Health insurance marketplace

June 2016

Why health insurance?

CMS Health Insurance Marketplace Data

Health Insurance Marketplace Public Use Files (Marketplace PUF)

- Plan-level data on essential health benefits, coverage limits, and cost sharing
- Plan-level data on individual rates based on an eligible subscriber's age, tobacco use, and geographic location
- Plan-level data on maximum out of pocket payments, deductibles, cost sharing, HSA eligibility, formulary ID, and other plan attributes
- Plan-level data on the application of rates, such as allowed relationships (e.g., spouse, dependents) and tobacco use
- Issuer-level data on the geographic coverage or service area (i.e., where the plan is offered) including state, county, and zip code
- Issuer-level data identifying provider network URLs
- Plan-level data mapping plans offered in 2014 to plans offered in 2015

Available Data:

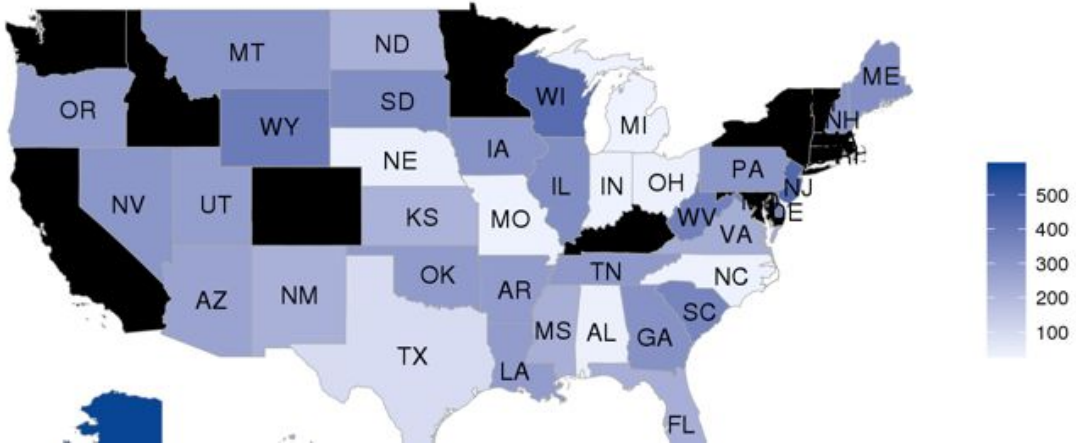
~ 12M lines of data
~200 columns

Data used:

- Age
- Tobacco usage
- State
- Metal level
- Issuer Id
- Year
- Individual premium rate

Variability among insurance plans across the U.S.

Avg. Premium ($ / mo) in 2016 for Individuals



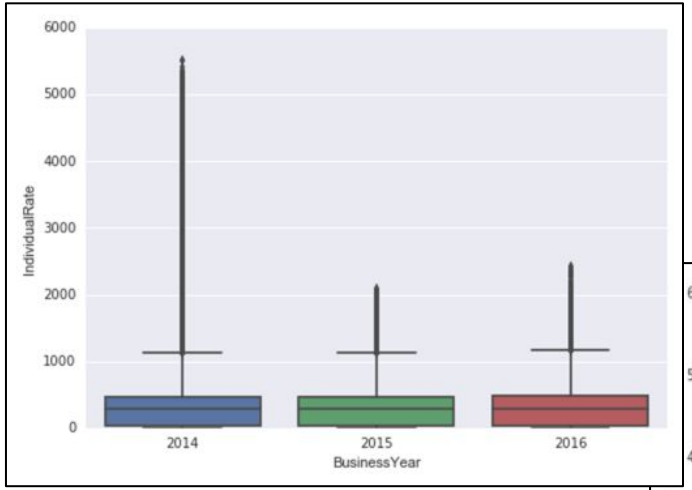
In 2014, 2015, 2016:

39 states

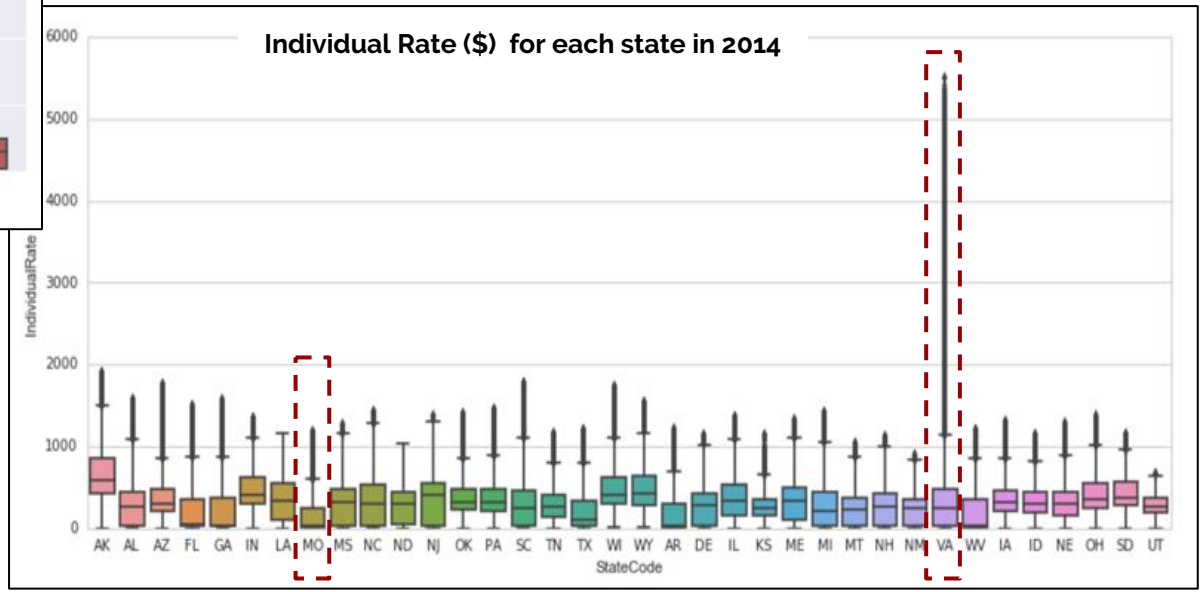
16,808 plans

- Not all states make use of the federal network
- Large variability between states in 2016

Rabbit holes!

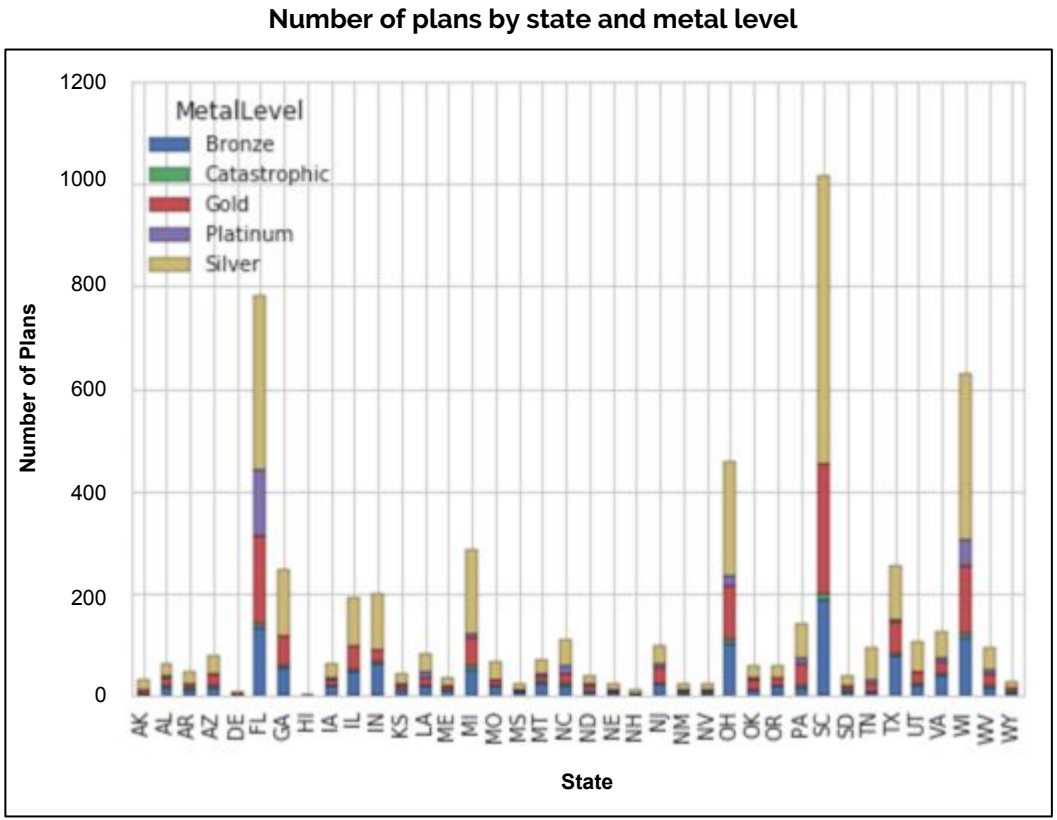


- 2014 had extremely high premiums, specifically VA
- Montana has the lowest individual rate in 2014

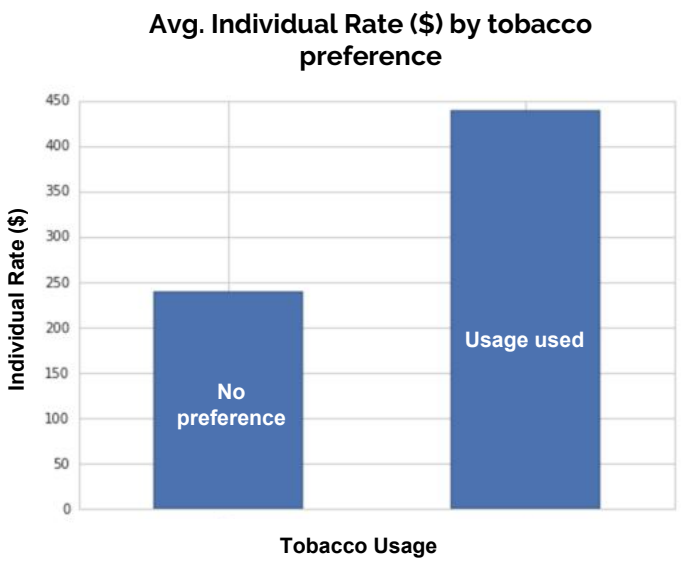


Number of plans in each state

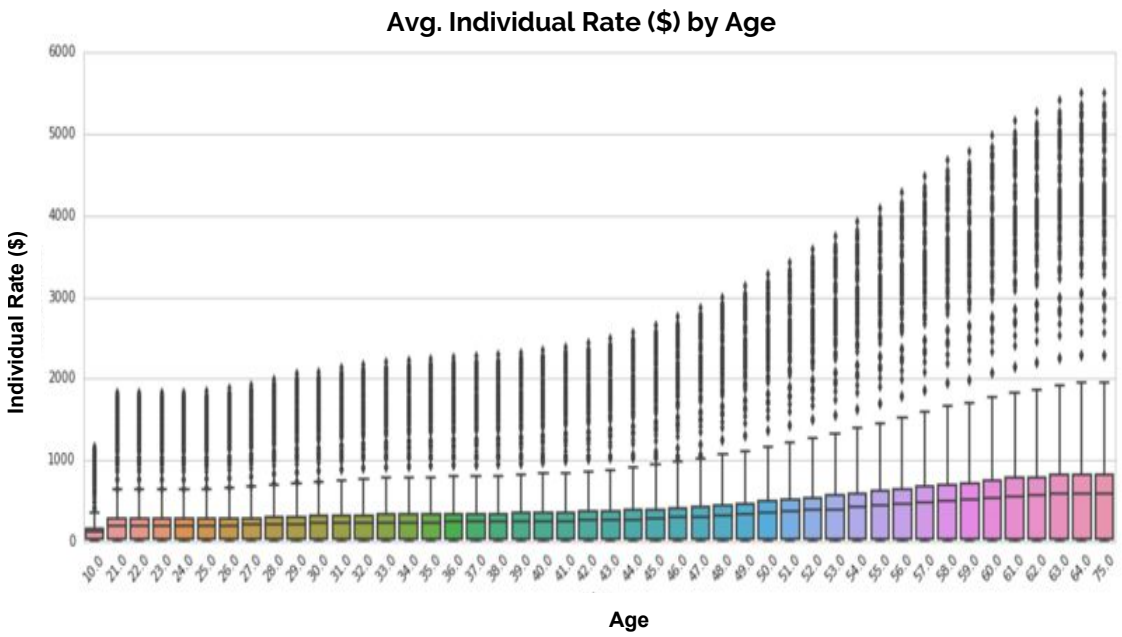
- Some states offer significantly more plans than others
- States offer varying amounts of types of plans, although the Silver plan is the most common



Individual rate by tobacco and age



- Plans that factor in tobacco usage have a significantly higher premium
- Clear increasing trend of rate with age, esp over the age of 40



Data exploration / analysis cycle

Data exploration & finding the right features

- Features?
- Change the bounds of the problem?

Regression analysis

- Are there any linear correlations between the features and rate?
- Improve R2?

What drives the monthly premium rate for individuals?

PCA / kmeans

- Is there an unseen cluster of insurers? States?
- Is there a particular feature contributes that I can't see?

Decision trees

- What are the most important features?
- How does subsetting the data change the importance?

Linear Regression

First time:

- Age
- Tobacco usage
- State
- Issuer Id

R2 = 0.267553966095

Second time, classification tree:

- Removed dental plans
- Included metal level

R2 = 0.641465715674

No significant p-values

Features	Coefficient
Age	11.99
Tobacco	-47.10
Issuer	0.000197
State	-0.951
Metal Level	57.76

Decision Trees vs. Classification Trees

First time:

- Age
- Tobacco usage
- State
- Issuer Id

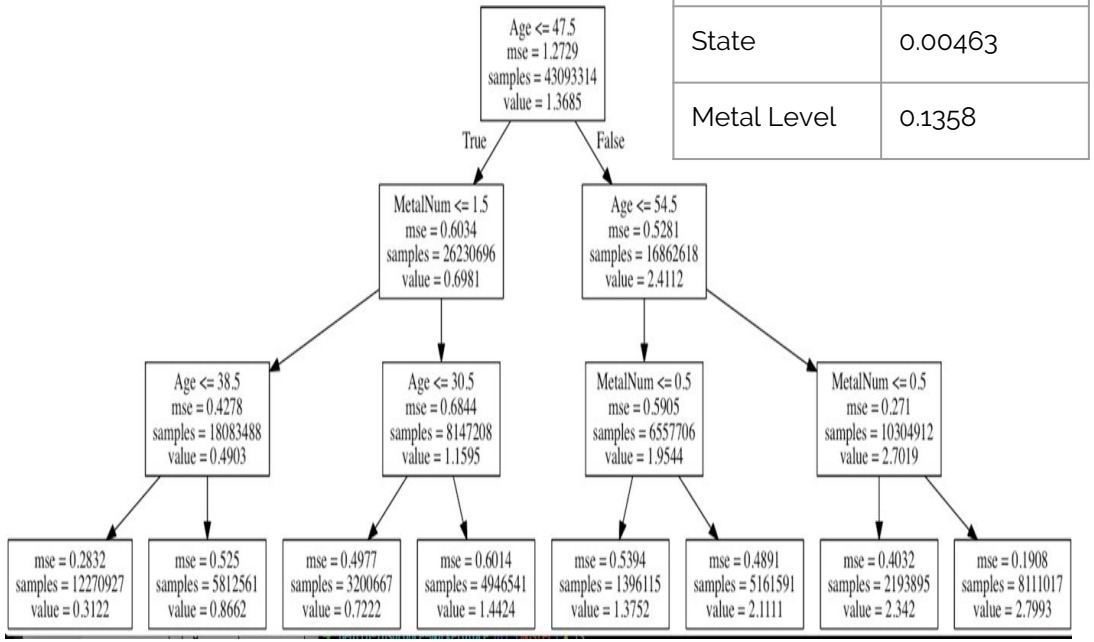
RMSE = 118.208060287

Second time, classification tree:

- Removed dental plans, included metal level
- Converted Individual Rate into 4 categories

RMSE = 0.625832715157

Features	Importance
Age	0.802
Tobacco	0.057
Issuer	0.00
State	0.00463
Metal Level	0.1358



Further Analysis

Main Findings:

- Much more variance by state and issuer than expected
- Age, tobacco usage, metal level, and state are drivers but not necessarily significant
- Categorizing individual rate can help with predictions

Areas for further analysis:

- Out of pocket / Total cost of health care
- Tobacco rates, family rates, dependent rates
- Explore rural vs. urban, conservative vs. liberal
- Explore variances by procedure
- Control for Metal level, variance in each metal level