Health Insurance Analysis and Prediction

- Aiswarya Sriram(as14988)
- Reuben Cherian(rc4610)
- Sindhu Bhoopalam Dinesh(sb8019)

Introduction

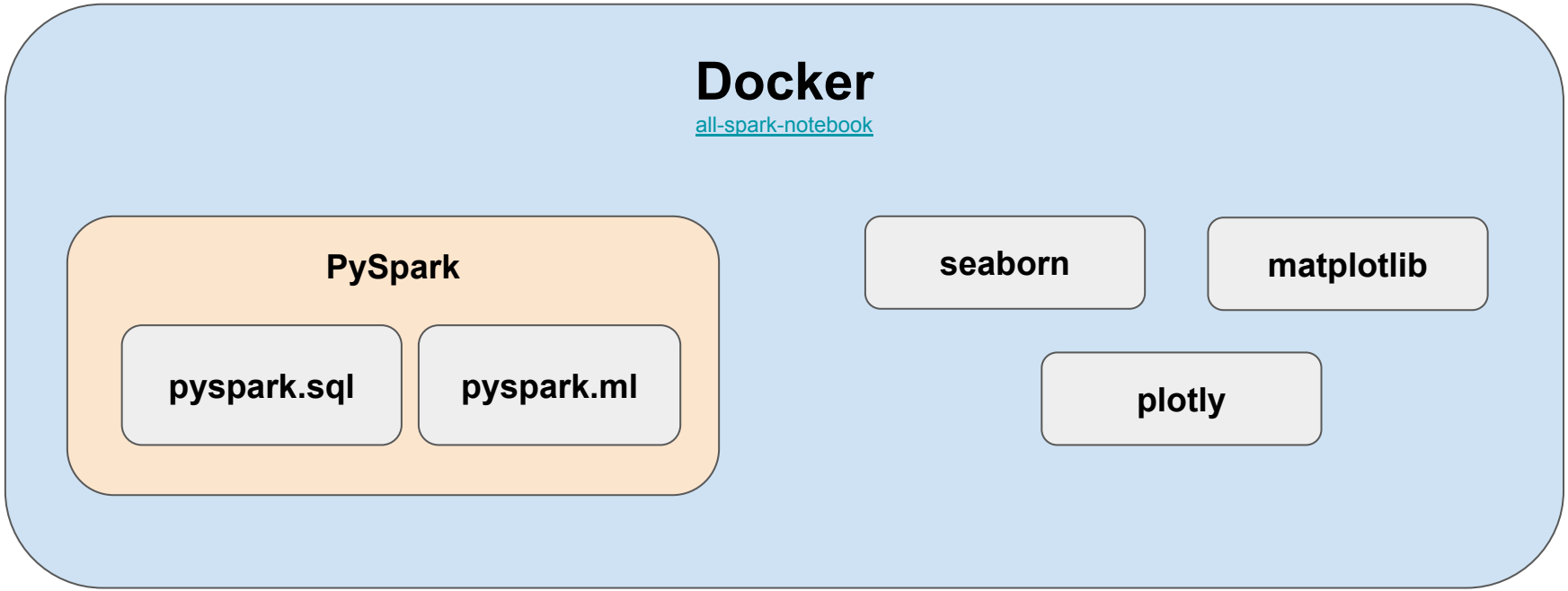
- We used the Health Insurance Marketplace Public Use Files.
- Contains information related to health and dental plans offered in the US Health MarketPlace.
- We wanted to analyze medical insurance data of different states, age groups, and insurance providers.
- This can provide us with indicators that help people make informed decisions to choose an optimal insurance plan.
- Dataset size - 3.4 GB (3 years)
- Why Big Data?
    - As the number of years increase, the scale of the data increases
    - To perform analysis with machine learning, with larger number of features, we would need Big Data infrastructure

Objective

Our objective is to analyse medical insurance plan parameters to observe -

1. How plan rates affect people from different states
2. Analysis of plan benefits across the states
3. Effect of health habits and age of a person on the plan rates
4. Distribution of plan rates across the insurance issuers
5. Predict plan rates using applicable features.

Architecture



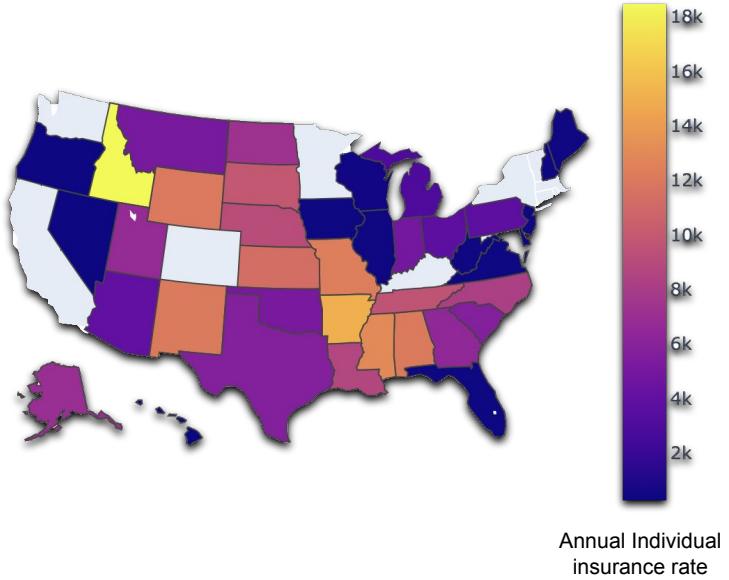
Part I - Rate vs State Analysis

A user or an insurance company may want to know how the overall rates for an insurance plan varies across states.

Why?

- Informs user whether living in a particular state has an effect on cost of their insurance plan.
- The insights obtained can help insurance companies structure new insurance plans.

Visualized average insurance rate for a particular state so insurance companies can better understand their user base.

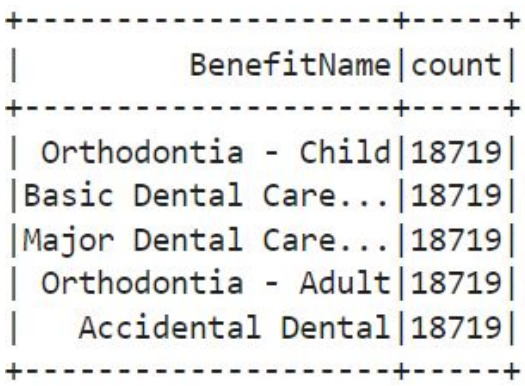


Conclusion - There is no clear dependence between an individual's insurance plan rate and the state that the individual is from (there was a low correlation ratio - 0.0032).

Part II - Analysis of Plan Benefits

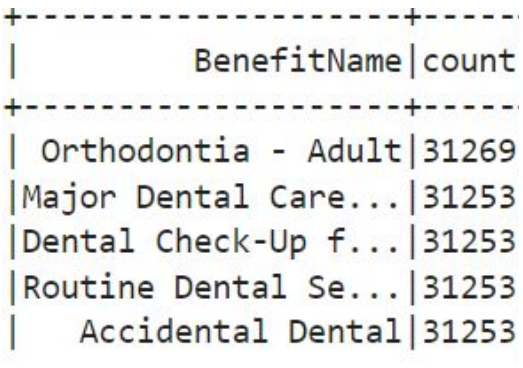
Plan Benefits across the Years -

Top 5 benefits in each of the years 2014, 2015, 2016 -



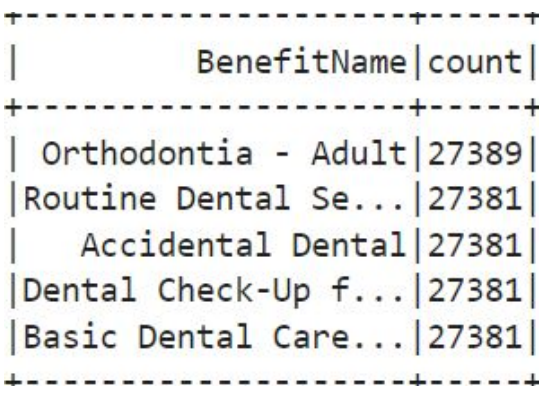
BenefitName	count
Orthodontia - Child	18719
Basic Dental Care...	18719
Major Dental Care...	18719
Orthodontia - Adult	18719
Accidental Dental	18719

2014



BenefitName	count
Orthodontia - Adult	31269
Major Dental Care...	31253
Dental Check-Up f...	31253
Routine Dental Se...	31253
Accidental Dental	31253

2015



BenefitName	count
Orthodontia - Adult	27389
Routine Dental Se...	27381
Accidental Dental	27381
Dental Check-Up f...	27381
Basic Dental Care...	27381

2016

Conclusion:

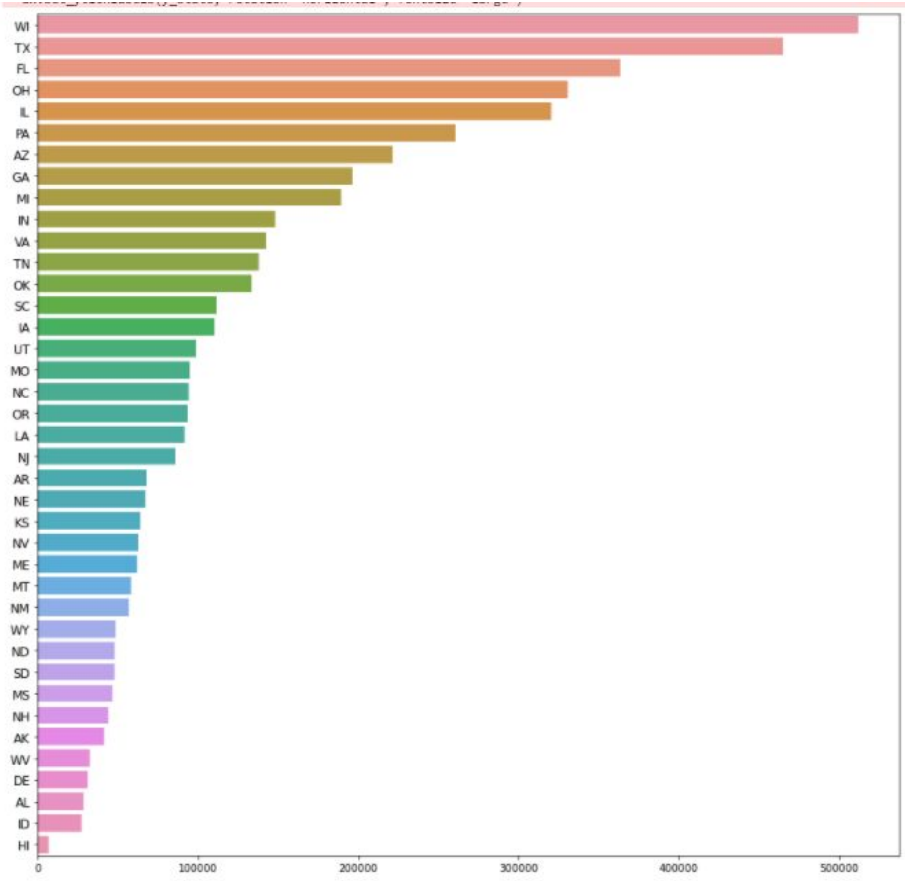
- For most of the states, the maximum used benefit is related to dental care/Orthodontia
- Thus dental care is the most sought after benefit
- Insurance providers can use this info and design their plans accordingly

Most availed Plan Benefit per State

- Find distinct states
- Preprocess state codes to remove junk values
- Group by state code, benefit name and find the benefit counts
- Use the window function to find the top benefit used in each state
- We can get an idea of which healthcare service is most needed in each state from this.

StateCode	BenefitName	count
AK	Orthodontia - Child	720
AL	Orthodontia - Adult	653
AR	Orthodontia - Adult	1077
AZ	Accidental Dental	3345
DE	Dental Check-Up f...	602
FL	Basic Dental Care...	5130
GA	Dental Check-Up f...	2893
HI	Orthodontia - Child	110
IA	Major Dental Care...	1727
ID	Basic Dental Care...	447
IL	Orthodontia - Adult	4299
IN	Routine Dental Se...	2347

Plan Benefits across States



- Group by StateCode
- Find the total number of benefits associated with each state.
- State code - x axis, no of benefits- y axis

Conclusion:

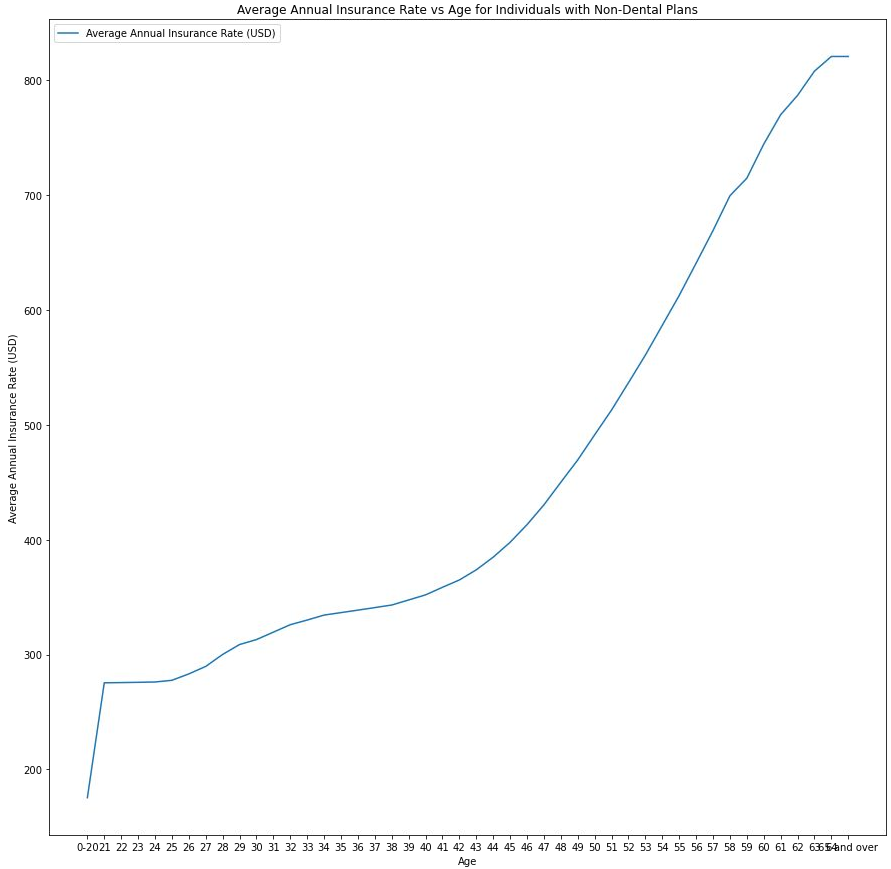
- We can see that state WI(Wisconsin) is using maximum number of benefits, followed by TX(Texas)
- The state using least number of benefits is Hawaii
- This means that insurance providers are offering a wider range of benefits in the top states

Part III - Effect of Age on Plan rates

- Comparing the average plan rate with age for non-dental plan, the average increase with age.
- Correlation is - 0.7599

Conclusion: Age and Plan rates are strongly dependent.

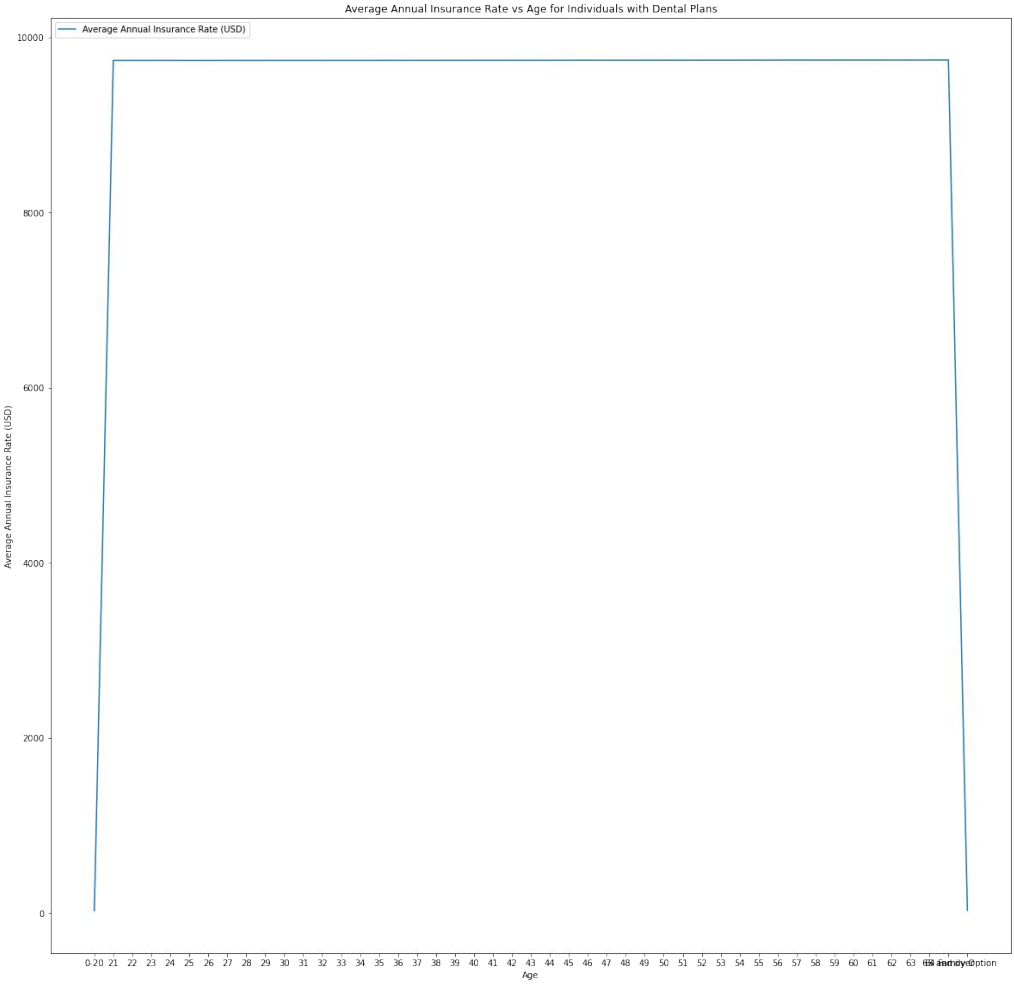
Insurance users can keep this in mind while setting aside money for an insurance plan.



- Comparing the average plan rate with age for dental plan, the average increase with age.
- Correlation is - 0.0037

Conclusion: Age and Dental Plan rates don’t seem to be dependent.

Also, the dental insurance plan rates are relatively higher. Thus dental insurance is sought after, irrespective of the age.



Part III - Effect of Health Habits on Plan Rates

- Some plan users have particular tobacco preferences and some have no preference.
- Considering the Individual Tobacco Rate for users with tobacco preference, the correlation between Individual Tobacco Rate and Plan Rates is 0.9737 with the plan rates increasing as tobacco rates increase.

Conclusion: Insurance providers could look to make more affordable plans for Tobacco users.

Part IV - Distribution of Rates across Issuers

- Analysing how the individual rates are distributed across each of the issuers tells us how diverse an issuers portfolio is.
- Calculating the mean, standard deviation and the count of individual rates for each issuer can tell us how the rate is distributed among these issuers.

IssuerId	count	mean	stddev
17859	3864	326112.3822981366 5	468821.8232593795
26075	2320	193988.1014956896 5	395475.8662060493 4
26904	7140	176490.6838739496	381236.974109308
11324	3336	161892.2988399281	368377.1737756411 3
42757	3336	161897.1721942446 3	368375.0318917477

only showing top 5 rows

Conclusion -

Displayed the final issuers which have the highest standard deviation since they have the most diverse portfolio.

Part V - Predicting Individual Insurance Rate

- Create a linear regression model to predict individual Rate
- Features (predictor variables) - Age, Individual Tobacco rate
- Criterion variable - Individual Insurance Rate
- Pyspark ML features - Vector Assembler, Pipeline
- Train, test split of 70-30
- R2 value - 0.94 Accuracy - 94.95%

Overall Conclusions

- No clear dependence between an individual's insurance plan rate and the state that the individual is from
- The most common benefit used in every state is related to dental care - useful for insurance providers - lets them know they should invest in dental care options.
- There is a good correlation between:
  - age and Individual rate
  - Individual Tobacco Rate and Individual Rate
- Our linear regression model to predict Individual Insurance plan rates has an R2 value of 0.94 (94% accuracy)

Future Work

- Analyze correlation between other variables to determine the best features to use to predict insurance rates more reliably.
- Use new features to build a recommendation system for the user.