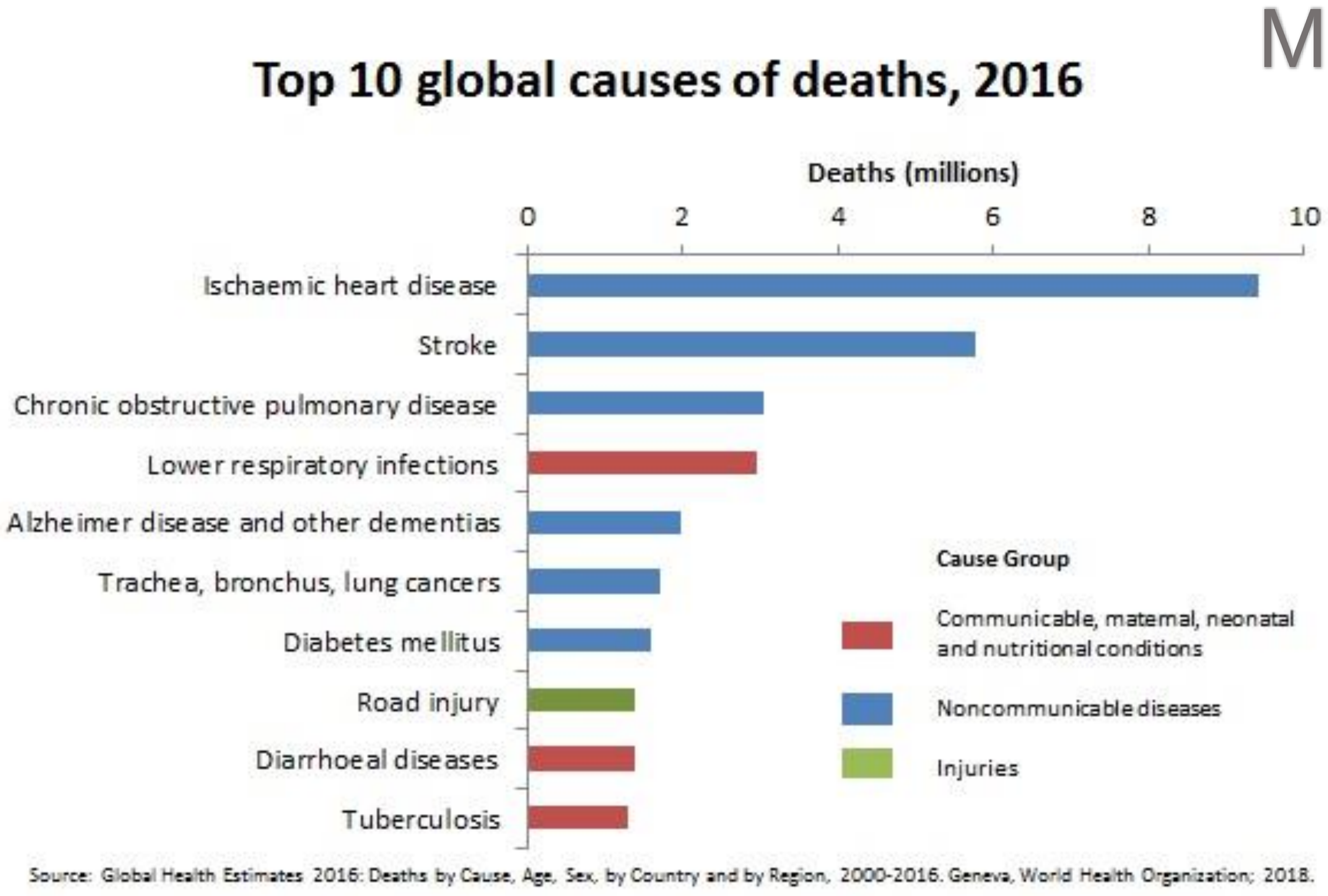
Visualization of relationship between chronic diseases and preventions in 500 US Cities

Bu Qianqian - Myles Lefkovitz - Salim Noorallah Ladak - Tri Nguyen

MOTIVATION



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

Diseases and health conditions represent 9 of the 10 most common causes of death globally. Of those 10 causes, chronic diseases (non-communicable diseases) represent 6 (blue bars). Understanding of related factors is of utmost importance to effective public health planning.

Chronic Diseases

THE DATA

Data Source. Published by the CDC (Center for Disease Control and Prevention)
- Contain prevalence of 13 diseases, 9 prevention practices and 5 unhealthy behaviors in 500 cities at National, State, City, and Census Tract levels.
- CSV format, 800,000 rows of data/235MB.
- Contain geographic data for each area.

Goals. Study which preventions and behaviors best predict a particular disease at state and national levels.

Processing. Data is cleaned and reformatted in Python, using Pandas and Numpy libraries.

ANALYSIS

Questions

Interpretable Relationship
- Between health outcomes and preventions/unhealthy behaviours?
- Prediction accuracy?
- Which model works best?
- What are the key features?

Experiment Design

ML with Scikit-Learn
- Multiple model comparison analysis
- Top feature selection at national and state levels
- Generate output file for visualization

Algorithms Evaluation

Multiple Regression Models
- Linear regression
- Ridge regression
- Lasso regression
- Support vector regression (SVR)

Results

Linear Relationships
- Found top 5 predictors for each health outcome
- Formulated SVR model with hyperparameter tuning
- Test accuracies averaged 0.89 at national level

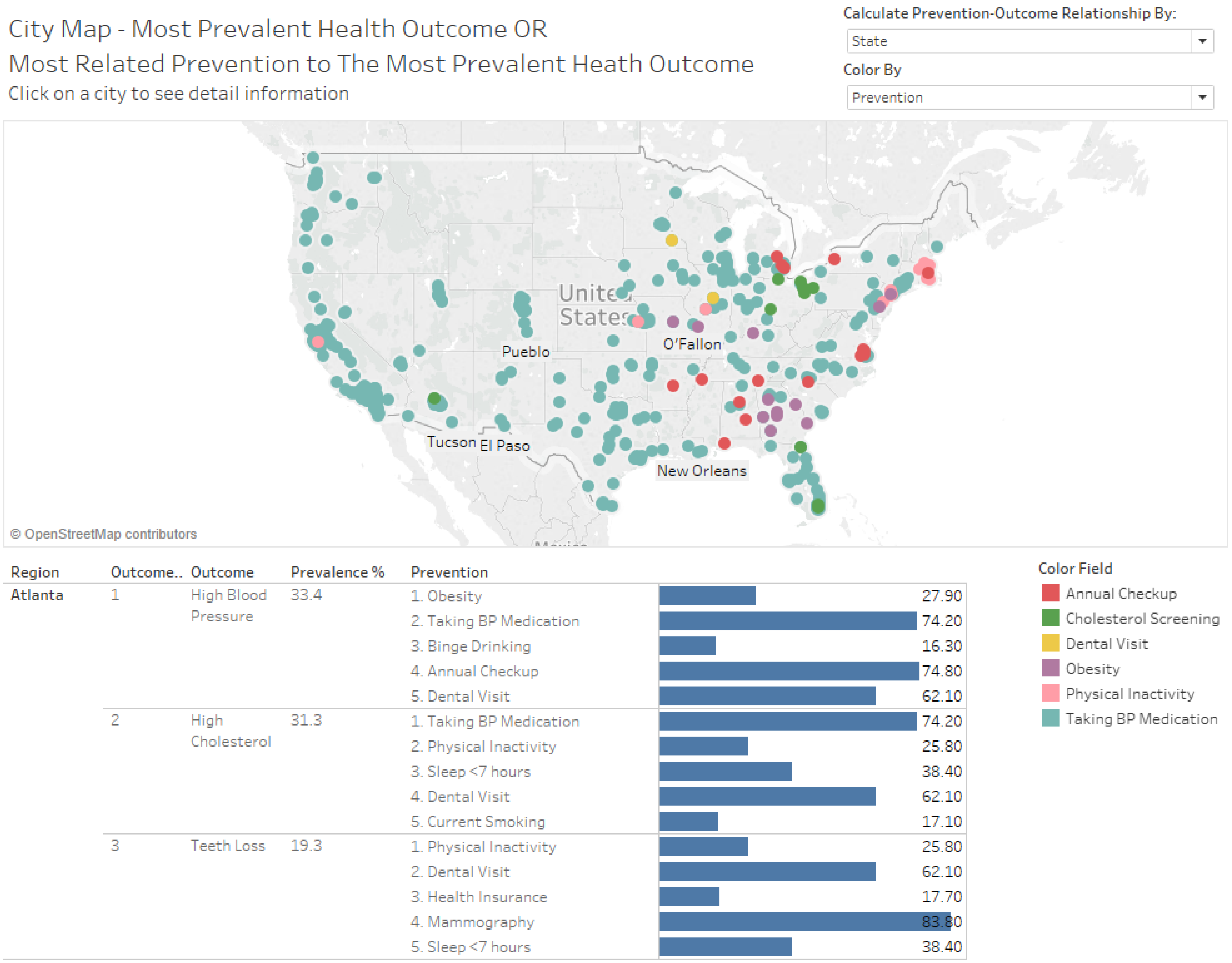
VISUALIZATION

INTERACTIVE MAPS

Developed in Tableau.
Show most prevalent diseases in each city/state.
Most related factors to each disease.
Calculation at **state** and **national** level.

RESULTS

Best predicting model: Support Vector Regression
Most prevalent diseases: high cholesterol and high blood pressure.
Most important factors: blood pressure medication, physical inactivity and obesity.
Visualization enabling users to explore the findings and make comparison across diseases and cities/states.



City Map - Most Prevalent Health Outcome OR
Most Related Prevention to The Most Prevalent Heath Outcome
Click on a city to see detail information

Calculate Prevention-Outcome Relationship By: State ▾
Color By: Prevention ▾

Region	Outcome..	Outcome	Prevalence %	Prevention	
Atlanta	1	High Blood Pressure	33.4	1. Obesity	27.90
				2. Taking BP Medication	74.20
				3. Binge Drinking	16.30
				4. Annual Checkup	74.80
				5. Dental Visit	62.10
	2	High Cholesterol	31.3	1. Taking BP Medication	74.20
				2. Physical Inactivity	25.80
				3. Sleep <7 hours	38.40
				4. Dental Visit	62.10
				5. Current Smoking	17.10
	3	Teeth Loss	19.3	1. Physical Inactivity	25.80
				2. Dental Visit	62.10
				3. Health Insurance	17.70
				4. Mammography	83.80
				5. Sleep <7 hours	38.40

Color Field
🟥 Annual Checkup
🟩 Cholesterol Screening
🟨 Dental Visit
🟪 Obesity
🟥 Physical Inactivity
🟩 Taking BP Medication