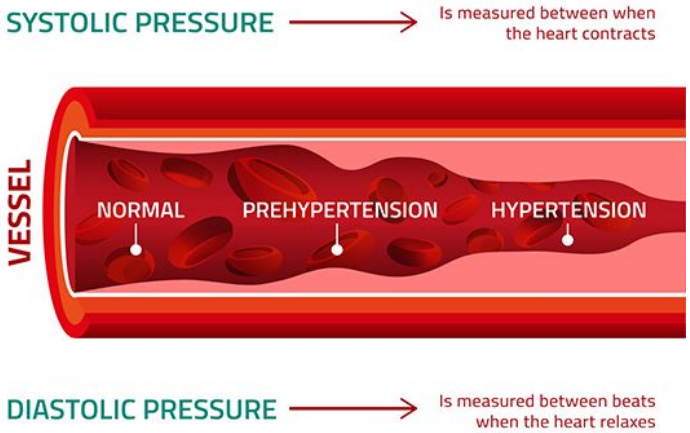
Predicting Hypertension

Capstone 1 Project

Alexia Marcous

Hypertension - Definition

- High blood pressure
  - Blood pressure is the force on artery walls
  - European guidelines = 140+ / 90+
  - United States lowered to 130+ / 80+ in 2017



- Often abbreviated as “HTN”

- **The** major risk factor for cardiovascular disease

Hypertension - an epidemic of global proportions

Globally

- Affects 1 in 4 men and 1 in 5 women
- More cases in low- and middle-income countries

United States

- One-third of the population has hypertension
- Untreated, 80% will die from a heart attack or stroke
- Attributed to 1,000 deaths per day

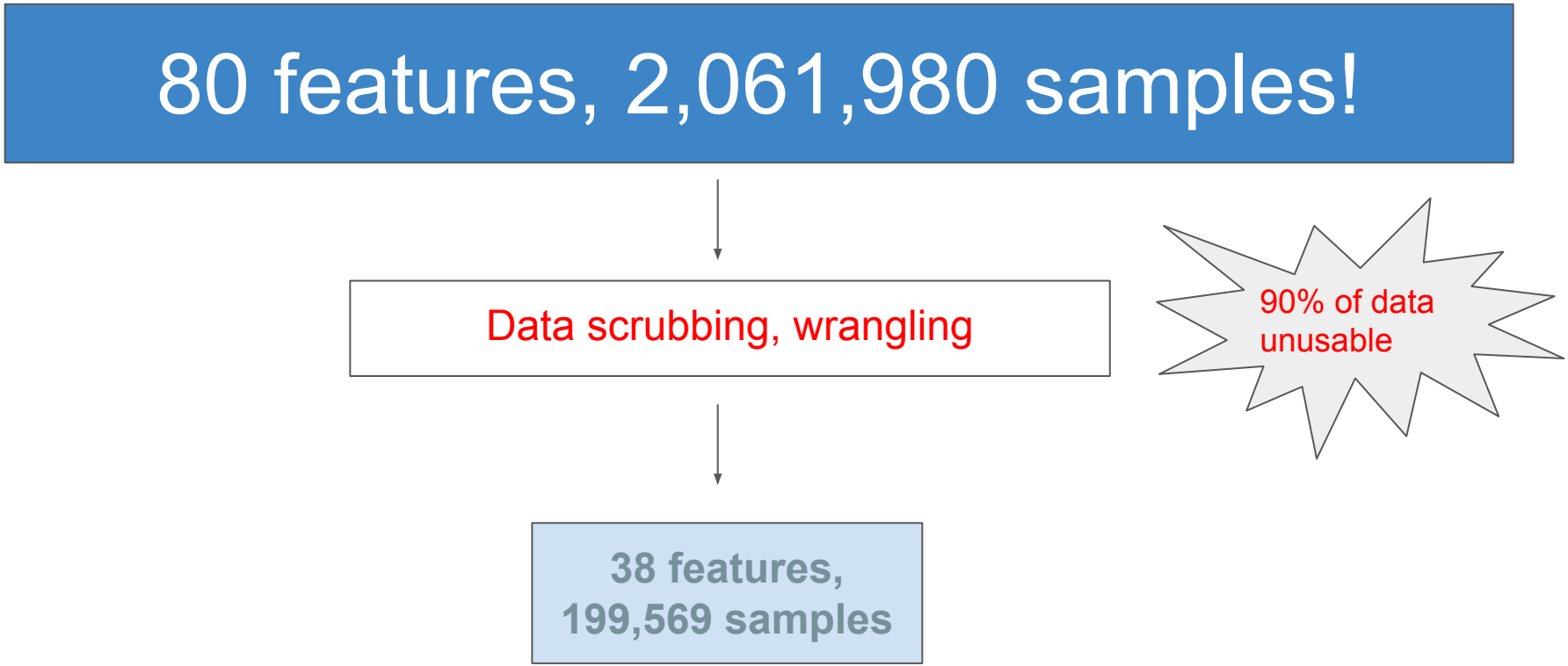
Need for prediction

- “Silent Killer” - goes undetected because often asymptomatic
- Many cases are preventable - screening, education and support for lifestyle changes can make a difference

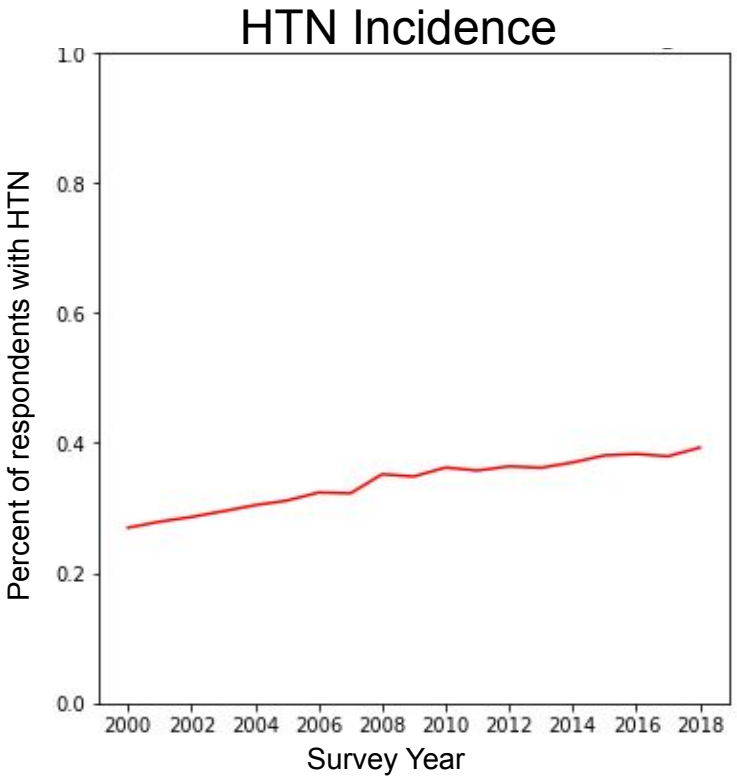
Data Source

- National Health Interview Survey (NHIS)
- 100,000 people annually
- Integrated Public Use Microdata Series (IPUMS) encodes
- Data is freely available at ipums.org

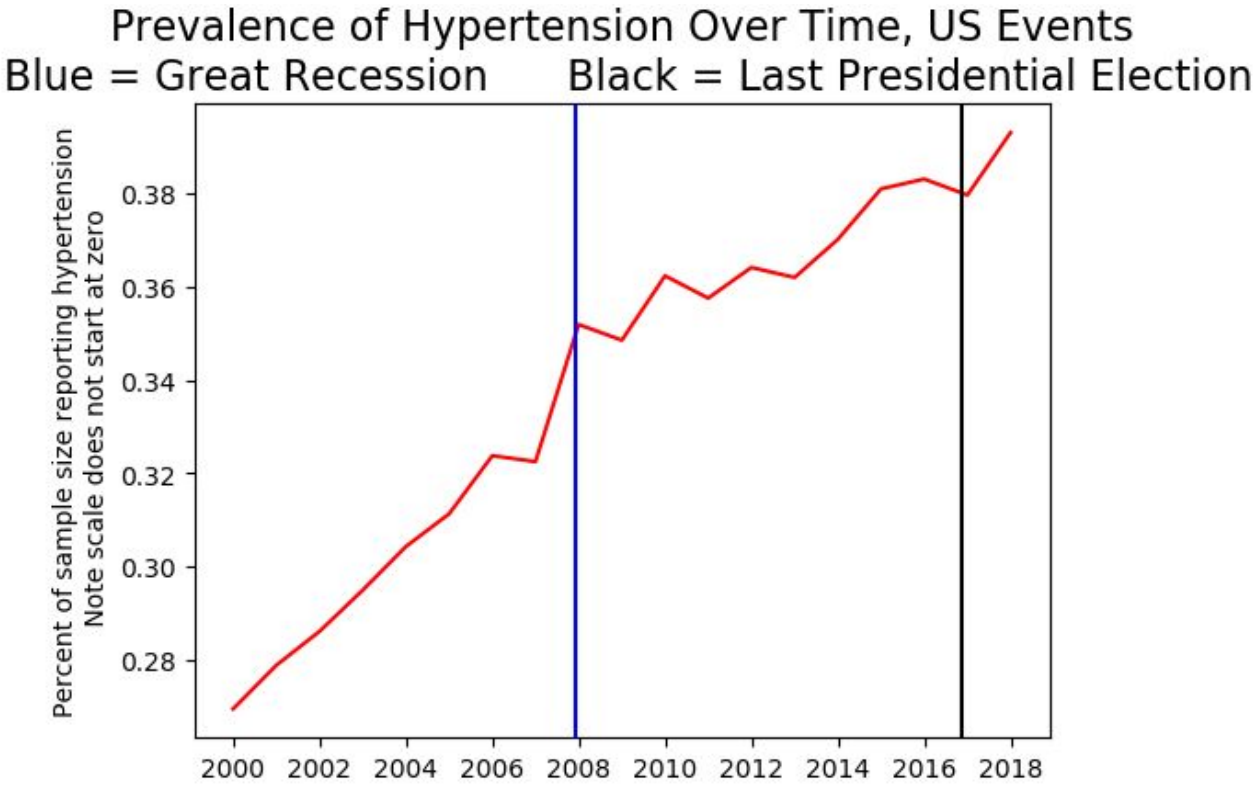
Data Wrangling



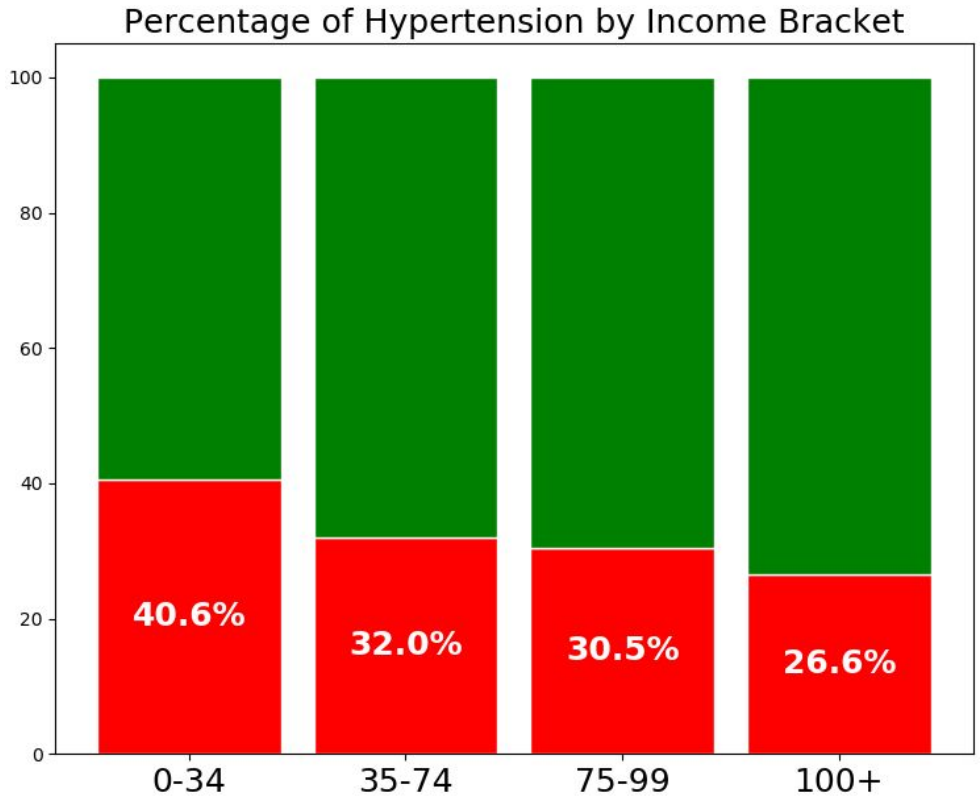
Exploratory Data Analysis - Prevalence Trend



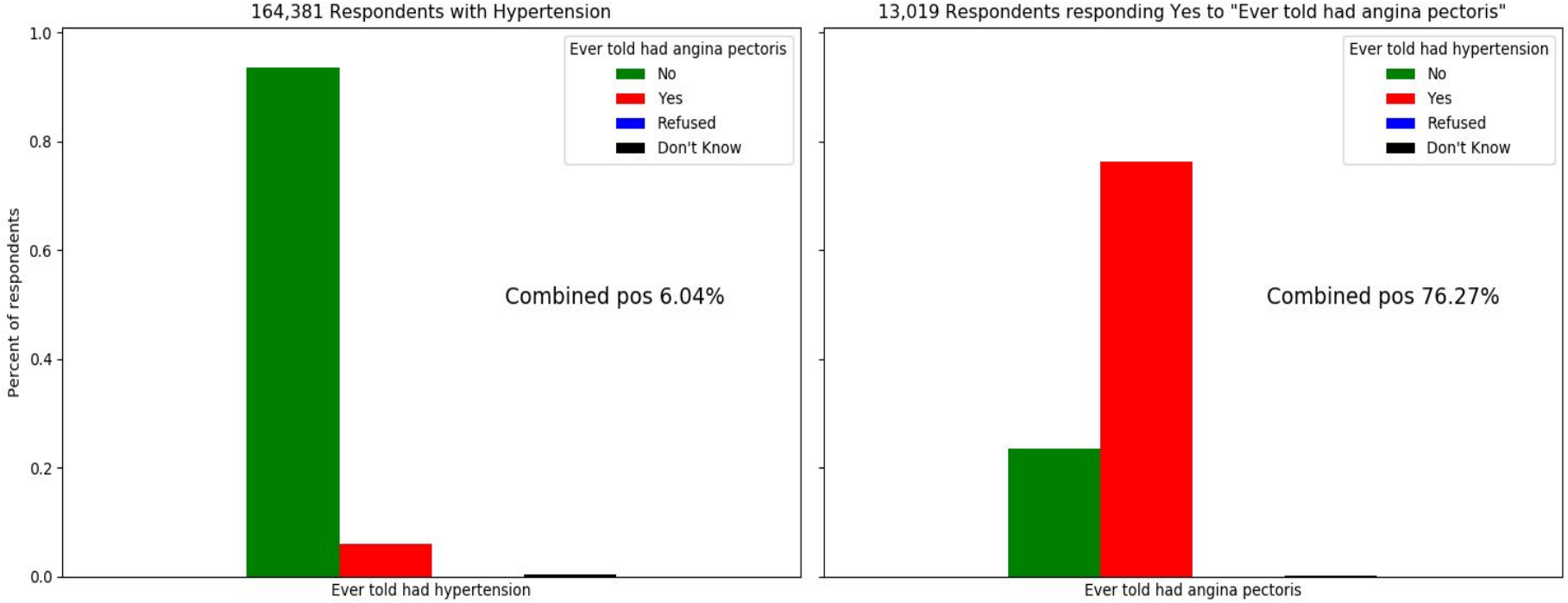
Exploratory Data Analysis - US Events



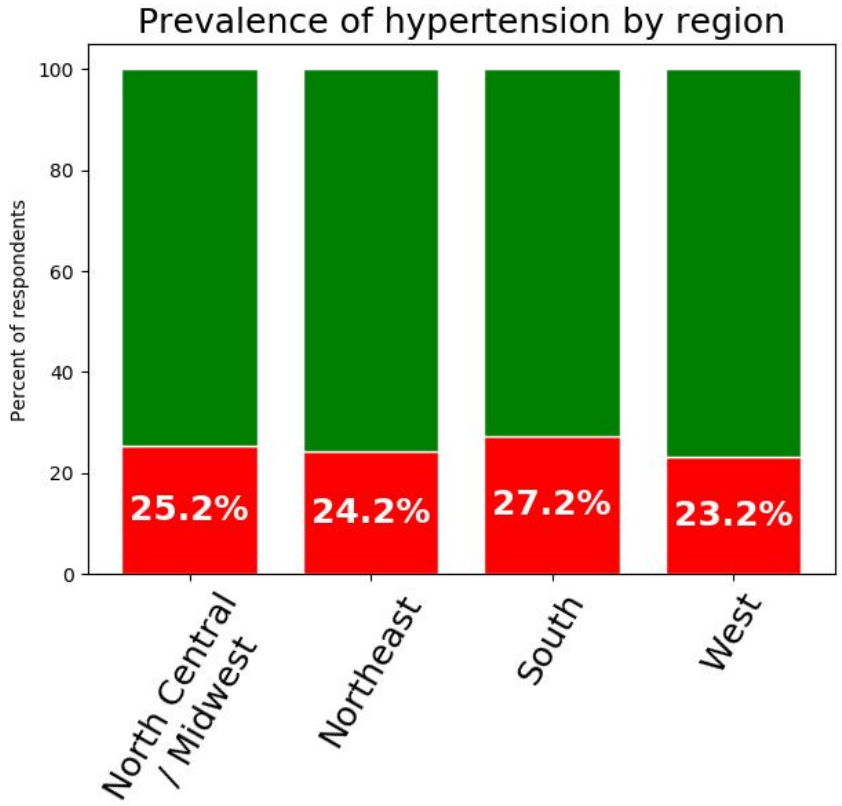
Exploratory Data Analysis - HTN and Income



Exploratory Data Analysis - HTN and Comorbidities



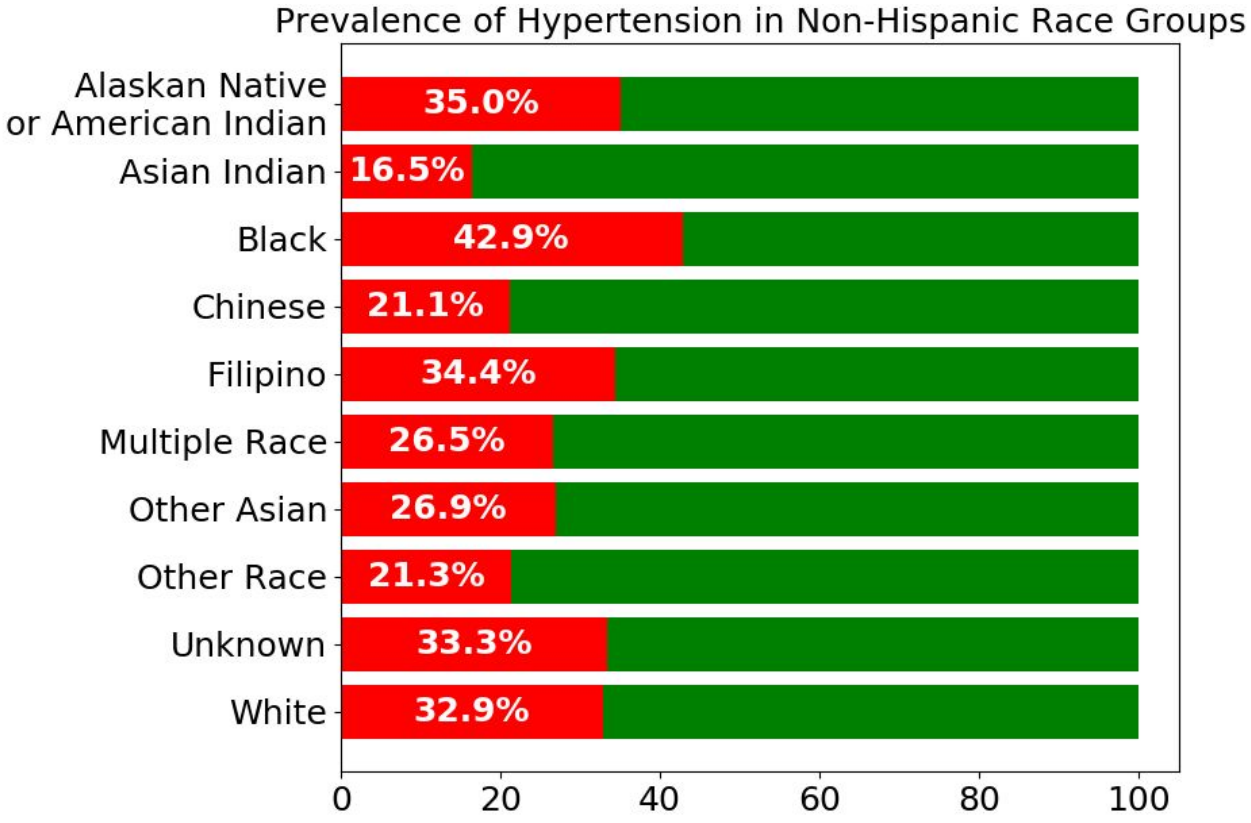
Exploratory Data Analysis - HTN and Region



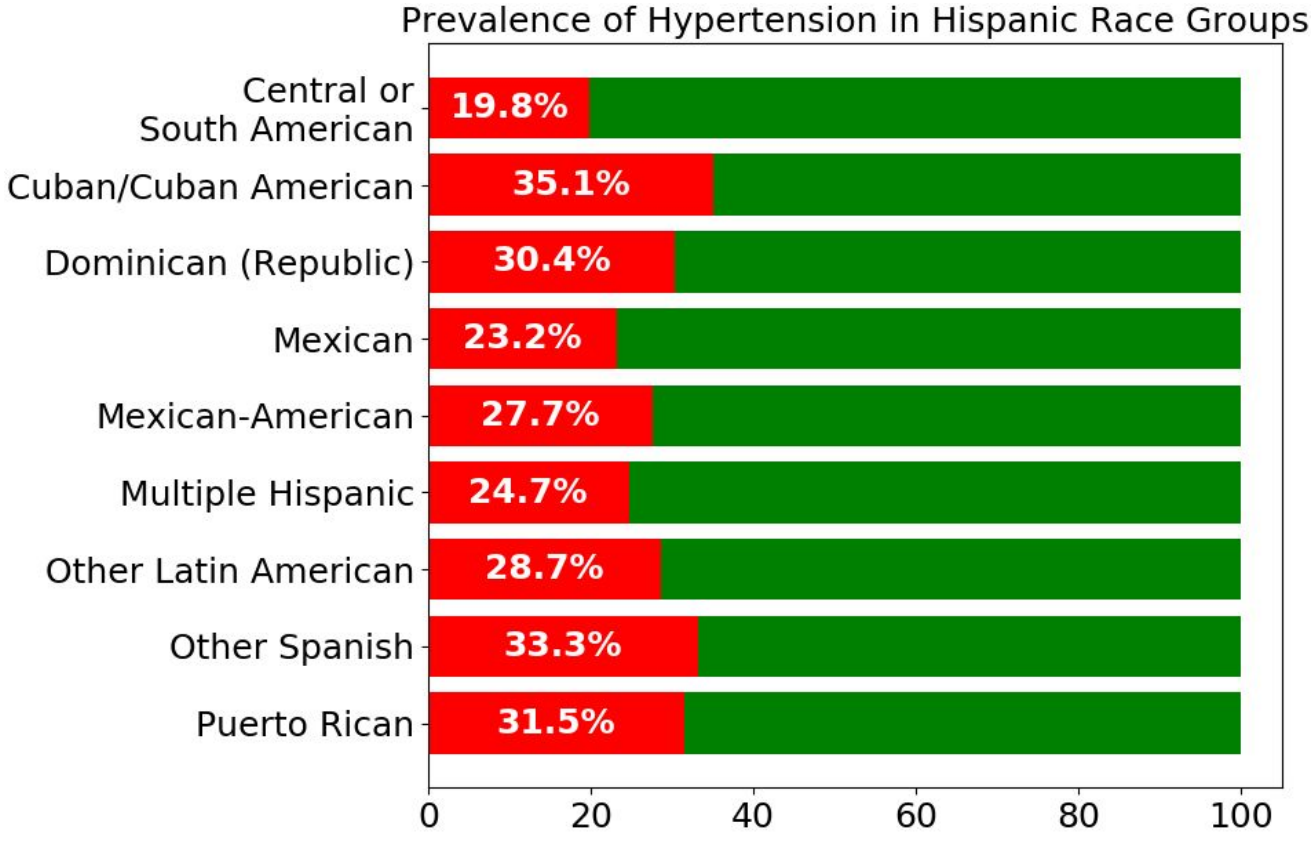
Exploratory Data Analysis - HTN and Gender

	Male	**Female**
Yes	35%	33%
No	65%	67%

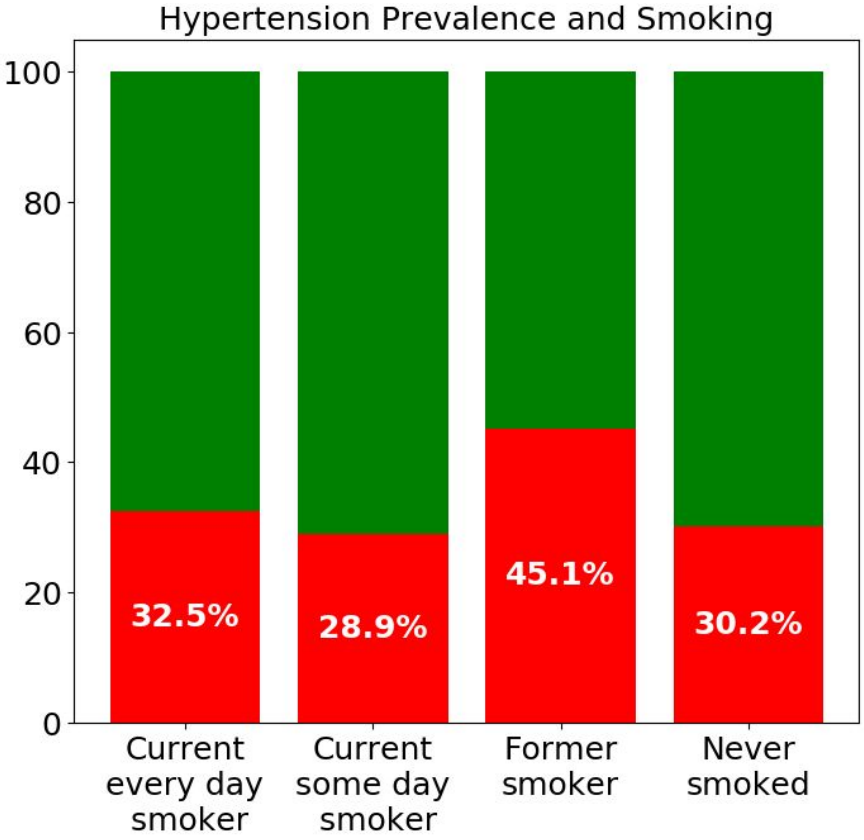
Exploratory Data Analysis - HTN and Race



Exploratory Data Analysis - HTN and Race



Exploratory Data Analysis - HTN and Smoking



Statistical Data Analysis - Correlation

- Feature-to-feature
  - Spearman's R
  - Cramer's V

- Feature-to-target
  - 0.28 highest - diabetes
  - Low correlation in EDA features

Feature	Cramer's V
Income	0.095459
Gender	0.011612
Race	0.093175
Region	0.261307
Smoking	0.132648

Statistical Data Analysis - Correlation & Comorbidities

- Theilman's U - the "uncertainty coefficient"

Sample results from Theil's U Test with Comorbidities		Theil's U
Ever told had hypertension	Ever told had angina pectoris	0.018354
Ever told had angina pectoris	Ever told had hypertension	0.087731
Ever told had hypertension	Ever told had cancer	0.015805
Ever told had cancer	Ever told had hypertension	0.033008
Ever told had **hypertension**	Ever told had diabetes	0.061704
Ever told had **diabetes**	Ever told had hypertension	**0.100838**
Ever told had hypertension	Ever told had heart attack	0.024073
Ever told had heart attack	Ever told had hypertension	0.091252

Machine Learning

- Algorithms for Classification Problems
  - K-Nearest Neighbors (KNN)
  - Logistic Regression
  - Random Forest Classifier
  - Adaptive Boosting (AdaBoost)
  - Extreme Gradient Boosting (XGBoost)
- Metric - F1
  - Unbalanced data
  - Cost of false negatives

Machine Learning - Iterative Improvement

F1 scores

	Default Parameters	Additional Data Wrangling	Hyperparameter Tuning & Cross Validation
KNN	0.38	0.47	n/a
Logistic Regression	0.45	0.54	No improvement
Random Forest	0.49	0.59	**0.60**
AdaBoost	0.47	0.57	0.58
XGBoost	0.52	0.58	**0.60**

Machine Learning - Iterative Improvement

Random Forest Hyperparameter Tuning & Cross Validation

Round	F1	True Positive Rate (Sensitivity)	False Negative Rate (Bad!)	True Negative Rate (Specificity)	Precision
1	0.59 **+0.1**	54%	45%	81%	65%
2	0.60 **+0.01**	55% **+0.01**	44% **-0.01**	81%	65%
3	0.60	55%	44%	81%	65%

Machine Learning - Iterative Improvement

XGBoost Hyperparameter Tuning & Cross Validation

Round	F1	True Positive Rate (Sensitivity)	False Negative Rate (Bad!)	True Negative Rate (Specificity)	Precision
1	0.60 **+0.08**	53%	46%	83%	68%
2	0.60	53%	46%	83%	67% **-1%**
3	0.60	53%	46%	83%	67%
4	0.60	53%	46%	83%	68% **+1%**

Machine Learning - Best Results

	F1	True Positive Rate (Sensitivity)	False Negative Rate (Bad!)	True Negative Rate (Specificity)	Precision
Random Forest	0.60	55%	44%	81%	65%
XGBoost	0.60	53%	46%	83%	68%

Summary

- Results not as good as hoped
  - Selection bias
  - Response bias
  - Data is self-reported
- High-specificity tests are good for ruling-IN
  - Model is good for screening tool in community outreach
- Importance of case-relevant metrics
- Have your blood pressure tested!