CAPSTONE PROJECT- Recommendation System

MENTOR:

Dr. Srabashi Basu

Professor, Analytics & Quantitative Methods

Great Learning

TEAM MEMBERS:

Suraj Suresh

Aniket Roy

Girish Kumar

K Varshini

ACKNOWLEDGEMENT

We sincerely thank Dr. Srabashi Basu, Great learning for her guidance and encouragement in carrying out this project. Her timely advice, meticulous scouting, scholarly advice and scientific approach have helped us to a very great extent to accomplish this task.

We thank great learning faculties who have given their valuable suggestions regarding this project which were essential in preparing our work. We appreciate all those who have helped us complete our project.

The completion of this work as a part of the capstone project in Great Lakes Institute of Management gives us immense pleasure and would definitely be a milestone in our Data Science careers.

ABSTRACT

Insurance analytics involves analyzing data which contains the history of the death rate, age, average deaths, education, insurance provided by public, private or employee private coverage, birth rate , races etc. cancer mortality rate in the U.S and recommending the government to increase the premium of the top 50 counties where the incidence of cancer mortality is high.

CONTENTS

INTRODUCTION

INSURANCE ANALYTICS

- The goal of this project is to recommend to the government, the top 50 counties to target first in order to increase the premium.

SCOPE OF THE PROJECT

- The scope of the project is to increase its insurance premium for counties where incidence of Mortality caused by Cancer is high, which 50 counties should the government target first

DATA DICTIONARY

TARGET_deathRate: Dependent variable. Mean *per capita* (100,000) cancer mortalities

avgAnnCount: Mean number of reported cases of cancer diagnosed annually

avgDeathsPerYear: Mean number of reported mortalities due to cancer

incidenceRate: Mean *per capita* (100,000) cancer diagoses

medianIncome: Median income per county

popEst2015: Population of county

povertyPercent: Percent of populace in poverty

studyPerCap: *Per capita* number of cancer-related clinical trials per county

binnedInc: Median income per capita binned by decile

MedianAge: Median age of county residents

MedianAgeMale: Median age of male county residents

MedianAgeFemale: Median age of female county residents

Geography: County name

AvgHouseholdSize: Mean household size of county

PercentMarried: Percent of county residents who are married

PctNoHS18_24: Percent of county residents ages 18-24 highest education attained: less than high school

PctHS18_24: Percent of county residents ages 18-24 highest education attained: high school diploma

PctSomeCol18_24: Percent of county residents ages 18-24 highest education attained: some college

Group 11

PctBachDeg18_24: Percent of county residents ages 18-24 highest education attained: bachelor's degree

PctHS25_Over: Percent of county residents ages 25 and over highest education attained: high school diploma

PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree

PctEmployed16_Over: Percent of county residents ages 16 and over employed

PctUnemployed16_Over: Percent of county residents ages 16 and over unemployed

PctPrivateCoverage: Percent of county residents with private health coverage

PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance)

PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage

PctPublicCoverage: Percent of county residents with government-provided health coverage

PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone

PctWhite: Percent of county residents who identify as White

PctBlack: Percent of county residents who identify as Black

PctAsian: Percent of county residents who identify as Asian

PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian

PctMarriedHouseholds: Percent of married households

BirthRate: Number of live births relative to number of women in county

DATA PREPROCESSING

DATA SET CHOSEN:

- The final dataset at which we are working is the data which was aggregated from a number of sources including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov.

MISSING VALUE TREATMENT:

- The **feature**, which had more than **70 percent** of the missing values of the total records, was removed from the analysis.
- The remaining features which had missing values less than **20 percent** of the total observations, was imputed with state wise median value.

ANOMALIES IN THE DATA:

- The states of Kansas, Minnesota and Nevada had anomalies in the column AvgAnnCount and Incidence rate.
- They had repeating values in all counties of these states.
- Here the death rates were higher than the actual population of the county, which clearly intimidated about the anomaly in the data.
- To solve this issue, we considered these states along with its neighbors, and binned the data based on the population size of all the counties present.
- Followed by this, we found the median of AvgAnnCount of these bins and imputed this value in the improper counties of the 3 states (Kansas, Minnesota and Nevada).

OUTLIER TREATMENTS:

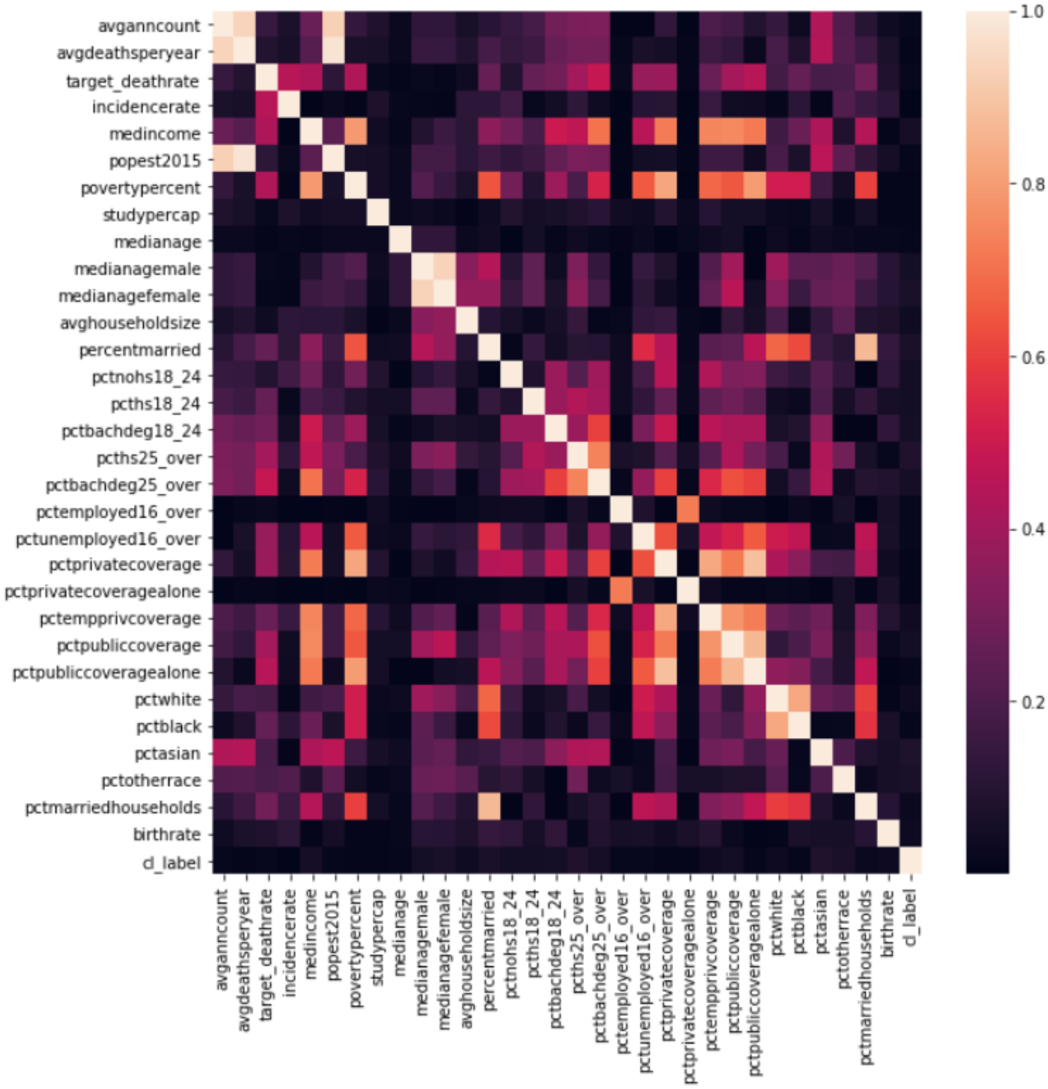
- Since, almost all the variables had extreme outliers we have chosen 3*IQR instead of 1.5*IQR and as each record is sensitive to the county of the state, the outlier treatment will not be suitable here

EXPLORATORY DATA ANALYTICS

INTRODUCTION:

- EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data.

CORRELATION AMONG FEATURES

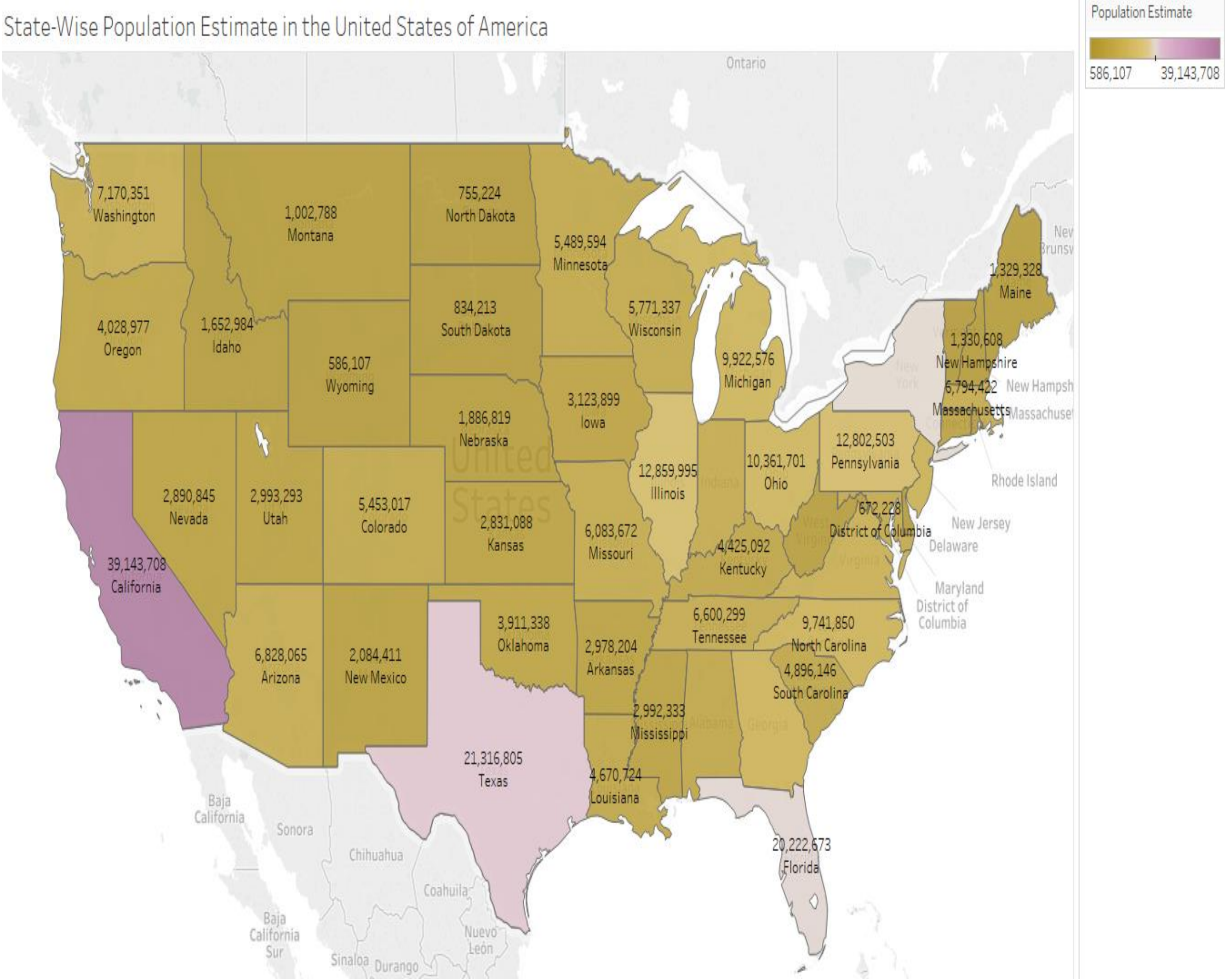


- The above correlation heatmap shows that there is multicollinearity among the independent variables.

	column	correlated	num_correlated
0	avganncount	[avgdeathsperyear, popest2015]	2
1	avgdeathsperyear	[avganncount, popest2015]	2
2	medincome	[povertypercent, pctbachdeg25_over, pctprivate...	6
3	popest2015	[avganncount, avgdeathsperyear]	2
4	povertypercent	[medincome, percentmarried, pctunemployed16_ov...	8
5	medianagemale	[medianagefemale]	1
6	medianagefemale	[medianagemale]	1
7	percentmarried	[povertypercent, pctwhite, pctblack, pctmarrie...	4
8	pcths25_over	[pctbachdeg25_over]	1
9	pctbachdeg25_over	[medincome, pcths25_over, pctprivatecoverage, ...	5
10	pctemployed16_over	[pctprivatecoveragealone]	1
11	pctunemployed16_over	[povertypercent, pctprivatecoverage, pctpublic...	3
12	pctprivatecoverage	[medincome, povertypercent, pctbachdeg25_over,...	7
13	pctprivatecoveragealone	[pctemployed16_over]	1
14	pctempprivcoverage	[medincome, povertypercent, pctprivatecoverage...	5
15	pctpubliccoverage	[medincome, povertypercent, pctbachdeg25_over,...	6
16	pctpubliccoveragealone	[medincome, povertypercent, pctbachdeg25_over,...	7
17	pctwhite	[percentmarried, pctblack]	2
18	pctblack	[percentmarried, pctwhite]	2
19	pctmarriedhouseholds	[povertypercent, percentmarried]	2

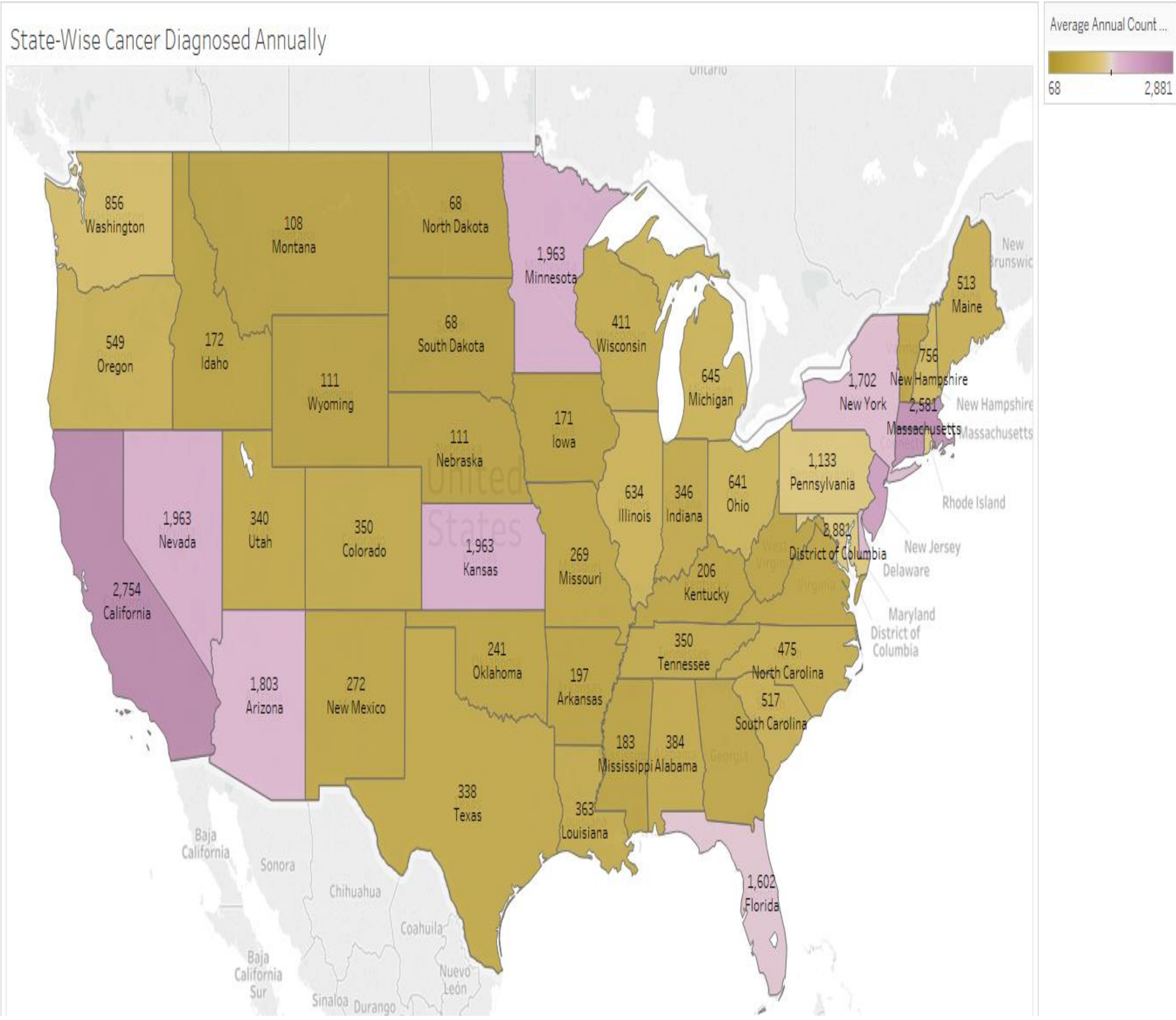
- The education related features are highly correlated to the insurance related features.
- People who are highly educated opt for private insurances.
- People who are not highly educated opt for public insurances.
- Also, people with high income are usually highly educated and therefore opt for private insurances. Whereas, people with low income are not highly educated and usually opt for public insurances.
- All these above inferences are relevant and make sense.

STATE-WISE POPULATION ESTIMATE IN THE US



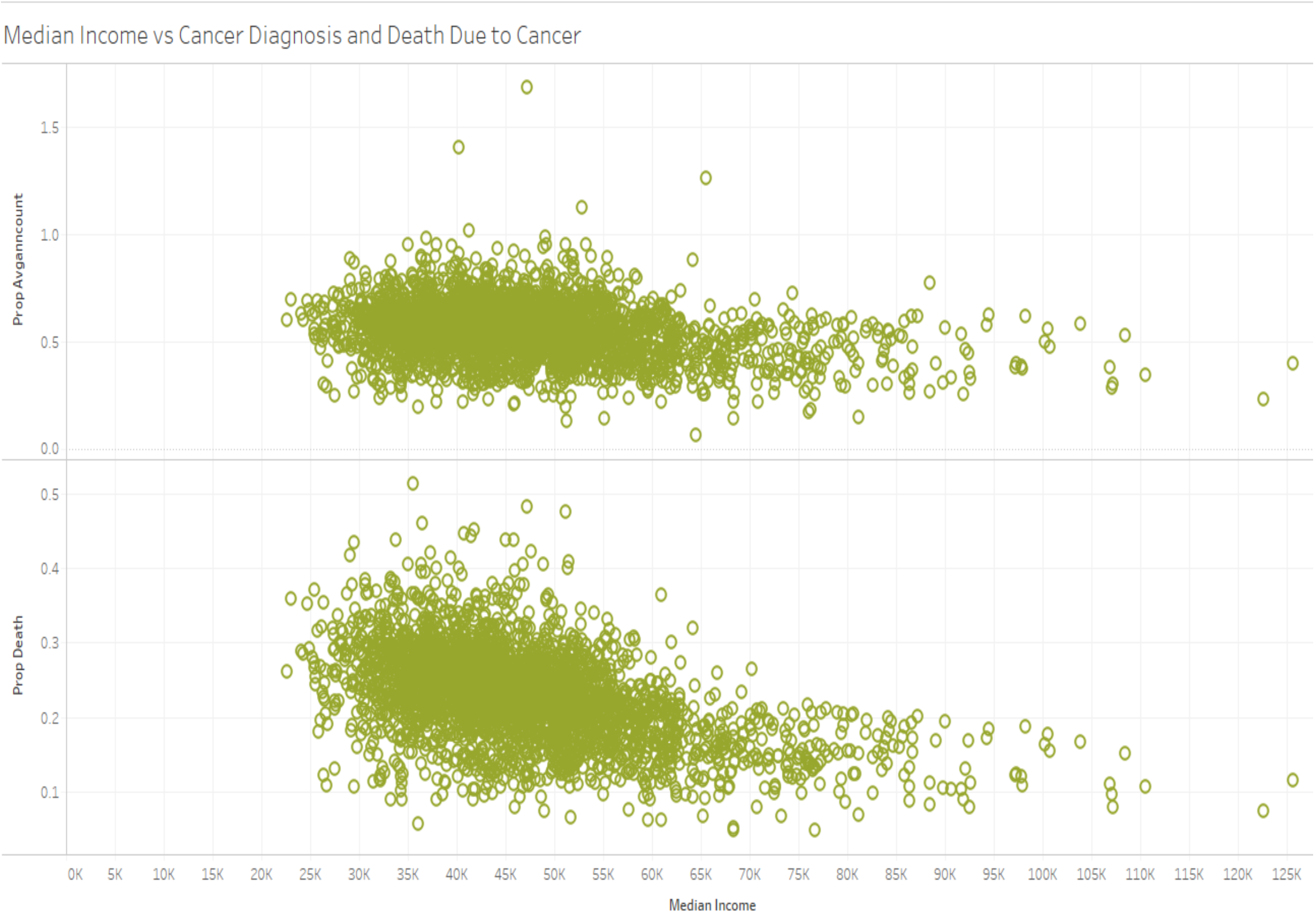
- The above map shows the Population Estimate of different states in the US.

STATE WISE CANCER DIAGNOSED



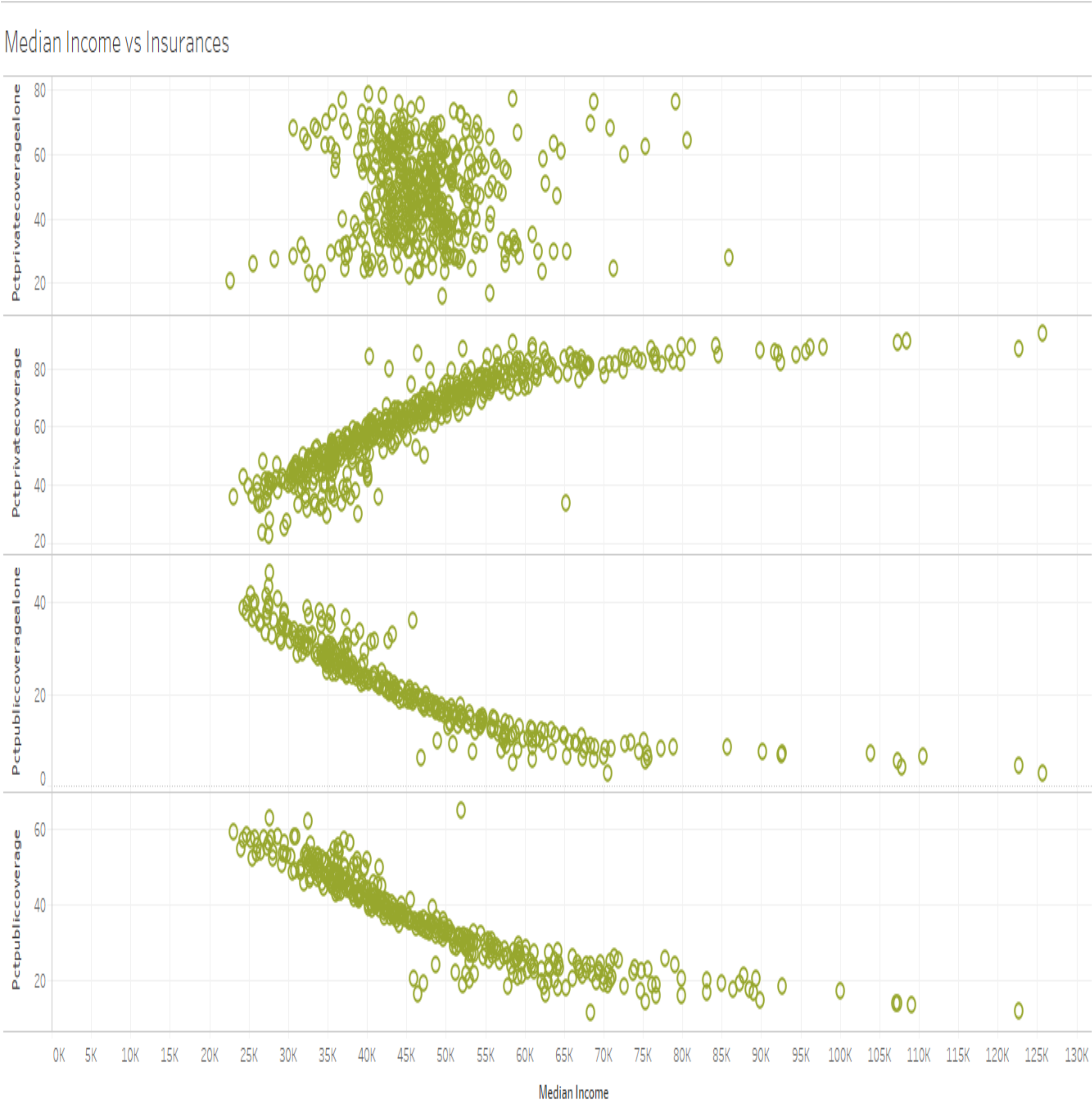
- The above is a map of the United State of America.
- It shows state-wise average count of people diagnosed with cancer annually.

INCOME vs CANCER DIAGNOSED & DEATH DUE TO CANCER



- The above scatter plots show the income and cancer diagnosis and death due to cancer.
- People with all levels of income are diagnosed with cancer.
- But, people with higher income tend to beat the cancer and the ones with lesser income cannot beat the cancer.
- This makes complete sense.

INCOME vs INSURANCES

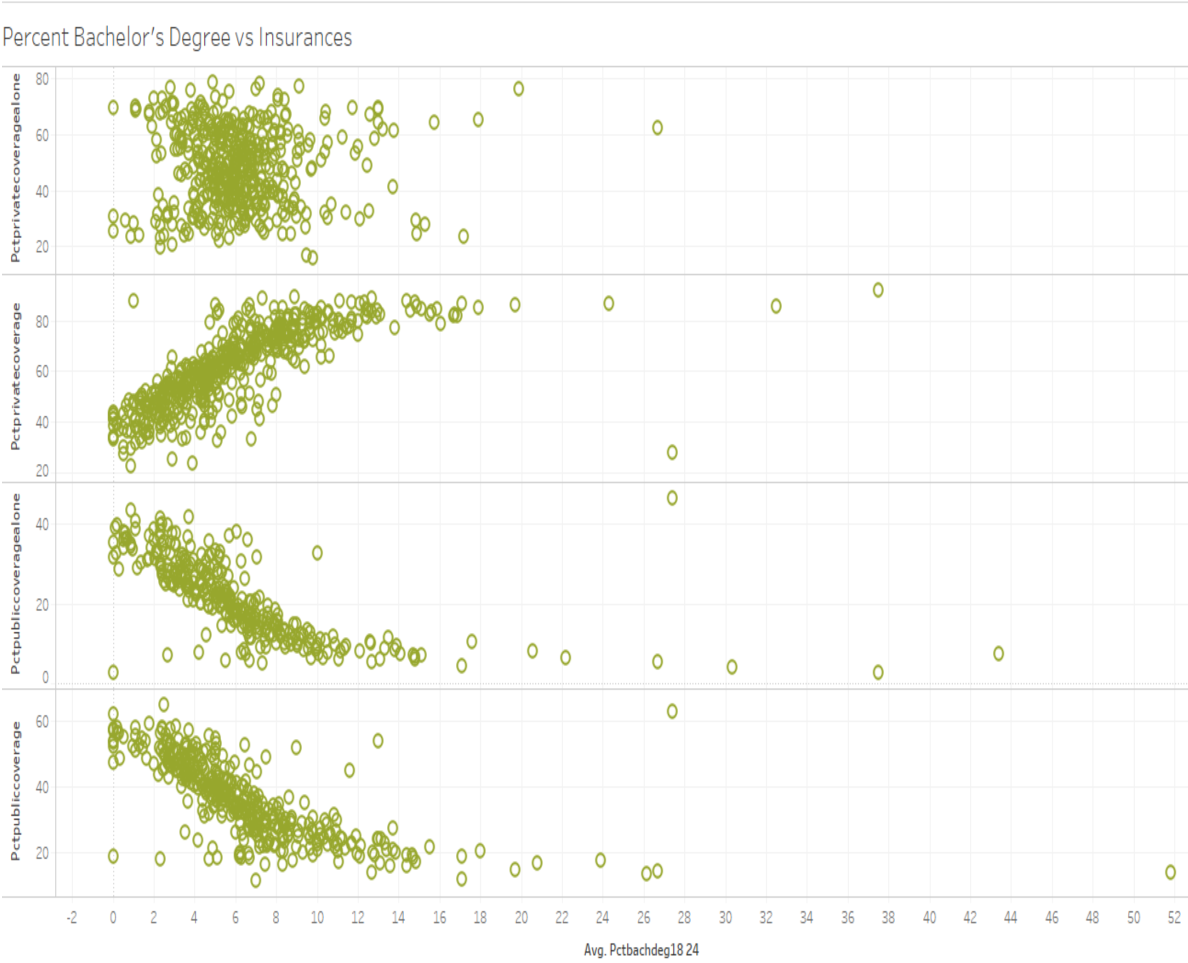


- The above scatter plots depict the effect of income on the insurances people buy.
- It clearly shows that higher the income, more private insurances are bought.

PERCENT UNEMPLOYED & EMPLOYED vs INSURANCES

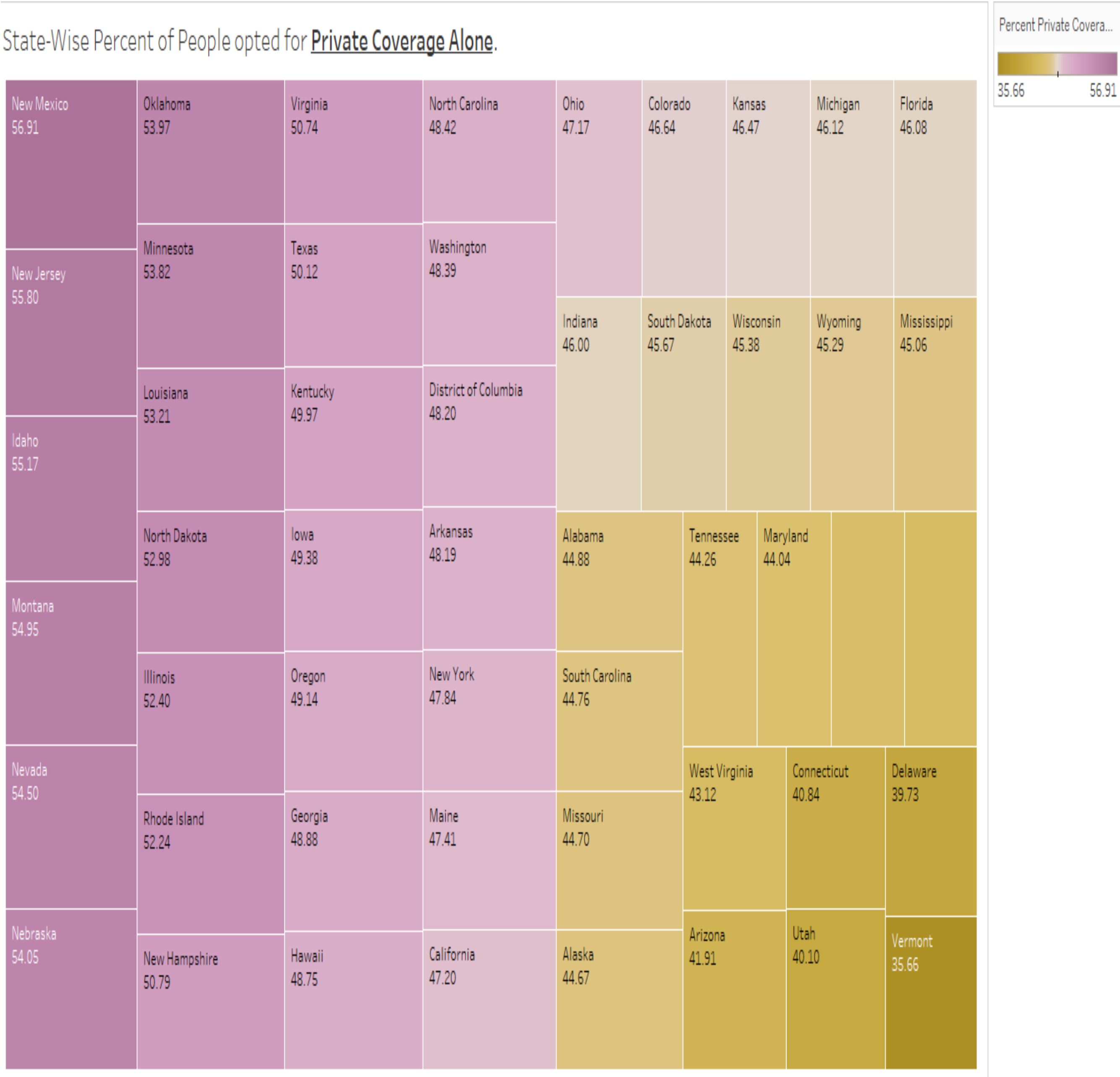
- The above scatter plots depict the type of insurances people take when they are employed or unemployed.
- The above plots clearly show that the unemployed opt for public whereas the employed opt for private insurances.

PERCENT BACHELOR'S DEGREE vs INSURANCES



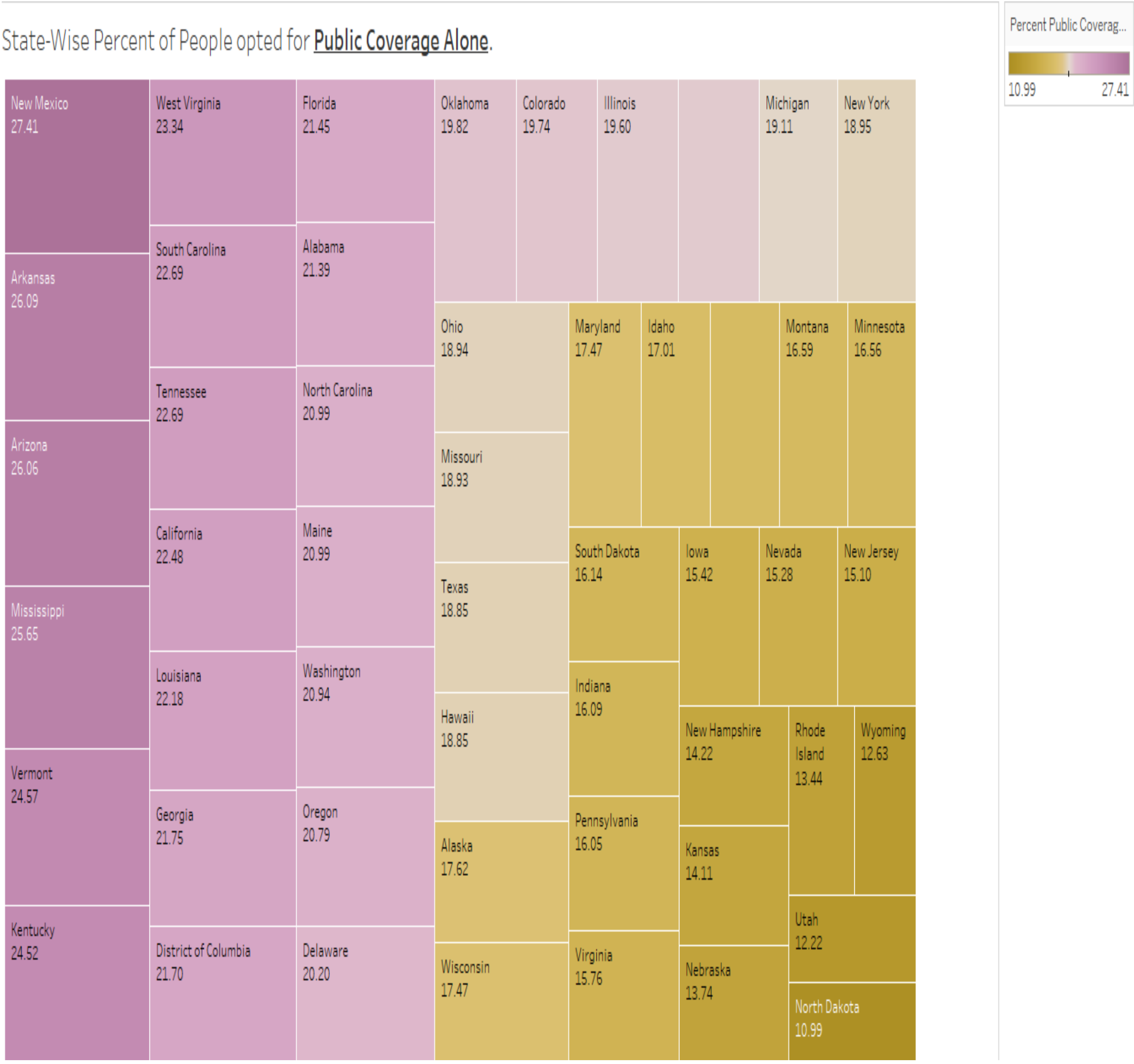
- The above scatter plots show the effect of having a bachelor's degree on the type of insurance people opts.
- It is clearly visible that people who have a bachelor's degree have opted to buy private insurances.
- This makes sense as people with bachelor's degree tend to be employed and they go for private insurance.

PERCENT PRIVATE COVERAGE ALONE



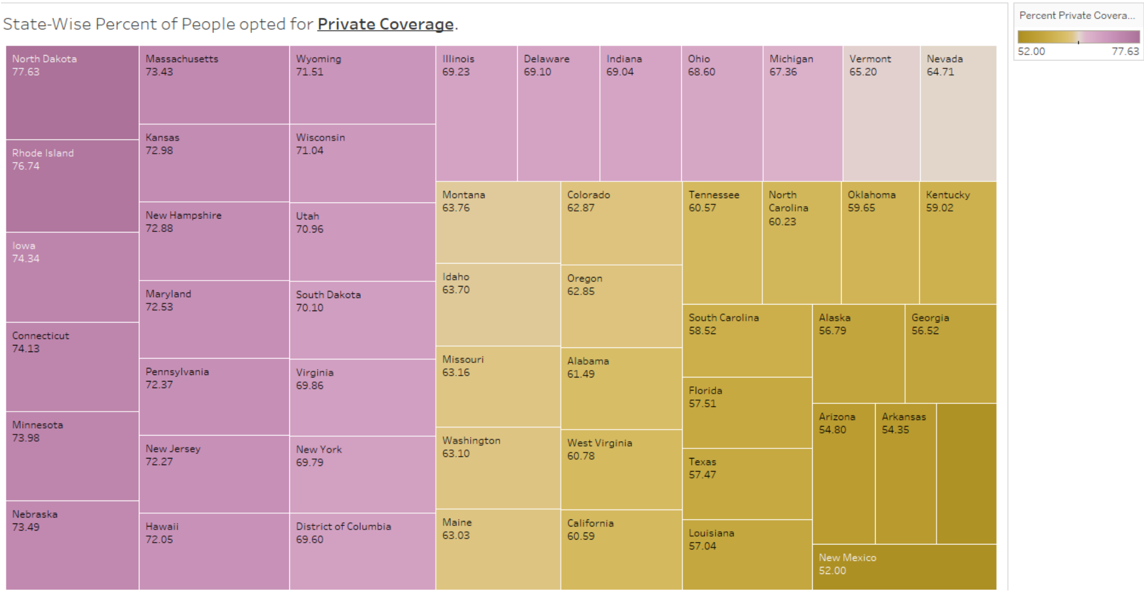
- The above treemap depicts the Percent of People opted for Private Coverage Alone.
- This also shows the richer and not so richer states of the US.
- As explained earlier, the richer people tend to buy private insurance compared to the not so richer people.

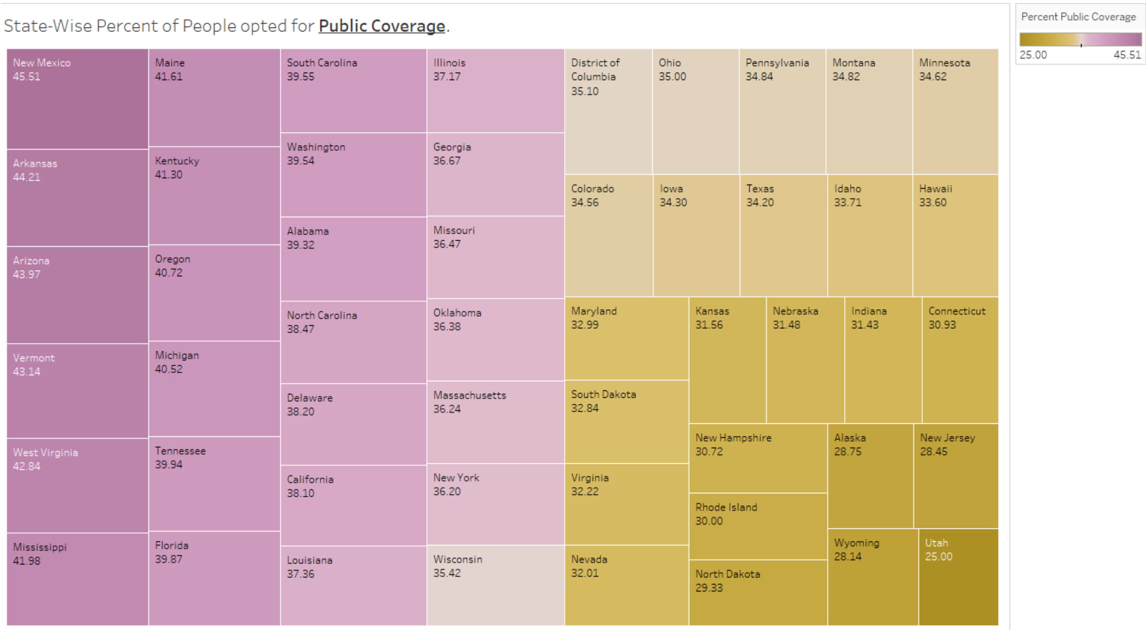
PERCENT PUBLIC COVERAGE ALONE



- The above treemap depicts the Percent of People opted for Public Coverage Alone.
- As we can see in the tree map, the not so richer states tend to opt for public coverage.

PERCENT PRIVATE COVERAGE AND PERCENT PUBLIC COVERAGE





- The first treemap depicts the Percent of People opted for Private Coverage.
- The first treemap depicts the Percent of People opted for Public Coverage.
- This shows the mid class states of the US who have opted for both private and public coverages.

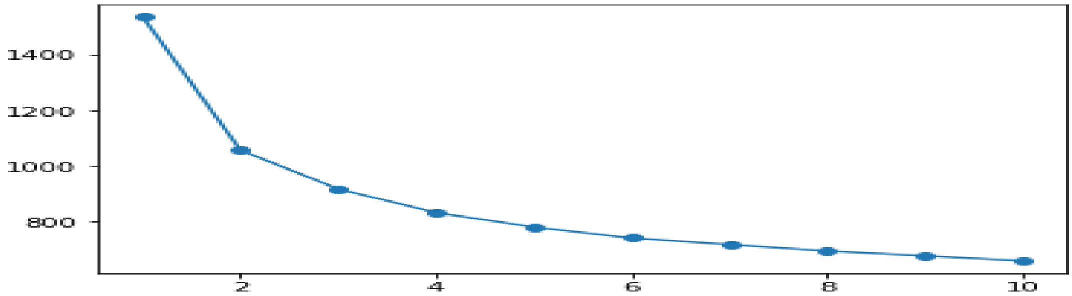
OTHER ATTEMPTS

Assuming this problem to be a regression problem, we built many regression models like Decision Tree, K Nearest Neighbors, Random Forest and Linear Regression only to find out that the models performance is very bad.

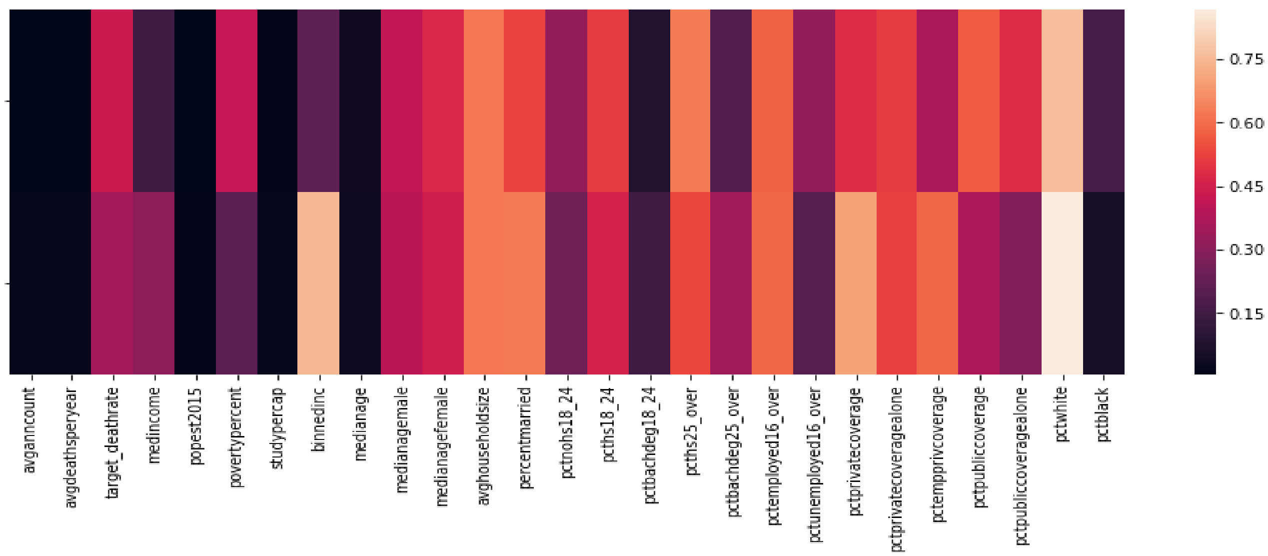
We engineered the following features and conducted many tests to see if they added any value to our analysis. But, were eventually not contributing to the project.

- Proportion of people who have cancer out of the population = **avgAnnCount/popEst2015**
- Proportion of Deaths due to cancer out of the entire population = **avgDeathsPerYear/popEst2015**
- Proportion of deaths due to cancer out of the people who were diagnosed with cancer = **avgDeathPerYear/avgAnnCount**
- Proportion of people who got tested for cancer out of the entire population = **studyPerCapita/popEst2015**
- Proportion of people who had cancer after getting tested = **avgAnnCount/studyPerCapita**
- Instead of the percentage, we took the actual count of the races present in the dataset. They did not aid in cluster formation.

SCORING FUNCTION



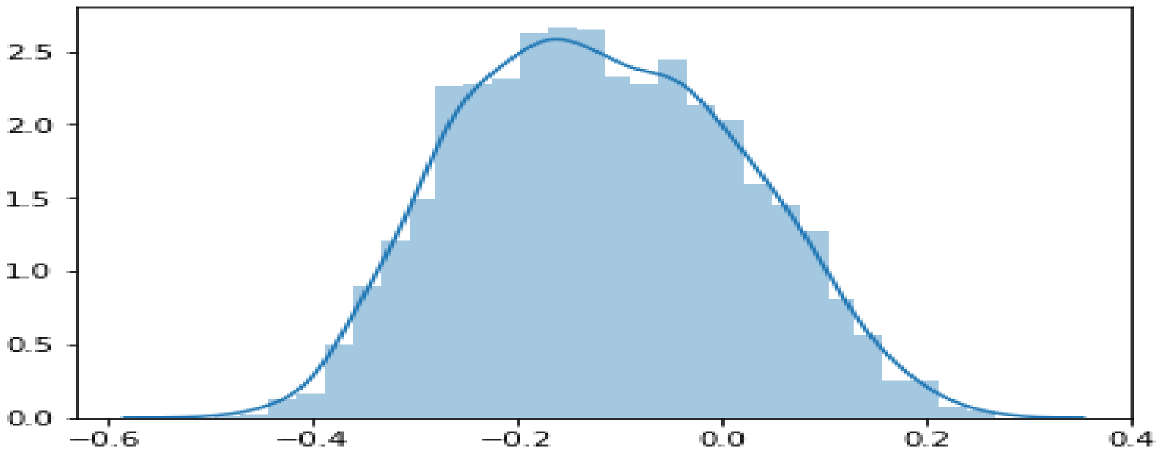
- From the elbow plot above, we found the ideal number of clusters to be 2.
- After clustering using KMeans Clustering, we found the following centroid differences for the clusters as shown in the heatmap.



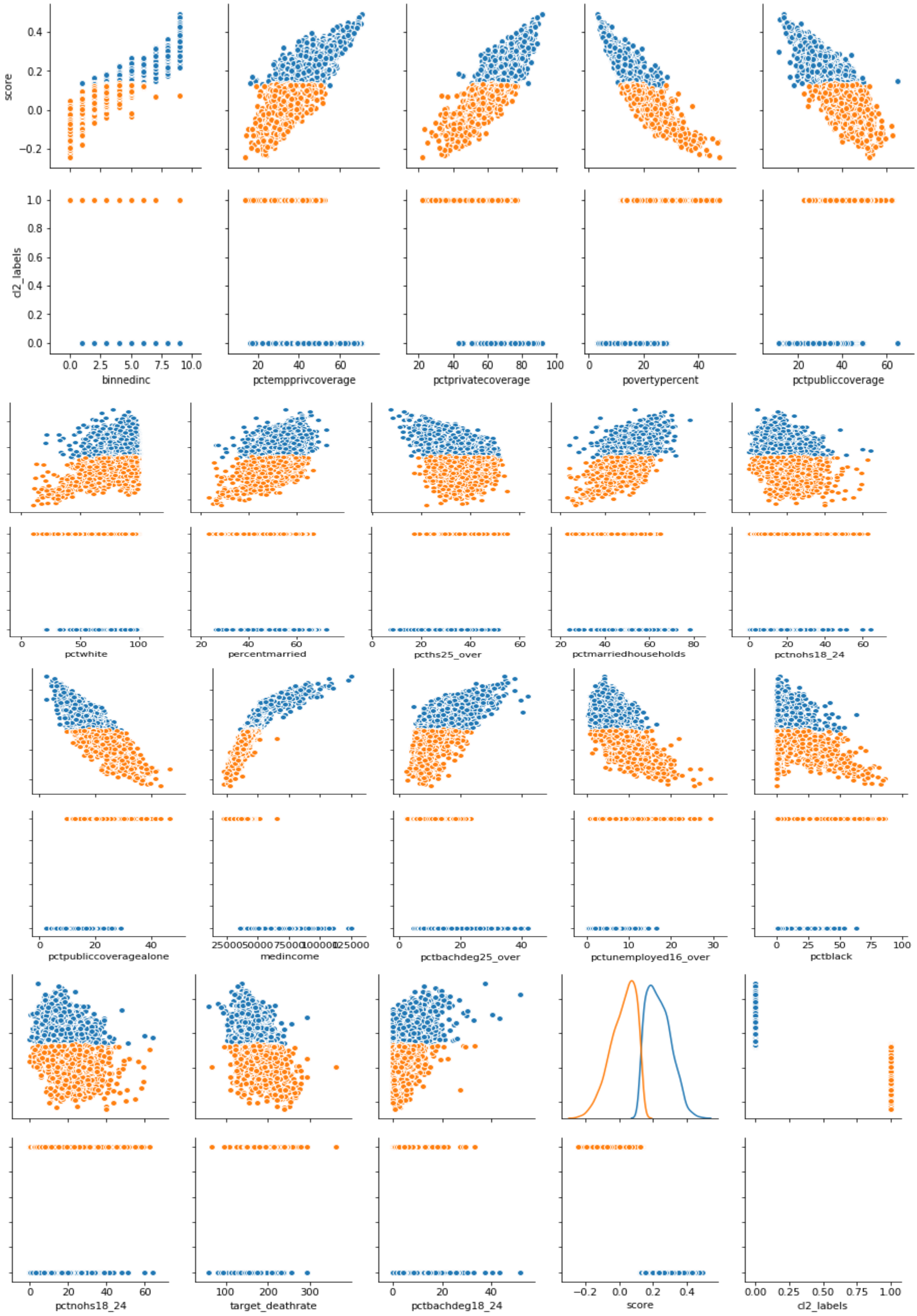
- From the above heatmap, we can infer that the clustering has happened based on the income of the counties.
- From the EDA and also the clustering here, the features that depended on the income i.e. insurances, poverty and education were also clustered based on the income.

	feature	one	two	differ	abs_diff	pct_exp
7	binnedinc	0.215039	0.749247	-0.534209	0.534209	0.190605
22	pctempprivcoverage	0.366168	0.587626	-0.221457	0.221457	0.079016
20	pctprivatecoverage	0.482980	0.704010	-0.221030	0.221030	0.078863
5	povertypercent	0.420307	0.212334	0.207973	0.207973	0.074205
23	pctpubliccoverage	0.569944	0.372667	0.197276	0.197276	0.070388
24	pctpubliccoveragealone	0.482626	0.286667	0.195959	0.195959	0.069918
3	medincome	0.149616	0.313814	-0.164198	0.164198	0.058586
17	pctbachdeg25_over	0.189305	0.343686	-0.154381	0.154381	0.055083
19	pctunemployed16_over	0.323326	0.198845	0.124481	0.124481	0.044415
26	pctblack	0.161480	0.057332	0.104148	0.104148	0.037160
25	pctwhite	0.763798	0.865265	-0.101467	0.101467	0.036203
12	percentmarried	0.528088	0.626310	-0.098222	0.098222	0.035045
16	pcths25_over	0.628072	0.532747	0.095324	0.095324	0.034012
13	pctnohs18_24	0.327186	0.246746	0.080439	0.080439	0.028701
2	target_deathrate	0.433454	0.356599	0.076855	0.076855	0.027422
15	pctbachdeg18_24	0.081887	0.151305	-0.069419	0.069419	0.024768
14	pcths18_24	0.513654	0.455741	0.057913	0.057913	0.020663

- The above dataframe is obtained from the centroid values of the 2 clusters for every feature.
- The 1st column shows the feature labels. Columns ‘one’ and ‘two’ are the centroid values for the two clusters.
- Then we found the difference of the centroid values for all the features.
- Then we retrieved only the magnitude of the differences by taking the absolute value.
- Then we saw the percentage of difference explained by every feature and sorted in descending order.
- We then dropped the features that explained less than 2% of the net difference.
- The final column created is the percentage of difference explained by each feature. They were considered as the coefficients in the scoring function which is a linear combination of the important features shown above.



- The scores of the counties are normally distributed.
- The score also has a linear relationship with the important features as shown below.


score
cl2_labels
binnedinc
pctempprivcoverage
pctprivatecoverage
povertypercent
pctpubliccoverage
pctwhite
percentmarried
pcths25_over
pctmarriedhouseholds
pctnohs18_24
pctpubliccoveragealone
medincome
pctbachdeg25_over
pctunemployed16_over
pctblack
pctnohs18_24
target_deathrate
pctbachdeg18_24
score
cl2_labels

	feature	coefficients
0	binnedinc	0.180547
1	pctempprivcoverage	0.074961
2	pctprivatecoverage	0.074814
3	povertypercent	-0.070427
4	pctpubliccoverage	-0.066791
5	pctpubliccoveragealone	-0.066496
6	medincome	0.055492
7	pctbachdeg25_over	0.051846
8	pctunemployed16_over	-0.042256
9	pctblack	-0.034938
10	pctwhite	0.034094
11	percentmarried	0.033610
12	pcths25_over	-0.031690
13	pctmarriedhouseholds	0.030052
14	pctnohs18_24	-0.027212
15	target_deathrate	-0.025908
16	pctbachdeg18_24	0.023097

- The above dataframe is the coefficients and its respective features.

.

RESULTS

```
Falls Church city, Virginia
Douglas County, Colorado
Loudoun County, Virginia
Williamson County, Tennessee
Hamilton County, Indiana
Delaware County, Ohio
Los Alamos County, New Mexico
Carver County, Minnesota
Summit County, Utah
Arlington County, Virginia
Lincoln County, South Dakota
Hunterdon County, New Jersey
Scott County, Minnesota
Dallas County, Iowa
Morris County, New Jersey
Broomfield County, Colorado
Howard County, Maryland
Morgan County, Utah
Fairfax County, Virginia
```

```
Forsyth County, Georgia
Johnson County, Kansas
Somerset County, New Jersey
Oldham County, Kentucky
Washington County, Minnesota
Ozaukee County, Wisconsin
Waukesha County, Wisconsin
Teton County, Wyoming
Putnam County, New York
Kendall County, Illinois
Davis County, Utah
Poquoson city, Virginia
St. Charles County, Missouri
Chester County, Pennsylvania
Warren County, Ohio
Rockingham County, New Hampshire
Oconee County, Georgia
St. Croix County, Wisconsin
Boone County, Indiana
```

```
     Calumet County, Wisconsin
      Hanover County, Virginia
      Carroll County, Maryland
        Sarpy County, Nebraska
   Washington County, Nebraska
 Norfolk County, Massachusetts
      Dakota County, Minnesota
   Livingston County, Michigan
     Stafford County, Virginia
     Hendricks County, Indiana
        Eagle County, Colorado
     Sussex County, New Jersey
```

The above are the top 50 counties we recommend.

REFERENCES

- The American Community Survey (census.gov)
- Clinicaltrials.gov
- Cancer.gov
- www.analyticsvidhya.com
- www.data.world/nrippner/ols-regression-challenge
- www.ncsl.org/research/health/health-insurance-premiums.aspx