

# All models are wrong, but some are useful...

An attempt to predict high cholesterol via multiple linear regressions



# All models are wrong, but some are useful... maybe not this one

An attempt to predict high cholesterol via multiple linear regressions

## Table of Contents:



- Problem
- Investigation
- Results
- Next Steps

# The problem: People hate needles

## Context

Cholesterol is a waxy substance found in our blood. High LDL cholesterol increases your **risk of heart disease, heart attacks, & stroke**. High cholesterol has **no symptoms**, and has been called a **silent killer**



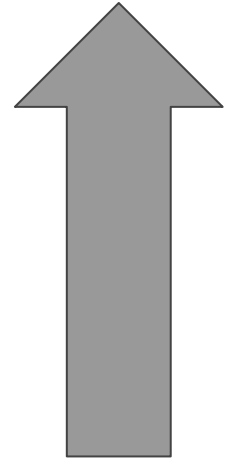
## Problem statement

Can we **predict high LDL (“bad”) Cholesterol** based on **self-measurable metrics** that don't involve needles or require visiting a doctor's office?

A topic that hits close to home...



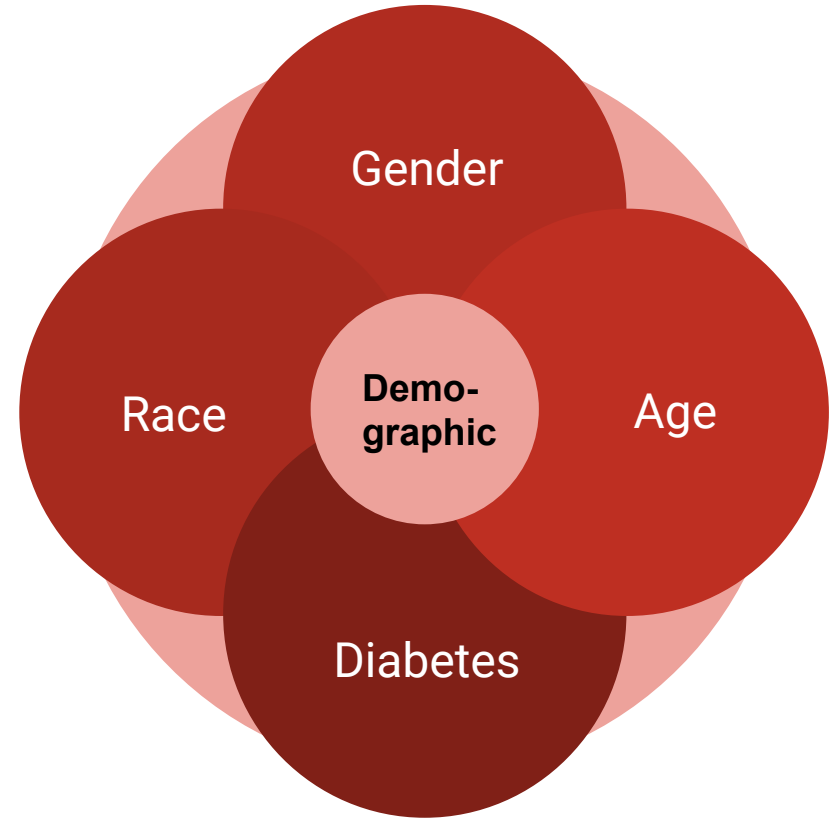
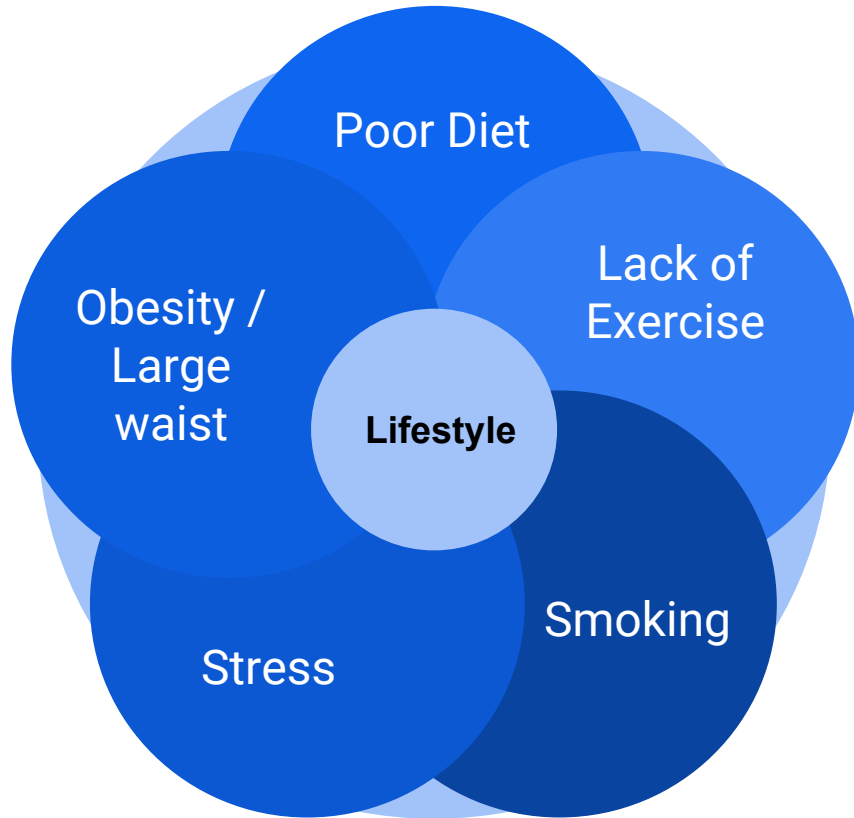
A topic that hits close to home.



## Table of Contents:

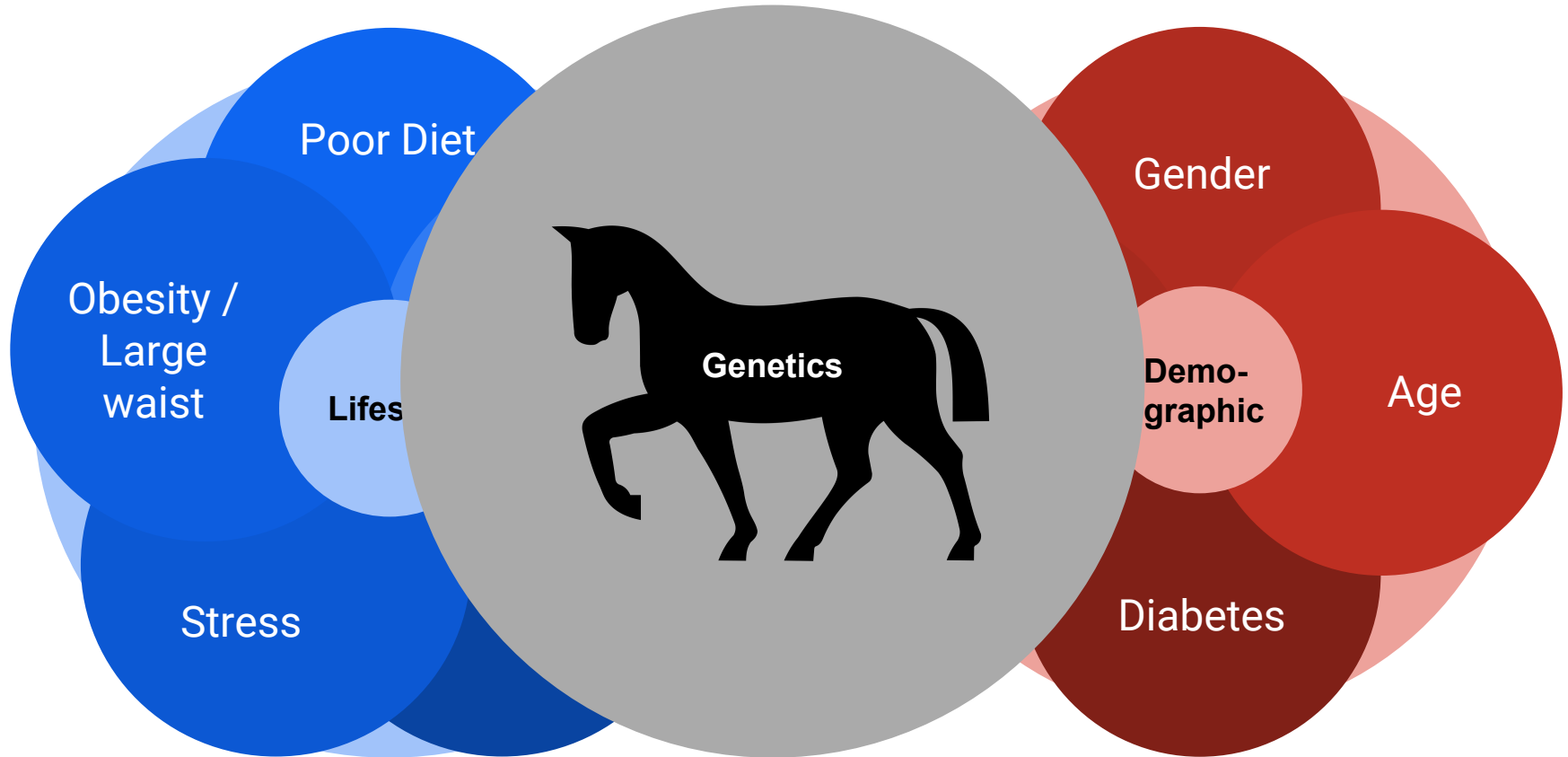
- Problem
- ➡ ● Investigation
- Results
- Next Steps

# Risk Factors include...

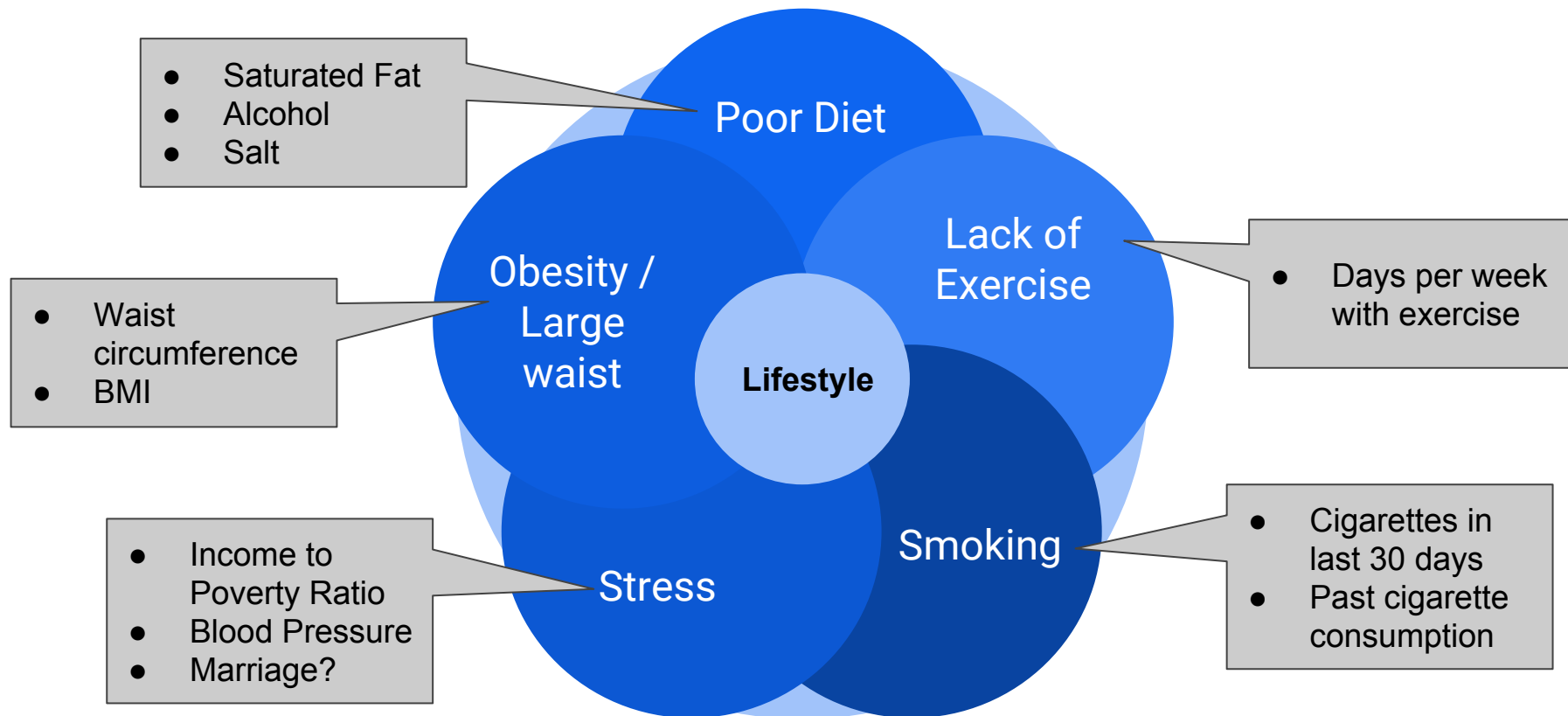




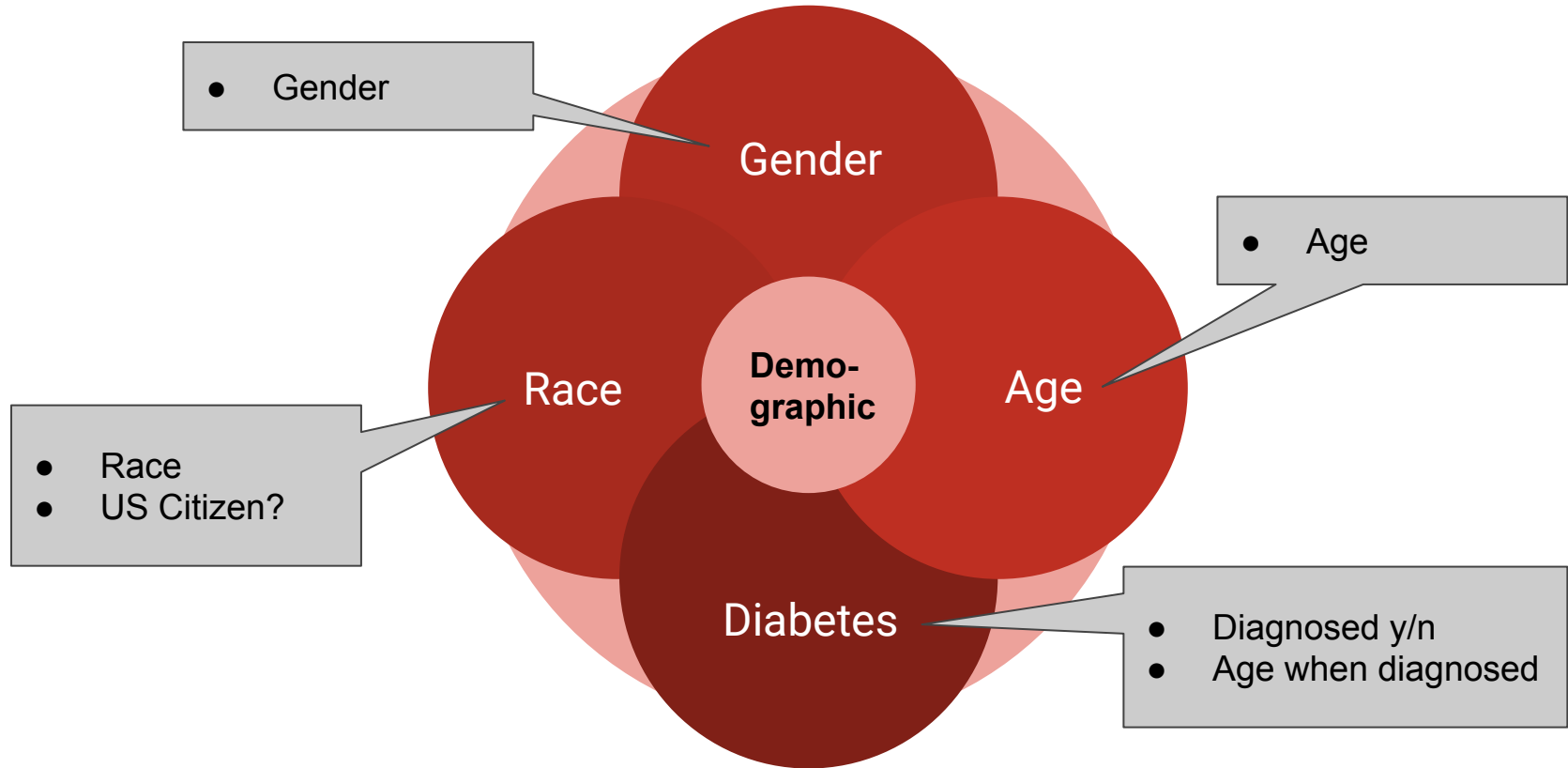
# Risk Factors include...



# National Health & Nutrition Examination Survey



# National Health & Nutrition Examination Survey

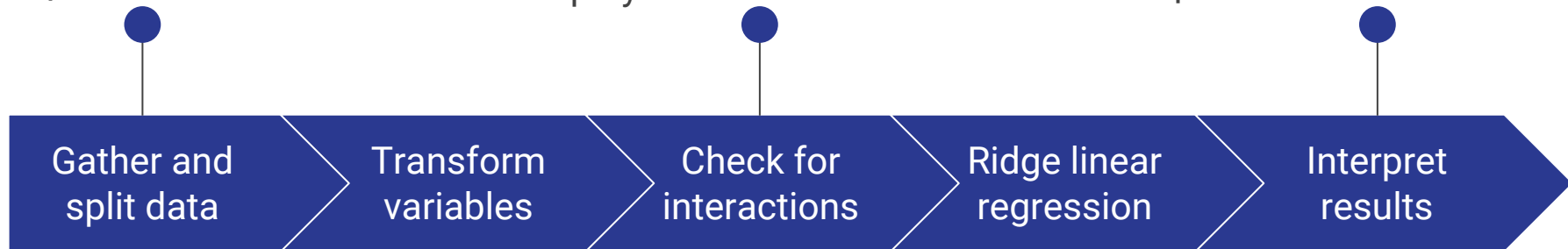


# Process

2431 Participants  
17 Predictor variables  
Split into “train” and “test”

MSE tests showed  
there was no need to  
generate higher order  
polynomials

“Test” how the model  
performs

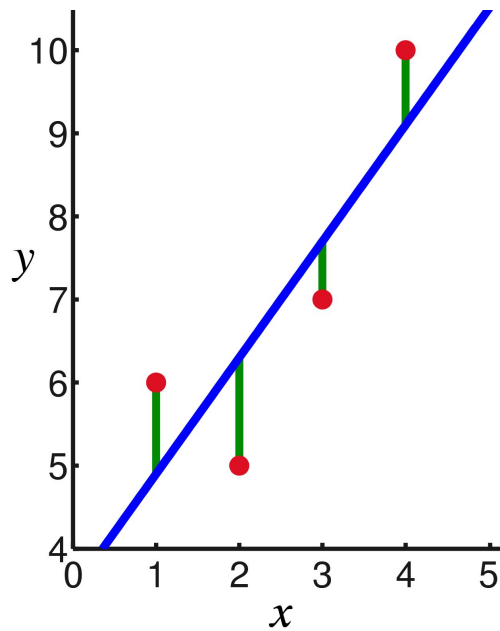


Several variables were  
non-normal, so these  
needed transforming<sup>1</sup>

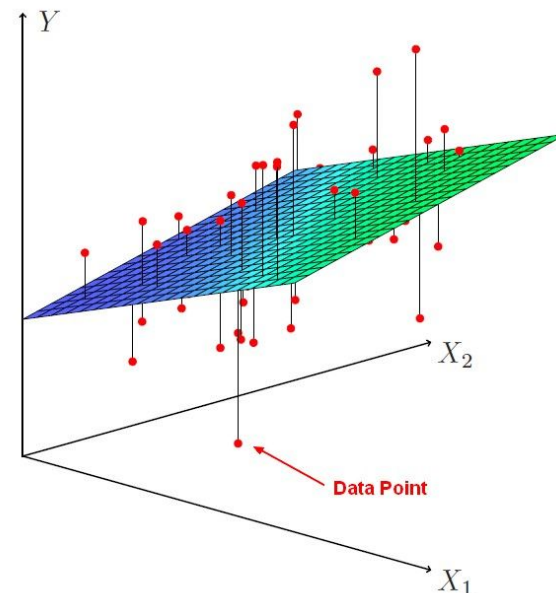
“Train” the model with  
regression technique  
that reduces both in  
and out of sample  
errors

1. Used BoxCox transformation

# Linear Regression: Ridge Regression



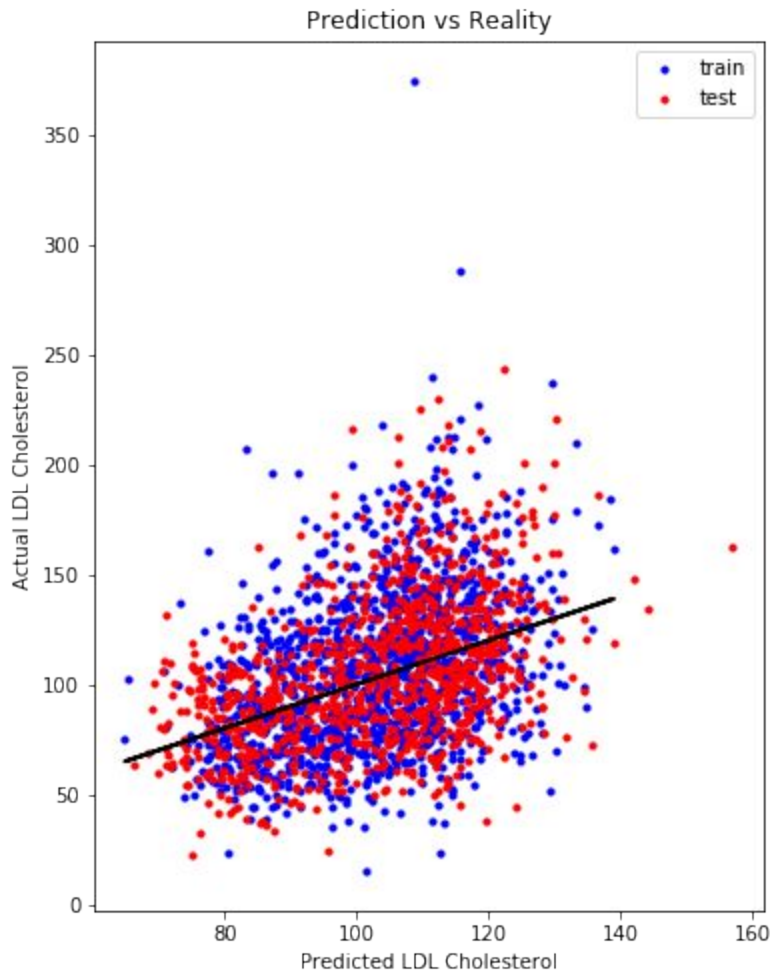
**Goal:**  
Minimize  
errors  
(SSE) and  
maximize  
predictive  
power



## Table of Contents:

- Problem
- Investigation
- Results
- Next Steps





Test  $R^2$ : 0.108

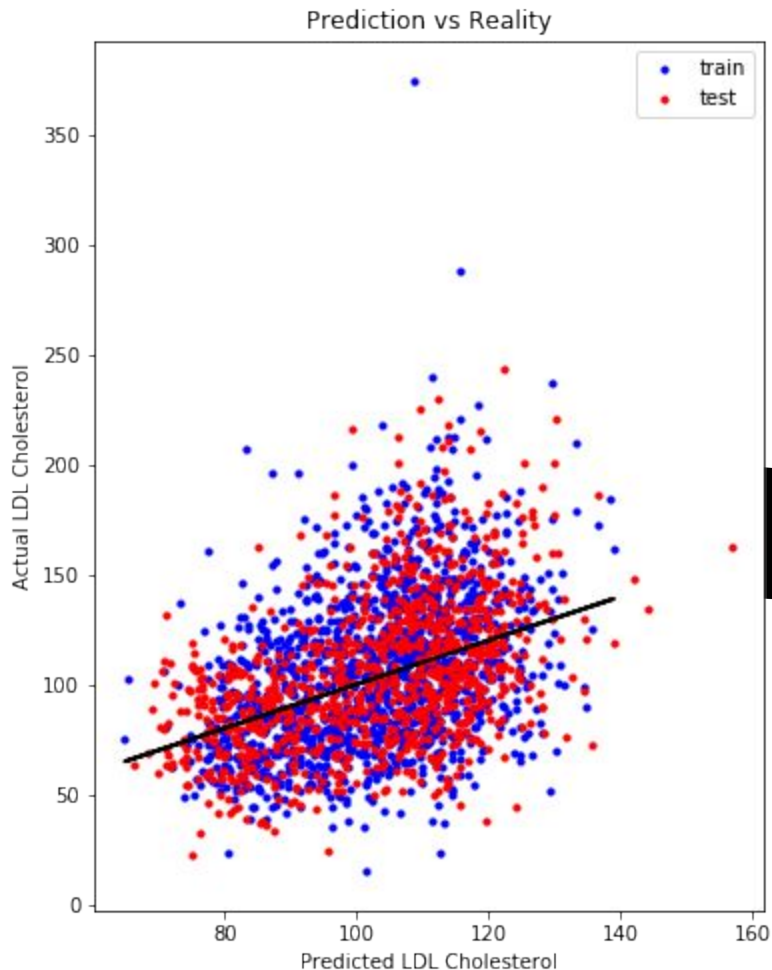
Difference from train  $R^2$ : 0.038

Test SSE: 3.04

Difference from train SSE 1.36

In other words, we expect our model could account for ~11% of variability seen in LDL Cholesterol

---



Test  $R^2$ : 0.108

Difference from train  $R^2$ : 0.038

Test SSE: 3.04

Difference from train SSE 1.36

In other words, we expect our model could account for ~11% of variability seen in LDL Cholesterol



## Table of Contents:

- Problem
- Investigation
- Results
- Next Steps



# Next Steps

Plenty of room for improvement, but linear regression may not be the right tool for the job

Ways we could improve linear regression:

- More Data - multiple years
- Smarter feature generation from lifestyle survey (e.g. exercise)
- Transformations of data with 0's (e.g. smoking, alcohol)