Justin MacDonald, PhD
justin@justinmacdonald.net

Data Use Case Presentation

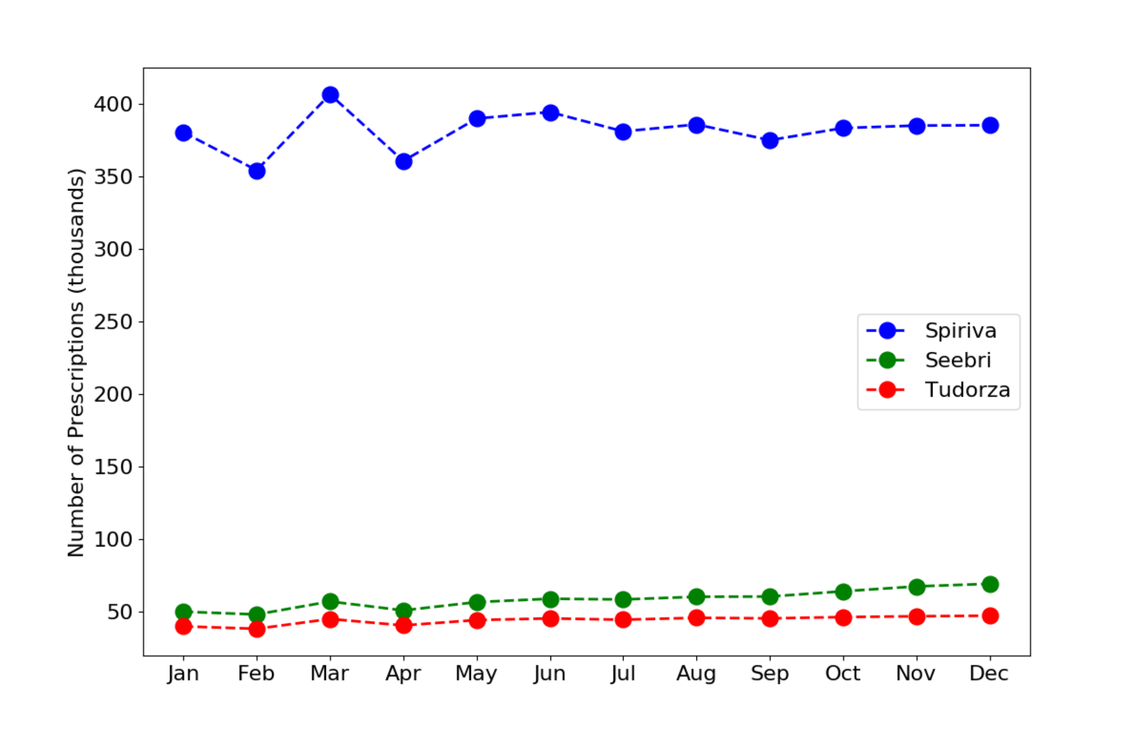
Project focus: Spiriva & Jardiance

These medications differ in several important aspects, making for a good comparison:

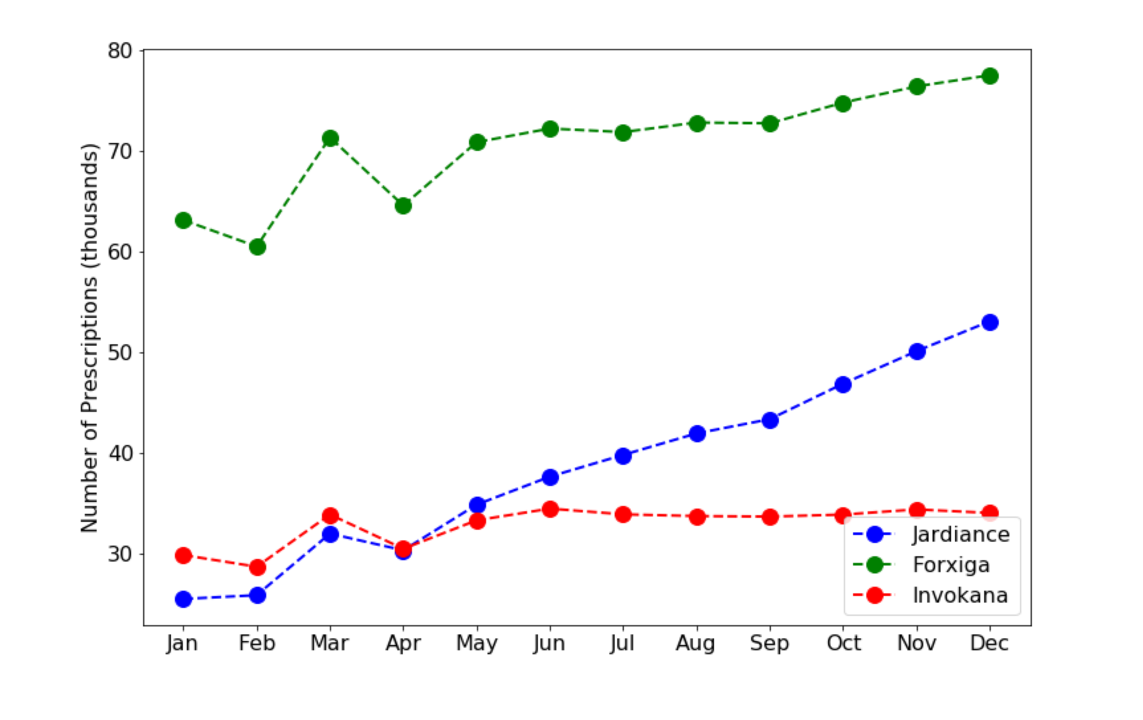
- Lifecyle
- Market dominance
- Revenue
- Drug class

Prescriptions written per month (2017)

Spiriva vs. competitors

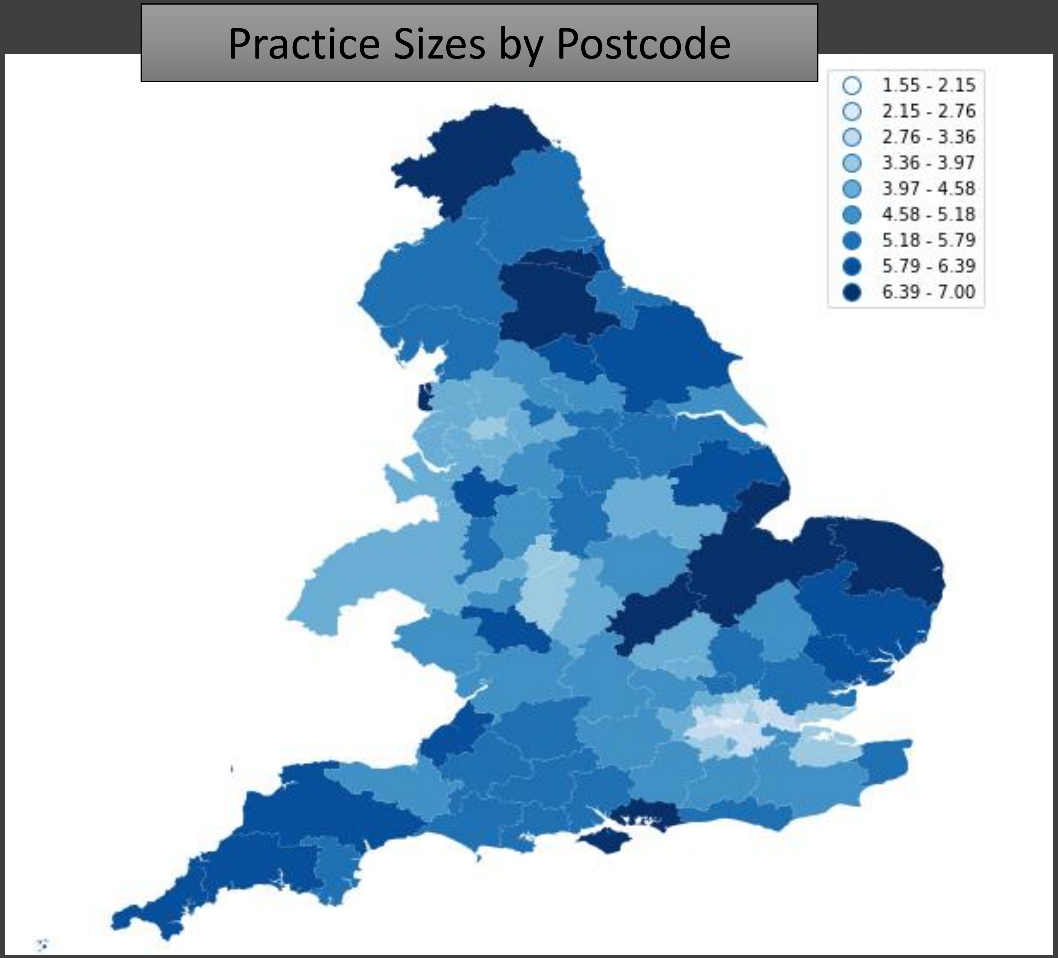


Jardiance vs. competitors



Investigation of GP practice sizes

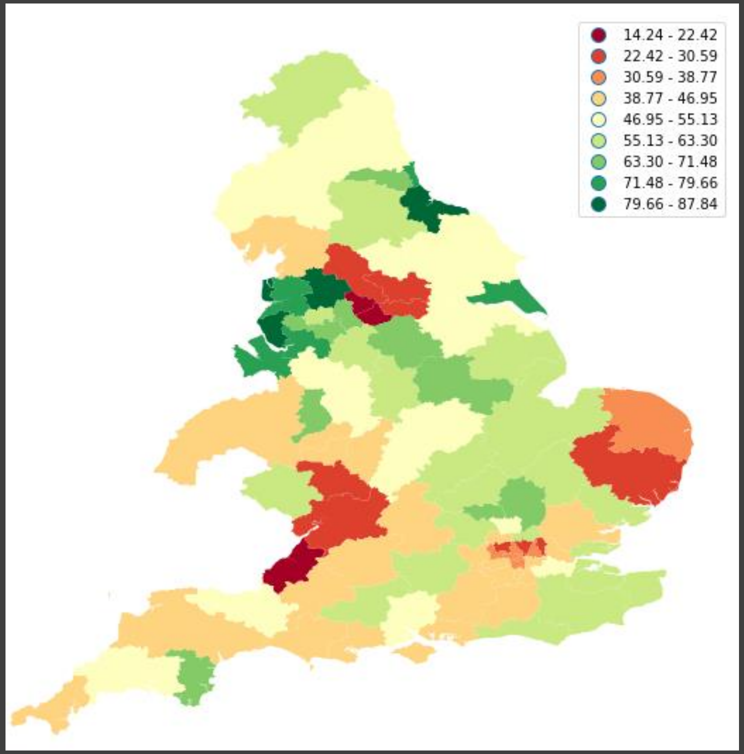
- Model estimates the number of GPs at each practice from 2017 prescription data
- Accurate estimates of practice size are required to allocate marketing resources effectively
- Conclusions:
    - Single-GP practices are less common than larges practices, allowing for less costly sales rep visits
    - As population density goes down, practice size goes up
    - Sales rep visits to more rural areas are absolutely necessary to reach large practices



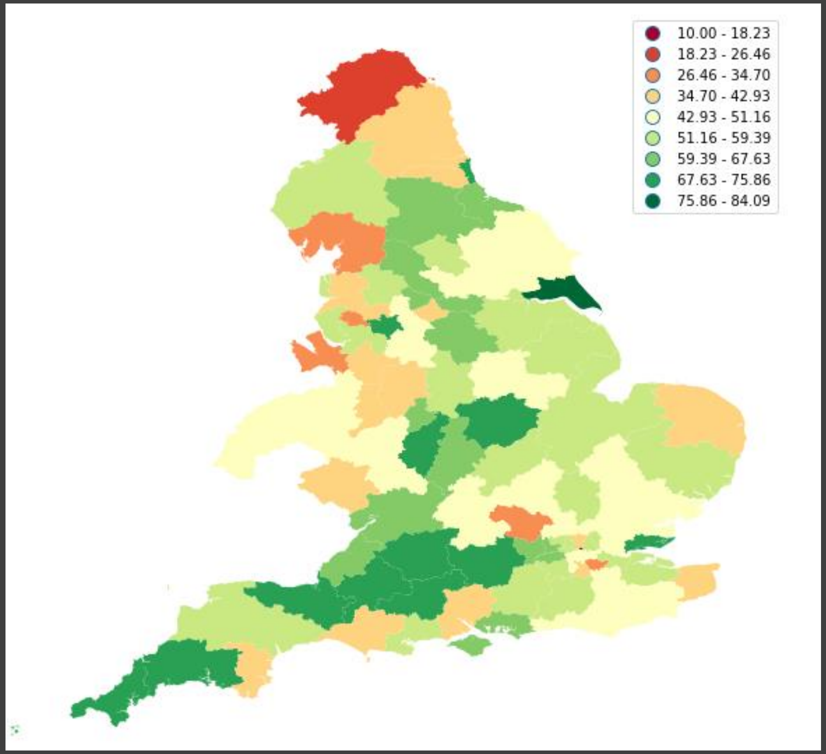
Top and bottom performers

- Regions with top-performing GPs (shown in green) vary with the medication considered
- Top Spiriva prescribers tend to be located in more rural areas
- Top Jardiance prescribers tend to be closer to urban areas
- This could be due to the relative difference in lifecycle between the two drugs

Percentiles by Postcode - Spiriva



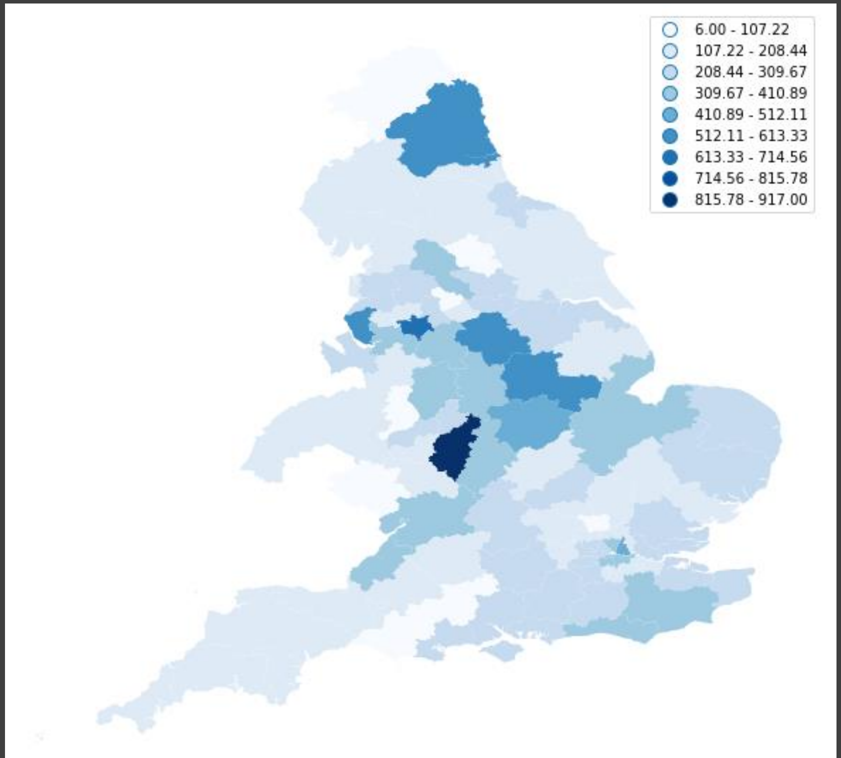
Percentiles by Postcode - Jardiance



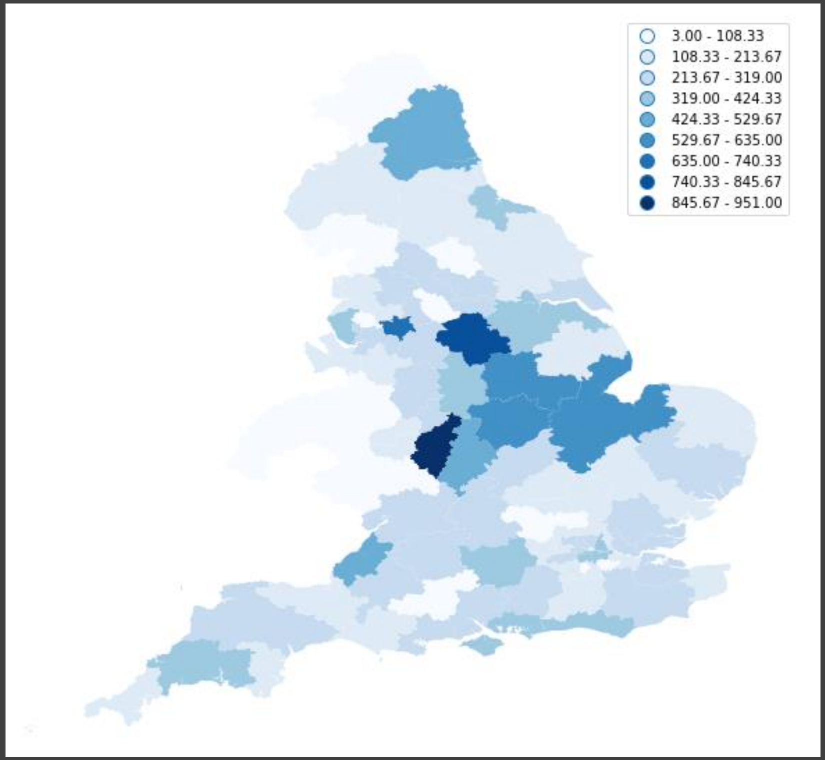
Face-to-face visit optimization

- Built a pair of machine learning models to predict new prescriptions written in 2018, one for each medication
- The models identify the top performers, allowing the more effective allocation of sales rep face-to-face visits
- Facing a 30% budget cut, the model intelligently allocates resources to produce an additional revenues of 1.1m Euros in Jardiance sales

GP Visitation Map - Spiriva



GP Visitation Map - Jardiance



Data Science Details

Dataset and Tools

- All analyses and models were completed using the GP practice prescribing data for all of 2017 (the 2016 dataset was also included in some analyses)
  - About 35 GB of data representing 2.4 billion prescriptions
- Software tools
  - Language of choice: Python
  - Data consolidation: PySpark/Hadoop on a Google Cloud Platform cluster
  - Data manipulation: Pandas, Numpy
  - Model construction: Scikit-learn, Scipy
  - Data visualization: Matplotlib, Geopandas
- Code is available at https://github.com/justinmacdonald/datausecase

Overall Rx Behavior across all of 2017

Question 1.1: Distribution of Monthly Rx Per Practice

Mixture modeling approach

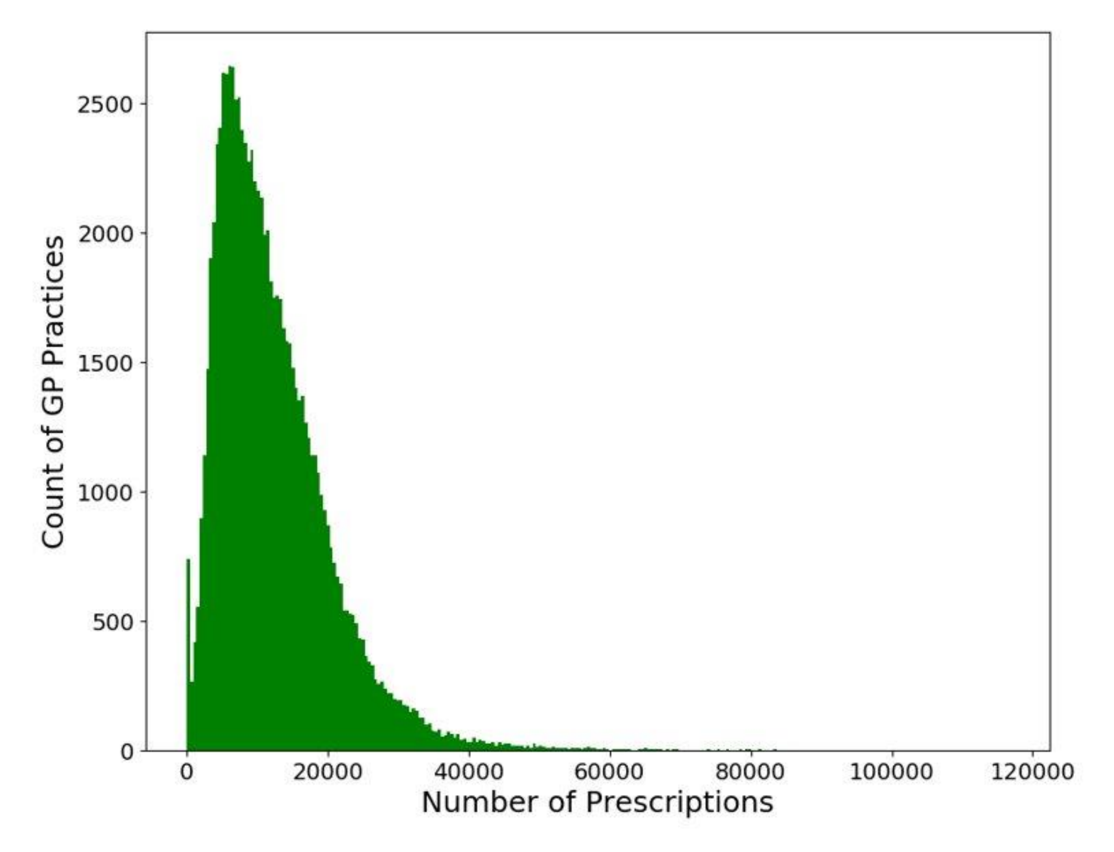
- The histogram to the right is is likely a representation of a mixture distribution
- If we assume that the prescribing behavior of GPs is iid with mean μ and variance σ^2, then n-GP practices are iid with mean $n\mu$ and variance $n\sigma^2$
- Each practice size can be treated as a component in the mixture:

$$\mu_{mix} = \sum_{n=1}^{\infty} w_n n\mu$$

$$\sigma_{mix}^2 = \sum_{n=1}^{\infty} w_n n\sigma^2$$

  w_n is the weight of the nth component
- Choose a common distribution for the components
  - Poisson (overdispersed)
  - Negative binomial
  - Gaussian

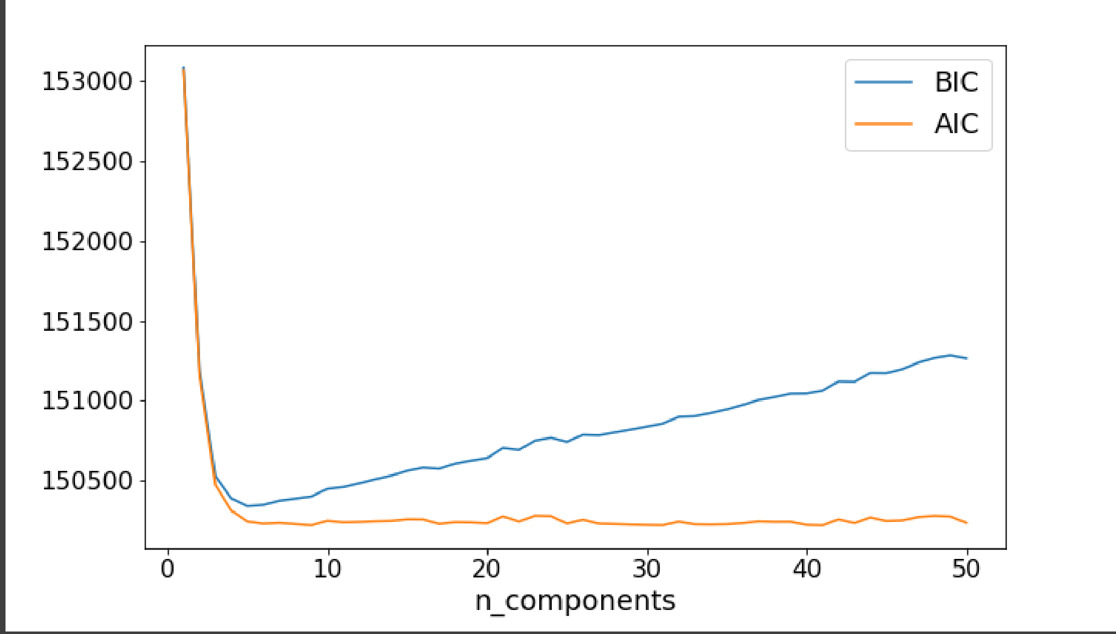
Monthly Prescriptions per Practice - Overall



Gaussian mixture model

- Implemented in Scikit-learn
- I debated about the number of components to include in the model
    - Option 1: Use the BIC to determine the number of components in the mixture
    - Option 2: Investigate the distribution of GP practice sizes to determine the number of components
- I went with Option 2

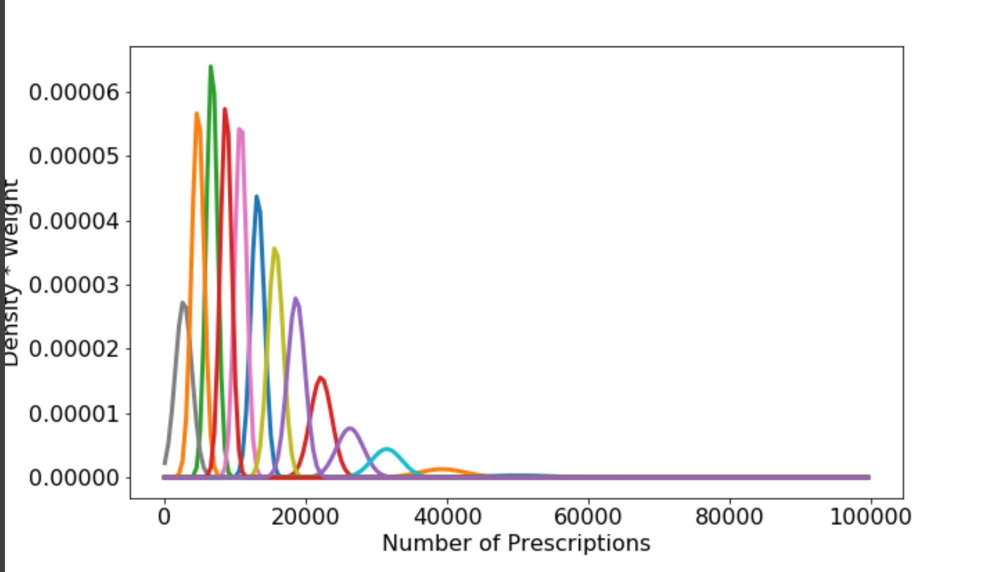
AIC/BIC by Number of Components



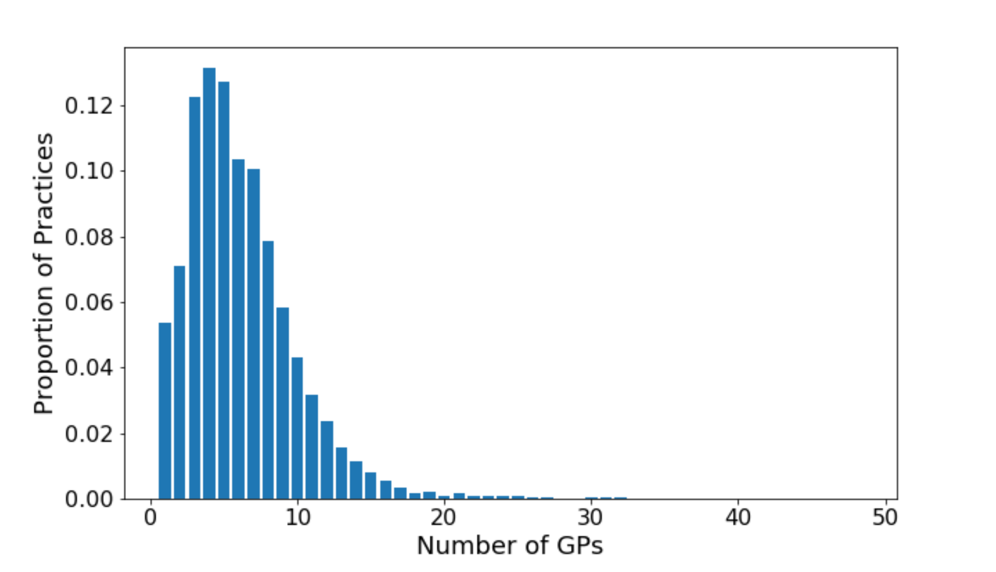
Gaussian mixture model

- 98% of GP practices have 15 GPs or fewer
- Accordingly, I chose 15 components for the mixture model

Component Distributions



Distribution of GP Practice Sizes

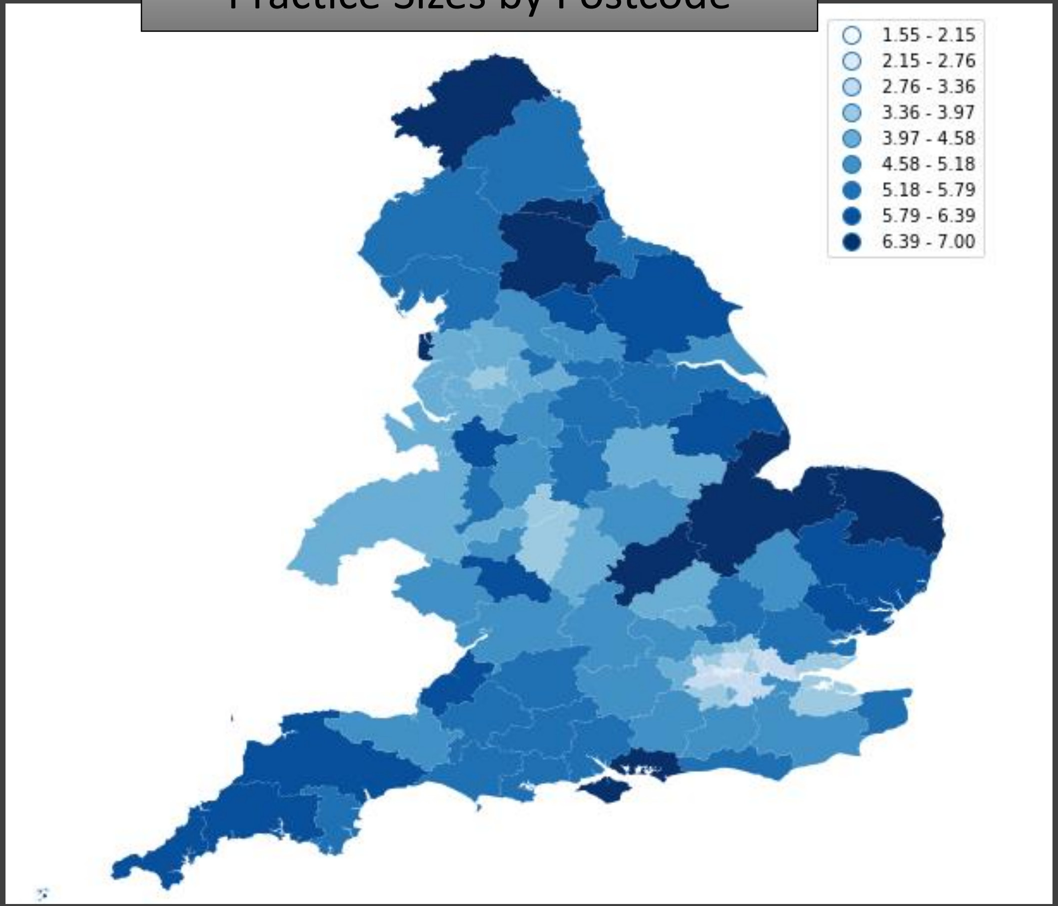


From: https://www.nhsbsa.nhs.uk/prescription-data/organisation-data/practice-list-size-and-gp-count-each-practice

Categorizing practices

- Assign practices to components (and therefore numGP categories) according to maximum likelihood
- The predict_proba method of GaussianMixture in Scikit-learn takes a sample and returns the likelihood for each of the component distributions
- Incorporate monthly data to improve predictions
- Calculate the product of the likelihoods across months for each of the components, choose the maximum as the category

Practice Sizes by Postcode



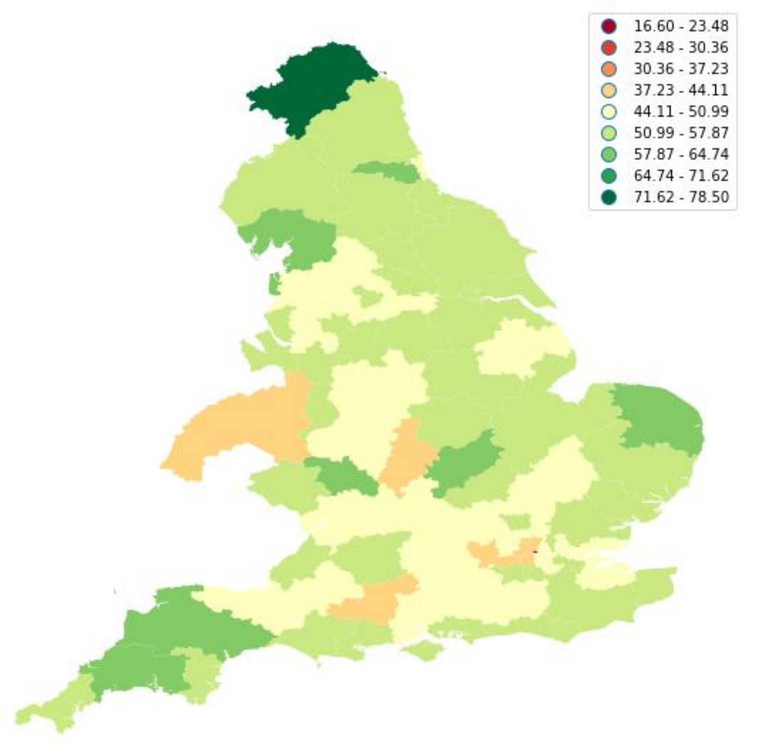
Practice size is inversely related to population density, please see Table 2.4: https://www.ifs.org.uk/uploads/publications/comms/R101.pdf

Spiriva, Jardiance, and Overall Rx Behavior

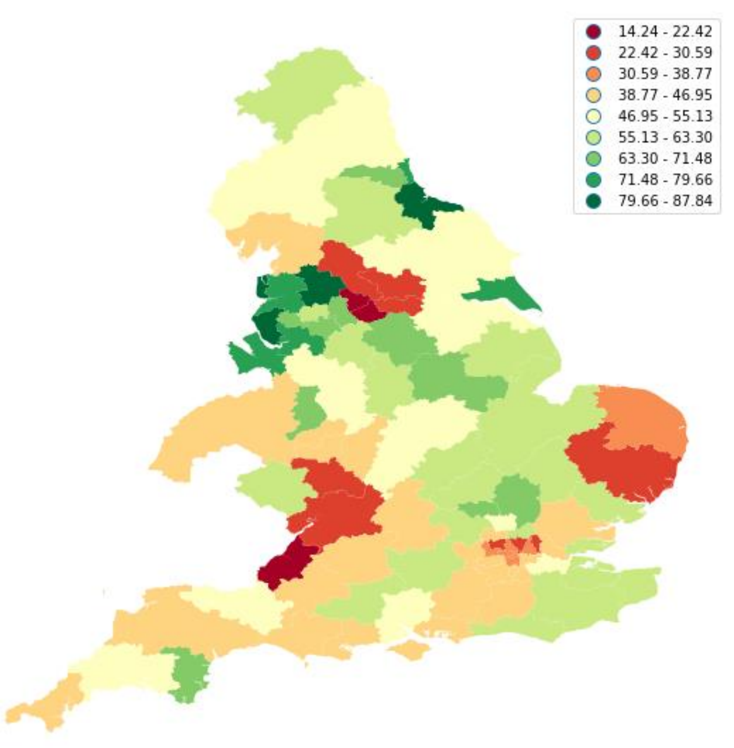
Question 1.2: Top and Bottom Performers

Geospatial attributes

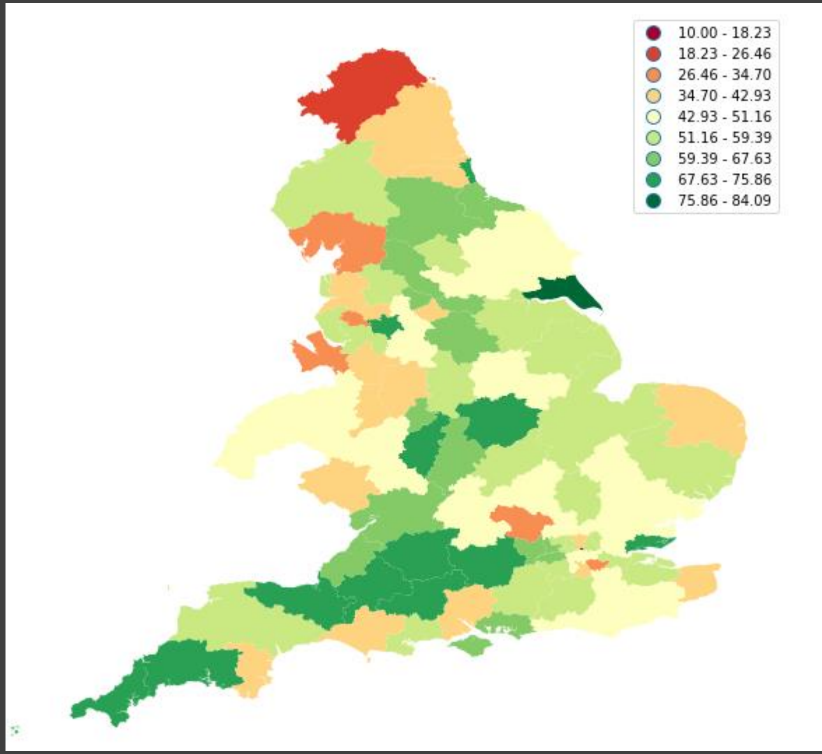
Percentiles by Postcode - Overall



Percentiles by Postcode - Spiriva



Percentiles by Postcode - Jardiance



Market share

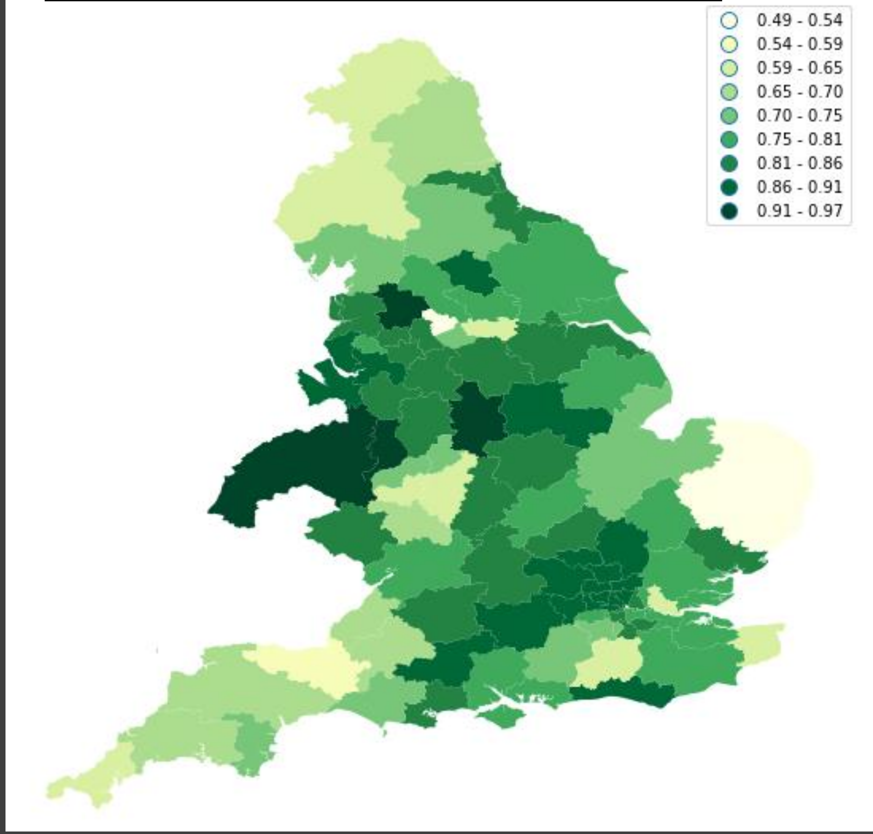
Spiriva market share =

$$\frac{Spiriva}{Spiriva + Seebri + Tudorza}$$

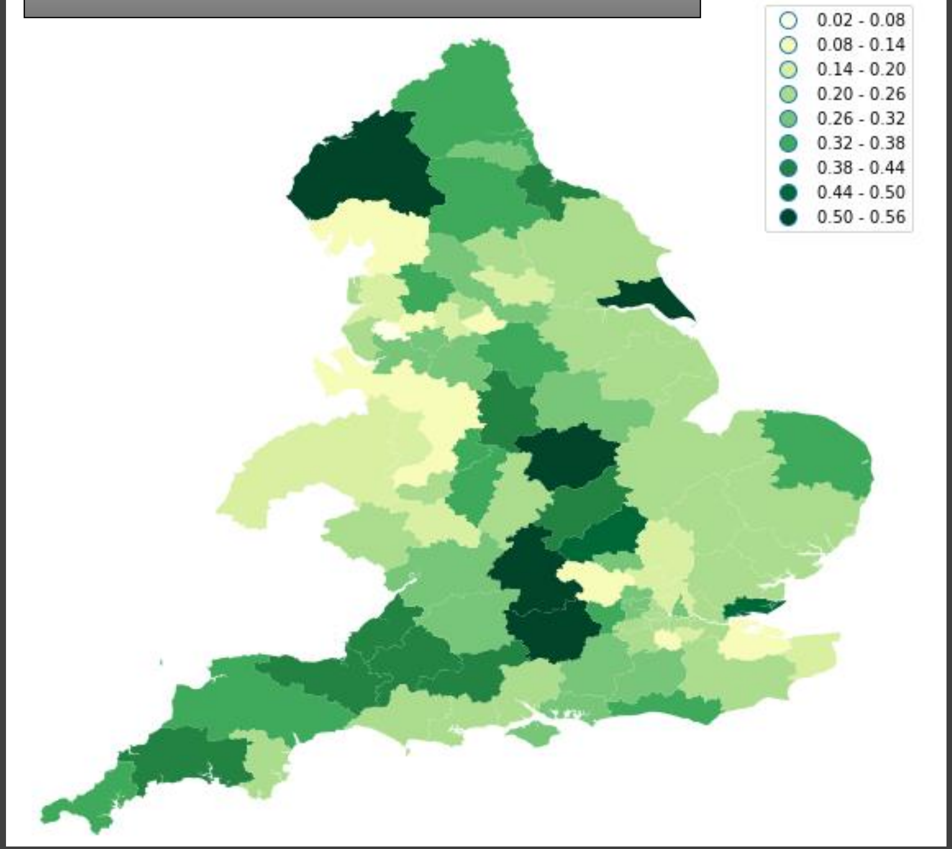
Jardiance market share =

$$\frac{Jardiance}{Jardiance + Forxiga + Invokana}$$

Market Share - Spiriva



Market Share - Jardiance

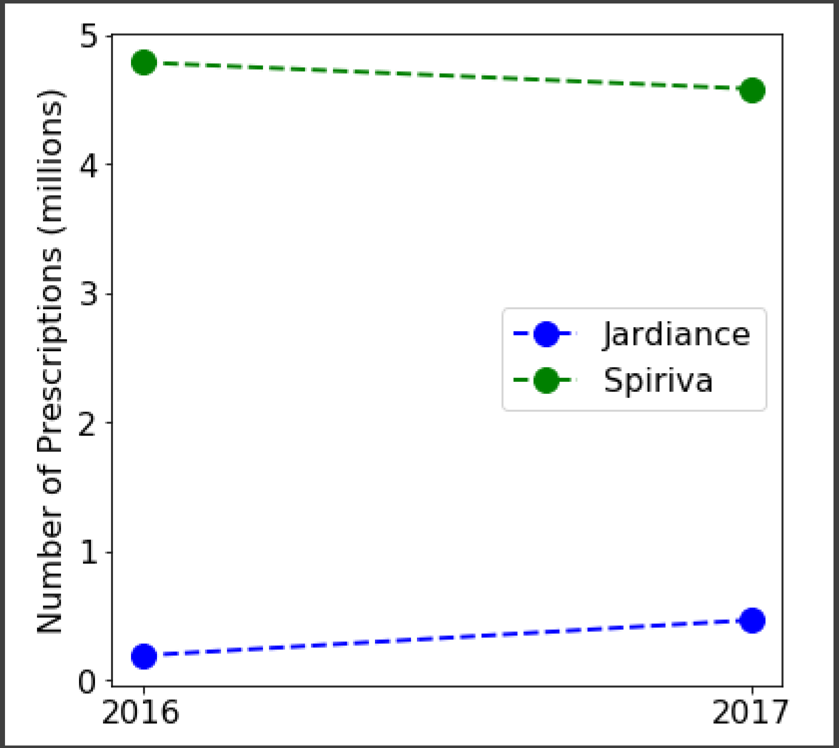


Spiriva and Jardiance

Question 2:
Face-to-face Visit Optimization

Metric and approach

Rx per Year in England

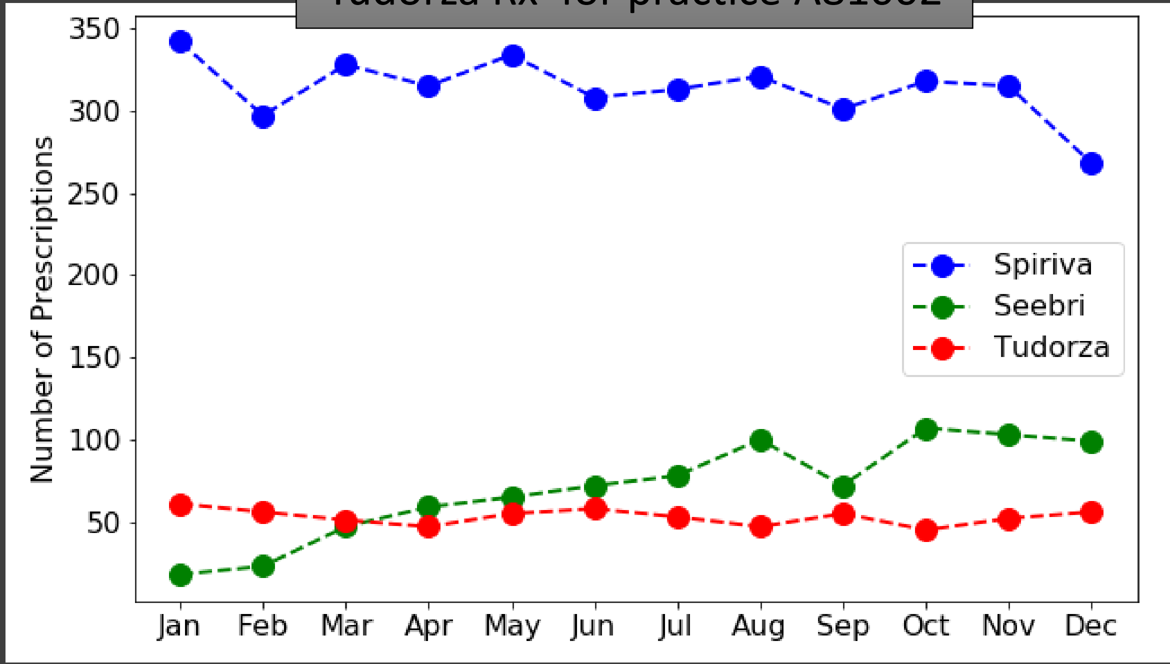


- Metric: the number of new prescriptions written, defined as the number of prescriptions written in 2017 minus the number written in 2016
- Population: GP practices in England
- Goal: train a machine learning model to predict the number of new prescriptions written in 2017 from the 2016 data on prescribing behavior
  - These predictions are made at the practice level since data on individual GPs are not available
- Build two models: Spiriva and Jardiance
- These predictions will inform how f2f visits are allocated going forward

Model features

- Used the 2016 dataset
- Prescribing data for all preparations in the same chapter of the British National Formulary
  - Spiriva: Chapter 3, Respiratory system
  - Jardiance: Chapter 6, Endocrine system
- Grouped by molecule, so different preparations of the same molecule were grouped together
- For each combination of GP practice and molecule, fit a line to the 12 months of Rx data
- The slopes of the lines were used as the features in the machine learning model
  - Spiriva model: 95 features
  - Jardiance model: 150 features

Example: Spiriva, Seebri, & Tudorza Rx for practice A81002



Model implementation & results

- Random Forest Regressor implemented in Scikit-learn
- Hyperparameters tuned using RandomSearchCV and GridSearchCV

Results

	Spiriva Model	Jardiance Model
R^2 - Train	0.74	0.75
R^2 - Test	0.30	0.54

Allocation of face-to-face visits

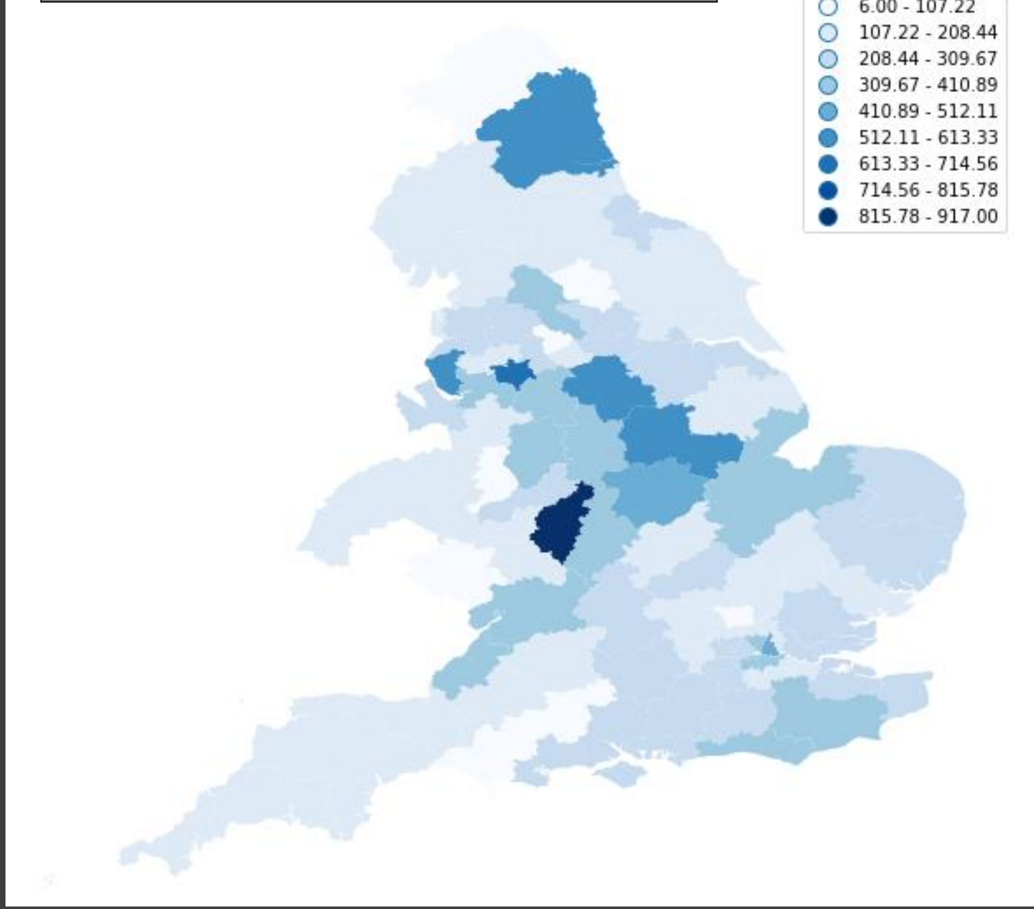
- Used the trained model to predict new 2018 prescriptions from the 2017 data
- Sorted the practices according to

$$\frac{estimated\ new\ Rx}{estimated\ number\ of\ GPs}$$

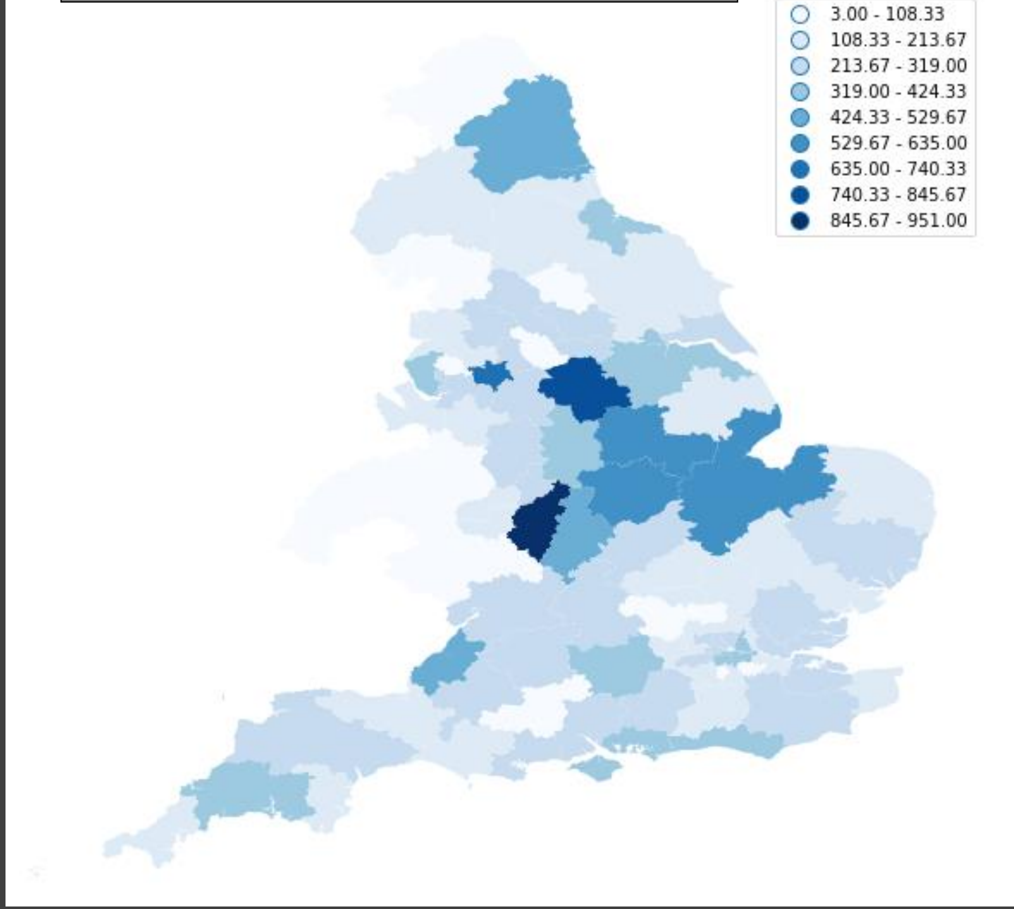
- GPs in the top 70% were put on the visit list for 2018

Allocation of face-to-face visits

GP Visitation Map - Spiriva



GP Visitation Map - Jardiance



Appendix: Assumptions

1. 1 face-to-face visit per week for the year increases new Rx by 5%
2. 70% of that visit rate increases new Rx by 3.5%
3. A new prescription is worth 300 Euros to BI

Appendix: Revenue calculations

	Spiriva	Jardiance
New Rx predicted by model (all GPs)	-118,152	364,253
3.5% positive change from f2f visits	+4,135	+12,749
Revenue from f2f visits	1,240,596	3,824,700

	Spiriva	Jardiance
New Rx predicted by model (top 70% of GPs)	55,419	328,650
5% positive change from f2f visits	+2,771	+16,432
Revenue from f2f visits	831,285	4,929,750