Predicting Chronic Kidney Disease

A Machine Learning Approach to Hyper-Local Prediction of Chronic Disease

Can census data predict hyperlocal prevalence of chronic kidney disease?

Why Chronic Kidney Disease?

- 30 million Americans
- $31 billion in annual treatment costs
- 89,000 deaths per year

- Can be prevented and easily managed if caught in early stages
- essential to find affordable and effective modes of prevention

About the Data

Data Collection

500 Cities: Local Data for Better Living (Centers for Disease Control)

- CSV with census-tract level data on chronic disease, poor health indicator, and preventative behavior rates

5-year American Community Survey (U.S. Census Bureau)

- Pulled data on more than 256 demographic features from Census Bureau API

28,004 census tracts (observations) from 500 largest cities in the U.S.

Data Cleaning: 500 Cities

- Limited to 1 target variable and 9 features:
  - 7 prevention measures
  - 2 unhealthy behaviors
- Paired down to census tract-level data
- Pivoted from long to wide data
- Extracted Tract ID from UniqueID column
- Converted to numeric and categorical (state and city) data types
- Imputed missing data with the variable mean

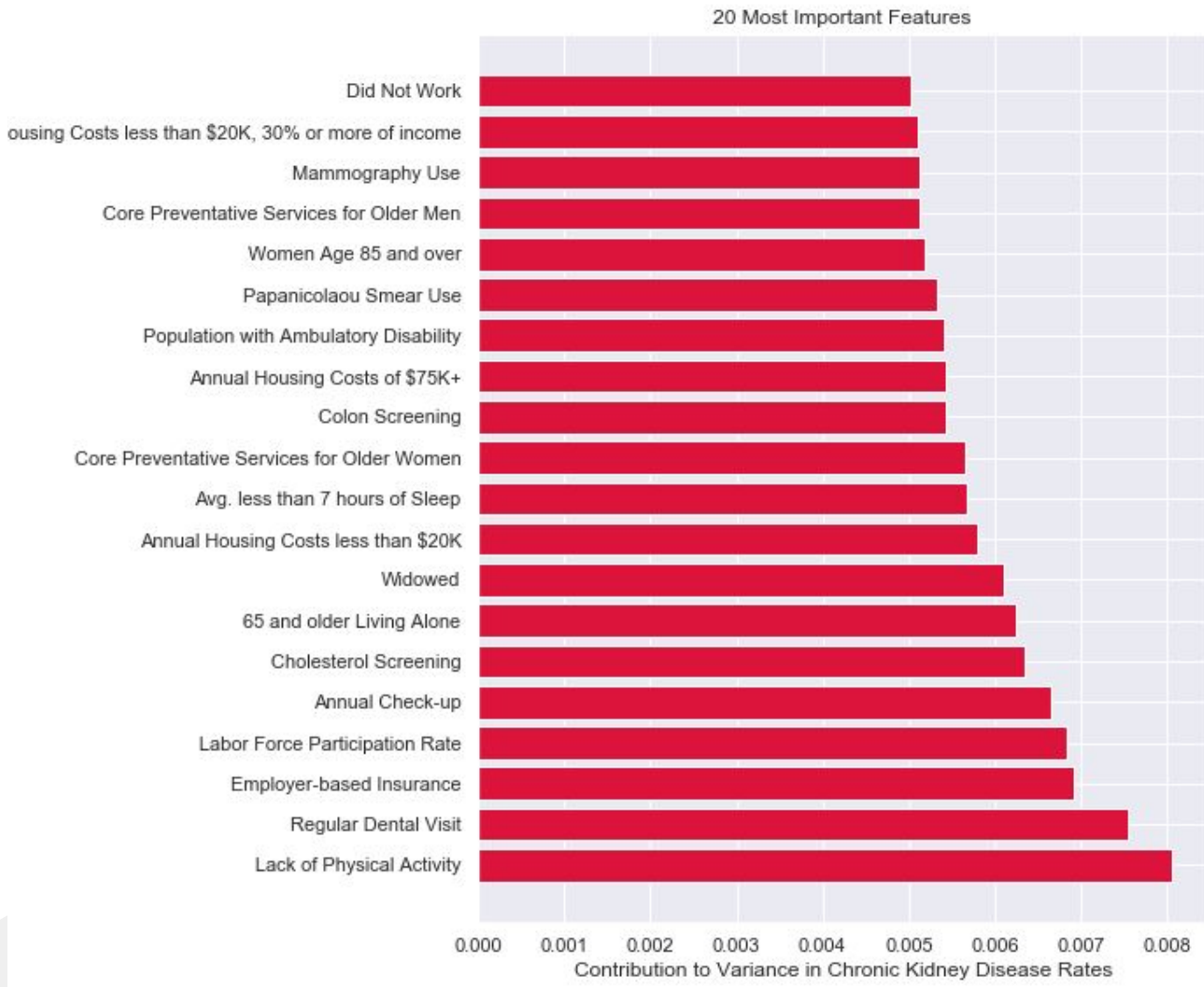
Data Cleaning: American Community Survey

- Dropped two empty columns
- Converted negative numbers to NaN
- Dropped all columns missing more than 80% of data
  - Brought number of features to 237
- Imputed missing values with variable mean

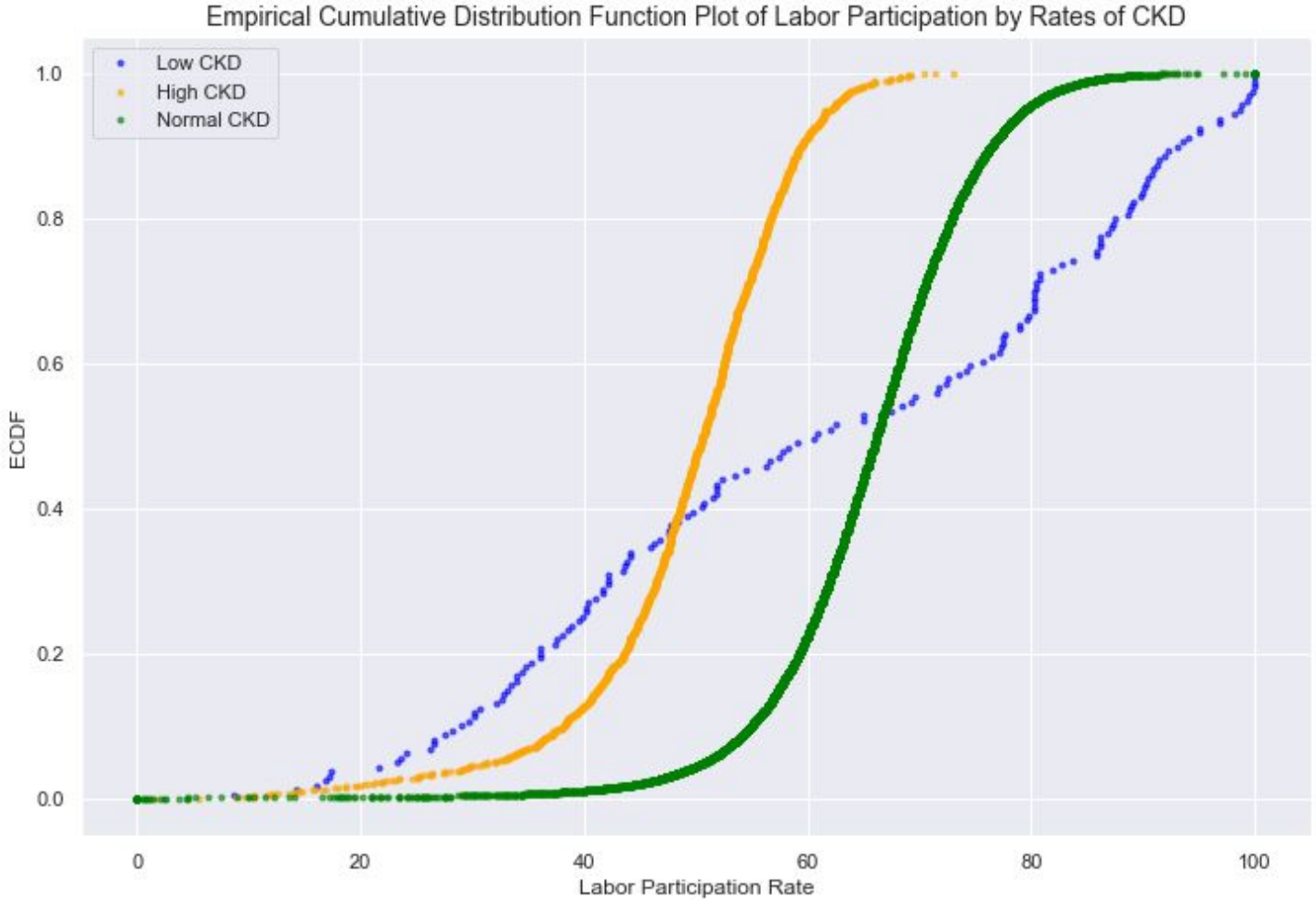
- Combined datasets on Tract IDs
  - Dropped all duplicate rows
- Final dataset:
  - 27,408 observations and 252 variables

Exploratory Data Analysis

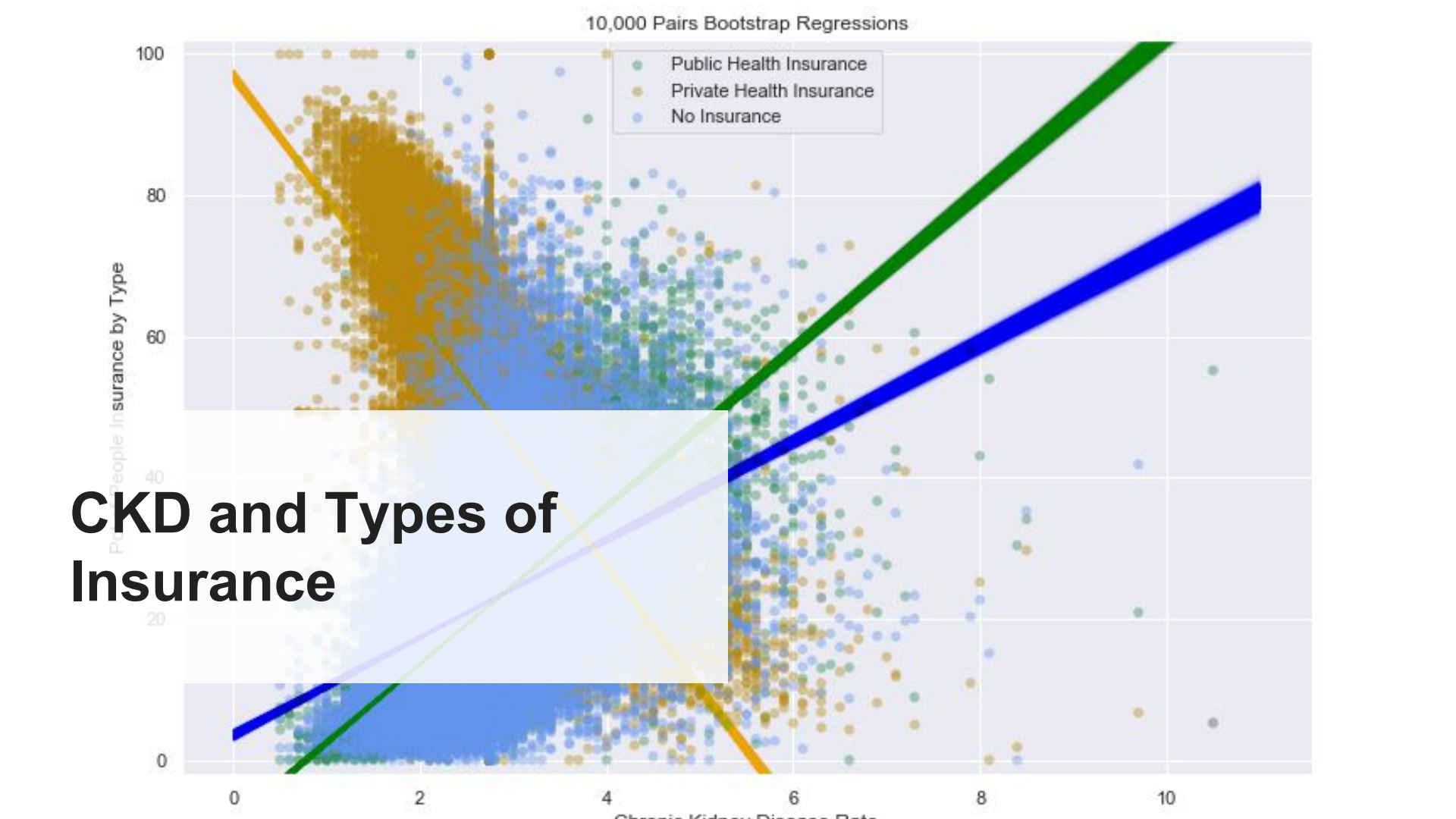
Extracting Important Features for Exploration



Labor Force Participation Rate and Chronic Kidney Disease



	Mean CKD Rate	Mean Labor Participation Rate	Standard Error of Labor Participation
Low Rates of CKD	0.939%	61.62%	25.18%
High Rates of CKD	4.917%	49.32%	9.65%

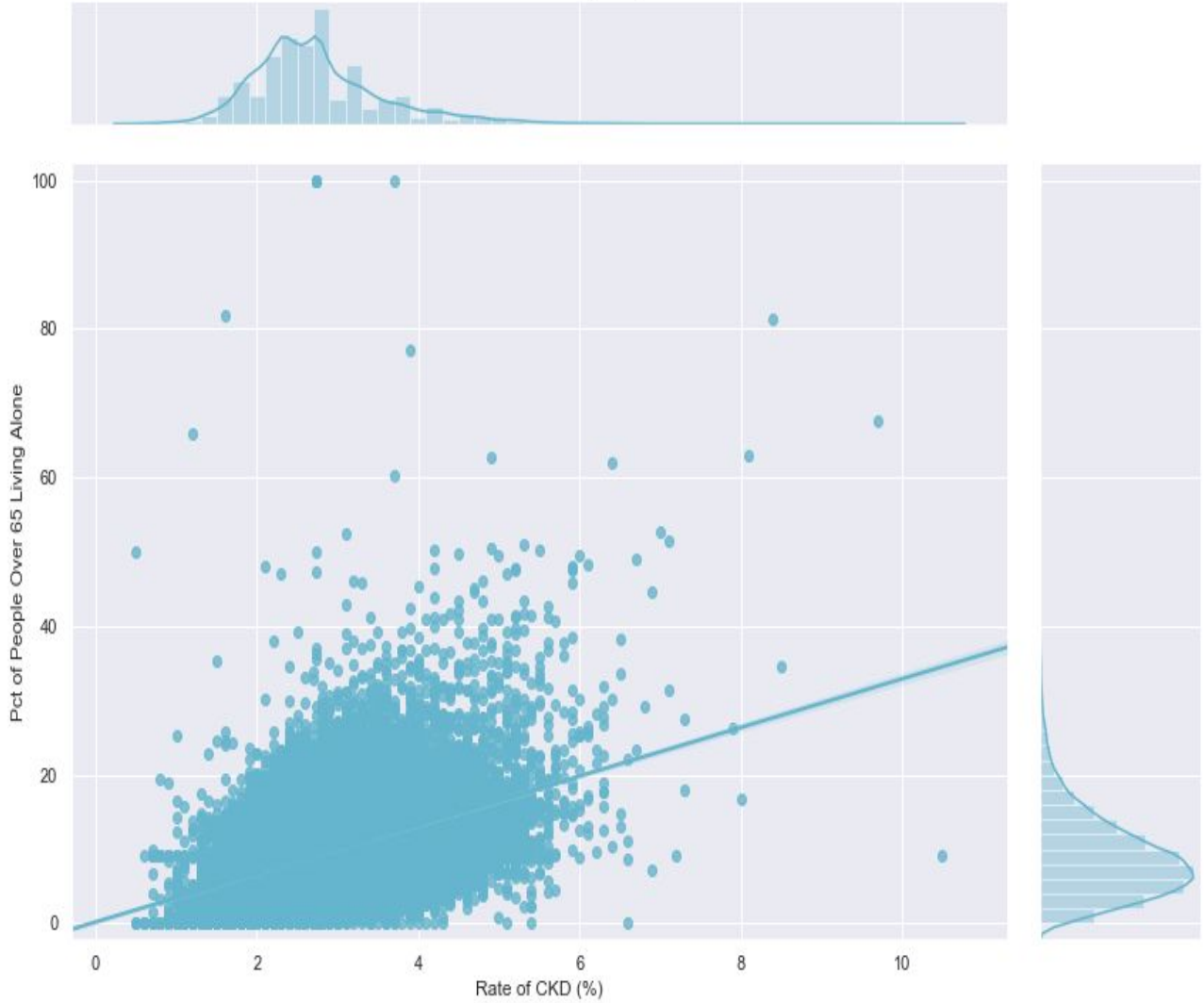


CKD and Types of Insurance

Comparing Rates of CKD Among Those with Public Insurance and Those with No Insurance

	Sample Difference	99% Confidence Interval	z-score	p-value
Difference of Mean Slope	4.23	(4.228,4.236)	2751..27	0.0
Difference of Pearson r	0.2814	(0.281, 0.282)	4285.0	0.0

The Role of Seniors Living Alone



The Combined Impact of Poverty and Sex