Community Health Analysis

DSIR 10/31 - Project 4
2022-01-20
Colby Tse | Jahnavi Kalpathy | Will Mohr

Overview

1) Background
2) Explanation of original data
3) EDA
4) Modeling
5) Results
6) Conclusions & Recommendations

Background

Goal: Impact Analysis of lifestyle on health
→ found **The China Study**

- animal products effect on cancer/cardiovascular disease
- whole-food, plant-based diet

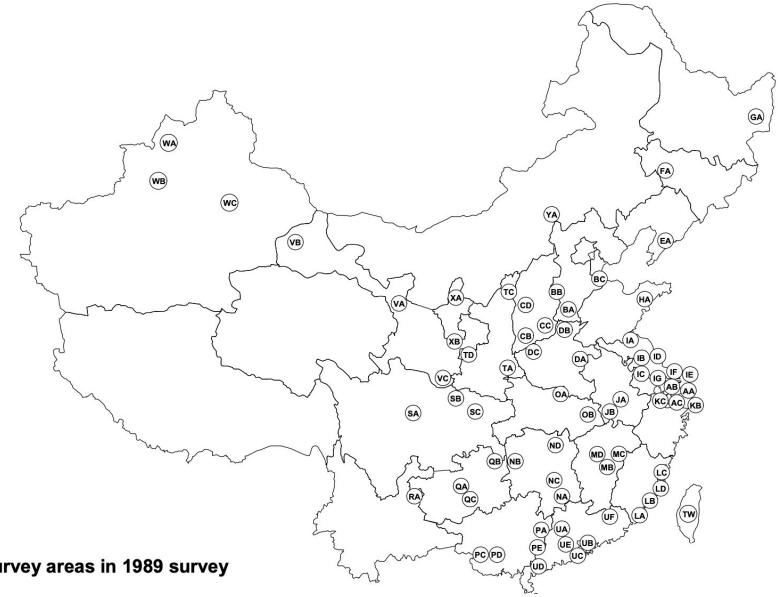
→ Critiques of conclusion, but data is solid

New Goal: use the original data

- Predict mortality from inputs
- Find variables with correlation to different causes of death

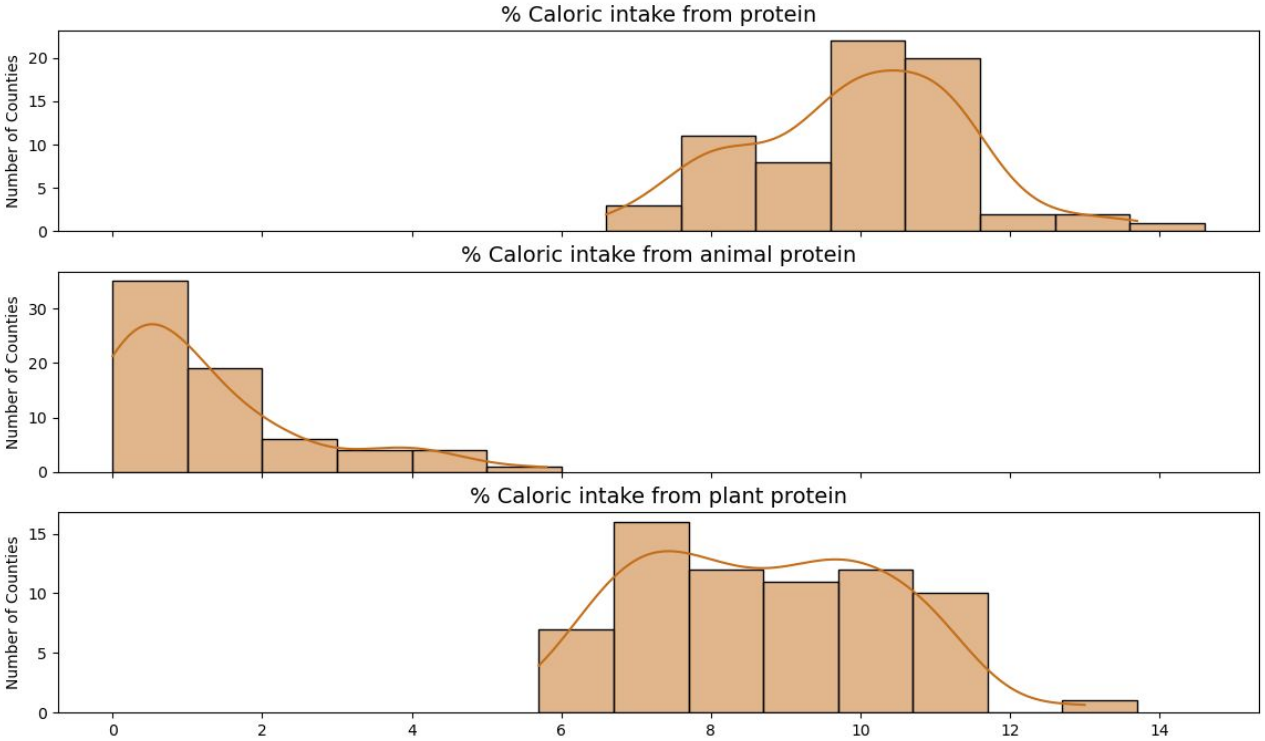
Data from The China Study

- Detailed survey in rural China over 20 years
  - 1973 - 1983; 1986 - 1989; 1993
  - Mainland China (69) and Taiwan (16)
  - 2 *xiangs* per county
  - 50-50 male-female adults
- 639 variables measured per county
  - 119 mortality rates
  - 161 diet measurements
  - 107 measurements for blood/urine
  - 247 questions related to lifestyle
- Limitations
  - Data aggregated by county
  - Data varies by survey
  - Only mortality by age

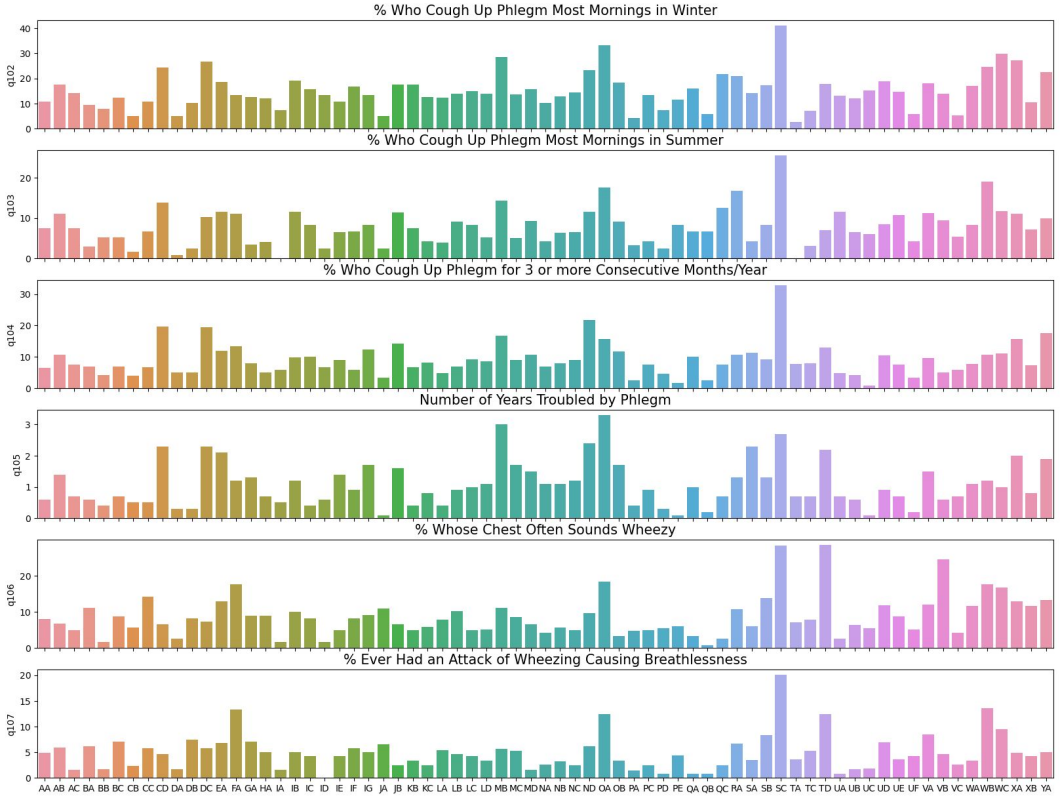


Survey areas in 1989 survey

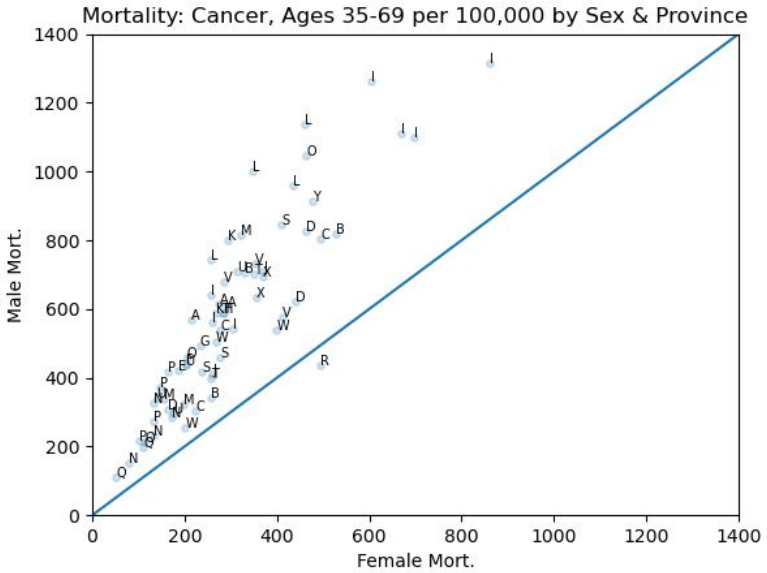
EDA - Diet & Geography

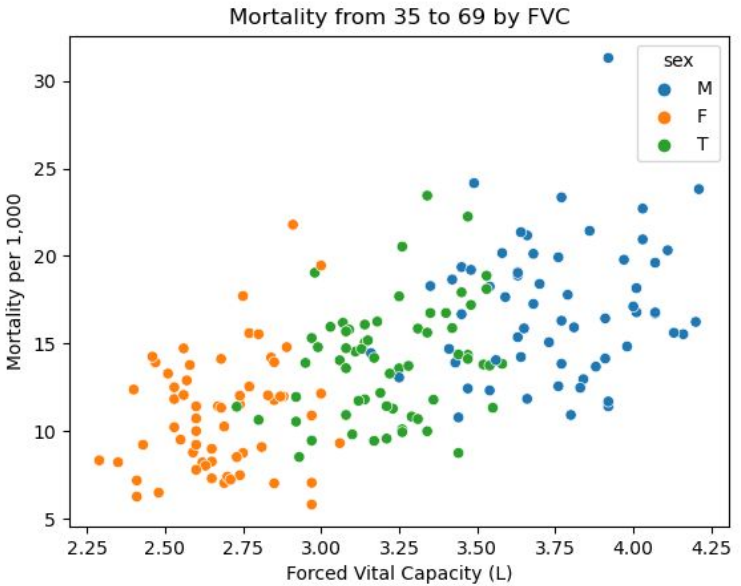


EDA - Lifestyle Questionnaire



EDA - Mortality





Modeling

Our data is continuous and numerical

- Linear Regression
  - Ridge
  - Lasso

	Model	Train Score	Val Score	X Val Score	RMSE Train	RMSE Val
0	m005_ALL_35-69.LinReg	1.000000	0.882940	0.616316	8.404957e-15	1.206978
1	m005_ALL_35-69.L2-1	0.999965	0.881611	0.614525	8.404957e-15	1.206978
2	m005_ALL_35-69.L1-1	0.999998	0.999996	0.999285	8.404957e-15	1.206978
3	m005_ALL_35-69.L1-2	0.999989	0.988186	0.931555	5.548124e-15	1.264333

Modeling

	Model	Train Score	Val Score	X Val Score	RMSE Train	RMSE Val
20	m021_SCHISTOc	0.905542	0.907366	-3.673384	2.974858e-14	14.432142
11	m012_INFECTc	0.784124	0.536770	-0.424291	1.409650e-13	46.694511
30	m031_LIVERCAc	0.686820	0.498407	-0.090399	1.735747e-13	72.214573
77	m081_TOTLIVRc	0.707458	0.490807	-0.295884	1.811700e-13	75.496279
69	m073_DIGESTIVb	0.863828	0.398663	-0.304247	2.366806e-14	7.630224
14	m015_PULMTBb	0.999998	-2.238436	-1.540799	1.135617e-14	3.815466
47	m051_MENTALc	0.999999	-2.383336	-0.398552	2.081377e-14	12.236984
58	m062_HYPTENSc	0.446013	-2.506764	-1.241458	9.788204e-14	26.377757
94	m101_HOMICIDEb	0.999993	-3.499497	-0.362157	9.308562e-15	3.143814
13	m014_INTESTINc	1.000000	-4.892427	-5.241444	3.698280e-14	14.156183

Best-Modeled Targets

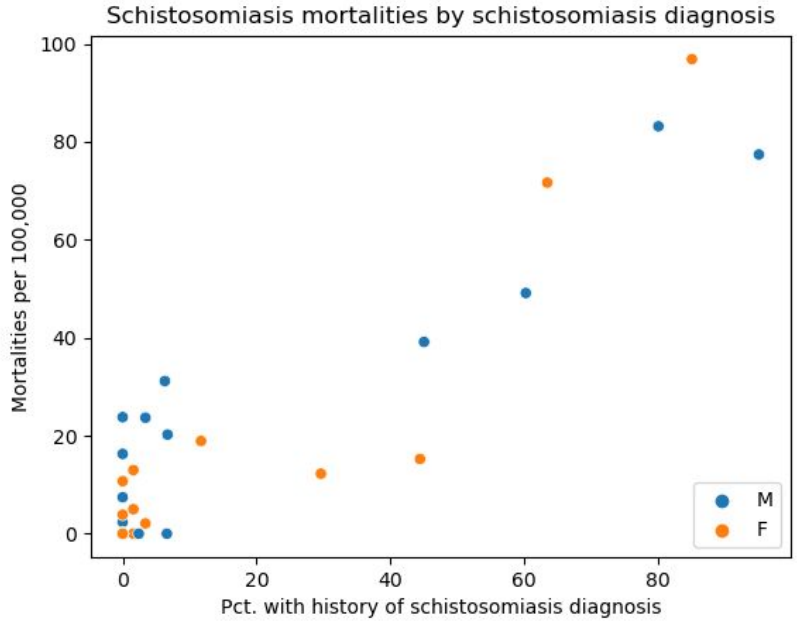
Male Model	Train Score	Val Score
m021_SCHISTOc	0.905542	0.907366
m012_INFECTc	0.784124	0.536770
m031_LIVERCAc	0.686820	0.498407
m081_TOTLIVRc	0.707458	0.490807
m073_DIGESTIVb	0.863828	0.398663

m012_INFECTc	description
7.742770	U032 TNOCa nitrosamine study TOTAL NITROS...
5.660898	Q157 dRICE questionnaire DAILY CONSUMPTIO...
5.516975	Q134 dSMOK<25m questionnaire PERCENT OF TOTAL...
4.841607	P025 VITC plasma VITAMIN C (ascorbic aci...
3.707410	P021 NEURSPOR plasma NEUROSPORENE (ug/dL) ...

Female Model	Train Score	Val Score
m021_SCHISTOc	0.920081	0.893889
m029_COLRECCAc	0.712298	0.347226
m006_ALL70_79	0.944305	0.345626
m097_DROWNb	0.513302	0.332789
m031_LIVERCAc	0.809288	0.301275

m029_COLRECCAc	description
3.031821	Q095 dSCHISTO questionnaire PERCENTAGE WITH ...
1.092013	Q245 fHTadj questionnaire HEIGHT OF SCHOOL...
0.979894	Q151 dBEERday questionnaire CURRENT DAILY CO...
0.806137	Q243 fWTadj questionnaire WEIGHT OF SCHOOL...
0.705052	P024 FOLATE plasma FOLATE (ng/mL) ...

Results - Feature Importances



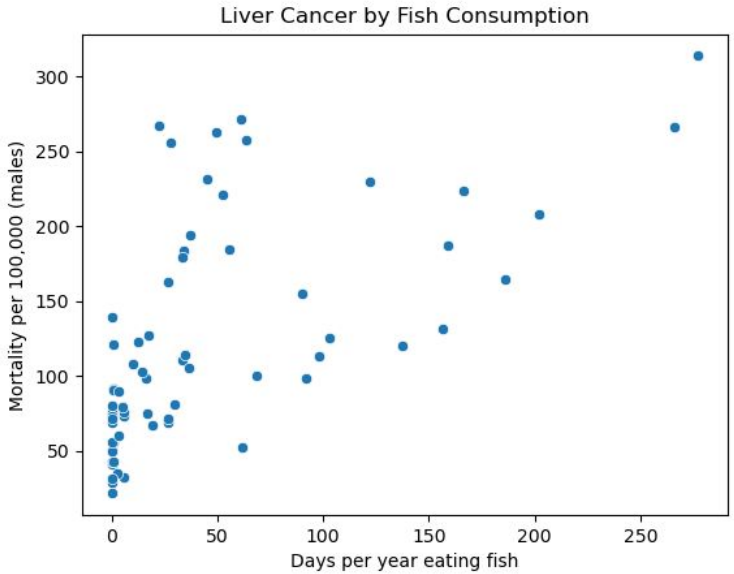
A coefficient of 9.8 for

PERCENTAGE WITH HISTORY OF SCHISTOSOMIASIS DIAGNOSIS

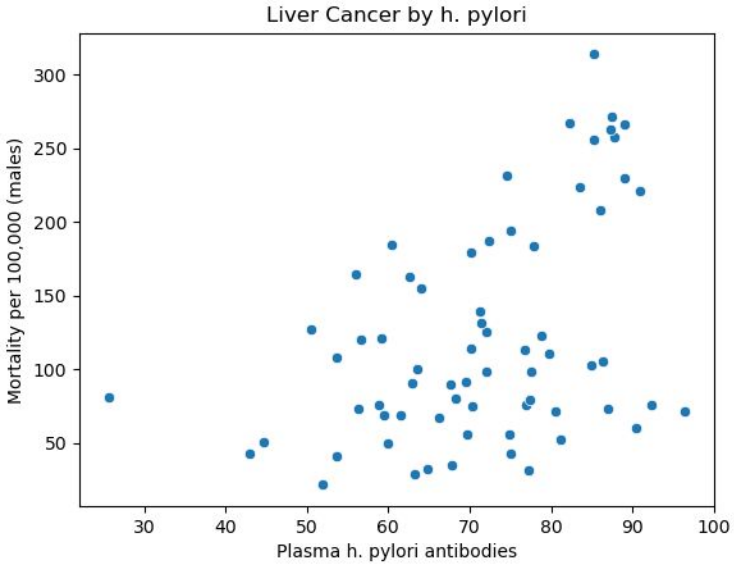
Is seen when predicting mortality due to the condition

Highest corr | Highest coeff

Correlation: .62

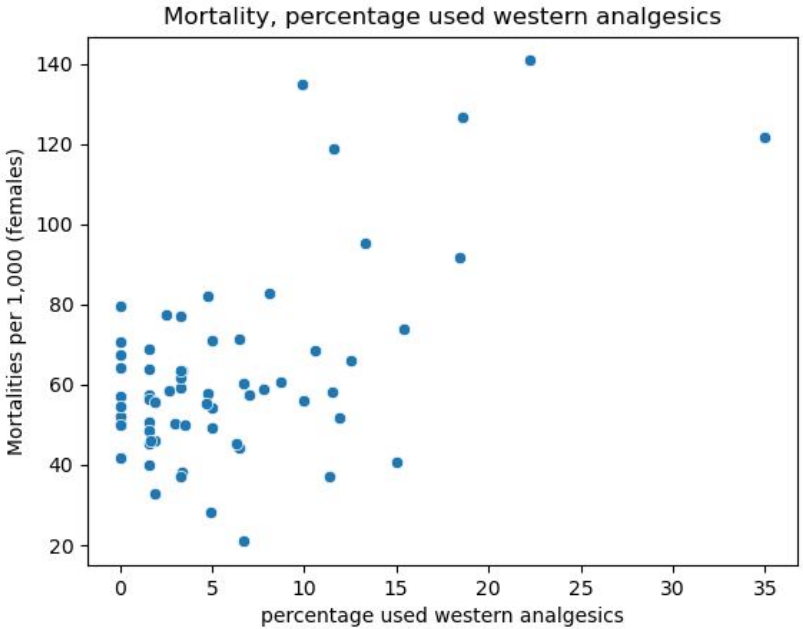


Coeff: 13.6

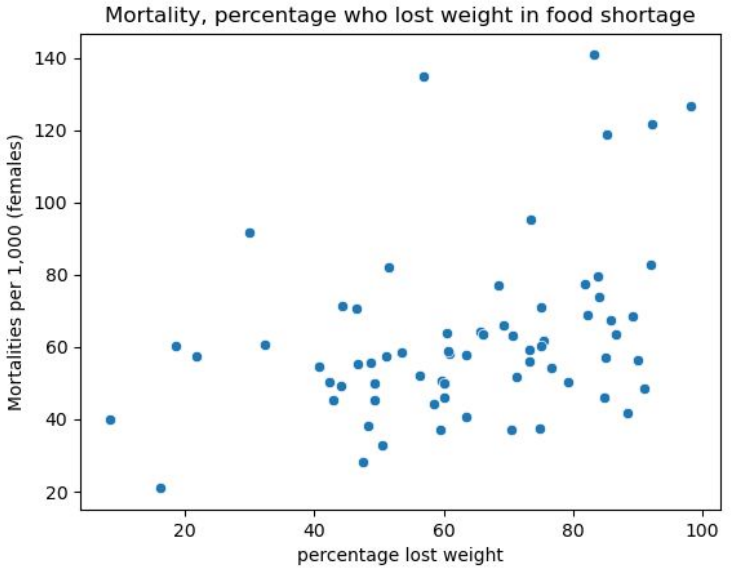


Results - Feature Importances

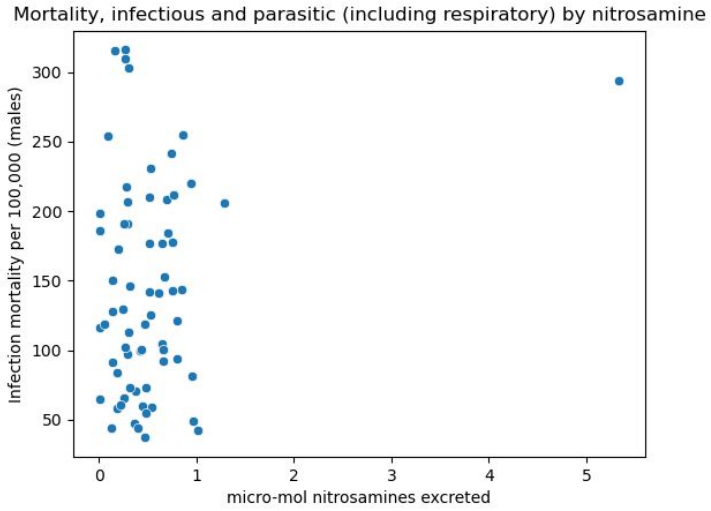
Corr: 0.59, Coeff: 2.77

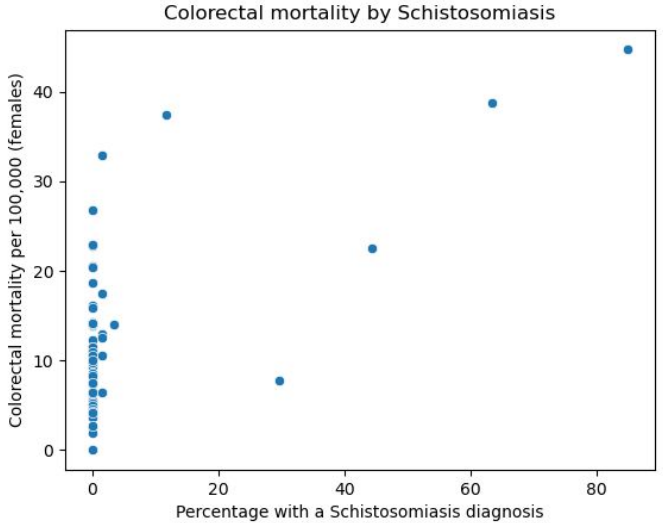


Coeff: 6.01, Corr: .36



Infection | CR Cancer





Conclusions

- Original predictive goal not feasible
  - Data not well suited
  - No evidence to support a model that predicts mortality from diet

- Social value is in the analysis and ability to parse through data
  - Pipeline lets us select a mortality target and identify most important feature variables

- Allows individuals to see a list of possible factors and make personal changes according to their lifestyles

Next Steps

- Streamline pipeline into a ‘search’ tool for individual use
  - Input target (current) to return list of important features
  - Or input a feature variable and the tool returns most correlated mortalities and the coefficient (planned)

- Improve feature elimination/ extraction - specific to each target

- Find way to add data from other survey years to training model

THANKS!

Do you have any questions?