Predicting 2020 US Life Expectancy at Birth:

Modeling Poverty, Education, Unemployment, and Income Effects

Problem Statement:

- Life expectancy at birth is an important metric of economic development.
- Many factors affect life expectancy at birth. Factors taken into consideration intentionally limited the for this project.
- Modeling life expectancy at birth using education, poverty, income, unemployment data in the US will provide insight on how these factors affect life expectancy.

Data Wrangling:

Sources:

1. 2020 Life Expectancy at Birth Datasets

   US Centers for Disease Control - US Life Expectancy Data

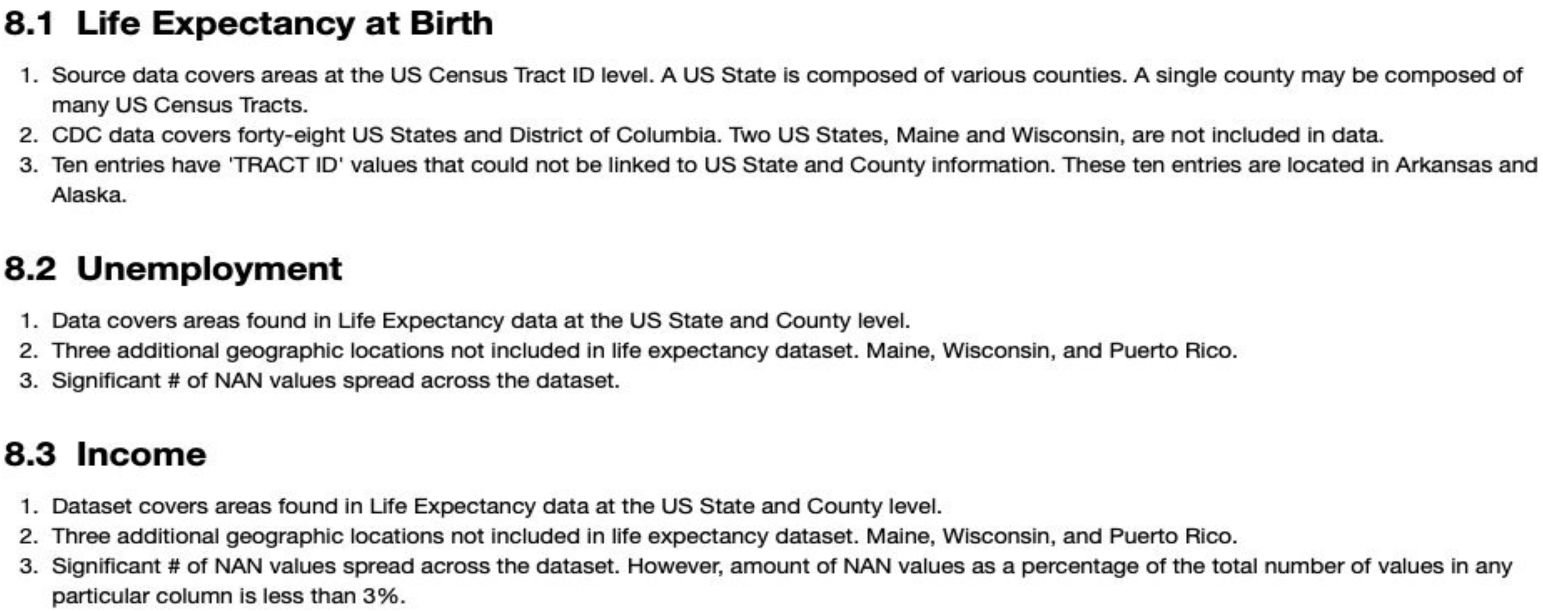
   https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html

2. Education, Poverty, Income, Unemployment Datasets

   USDA Economic Research Service

   https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

Data Wrangling Summary



8.1 Life Expectancy at Birth

1. Source data covers areas at the US Census Tract ID level. A US State is composed of various counties. A single county may be composed of many US Census Tracts.
2. CDC data covers forty-eight US States and District of Columbia. Two US States, Maine and Wisconsin, are not included in data.
3. Ten entries have 'TRACT ID' values that could not be linked to US State and County information. These ten entries are located in Arkansas and Alaska.

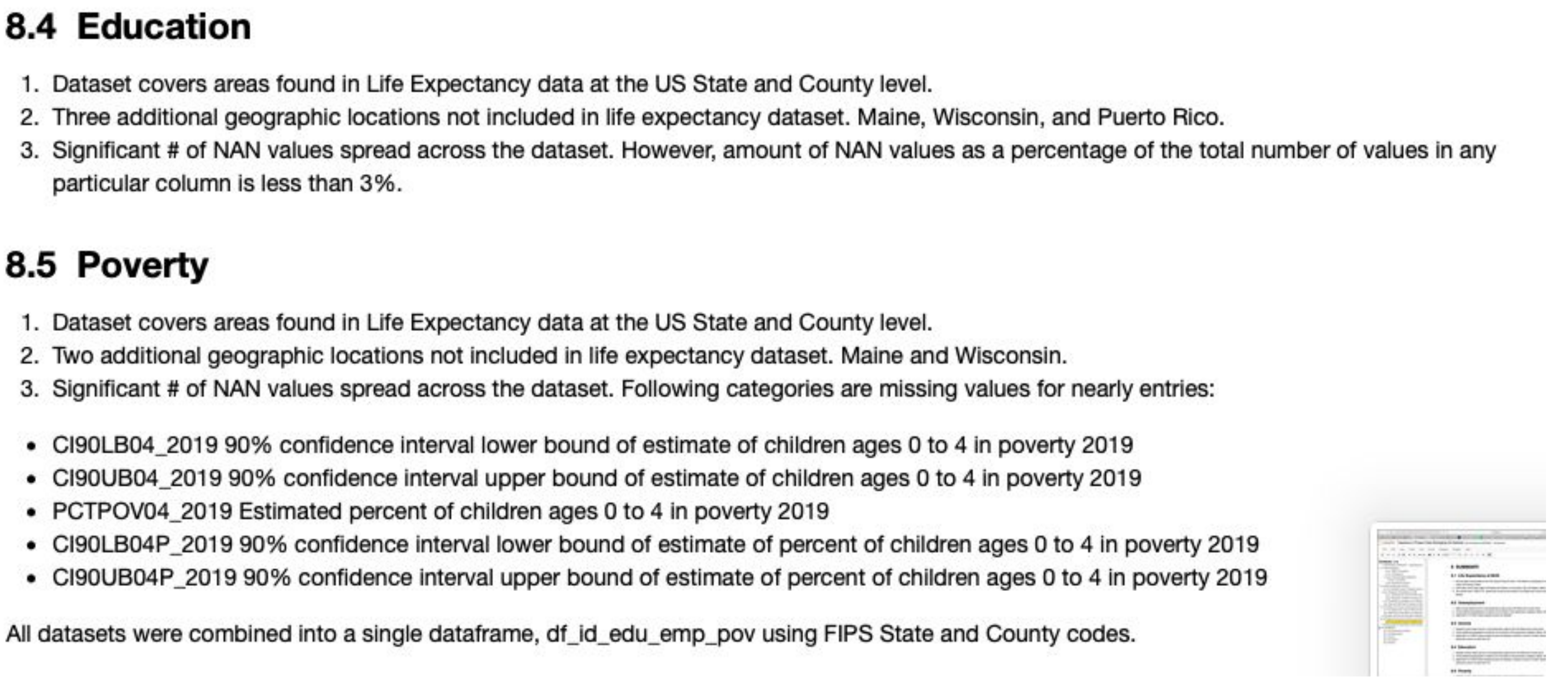
8.2 Unemployment

1. Data covers areas found in Life Expectancy data at the US State and County level.
2. Three additional geographic locations not included in life expectancy dataset. Maine, Wisconsin, and Puerto Rico.
3. Significant # of NAN values spread across the dataset.

8.3 Income

1. Dataset covers areas found in Life Expectancy data at the US State and County level.
2. Three additional geographic locations not included in life expectancy dataset. Maine, Wisconsin, and Puerto Rico.
3. Significant # of NAN values spread across the dataset. However, amount of NAN values as a percentage of the total number of values in any particular column is less than 3%.

Data Wrangling Summary



8.4 Education

1. Dataset covers areas found in Life Expectancy data at the US State and County level.
2. Three additional geographic locations not included in life expectancy dataset. Maine, Wisconsin, and Puerto Rico.
3. Significant # of NAN values spread across the dataset. However, amount of NAN values as a percentage of the total number of values in any particular column is less than 3%.

8.5 Poverty

1. Dataset covers areas found in Life Expectancy data at the US State and County level.
2. Two additional geographic locations not included in life expectancy dataset. Maine and Wisconsin.
3. Significant # of NAN values spread across the dataset. Following categories are missing values for nearly entries:

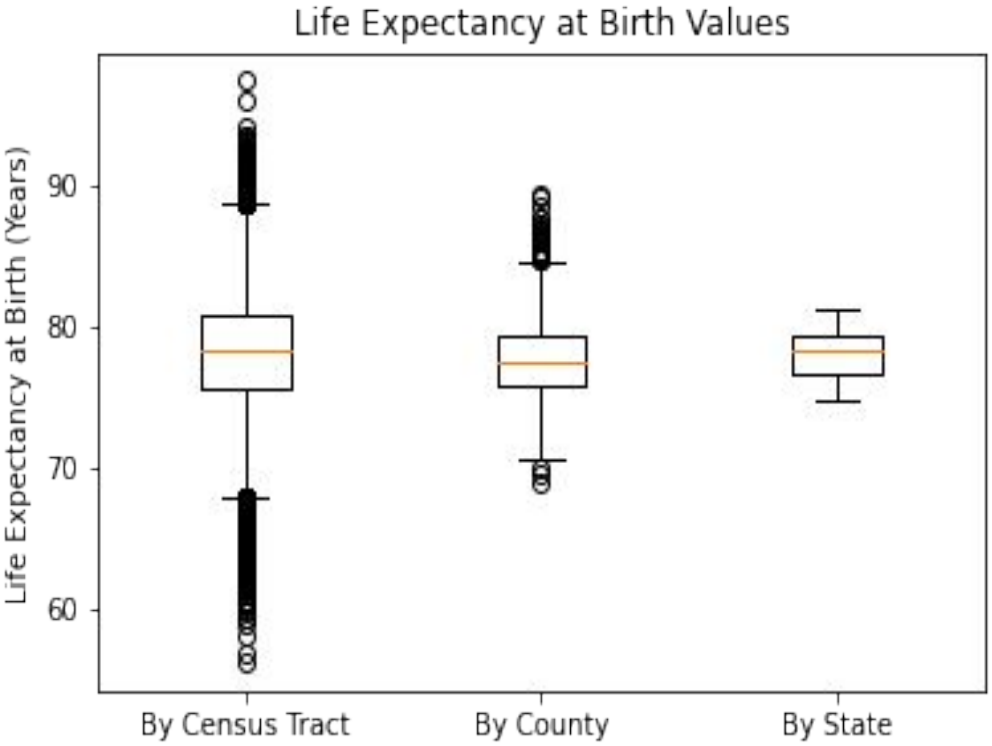
- CI90LB04_2019 90% confidence interval lower bound of estimate of children ages 0 to 4 in poverty 2019
- CI90UB04_2019 90% confidence interval upper bound of estimate of children ages 0 to 4 in poverty 2019
- PCTPOV04_2019 Estimated percent of children ages 0 to 4 in poverty 2019
- CI90LB04P_2019 90% confidence interval lower bound of estimate of percent of children ages 0 to 4 in poverty 2019
- CI90UB04P_2019 90% confidence interval upper bound of estimate of percent of children ages 0 to 4 in poverty 2019

All datasets were combined into a single dataframe, df_id_edu_emp_pov using FIPS State and County codes.

Exploratory Data Analysis:

- ‘Tidy Data’: 65662 entries, 182 features

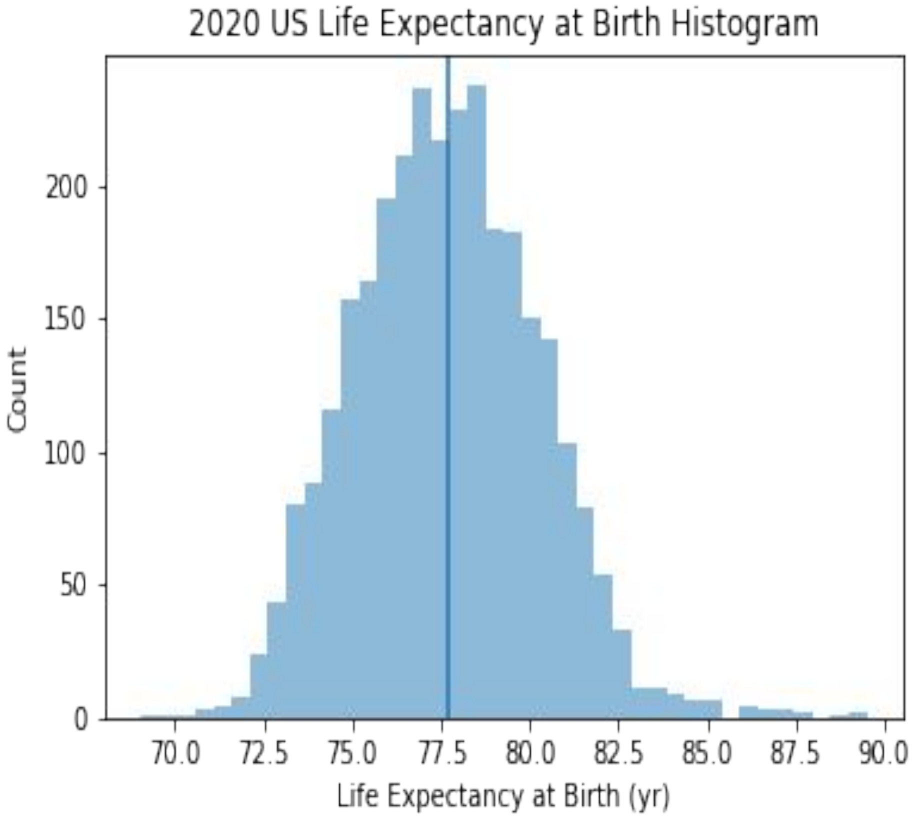
- Groupby/Average Census Tract values to produce County life expectancy values



Exploratory Data Analysis -

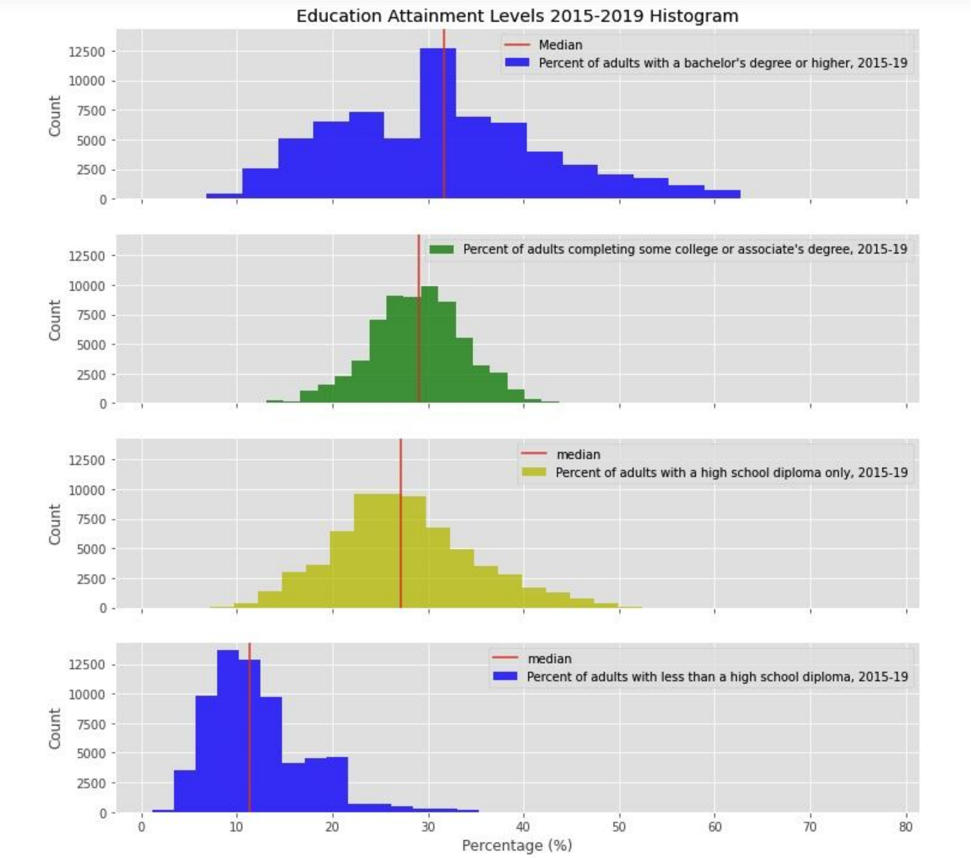
Life Expectancy Data:

77.6 yrs



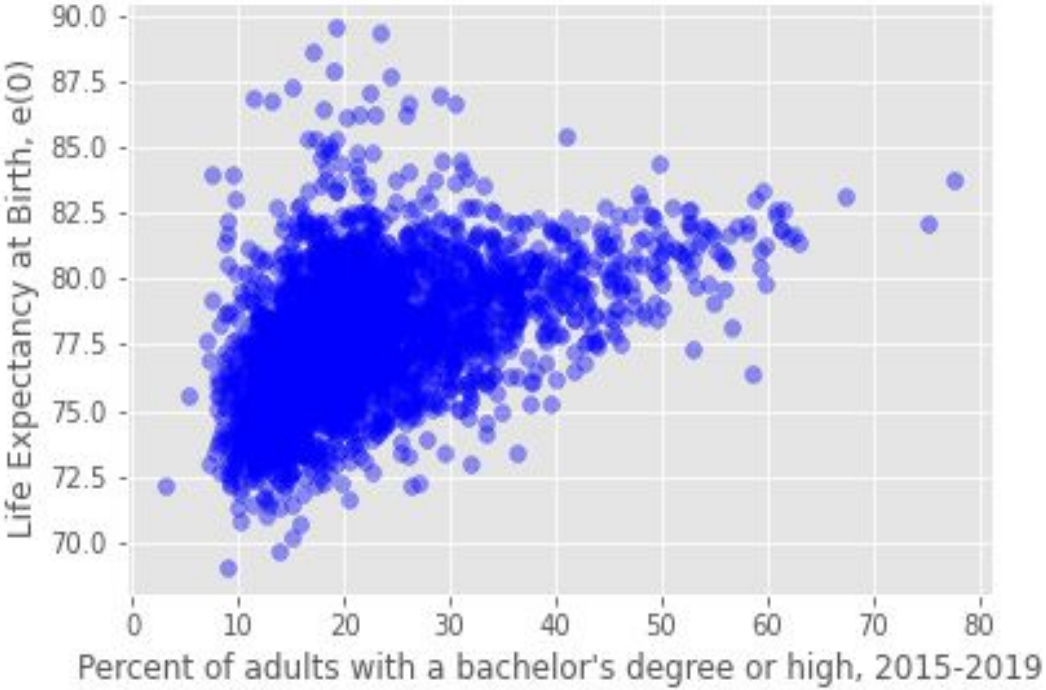
Exploratory Data Analysis

Percentage of population is greatest for "Bachelor's Degree or Higher"



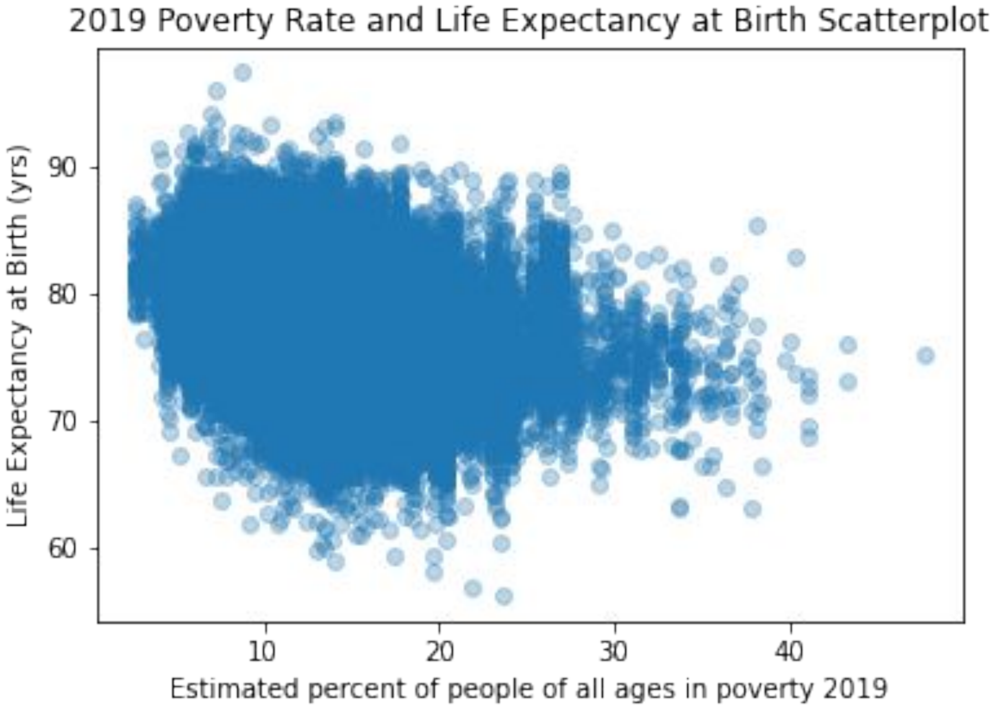
Exploratory Data Analysis

Scatterplot indicates positive correlation with life expectancy at birth.



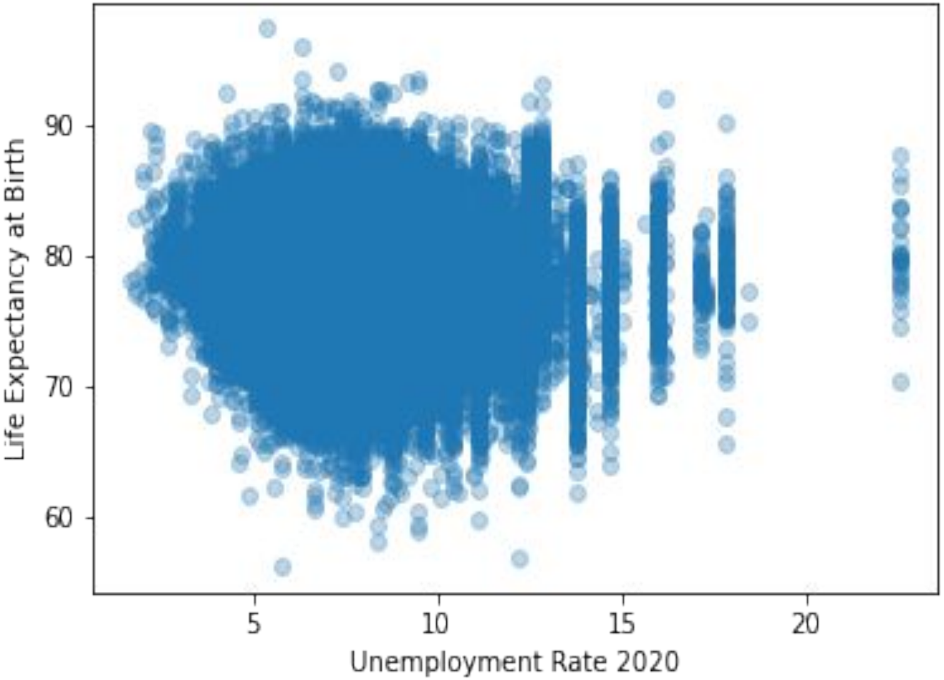
Exploratory Data Analysis

Negative correlation between 2019 Poverty Rate and Life Expectancy at Birth



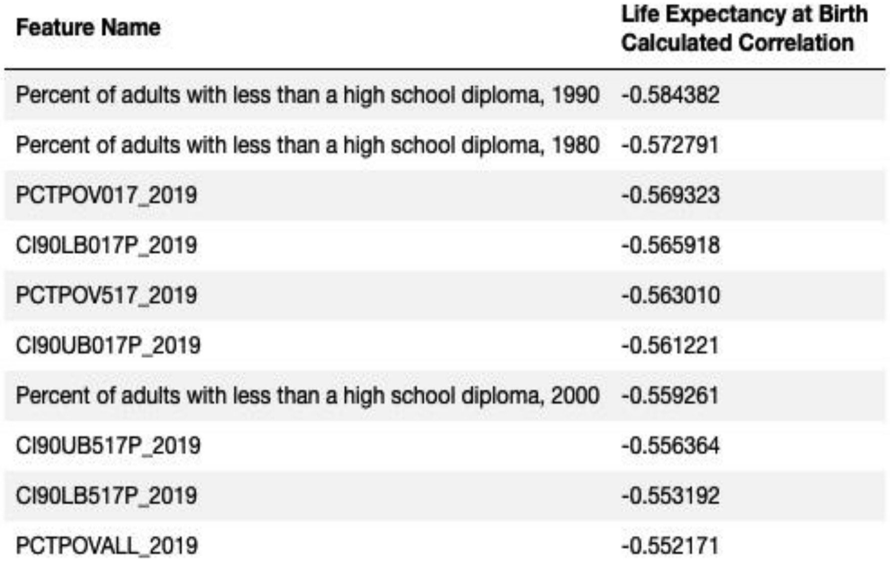
Exploratory Data Analysis

Unemployment Rate in 2020 shows no correlation to Life Expectancy at Birth



Exploratory Data Analysis Summary

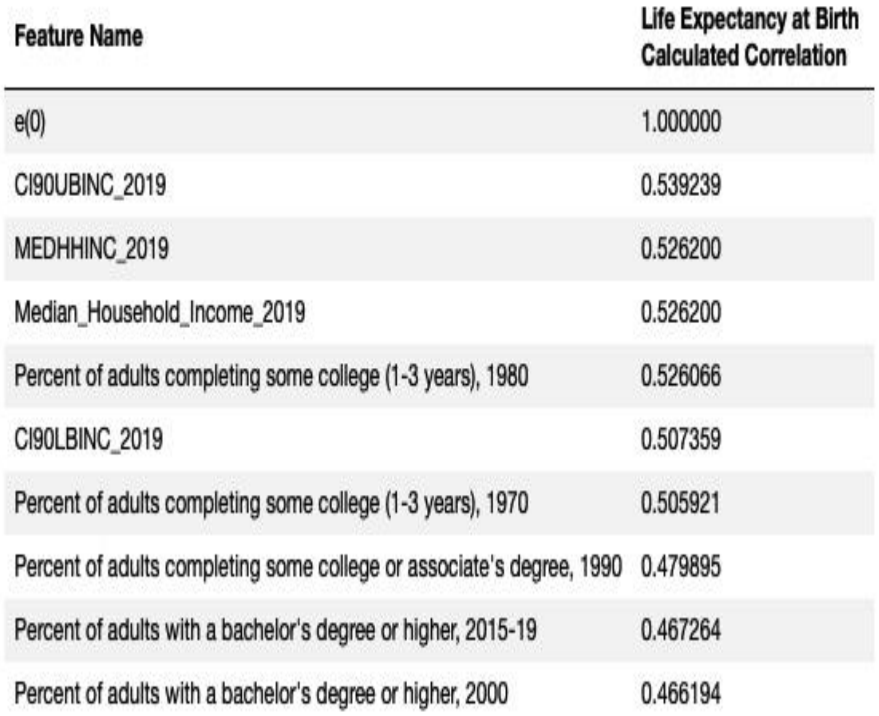
- Percent of adults with less than a high school degree in 1990 have largest negative correlation



Feature Name	Life Expectancy at Birth Calculated Correlation
Percent of adults with less than a high school diploma, 1990	-0.584382
Percent of adults with less than a high school diploma, 1980	-0.572791
PCTPOV017_2019	-0.569323
CI90LB017P_2019	-0.565918
PCTPOV517_2019	-0.563010
CI90UB017P_2019	-0.561221
Percent of adults with less than a high school diploma, 2000	-0.559261
CI90UB517P_2019	-0.556364
CI90LB517P_2019	-0.553192
PCTPOVALL_2019	-0.552171

Exploratory Data Analysis Summary

- 90% confidence interval Upper Bound Poverty (CI90UBINC_2019) has highest positive correlation with life expectancy at birth



Feature Name	Life Expectancy at Birth Calculated Correlation
e(0)	1.000000
CI90UBINC_2019	0.539239
MEDHHINC_2019	0.526200
Median_Household_Income_2019	0.526200
Percent of adults completing some college (1-3 years), 1980	0.526066
CI90LBINC_2019	0.507359
Percent of adults completing some college (1-3 years), 1970	0.505921
Percent of adults completing some college or associate's degree, 1990	0.479895
Percent of adults with a bachelor's degree or higher, 2015-19	0.467264
Percent of adults with a bachelor's degree or higher, 2000	0.466194

Exploratory Data Analysis Summary

6 principal components responsible for 89.2% of variance in dataset

Features Selected Using Principal Component Analysis-

- Percent of adults with less than a high school diploma, 2015-19
- Percent of adults with a high school diploma only, 2015-19
- Percent of adults completing some college or associate's degree, 2015-19
- Percent of adults with a bachelor's degree or higher, 2015-19
- PCTPOV017_2019
- Unemployment_rate_2020
- Median_Household_Income_2019
- Med_HH_Income_Percent_of_State_Total_2019

Modeling -

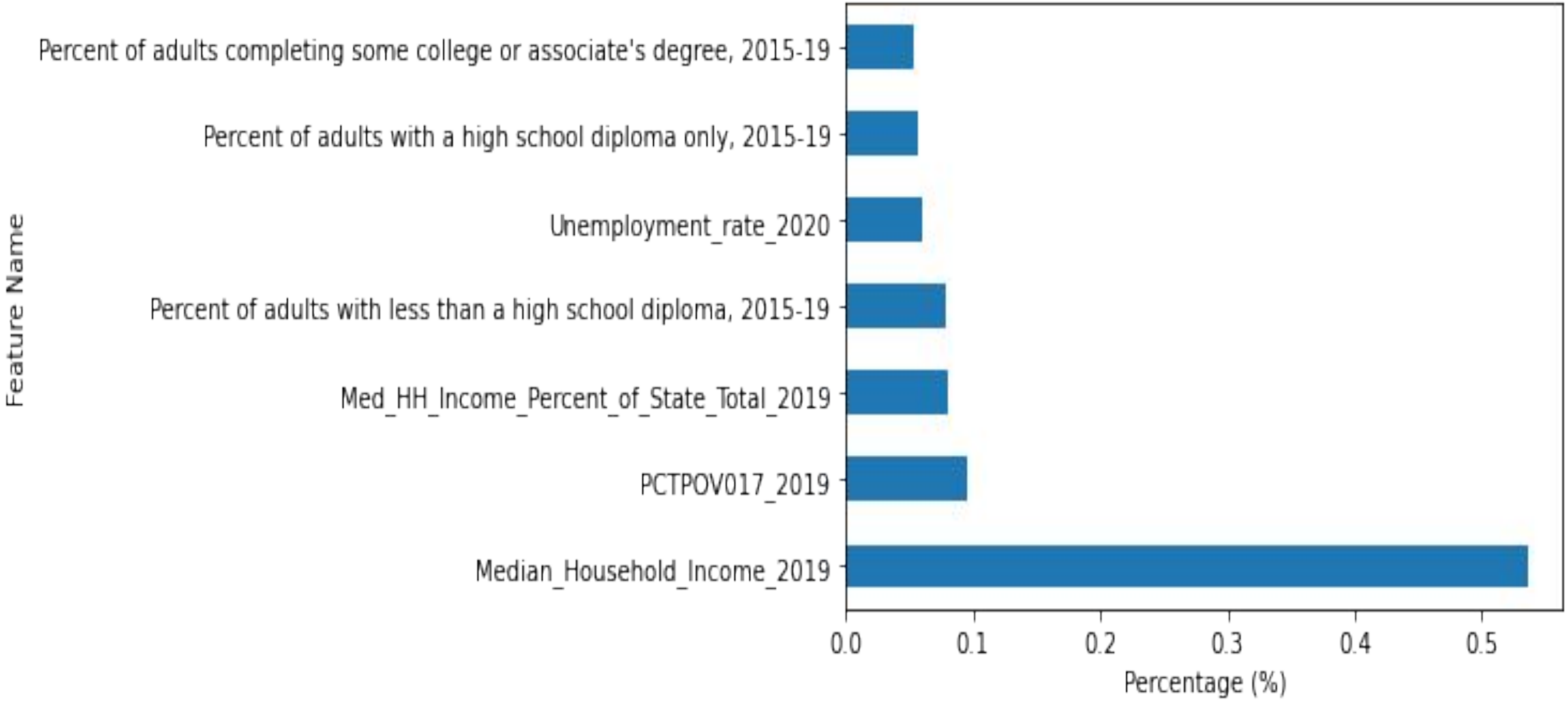
Regression Models to be Compared

- Random Forest Regressor
- XGBoost Regressor

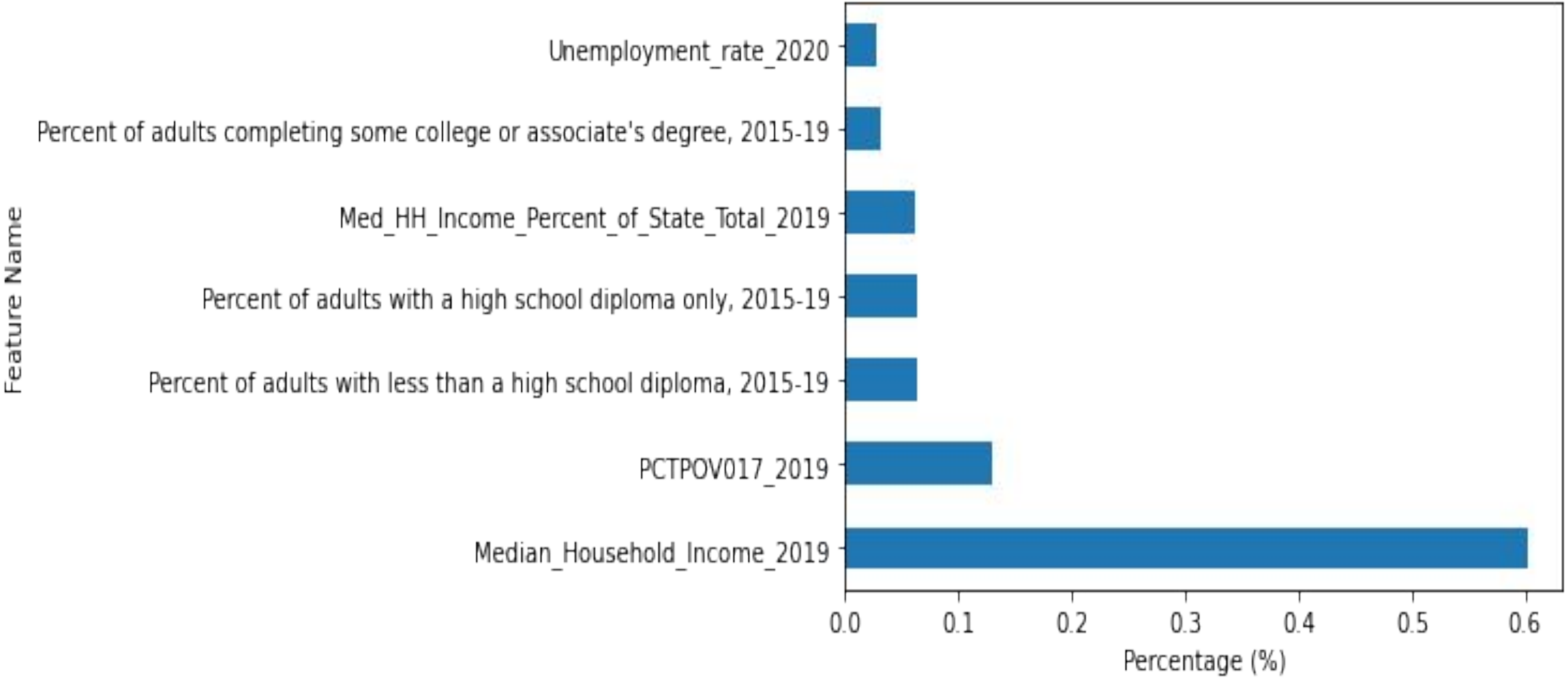
Modeling Pipeline Format

- StandardScaler for all numerical features
- GridSearchCV with 5-fold Cross-Validation for hyperparameter Tuning

RandomForestRegressor - Feature Importance



XGBoost Modeling - Feature Importance



Conclusions:

- Random Forest

Train set score: 0.322864

Test set score: 0.25

Best parameters: {'forest__max_features': 6, 'forest__n_estimators': 100}

RMSE score: 3.4340822236376454

Best cross_validation score: 0.24

- XGBoost

Train set score: 0.291235

Test set score: 0.26

Best parameters: {'gbrt__learning_rate': 0.1, 'gbrt__n_estimators': 500}

RMSE score: 3.434725486153979

Best cross_validation score: 0.26

Conclusion:

- Median Household Income Feature is largest influence
- Nearly 5x the amount of the next closest feature, Percent of Population Over 17 Living in Poverty.