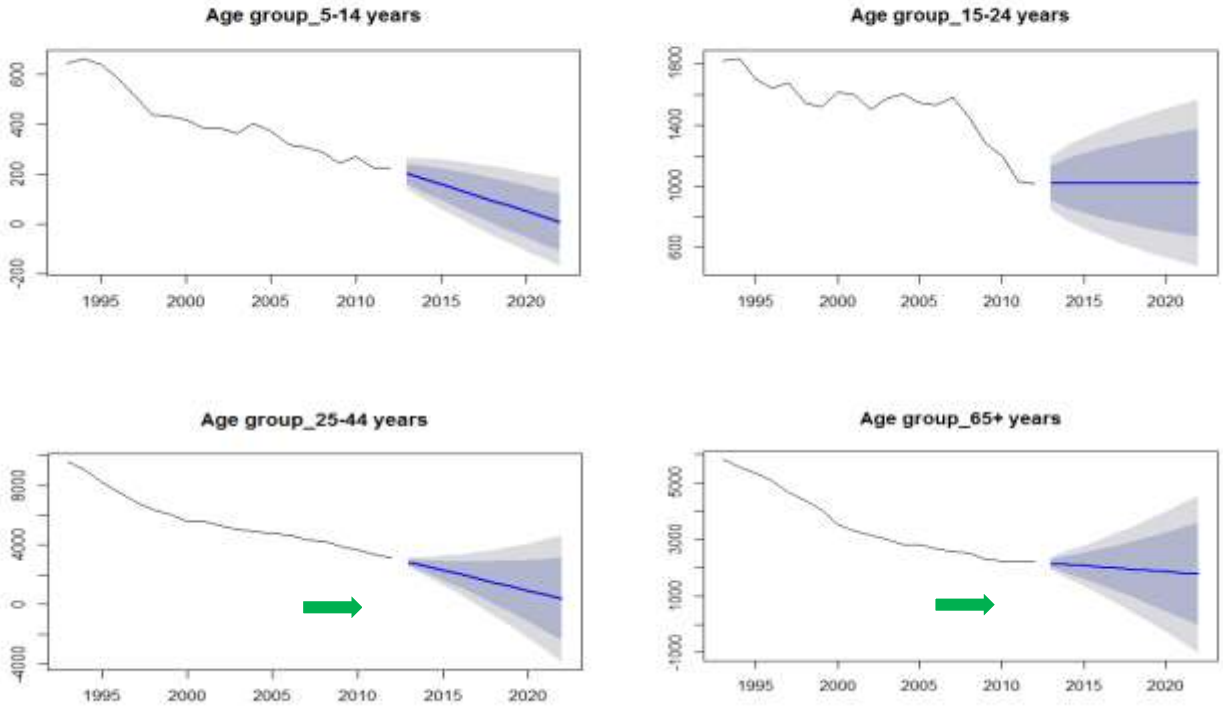
Model 3 with Risk factors is the best fit model among all according R-squared value.

6.2 Time Series Analysis

Now we have significant predictors among all categories, we decided to predict future trend with Time Series Analysis model. As, we have a long span of historical data (1993 to 2014), we are predicting values up to year of 2020. This would give authorities enough time to derive effective policies and action plans.

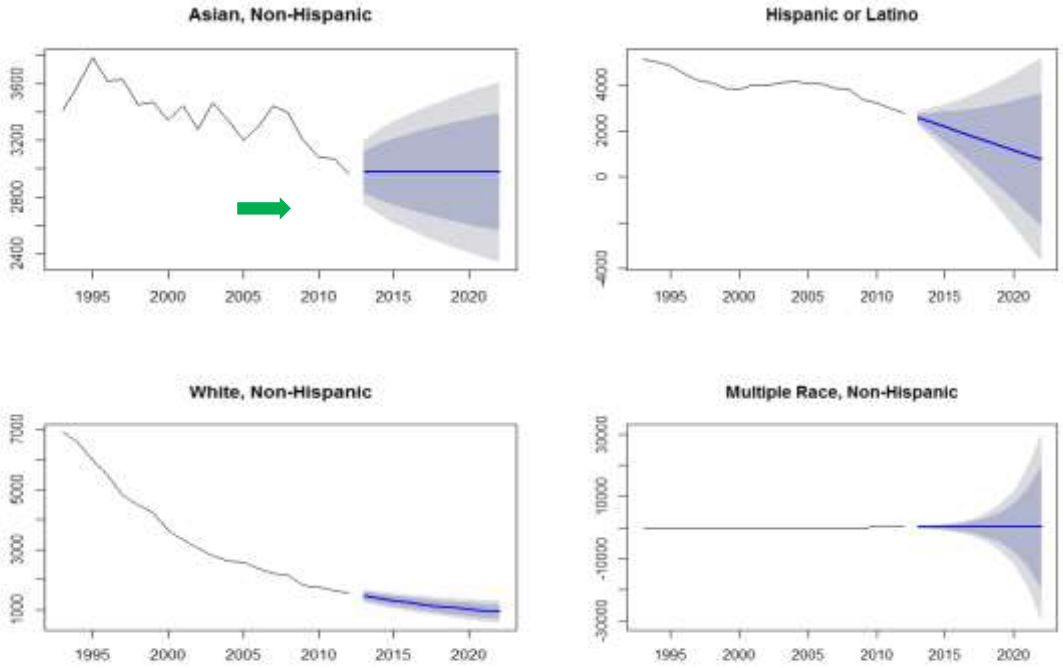
We are considering data up to 2010 to build model and data from 2011 to 2014 is used to validate our model. Model is built using 80% and 95% Confidence Interval (CI).

Age Groups

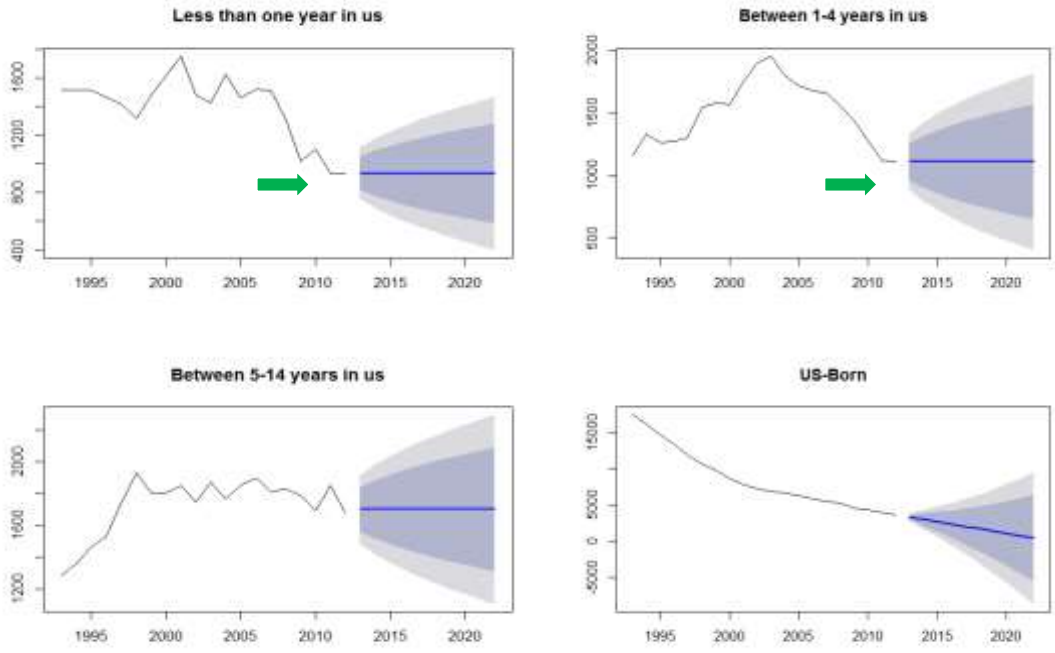


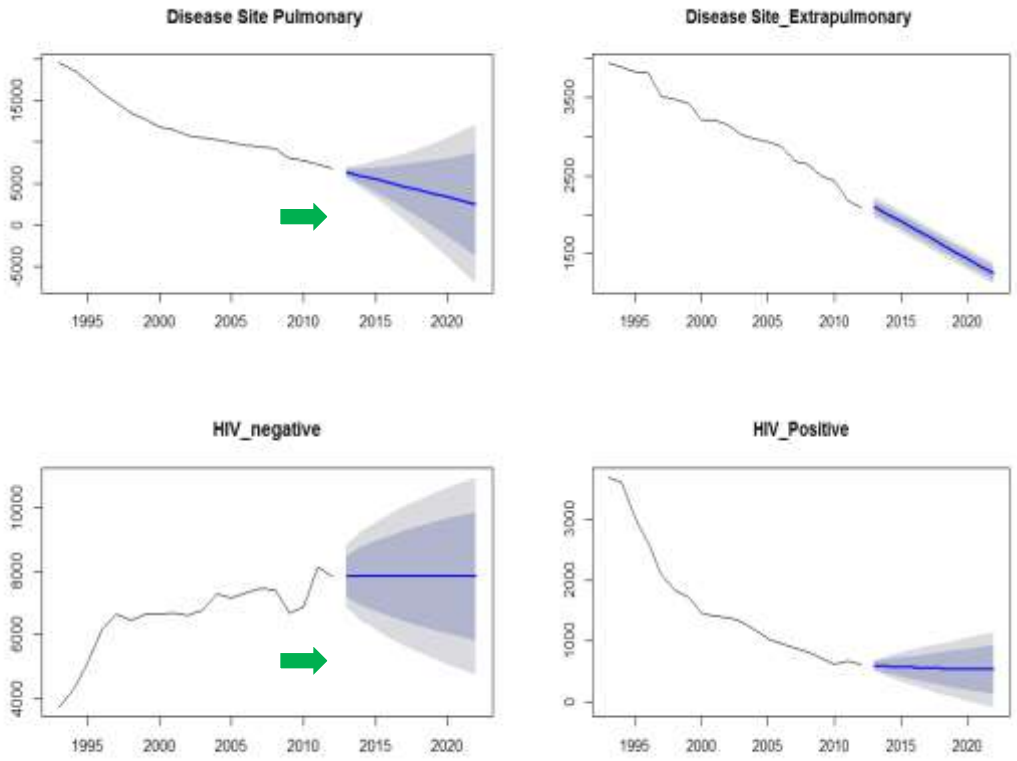
The above plot shows that people with age group 25-44 and 65+ have high number of cases in next five years.

Similarly, below plot gives Asian and Hispanic group as the most targeted people. Here, other noticeable observation would be for Multiple Race people. It shows zero fluctuation on the trend line, validating our estimates derived in Linear Regression model.



The same inferences can be derived from below time series plots. The final idea is given in conclusion section.





Here it is clear that pulmonary tuberculosis is very much prevalent.

7 CONCLUSION

Finally based the regression coefficients and time series analysis we conclude that for the following categories specified groups are more prone to get infected with tuberculosis

Category	Group
Age group	25-44
Race	Asian and Hispanic
Years in USA	Up to 4 years in USA
Risk factor	Pulmonary

The results are very much expected and practical. We can see that 25-44 is most prone age group as people of this age group move a lot which exposes them to bad air and can transmit Tb infection easily. Also from the exploratory data analysis it was proved at the immigrants who have just arrived in the Usa are more prone to the Tb Infection. And the last pulmonary playing an important role because pulmonary implies an infection to lung which is easily transmitted.

8 REFERENCES

- Woodruff RSYelk, Winston CA, Miramontes R (2013) Predicting U.S. Tuberculosis Case Counts through 2020. PLoS ONE 8(6): e65276. doi:10.1371/journal.pone.0065276
- Binkin NJ, Vernon AA, Simone PM, McRay E, Miller BI, et al. (1999) Tuberculosis prevention and control activities in the United States: an overview of the organization of tuberculosis services. Int J Tuberc Lung Dis 3(8): 663–674.
- Dye C, Williams BG (2008) Eliminating human tuberculosis in the twenty-first century. J R Soc Interface 5(23): 653–662.
- Dye C, Garnett GP, Sleeman K, Williams BG (1998) Prospects for worldwide tuberculosis control under the WHO DOTS strategy. Lancet 352(9144): 1886–1891.
- https://www.cdc.gov/