Insurance Costs Prediction

B1: 1)

In the datafile of Insurance rates, there are 12694445 rows and 7 columns.

2)

The data is containing three years in total which are 2014, 2015 and 2016.

3)

There are several possible values of Age in data which are as follow. Column-1 is for Age and another one is for how many times that number is repeated.

Age	Number of times it is repeated.
0-20	275489
57	275067
35	275067
38	275067
44	275067
64	275067
62	275067
60	275067
43	275067
37	275067
27	275067
29	275067
55	275067
40	275067
65 and over	275067
51	275067
50	275067
52	275067
26	275067
22	275067
30	275067
21	275067
42	275067
61	275067
28	275067
54	275067
32	275067
24	275067
53	275067
45	275067
48	275067

```
49                275067
25                275067
31                275067
36                275067
23                275067
41                275067
46                275067
63                275067
59                275067
34                275067
39                275067
33                275067
47                275067
56                275067
58                275067
Family Option      40941
```

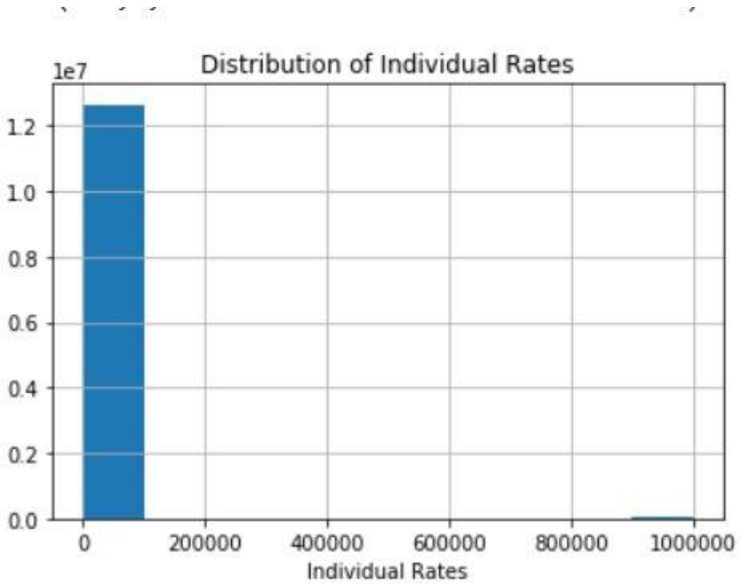
4)

According to all state codes, there must be 39 stats in the data 5)

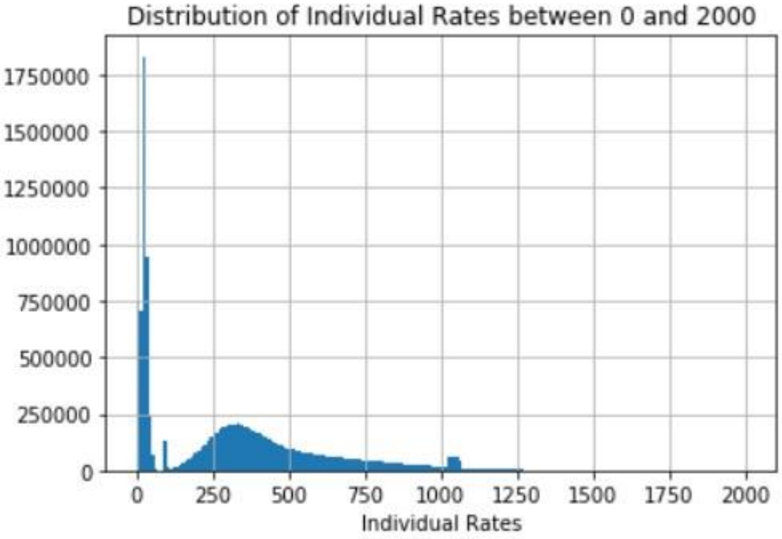
As stated in the data, there should be 910 insurance provider who can issue Insurances.

B2:

1)



As we may see in the graph above, the histogram does not seem to provide a proper distribution in all the data. According to histogram, it is showing all the values at once which is not indicating any groups. All of the values are in the only one place which is just making the graph a mess and even bins size is only 10 so it is forcing the data to store histogram only in one column therefore it is not accurate enough to show the difference between other individual rates. Most of the individual rates are in between 0 and 2000 so there is no point of plotting the data for all individual rates only to include the data of some few values more than 2000. It will end up giving us distributing the data into different parts and would give us only data in the first column with rest being empty and it will not separate the data into grouping. However, it we remove those outside values and just go with first column, it will partition the data into different groups.



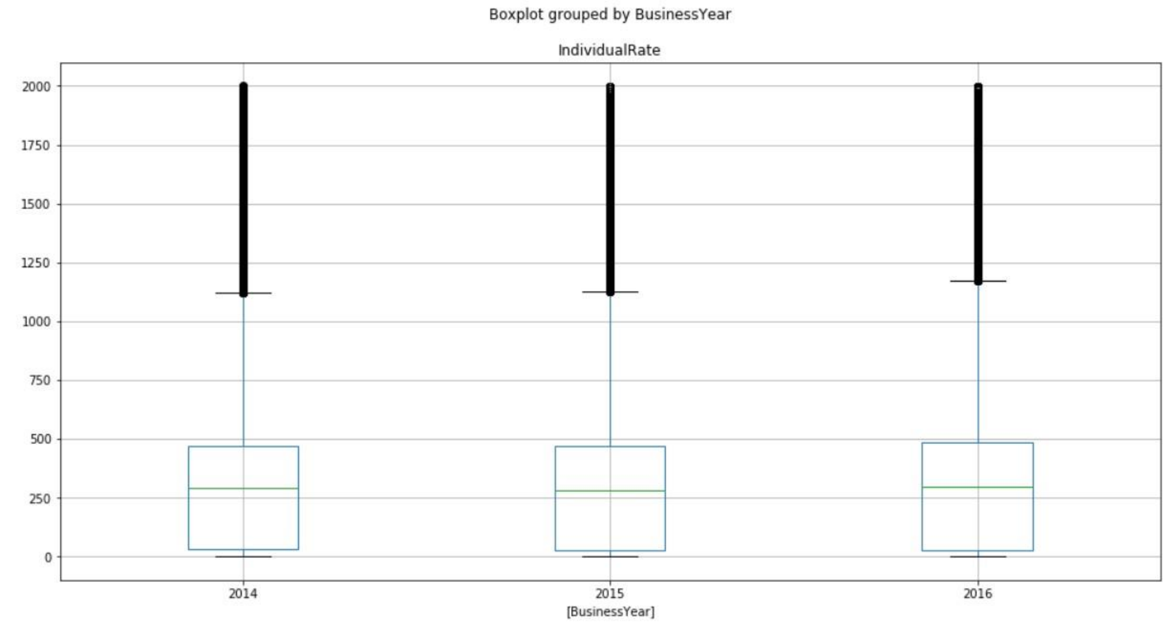
2)

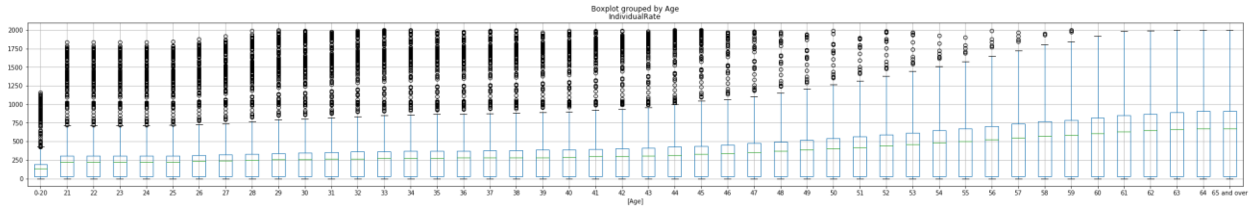
As it is said in the end of previous question and the graph above, this data looks more sensible, relevant, and accurate enough to predict the group of individual rates values. The recent data has the greatest number of individual rates in the starting period. These individual rates are maximum at that point which is one of all groups. In the halfway before 250, the trend of all rates is increasing linearly and then decreasing the same way until it reaches the individual rate of 1000. Now the data is stable and tend stops growing. From all that info, I would see three groups. The first one is when the data was maximum in the beginning. The second, the trend is growing up linearly and last but not least, it is happening when the trend grows downwards and reaches the end.

Those are three groups I could see from the distribution of individual Rates.

B3:

There is a boxplot of insurance costs against year and age generated in these below pictures.





1)

From the graph above, it seems like the median in year 2015 is higher than in the year 2014 and lower than the year in 2016. But the median values seem very close to each other. However, it can be predictable that insurance rates have decreased in year 2015 and have grown upwards in year 2016. It seems to be getting expensive in coming years.

2)

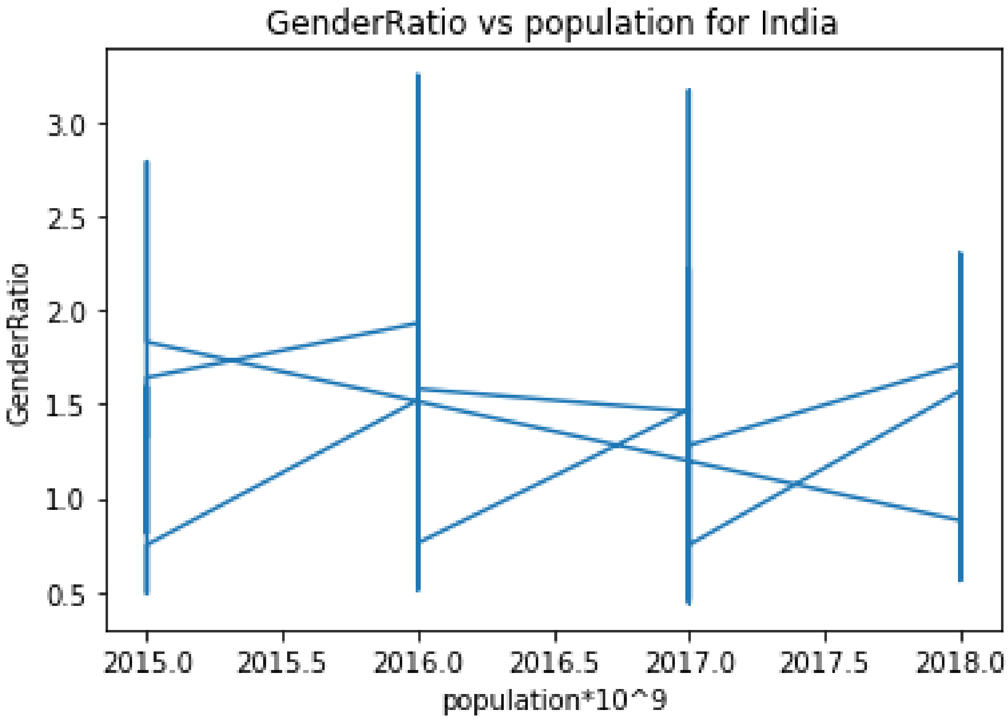
The graph of Individual insurance costs against age after filtering out 'Family option' is issued as shown above. Apparently, it is believable that the insurance rates are increasing as Age is getting above. For less than 20 years, it is really less (which is around 125, to be specific) and it keeps increasing all the way except some situations where it is the same. For example, insurance rates are same for age of 21 to 25. And it increases straight after. Although, rates are same for 29- and 30-years old people. It is the same scenario for rest of them. The only one thing is for sure that the trend never gets downwards. It is either same or gets increased for any higher number of ages. There is no drastic increment in any age either.

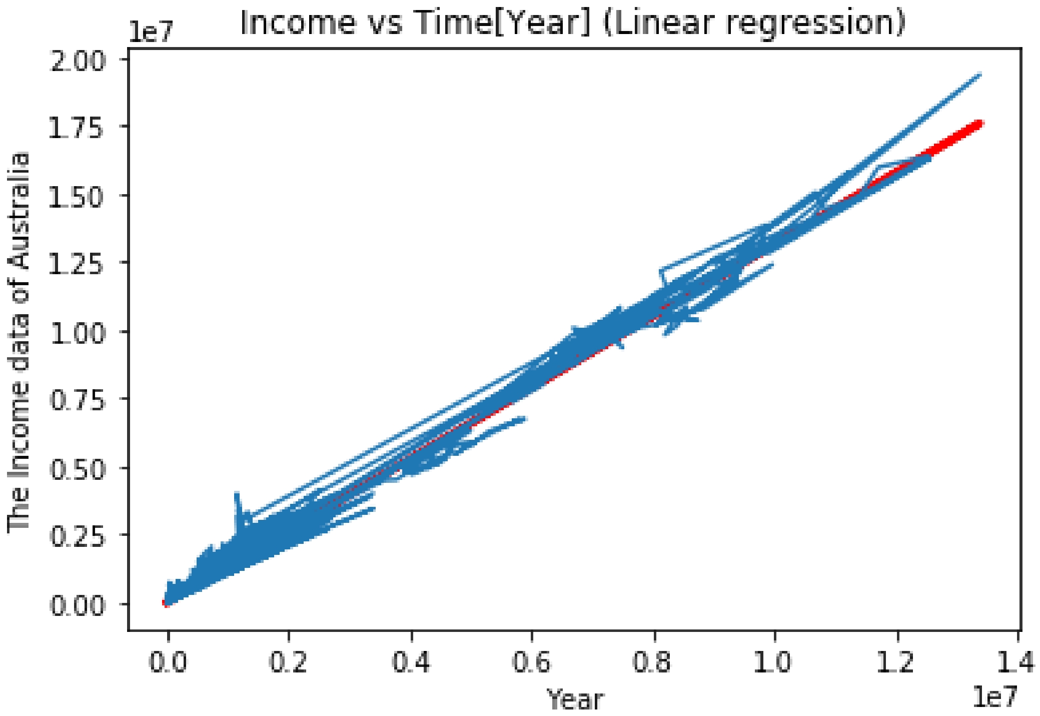
It means older people always never pay less than younger people. In some cases, they pay as much as younger people pay but in rest cases, it is always expensive to pay for insurance for older people than younger ones. Looking at the boxplot, oldest group must be paying around 550-unit individual rates more than the youngest group of people. That is why insurance is more expensive for old people than younger ones.

Task-C: [Additional Research Study]

There is one file named avocado.csv which contains unstructured data with having values for different size of bags, type of bags, number of bags, the volume of bags, the region(place), date and year it was produced and average price of that bag.

Among all these data, we want to see how much the bag’s worth increases at each time. So, the values of bag are plotted over each 6 months. As you can see in the graph below, the graph looks really stable, but it is containing a lot of data points because of all different values from different years. However, from linear regression, it can be seen that the slope is 0.0399262735741739 and intercept is -79.09129415232945. It means it is not linear and the value is very close to 0 so it can be non-linear. So, we cannot expect for Average bag values to increase at each year.





```
We can predict the values for 2020 and 2030 which are as follow.
```

For year 2020, bag value =

slope*Time + Intercept

= 0.0399262735741739 *2020 + (-79.09129415232945)

= 1.55977

bag value = 1.55977

For year 2030, bag value =

slope*Time + Intercept

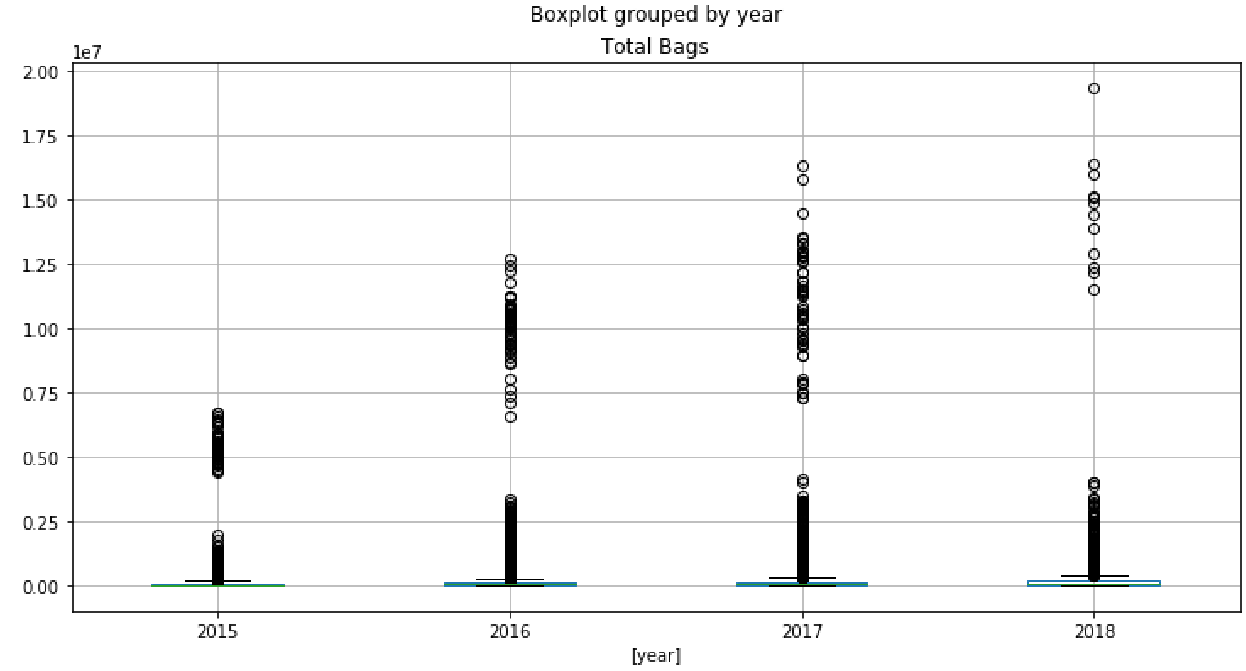
= 0.0399262735741739 *2030 + (-79.09129415232945)

= 1.958892

bag value = 1.958892

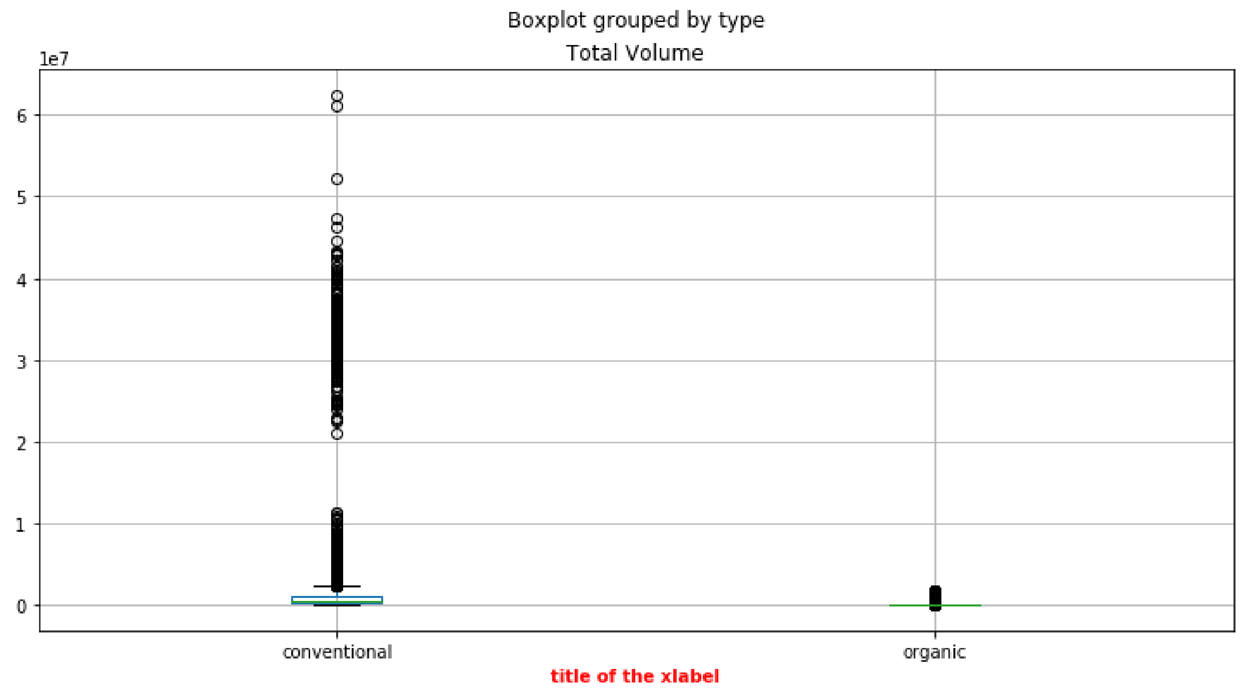
So, as you can see in the predicted values of bag, the price of bag in 2020 can be 1.5597 which is decreasing from current trend; while in next 10 years, it is predicted to be increased up to 1.958892 which is not a big difference. So, it does not change drastically so the bag values are just changing around by really lower difference.

We want to check how much portion of small bags are in the total bags. So, as you can see in the graph below, there is a picture of plotting between total bag and small bag. And it seems to be almost linear which means the trend of total vs small bags are increasing the exact same way and it feels like there is a large number of small bags out of all total bags. So, there must be a shortage of large and x-large size bags.



Now we want to compare how many total bags were produced during those four years so there is a boxplot attached here which a strongly indicate that the production of the bags is increasing each year. It was really very low at first in 2015 and then it got increased and it has grown highway too much in 2018. We can predict that from median value in a boxplot which is going high and high until 2018. So, it is picking up momentum. Hence, it is predictable that it would increase the trend of the graph drastically and go above and above each year.

Another boxplot we plotted which is shown below is to compare how much part of total volume is contributed by both types of bag. It can be seen from the boxplot that conventional bag has most total volume of bag. The median value of organic bag is higher than the one for organic bag. It means organic is not produced that often and conventional bag is more required in the market. People do not prefer to use organic bag and that is the reason conventional has been produced way too many times. It is estimated that conventional bag would be asked by almost everyone in the future someday compared to organic bag.



The link to the task-C CSV file:
https://drive.google.com/open?id=1eRdynBNpv7UGZhcJVZlg4pdFsWeB1vJ