HEALTH INSURANCE DATASET FOR ADVANCED ALGORITHMICS

Presented By:

Zanis Ali Khan (B77407)

Introduction:

I have taken the Dataset of health insurance market place, which contain multiple files, but I am taking rate file for individual family rates to analysis and plan attributes to identify the plan types. Our rate file contain the rows - 1048576 and dimensions - 24 (which makes rate file = 25, 165, 824). Plan attributes file contains the rows - 77354 and dimensions- 176 (Which makes plan file -13,614,304). Thus, we worked with huge amount of data. The data set is also divided based on year and the data of each year is given.

I have done the Analysis of the health insurance marketplace, dataset from the U.S.A. The dataset consists of health insurance and dental plans offered through healthcare (Government) between 2014 and 2016.It encompasses rates for smokers and non-smokers, separately listed for each age group, benefits included in the different plans, states in which the plans were offered, and other information. The dataset does not contain any information on actual purchases. This application (Jupyter notebook) shows visualization of the dataset.

Description and Purpose:

The data, which is used, is from the Centers for Medicare & Medicaid Services (CMS). In the Rate csv file - This csv describes the variables contained in the Rate-PUF. Each record relates to one issuer's rates based on plan, geographic rating area, and subscriber eligibility requirements. The Rate PUF is available for plan year 2014, plan year 2015, and plan year 2016.

Goals of the Application:

- **VISUALIZATION:**
  - Map for the registered states in Health Insurance Marketplace,
  - Box Plot for the individual rates by state
  - Year rate analysis
- **MACHINE LEARNING:**
  - PCA (Principle Component Analysis)

System Requirements:

Software's:

Platform

- Anaconda

Jupyter Notebook

Libraries:

[Plotly and seaborn]

Language

- Python

Hardware:

RAM – 8GB

Hard disk – 20GB

Operating System – Windows 7/8/10 , IOS

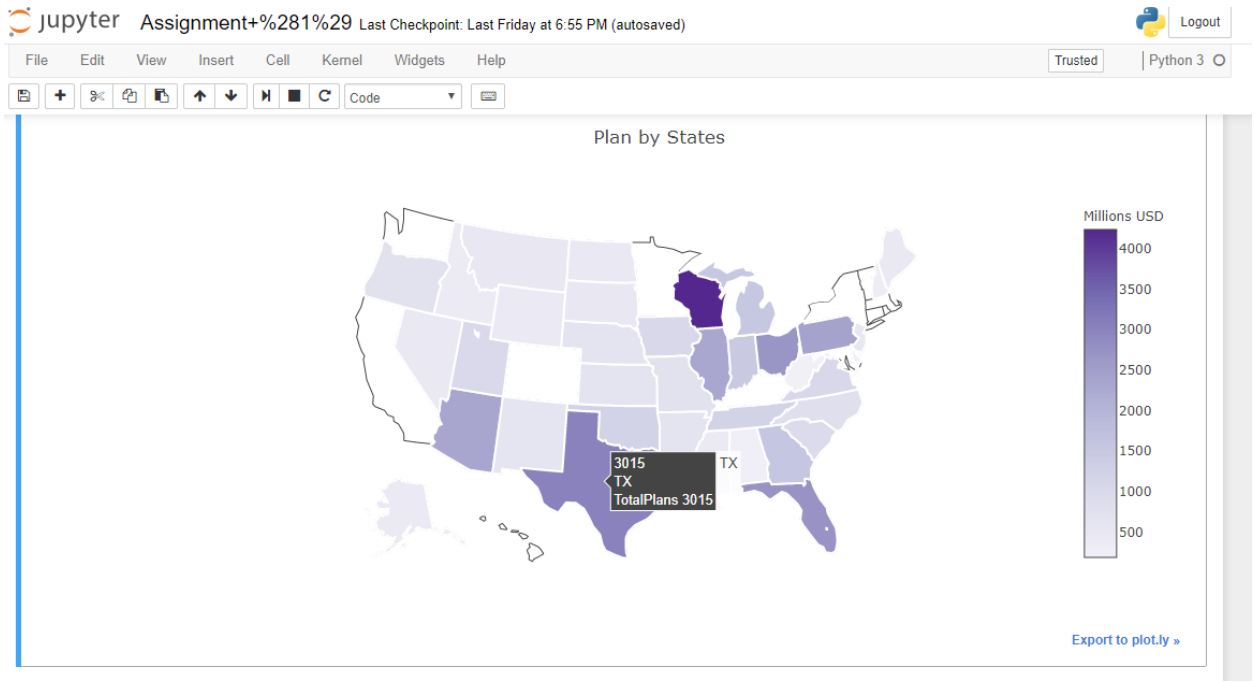
VISUALIZATION AND INTERACTION PART:

- **Exploration of Number of plans by States**
  - Number of plans for each states by counts the plans.

- **Analysis:** The below choropleth map shows the number of Medical plans state wise. Darker the shade more is the number of plane of the state and vice versa. Thus on exploration we can see that States like WI, TX have a plan more than 4000.

- **Python**: used pandas to group by states and count the plans.

Output: (Choropleth Plot)



BOX Plot:

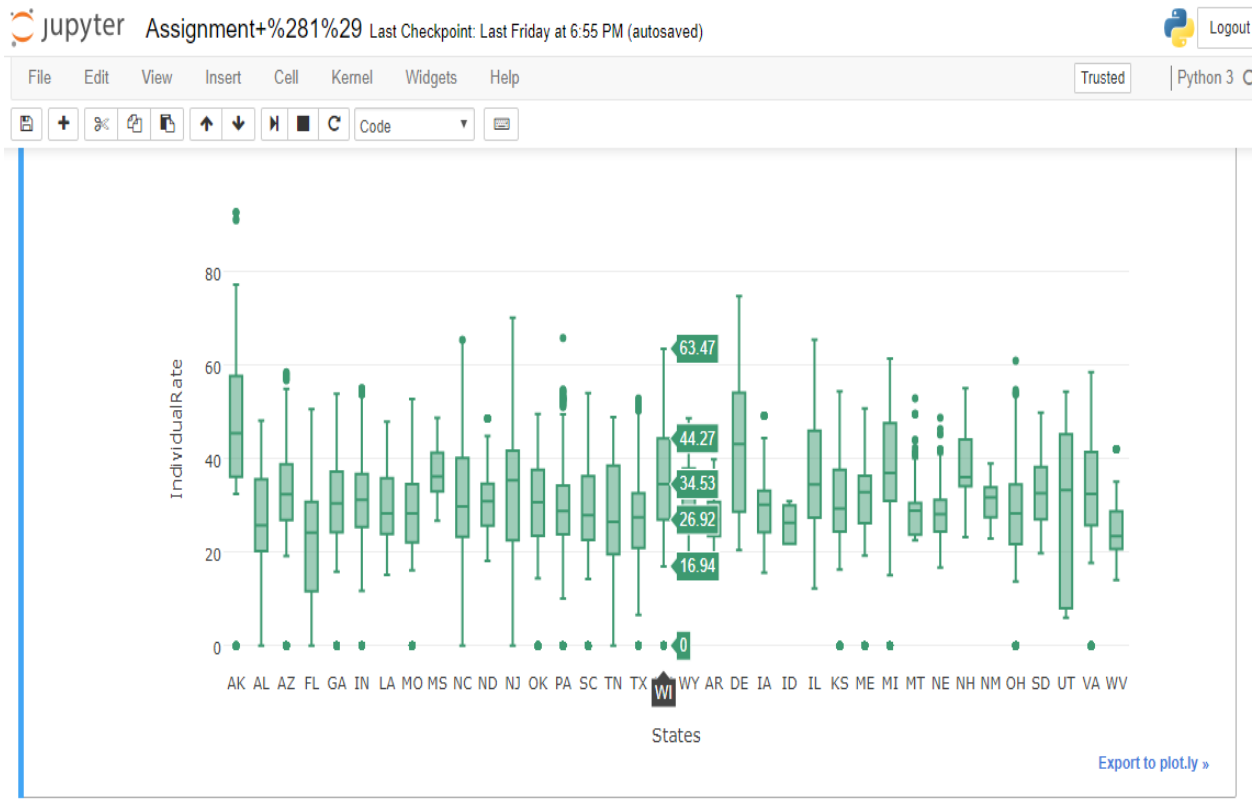
Showing distribution along with median, maximum, minimum, standard etc. This also displays a maximum that is quite supposed.

- **Individual Rates by State:**
  - Plot the Individual rates for each states to get most expensive states for insurance.

- **Analysis:** The below box plot map shows individual rates state wise. Higher the box then higher the rate for that state. Thus on exploration we can see that States like AK,DE have most expensive plans for insurance but states like UT have maximum number of plans.

- **Python**: used pandas Data frame for group by and other operations.

Output:



Pie Chart:

Showing total number of plans offered by United state Healthcare. These graph will also include user interactions using python library (plotly). Data getting from Kaggle. On kaggle this data is provided by CMS.

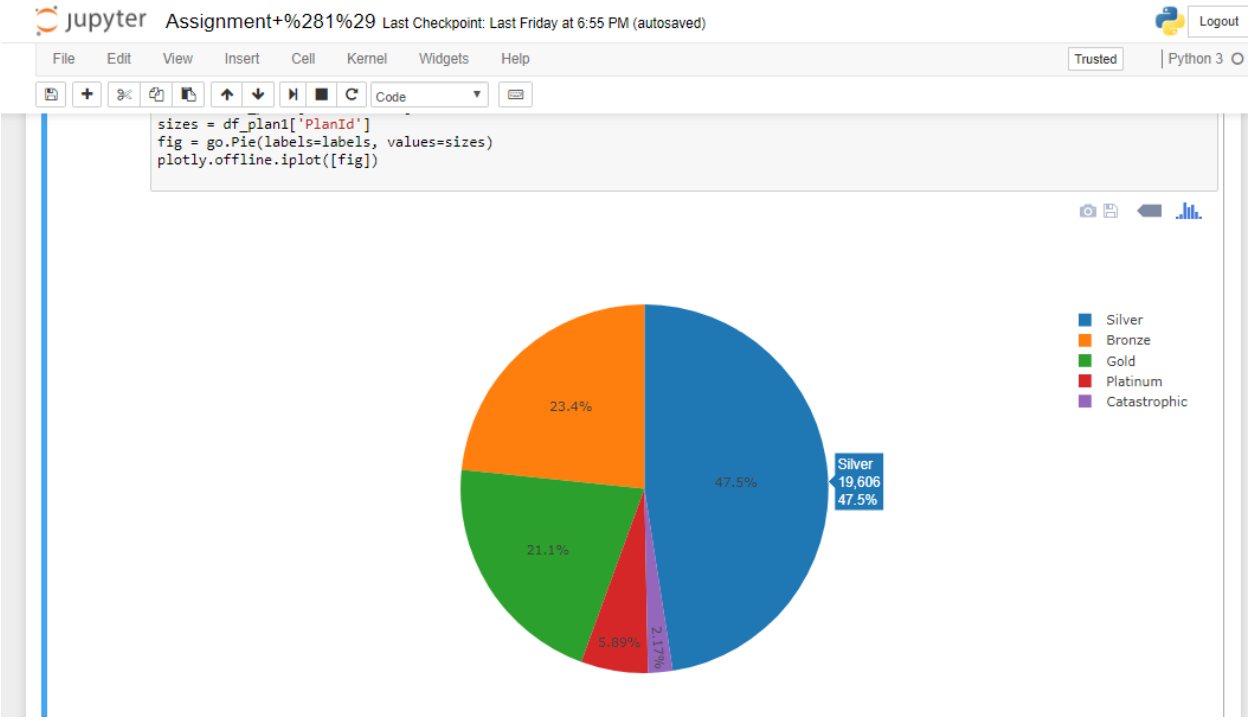
Types of Insurance Plans

Plot types of plan and distribution plan which plan is taken by mostly peoples.

- **Analysis:** The below Pie chart shows types of insurance plans. Mostly peoples takes silver plans and Bronze. This shows that these plans are cheap so mostly peoples takes these types of plans.

- **Python**: Used pandas to group by Data frame and for other operations.

Output:



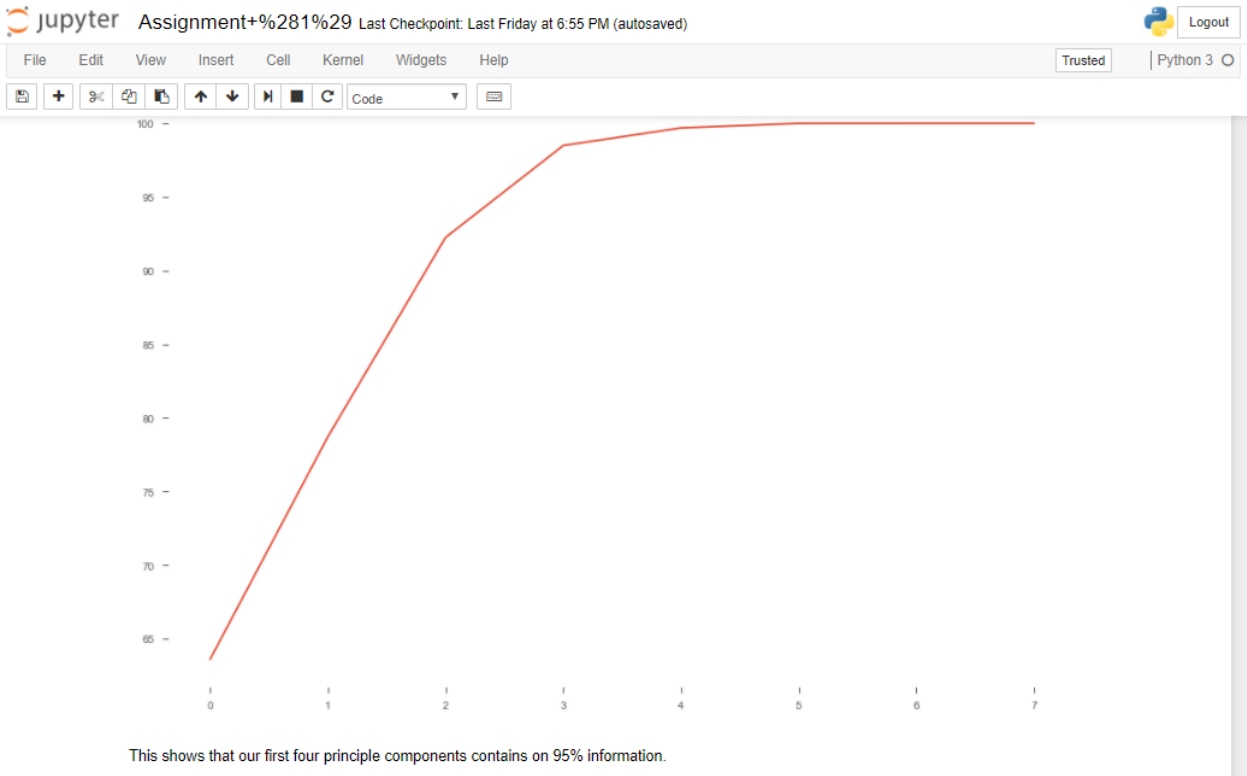
Algorithm (PCA):

Algorithm : (Principle Component Analysis)

- **Analysis:**
  - After sorting the Eigen pairs, the next question is "how many principal components are we going to choose for our new feature subspace?" A useful measure is the so-called "explained variance," which can be calculated from the eigenvalues. The explained variance tells us how much information (variance) can be attributed to each of the principal components.

Input: I used rate file to get variance between rate and to find the most import fields of rates.

Output:



REFERENCES:

Dataset:

- https://www.kaggle.com/hhsgov/health-insurance-marketplace

Analysis Research/Understanding:

- Three scholarly references (journal article, text) https://www.healthcare.gov/glossary/ - The glossary has a list of health insurance present.
- https://www.cms.gov/CCIIO/Resources/Data-Resources/Downloads/2-General_Information_Factsheet- 05032016_draft.pdf - This document outlines important information about the Health Insurance Marketplace Public Use Files (Marketplace PUF), including source data, file size, variables, key assumptions, analytic utility, and support information. A data dictionary is also available for each of the separate files within the Marketplace PUF.

- http://www.ixshealth.com/ - In this website we got to know what kind of insurance are present in the market and who can use which health insurance plans.