Task B: Exploratory Analysis on Big Data

B1. Summarising the Data

Load the InsuranceRates.csv data in Python (or R) and answer the following questions:

1. How many rows and columns are there?
2. How many years does the data cover? (Hint: pandas provide functionality to see 'unique' values.)
3. What are the possible values for 'Age'?
4. How many states are there?
5. How many insurance providers are there?
6. What are the average, maximum and minimum values for the monthly insurance premium cost for an individual? Do those values seem reasonable to you?
7. How much more on average do plans for smokers cost?

Answer:

1) There are 12694445 rows and 7 columns

2) The data covers 3 years: 2014, 2015 and 2016

3) The possible values of ages are: '0-20', 'Family Option', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38', '39', '40', '41', '42', '43', '44', '45', '46', '47', '48', '49', '50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60', '61', '62', '63', '64', '65 and over'

4) There are 39 states

5) There are 910 insurance providers

6) The aggregate values are:

Key	Insurance Cost
Mean	*4098.026458581588*
Max	*999999*
Min	*0.0*

The Max and Min values are not plausible as the values are too extreme on both ends. Probably junk records.

7) Plans for smoker costs 88.90566067009055 more on average

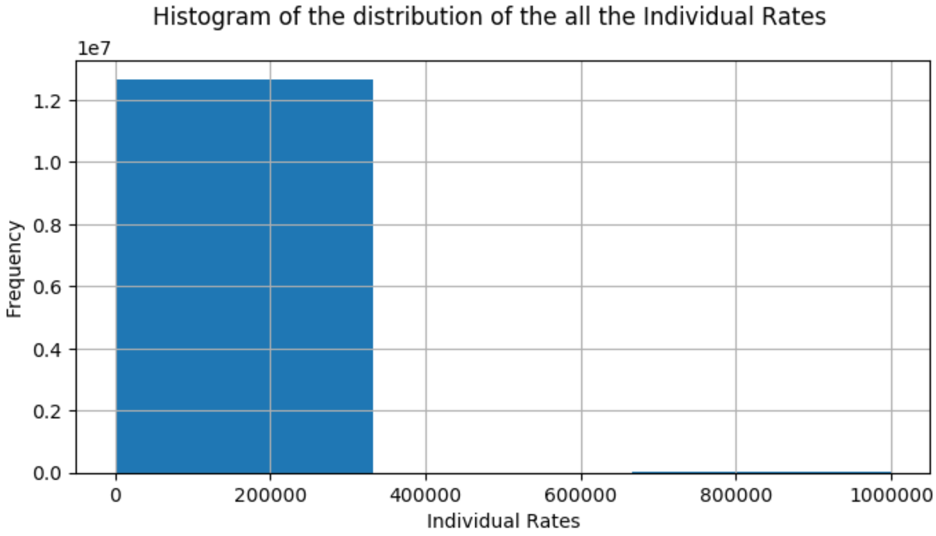
B2. Investigating Individual Insurance Costs

Now let's look more in detail at the individual insurance costs.

1. Show the distribution of 'IndividualRate' values using a histogram.

a) Does the distribution make sense to? What might be going on?

Answer: The distribution of Individual Rate is shown below using a histogram:



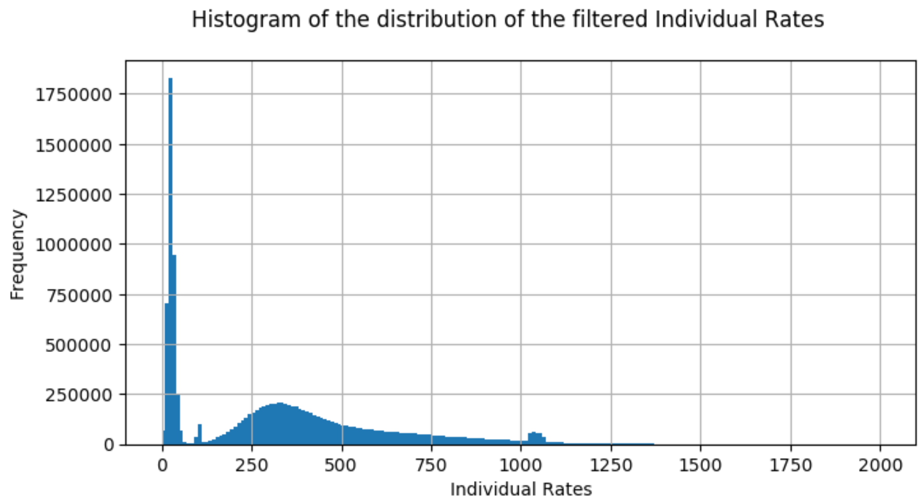
The above histogram doesn't make much sense due to the fact that the data for the distribution consists of all the Insurance Rates. The majority of the Insurance rates are paid in the first bar while a seemingly invisible outlier is observed at the end. The outlier cannot be a plausible value as the Insurance Rates are too high to be true. To get a proper insight we must delve into the data of the first bar.

2. Remove rows with insurance premiums of 0 (or less) and over 2000. (Use this data from now on). Generate a new histogram with a larger number of bins (say 200).

a) Does this data look more sensible?
b) Describe the data. How many groups can you see?

Answer: The distribution of Individual Rate is shown below using a histogram:

Next Page (Contd.)



The histogram data makes more sense now as we can clearly see the distribution of different Insurance Rates excluding the extreme values.

There are three groups of data in the histogram, which can be categorised into: Low, Medium and High insurance rates. There are significantly large number of users who are paying a Low insurance rates but have less options to choose from. For the Medium insurance rates, there is considerable a widest variety of rates to choose from. There is a small spike in High insurance rates indicating that there is a very small section of people paying at higher rates.

B3. Variation in Costs across States

How do insurance costs vary across states?

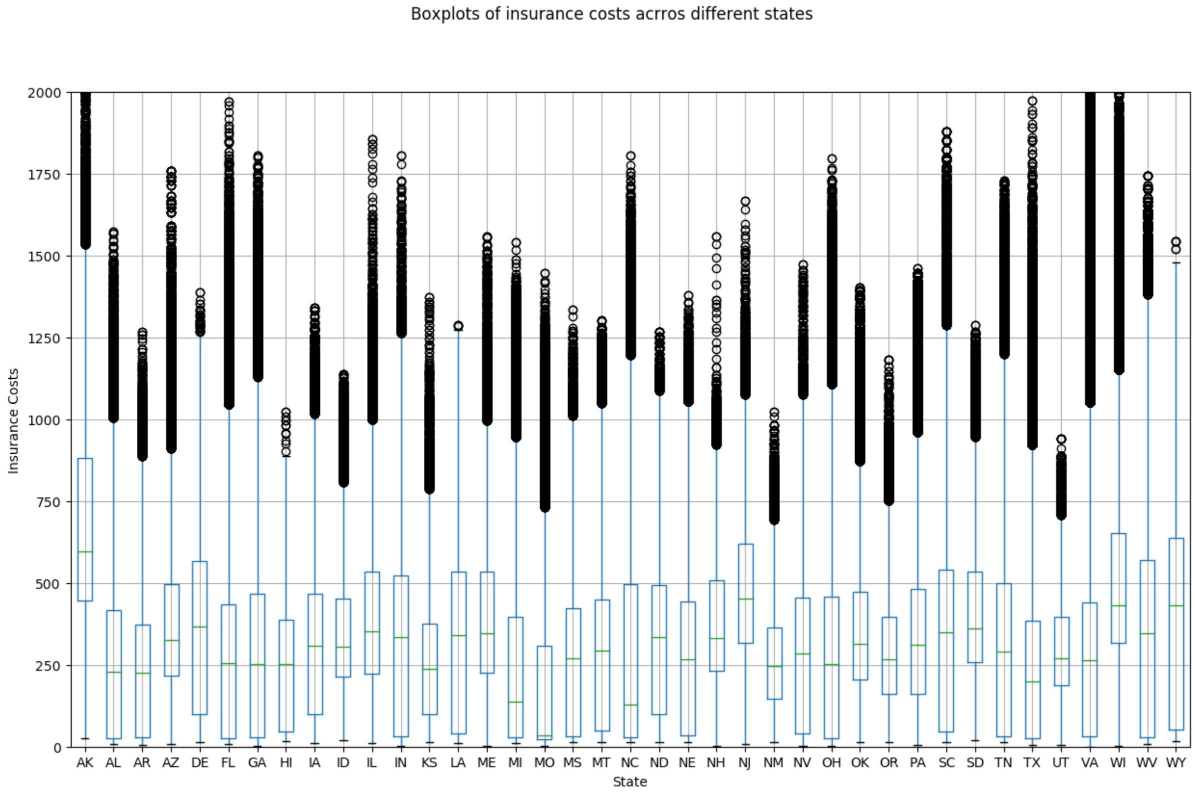
1. Generate a graph containing boxplots summarising the distribution of values for each state.

a) Which state has the lowest median insurance rates and which one has the highest? (Hint: you may need to rotate the state labels to be able to read the plot.)

b) Is there much variation in costs across states?

Answer: The insurance rates for the various states are shown in the below graph via box plots.

Next Page (Contd.)



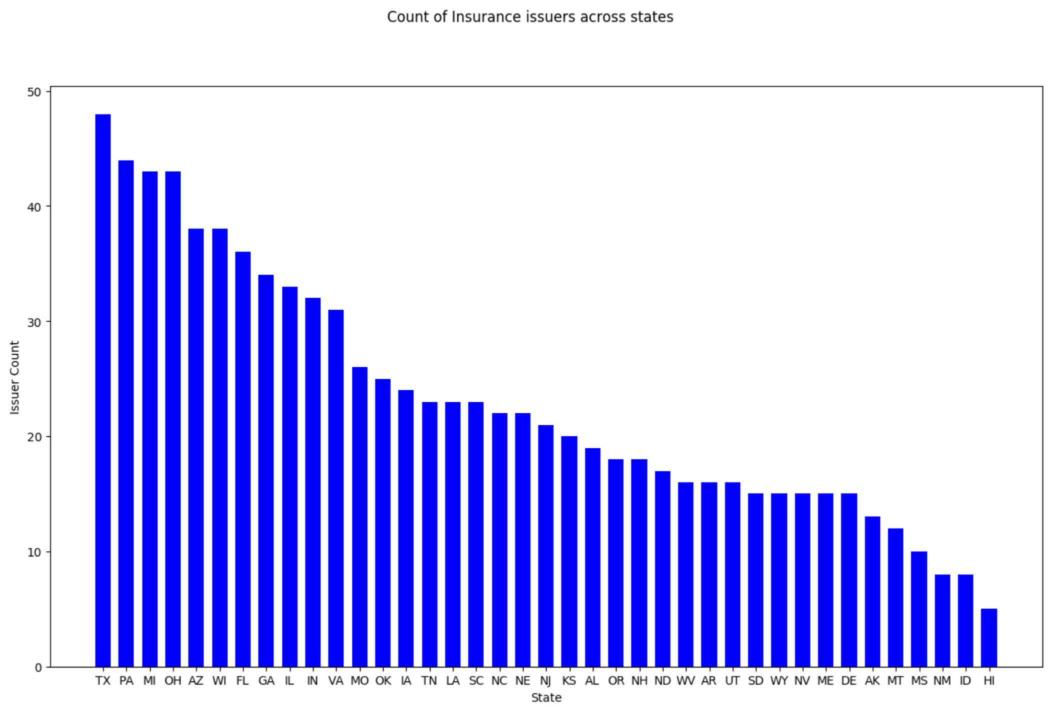
The state of ‘MO’ has the least median insurance rates while ‘AK’ has the highest median insurance rate. There is not much variation in the median insurance rates across each state. Most of the states have similar median insurance rate, close to between 250 and 350 [approximated]. However, on inspecting the outliers it can be seen that there is a wide variation in the price of highest insurance rate across different states. For example, the highest insurance rate in the state of ‘HI’ is around 1000 and that of NC is around 1800.

2. Does the number of insurance issuers vary greatly across states?

a) Create a bar chart of the number of insurance companies in each state to see. (Hint: you will need to aggregate the data by state to do this.)

Answer: The number of insurance companies are plotted in the graph below:

Next Page (Contd.)

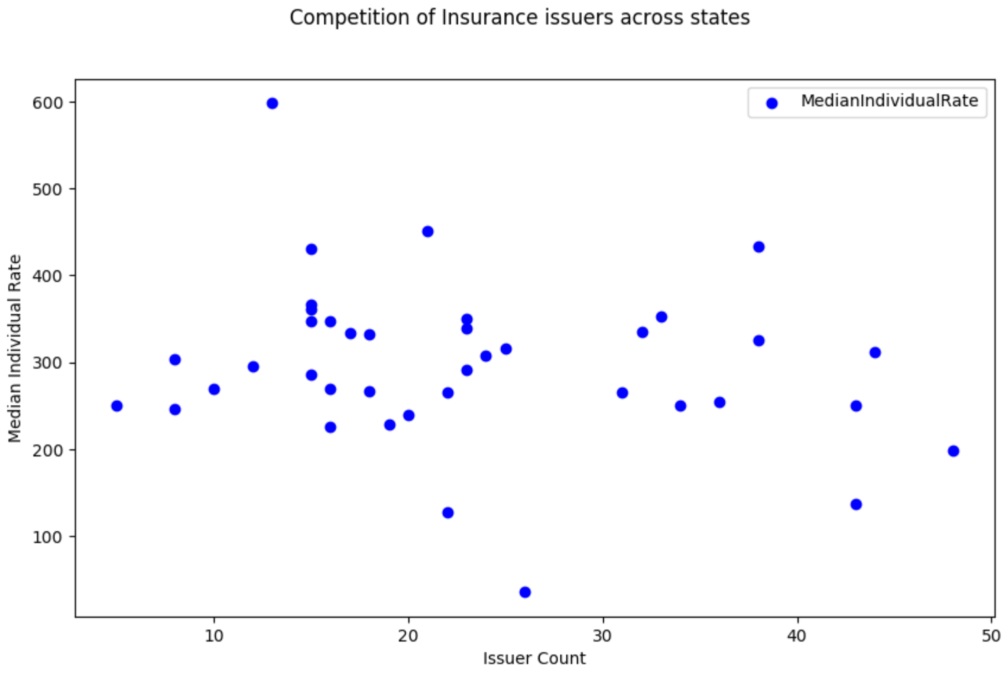


The bar graph clearly shows that the state of ‘TX’ has the highest number of issuers and the state of ‘HI’ has the least number of issuers. The graph depicts that the number of issuers across states in the descending order does not vary greatly against each other.

3. Could competition explain the difference in insurance premiums across states?

a) Use a scatterplot to plot the number of insurance issuers against the median insurance cost for each state.
b) Do you observe a relationship?

Answer: The scatter plot is plotted between median insurance rates and issuer count. The relation is as below:



In every state, there is a strong competition amongst insurance issuers where the insurance rate is close to between 250 and 350 [approximated]. Most insurance issuers are providing insurances in the previous mentioned rates with minute differences than that in the other state, attracting various customers as per their need. Insurance rates above 350 and below 250 holds minimum competition across insurance issuers across various states.

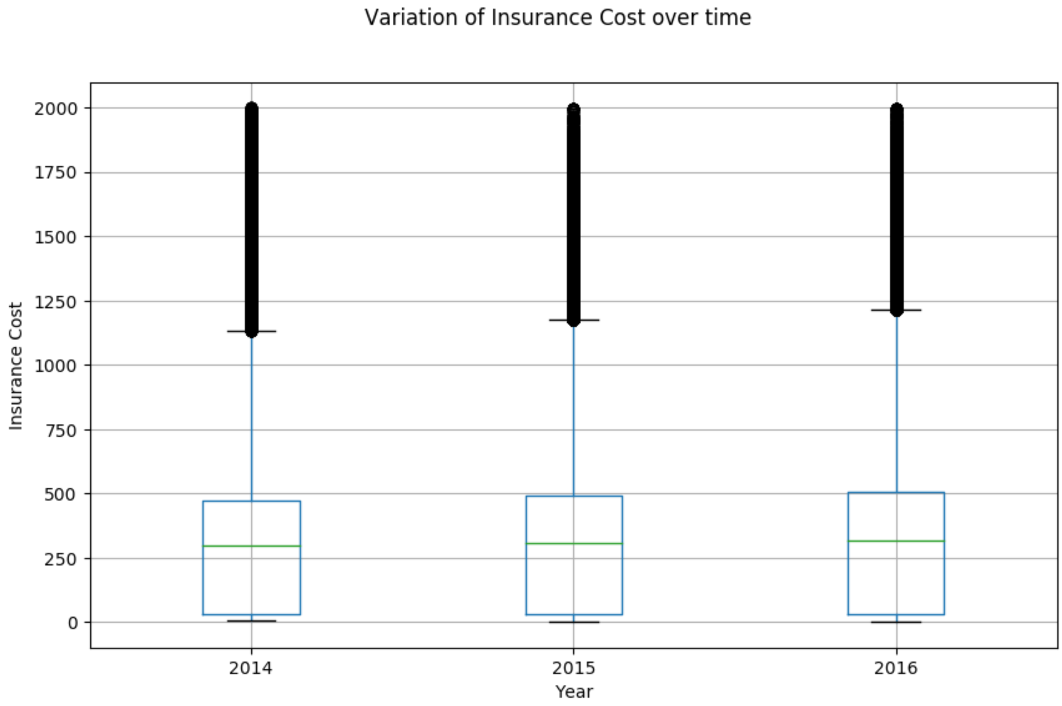
B4. Variation in Costs over Time and with Age

Generate boxplots (or other plots) of insurance costs versus year and age to answer the following questions:

1. Are insurance policies becoming cheaper or more expensive over time?

a) Is the median insurance cost increasing or decreasing?

Answer: The insurance cost is plotted over the year, yielding the below boxed graph:



The box plot shows that the median of the insurance cost is more or less same over the years. Also, it can be seen that there is a gradual increase in the number of high insurance rate policies over the years. However, on closer analysis, the median can be found to be gradually increasing as well by a little margin. The values are as follows:

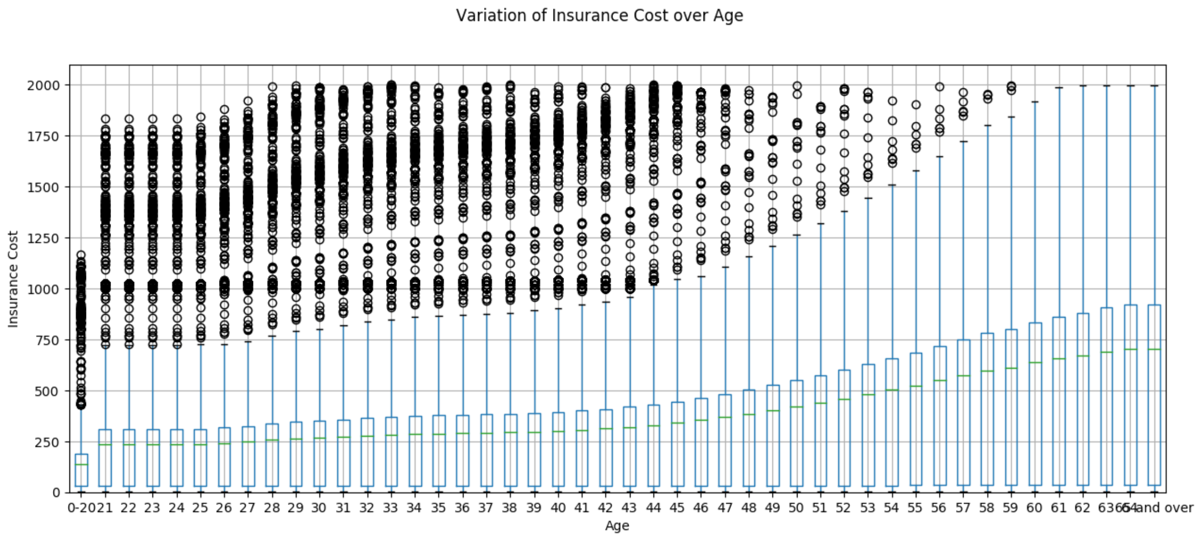
Year	Median Rate
2014	*299.31*
2015	*307.51*
2016	*317.37*

Hence it can be assumed from the above data that the insurance policies are becoming expensive over time.

2. How does insurance costs vary with the age of the person being insured? (Hint: filter out the value 'Family Option' before plotting the data.)

a) Do older people pay more or less for insurance than younger people? How much more/less do they pay?

Answer: The insurance cost is box plotted against each age and the below graph is obtained:



From the graph, it is clearly evident that the older people pay at a higher insurance rate that the younger people. The younger people [age: 0-20] pay an average insurance rate of 122.333209 while the older people [age: 65 and over] pay an average insurance rate of 584.594017. Thus, on an average the older people pay 462.26 more than the younger people.