Applying Big Data Analytics to Open Health Care Data: Exploring Relationships Between Seniority and Performance in Healthcare Providers

Daniel Clarke, Dr. Ravi Rao, Prof. Maryelena Vargas
Fairleigh Dickinson University

RUTGERS

GS-LSAMP

Garden - State Louis Stokes Alliance for Minority Participation

Abstract / Introduction

With numerous initiatives to collect and make publicly available data sets in healthcare, a framework based on open source data mining technologies, primarily extending Python's SciPy ecosystem, is being designed to process them in search of novel correlations. [1] Using almost 900,000 practitioners connections and 3,000 hospitals, there is much to be found in the 50 fields present in the Center for Medicare and Medicaid's Hospital Value-Based Purchasing data. [2]

We use our framework to test a hypothesis about a correlation between age and performance. Visualizations were built to illustrate the findings. Linear regression analysis and Spearman's rank correlation analysis were used to find results for all hospitals and in individual specialties which suggest that there is no significant correlation between age and reported total performance score.

Background

With a number of public datasets and public data analytic toolkits in Python, we've been able to collect, join, sanitize, and analyze a large amount of healthcare data accounting for approximately 7% of the US Healthcare Workforce. As we process these datasets we develop a high level interface to the open source components with a focus on automating common forms of processing in the healthcare field.

Datasets and Views

- Medicare Hospital Value-Based Purchasing [2]: 3,090
- Medicare Practitioners [2]: 868,384
- Medicare Hospital Practitioner Connections: 1,067,937
- NY SPARCS Outcomes [3]: 1,907
- NY Practitioner Licenses [4]: 156
- NY Outcome-License connections: 1,865

Tools in use

- SciPy framework
  - Pandas [5]
  - Matplotlib [6]
- Scikit-Learn [7]
  - K-Means

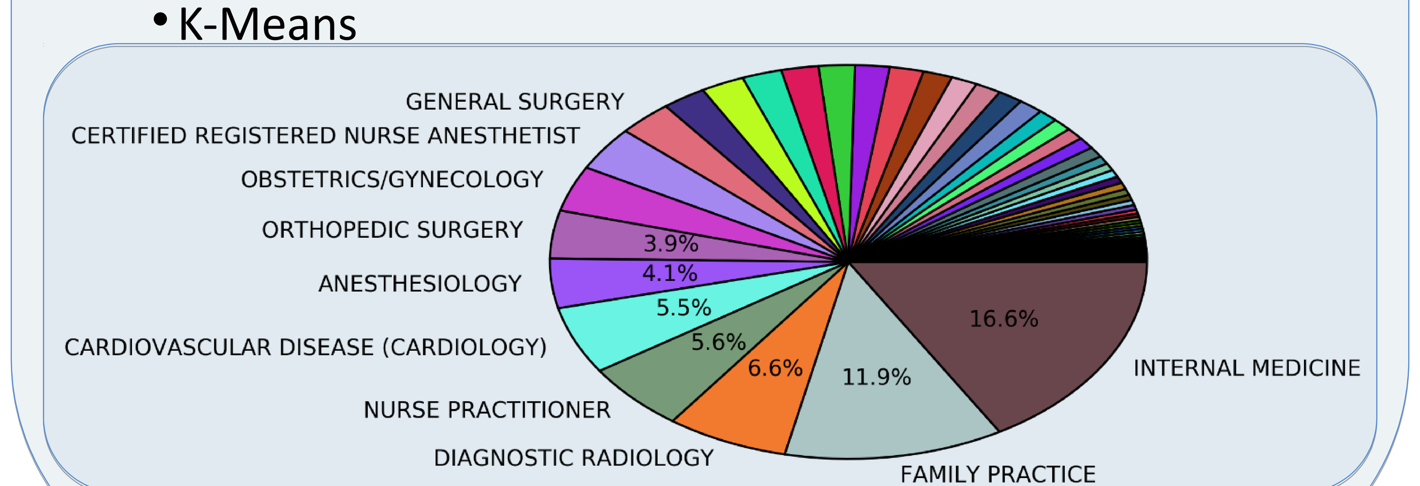

Figure 1. Pie chart showing top specialties in dataset.

Objectives

- Coherent wrapper around existing Python toolkits
  - Some level of automatic data sanitation, encoding, scaling, and sampling
  - High level visualizations of common models
  - Abstract analysis for identifying significant trends or correlations in data
- Application of framework to public heathcare data
- Try to understand the relationship between hospital performance score and seniority

Methods

Initially finding a weak correlation by linear regression we dug deeper to validate the finding. We split the data by specialty and by graduation year to help understand any skewness in the data and performed a number of tests..

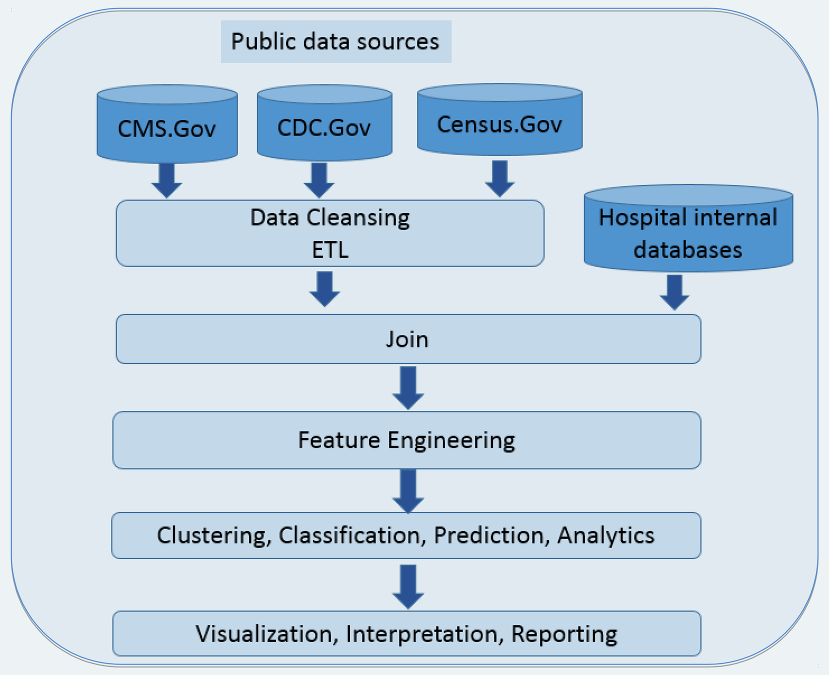

Figure 2. Diagram illustrating data analytic pipeline.

- Test for Normality with Skewness and Kurtosis
  - Performance Scores
    - Per year
    - Per specialty
  - Number of practitioners
  - Per hospital
    - Per graduation year
- K-Means trend categorization
  - Number of practitioner
    - Per graduation year

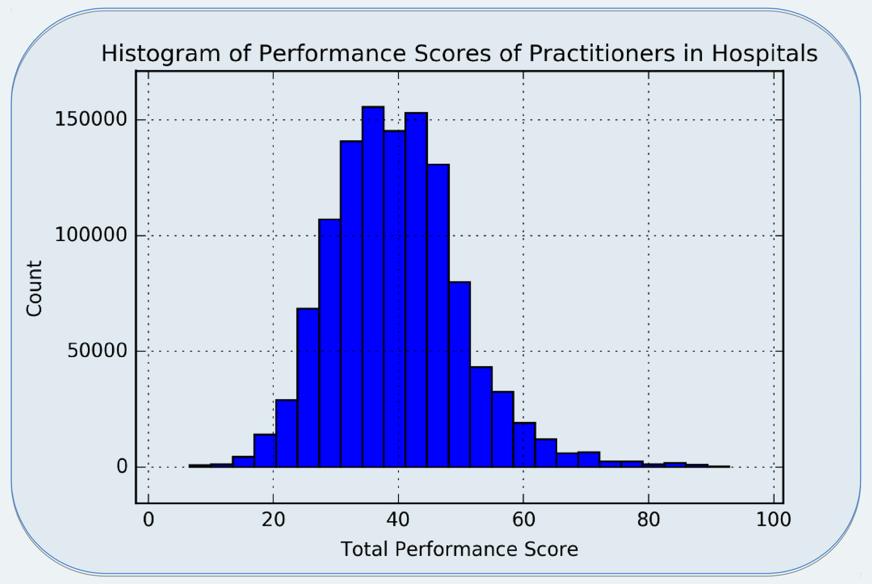

Figure 3. Histogram of all performance scores In the dataset.

- Linear Regression and Spearman Correlation Coefficient
  - Graduation year vs Performance Score
    - Per Specialty

Results

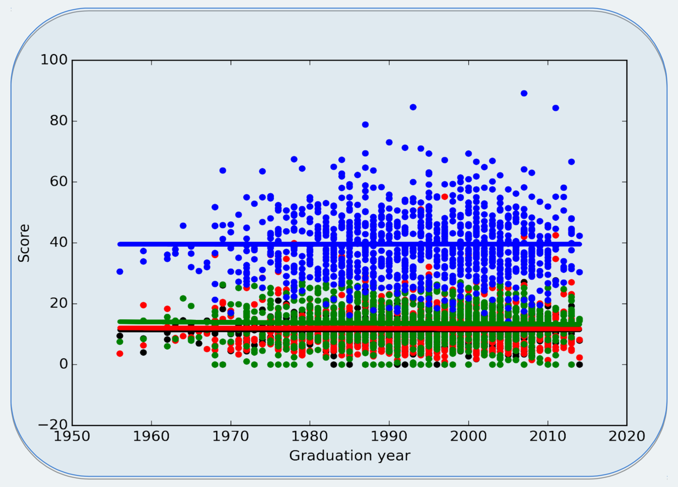

Figure 4. Simple linear regression plots of graduation year vs score.

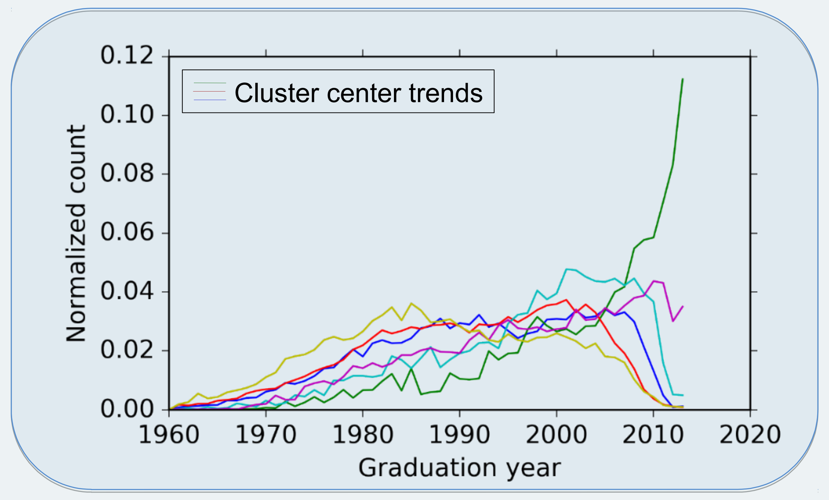

Figure 5. Plot showing k-means clustered trends of specialties for graduates in time.

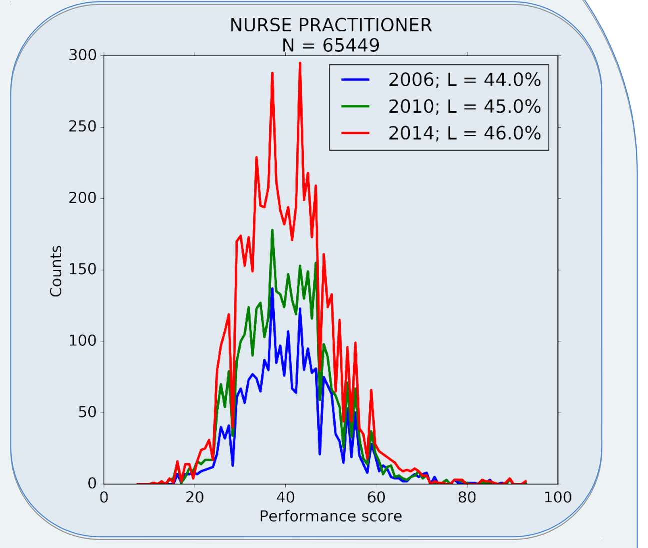

Figure 6. Plot showing recent quantitative trends in time for Nurse Practitioners.

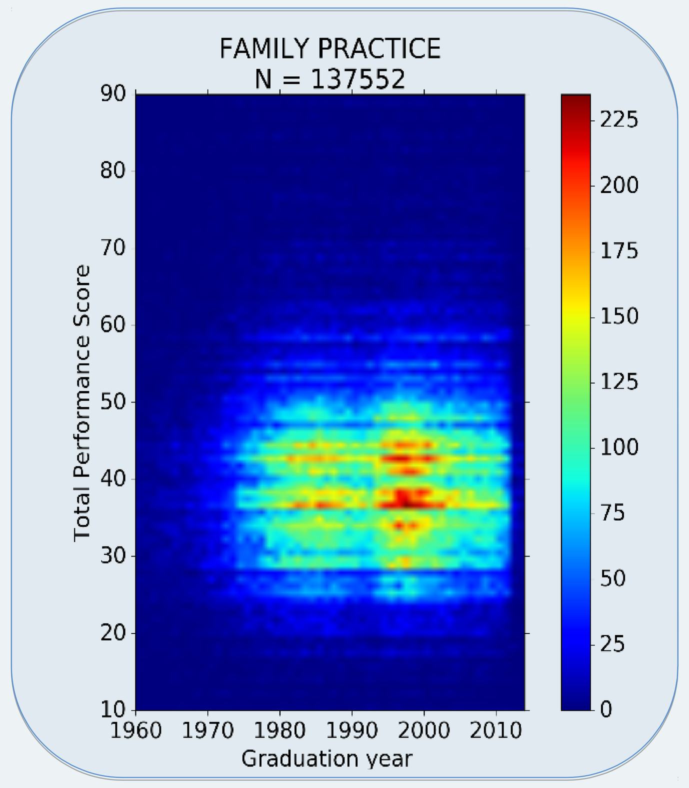

Figure 7. Heatmap capturing distribution of seniority and scores for Family Practice.

Specialty	Spearman correlation coefficient	Spearman P-value	R^2	m	μ	σ	N
GERIATRIC PSYCHIATRY	0.164	0.04507	0.049	0.164	38.638	8.254	150
ANESTHESIOLOGY ASSISTANT	0.11	0	0.006	0.075	37.643	9.151	1742
HEMATOLOGY	0.077	0.00781	0.004	0.055	39.558	9.811	1191
ALLERGY/IMMUNOLOGY	0.063	0.00087	0.003	0.041	39.625	9.368	2820
MEDICAL ONCOLOGY	0.058	0.00001	0.003	0.049	39.724	9.979	5864
PULMONARY DISEASE	0.051	0	0.003	0.05	39.434	10.317	21587

Table 1. Table showing top specialty correlation results.

Conclusions / Future Directions

- Correlation between seniority and hospital performance score is highly insignificant
  - 29 out of 94 specialties had spearman rank correlation <= 0.164 with $p < 0.05$
- Many more avenues for analysis planned
  - Multivariate analysis with xgboost for higher dimensional trends [8]
  - Using location attributes (e.g. income) as predictors to performance
- Continue to expand the framework on github at https://github.com/fdudatamining/framework
- These results available on github at https://github.com/fdudatamining/healthcare_specialties

References / Acknowledgments

[1] Rao, A.R., et al., "A framework for analyzing publicly available healthcare data", Proceedings of IEEE Healthcom Conference, pp. 1-5, 2015.
[2] https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6
[3] https://www.health.ny.gov/statistics/sparcs/
[4] http://www.op.nysed.gov/opsearches.htm
[5] Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)
[6] John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007)
[7] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." The Journal of Machine Learning Research 12 (2011): 2825-2830.
[8] Chen, Tianqi, and Tong He. "xgboost: eXtreme Gradient Boosting." R package version 0.4-2 (2015).

The Garden State Louis Stokes Alliance for Minority Progress is supported by a grant from the National Science Foundation.