Factors associated with differences in Life Expectancy across the United States

Tonia Chu

Under the mentorship: Srdjan Santic

For the course: Data Science Career Track (Springboard)

Content

- Introduction
- Dataset
- Data Analysis
- Data Modeling
- Analysis Results
- Future Work

Introduction

In this project, I want to solve following problems:

- What is the shape of the income–life expectancy gradient?
- How are gaps in life expectancy changing over time?
- How do the gaps vary across local areas?
- What are the factors associated with differences in life expectancy?

DATASET

The dataset include 14 csv files of data tables.

Data Table 1: National life expectancy estimates (pooling 2001-14) for men and women, by income percentile

Data Table 2: National by-year life expectancy estimates for men and women, by income percentile

Data Table 3: State-level life expectancy estimates for men and women, by income quartile

Data Table 4: State-level estimates of trends in life expectancy for men and women, by income quartile

Data Table 5: State-level by-year life expectancy estimates for men and women, by income quartile

Data Table 6: CZ-level life expectancy estimates for men and women, by income quartile

Data Table 7: CZ-level life expectancy estimates for men and women, by income ventile

Data Table 8: CZ-level estimates of trends in life expectancy for men and women, by income quartile

Data Table 9: CZ-level by-year life expectancy estimates for men and women, by income quartile

Data Table 10: CZ-level characteristics described in eTable 9

Data Table 11: County-level life expectancy estimates for men and women, by income quartile

Data Table 12: County-level characteristics described in eTable 11

Data Table 13: International estimates of mean life expectancy at age 40, by country for men and women

Data Table 14: Comparison of population and death counts in tax data and NCHS data

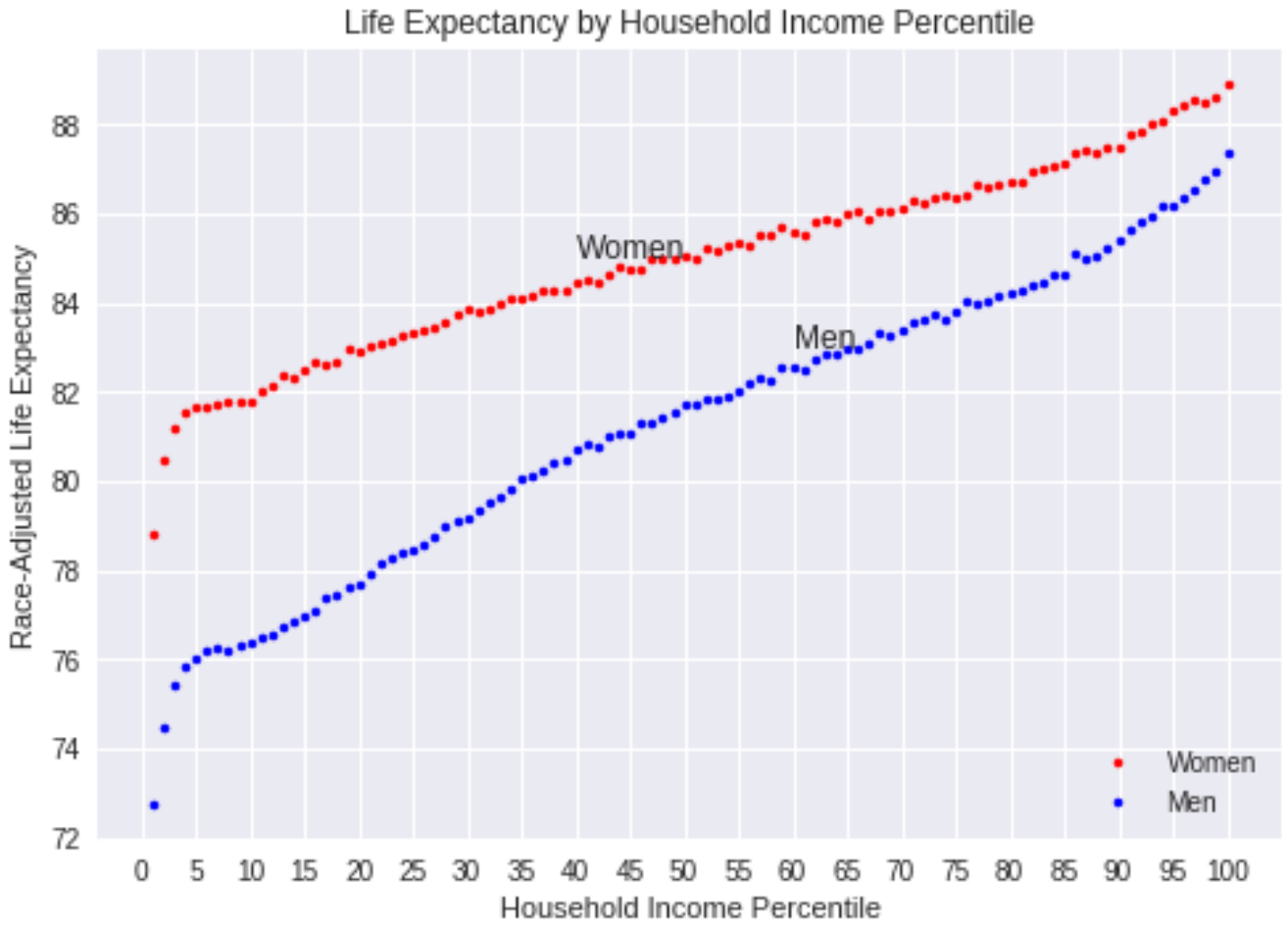
Data Wrangling

- Remove the unadjusted and Standard Error columns in the tables
- Fill missing values in table 10 and table 12
- There are 3 steps to fill missing values in table 12:
  - A county is removed if all the values of a column are missing.
  - A column is removed if there are more than 20% missing value.
  - Fill missing values with the mean value of that that county.

Exploratory Data Analysis

- 1. National Levels of Life Expectancy by Income
- 2. National Trends in Life Expectancy by Income in year 2001~2014
- 3. Local Area Variation in Life Expectancy gap by Income

National Levels of Life Expectancy by Income



National Levels of Life Expectancy by Income

Women, Bottom 1%: 78.8

Women, Top 1%: 88.9

Women, Life expectancy gap: 10.1

Men, Bottom 1%: 72.7

Men, Top 1%: 87.3

Men, Life expectancy gap: 14.6

National Levels of Life Expectancy by Income

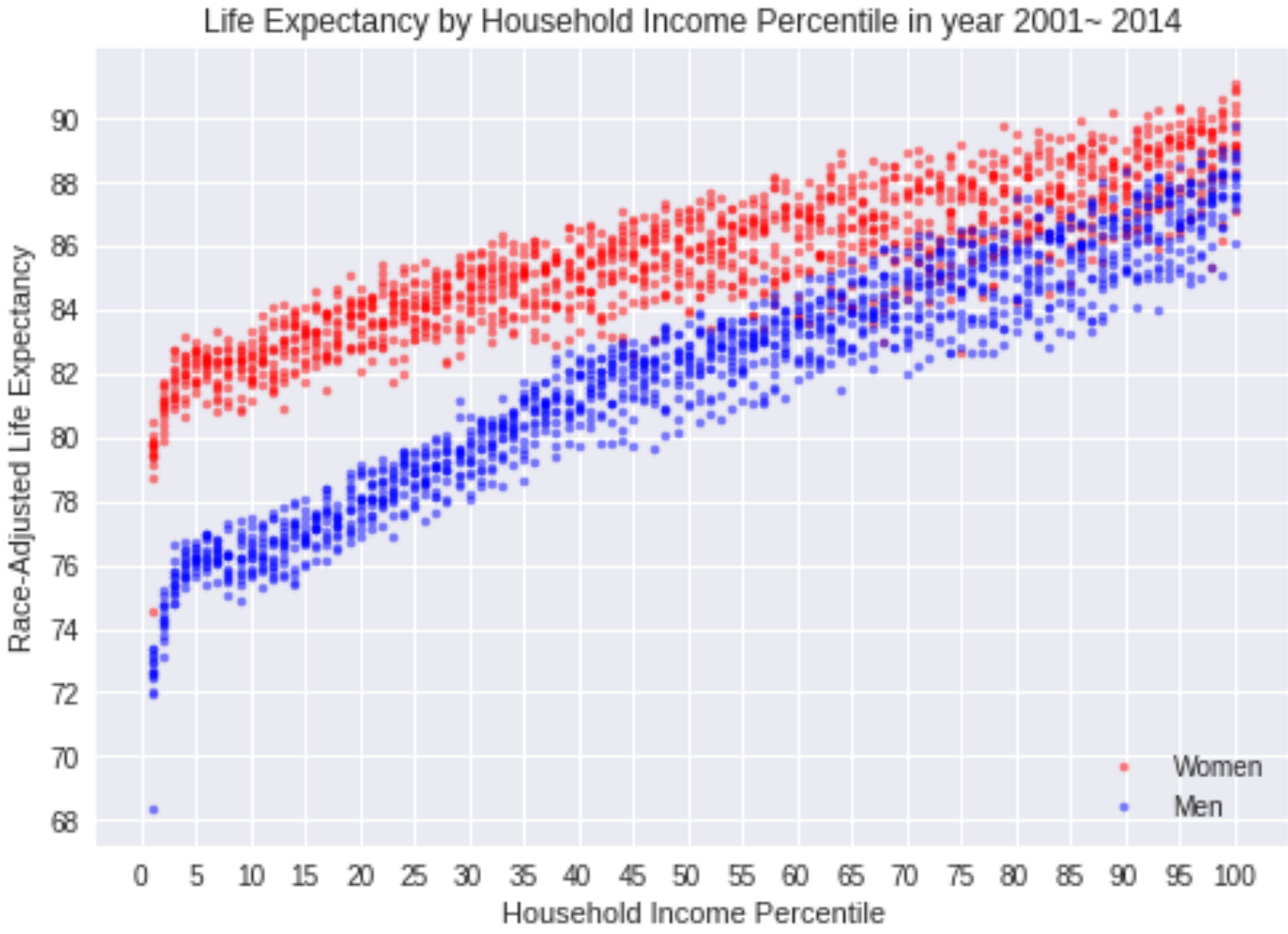
Gender gap, Bottom 1%: 6.0

Gender gap, Top 1%: 1.5

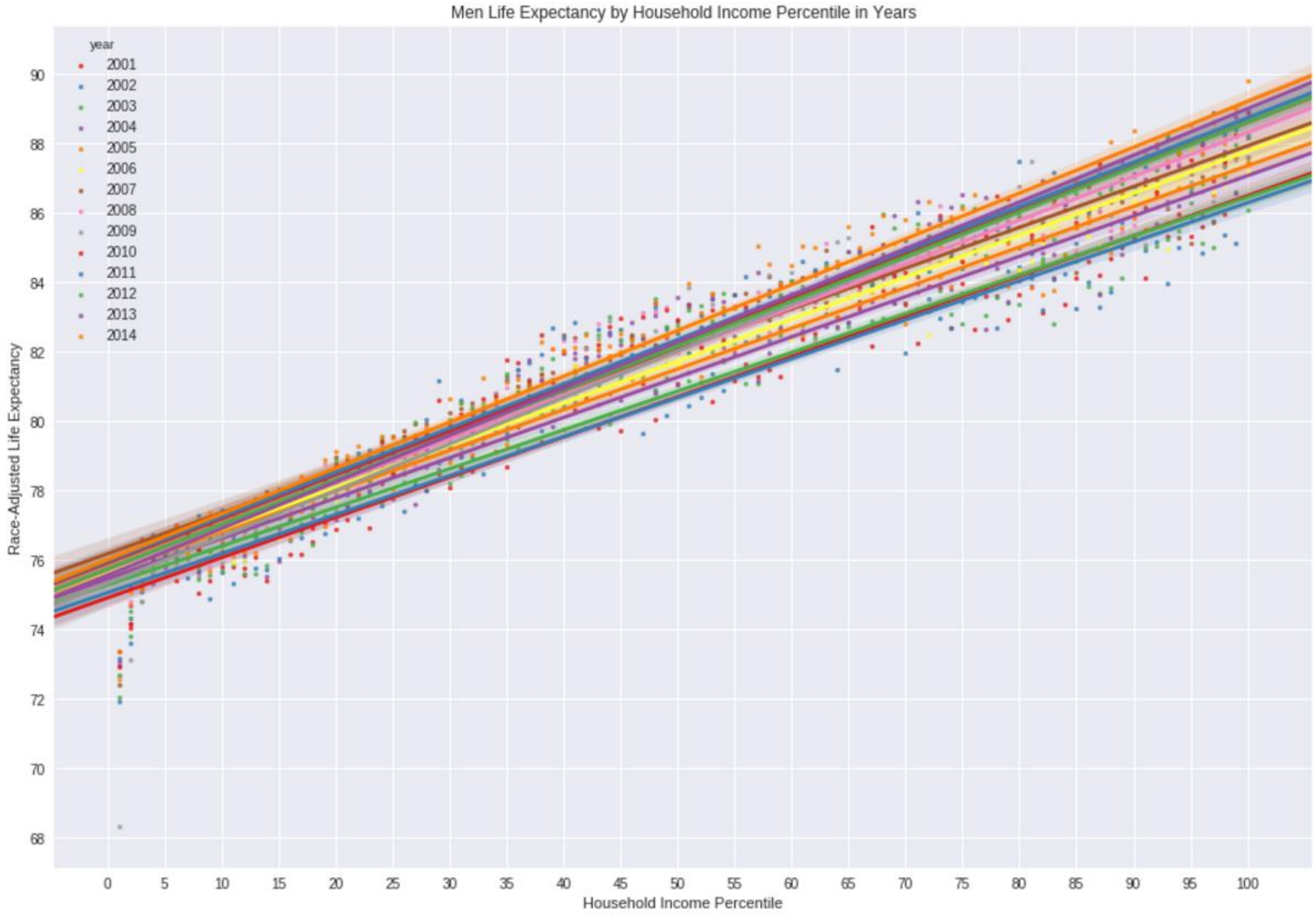
Women, Slope of linear regression: 0.07

Men, Slope of linear regression: 0.11

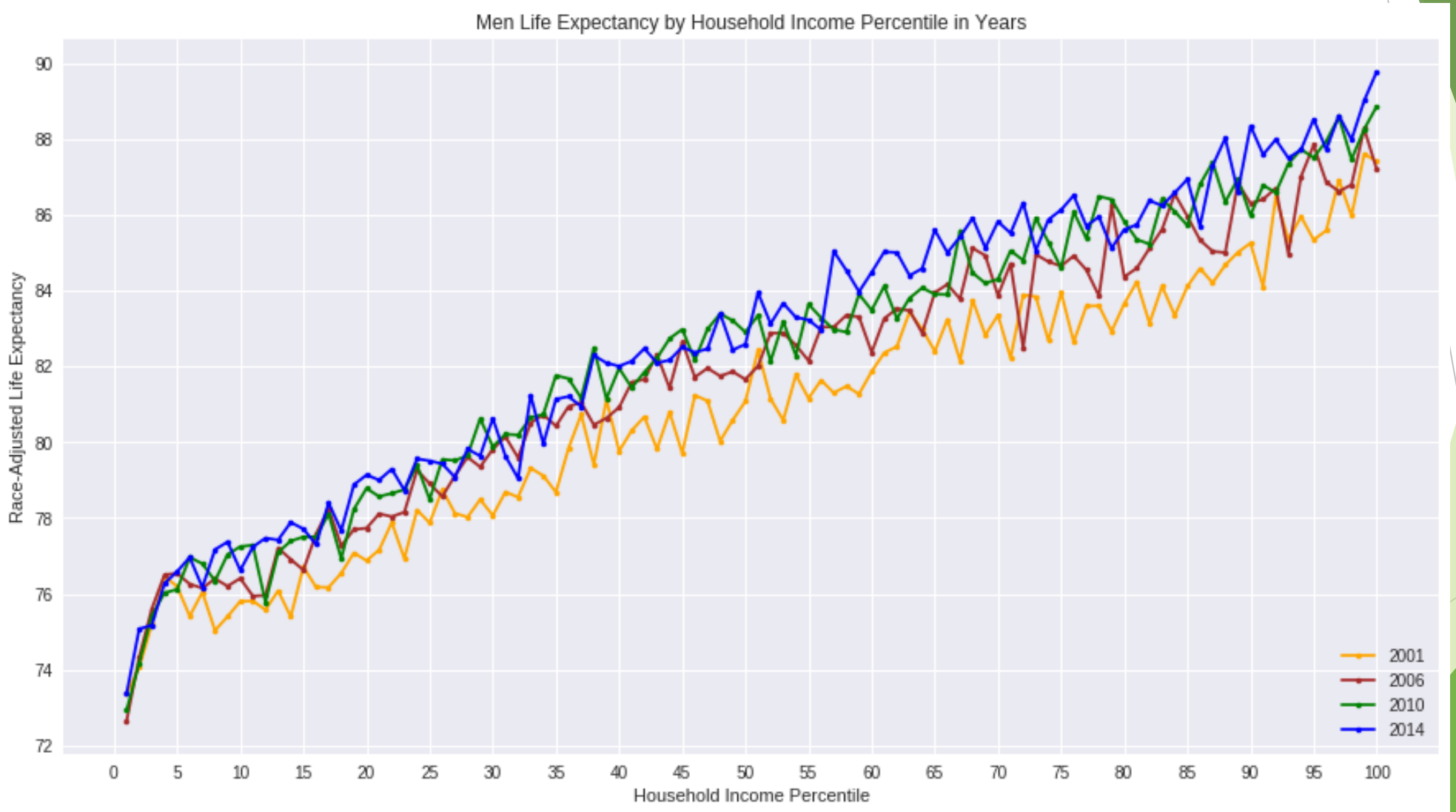
National Trends in Life Expectancy by Income in year 2001~2014



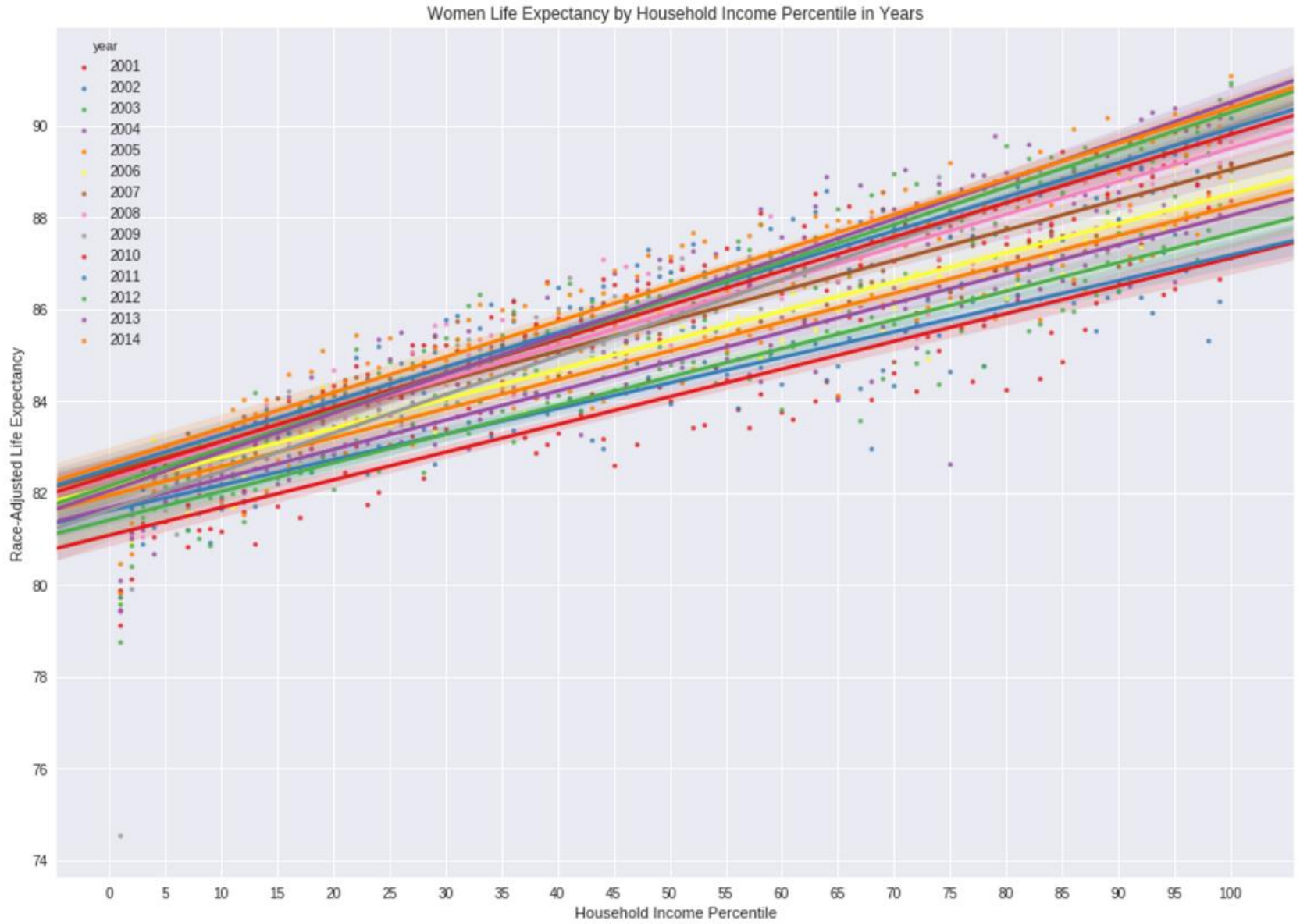
Men Life Expectancy by Household Income Percentile in year 2001~2014



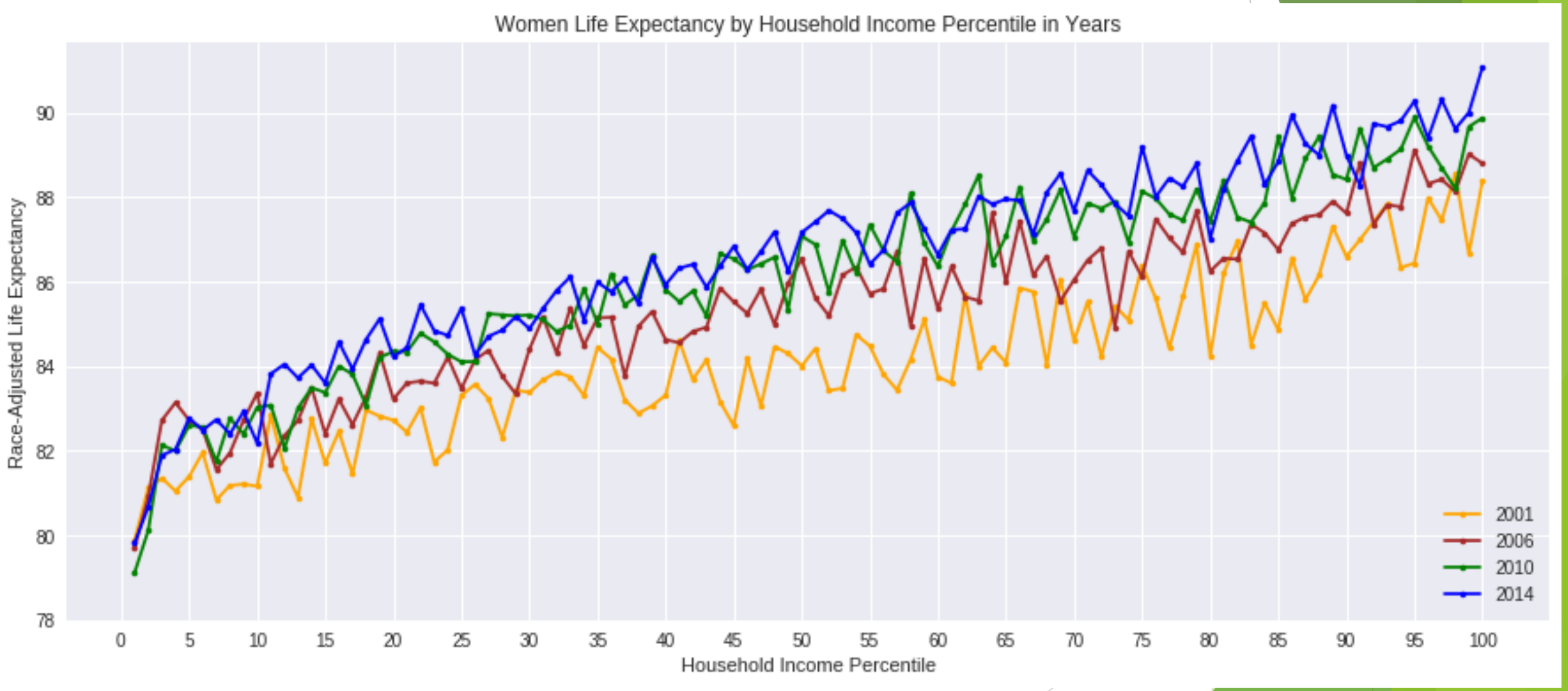
Men Life Expectancy by Household Income Percentile in Years



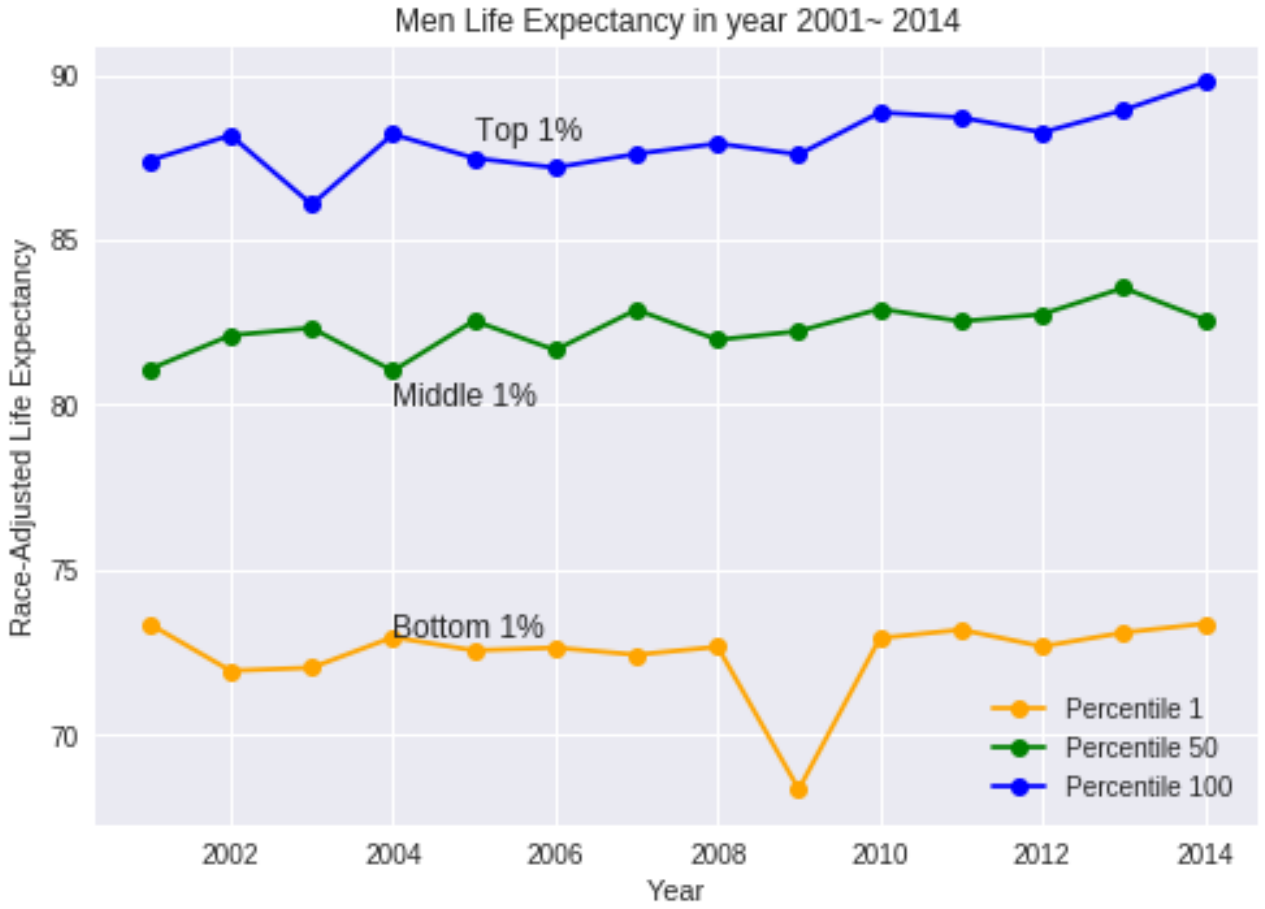
Women Life Expectancy by Household Income Percentile in year 2001~2014



Women Life Expectancy by Household Income Percentile in Years



Men Life Expectancy Trend in Years 2001~2014



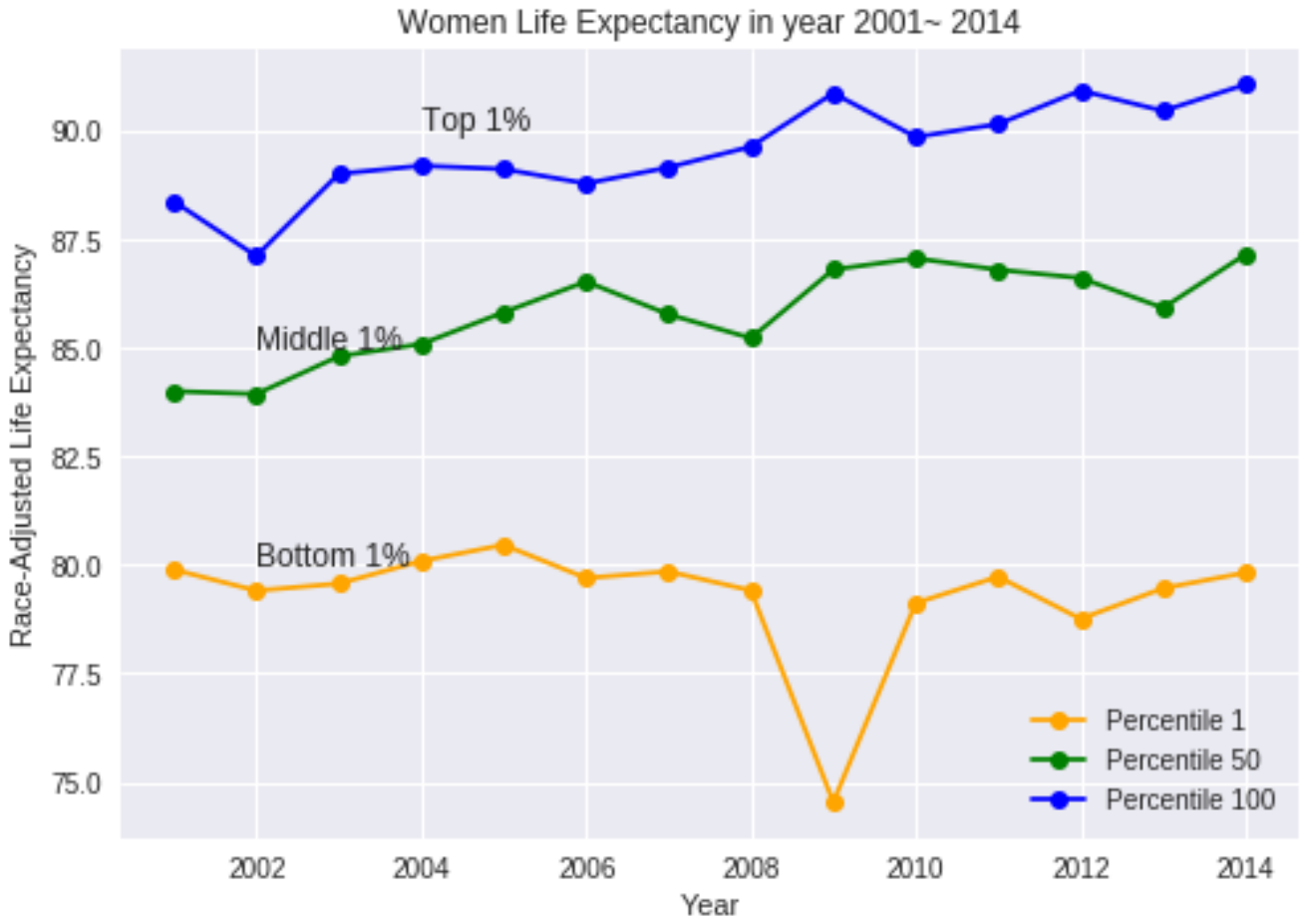
Men, Bottom 1%,

Life expectancy change: -0.1

Men, Top 1%,

Life expectancy change: 2.4

Women Life Expectancy Trend in Years 2001~2014



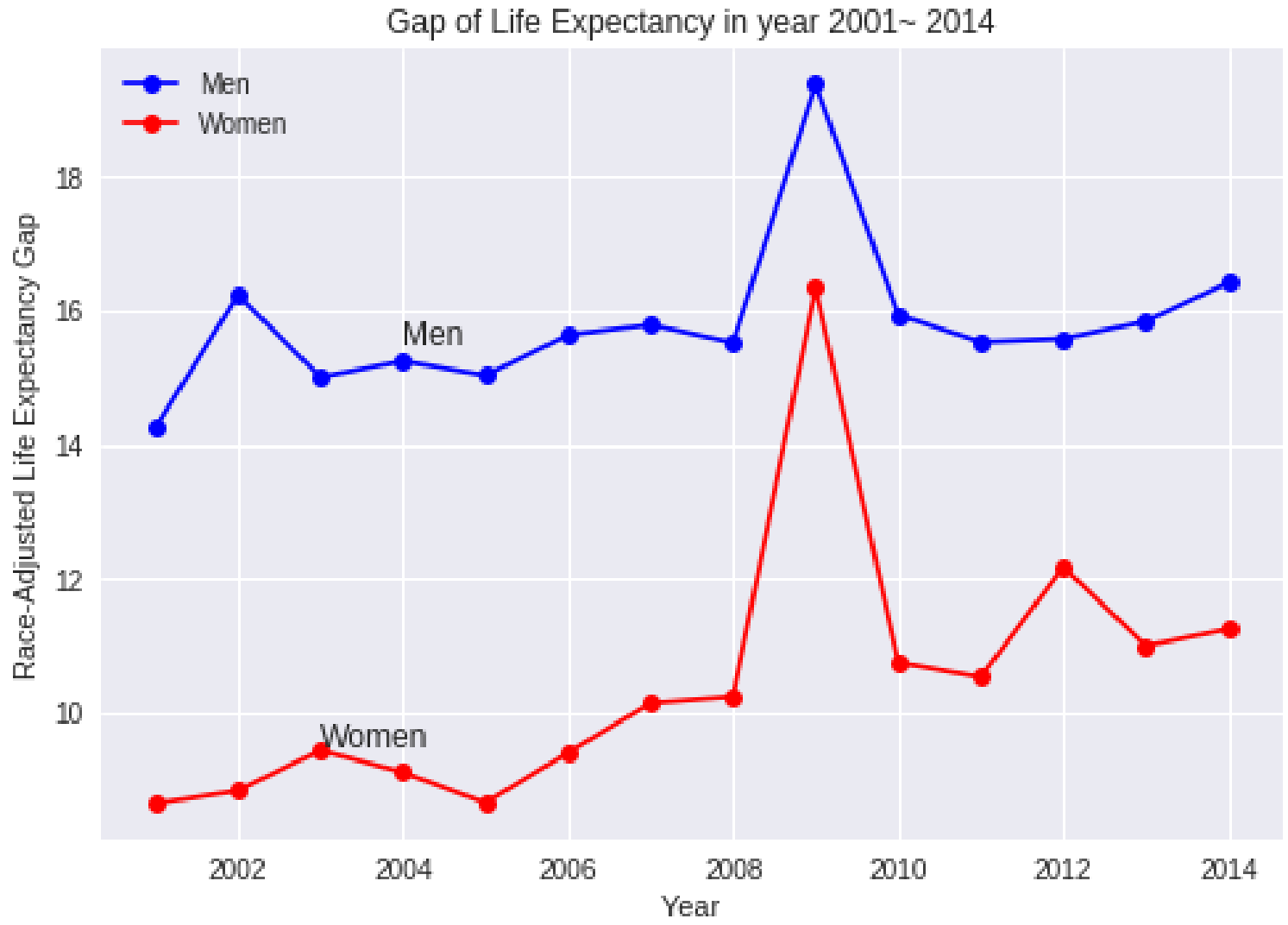
Women, Bottom 1%,

Life expectancy change: -0.1

Women, Top 1%,

Life expectancy change: 2.7

Life Expectancy Gap Trend in Years 2001~2014

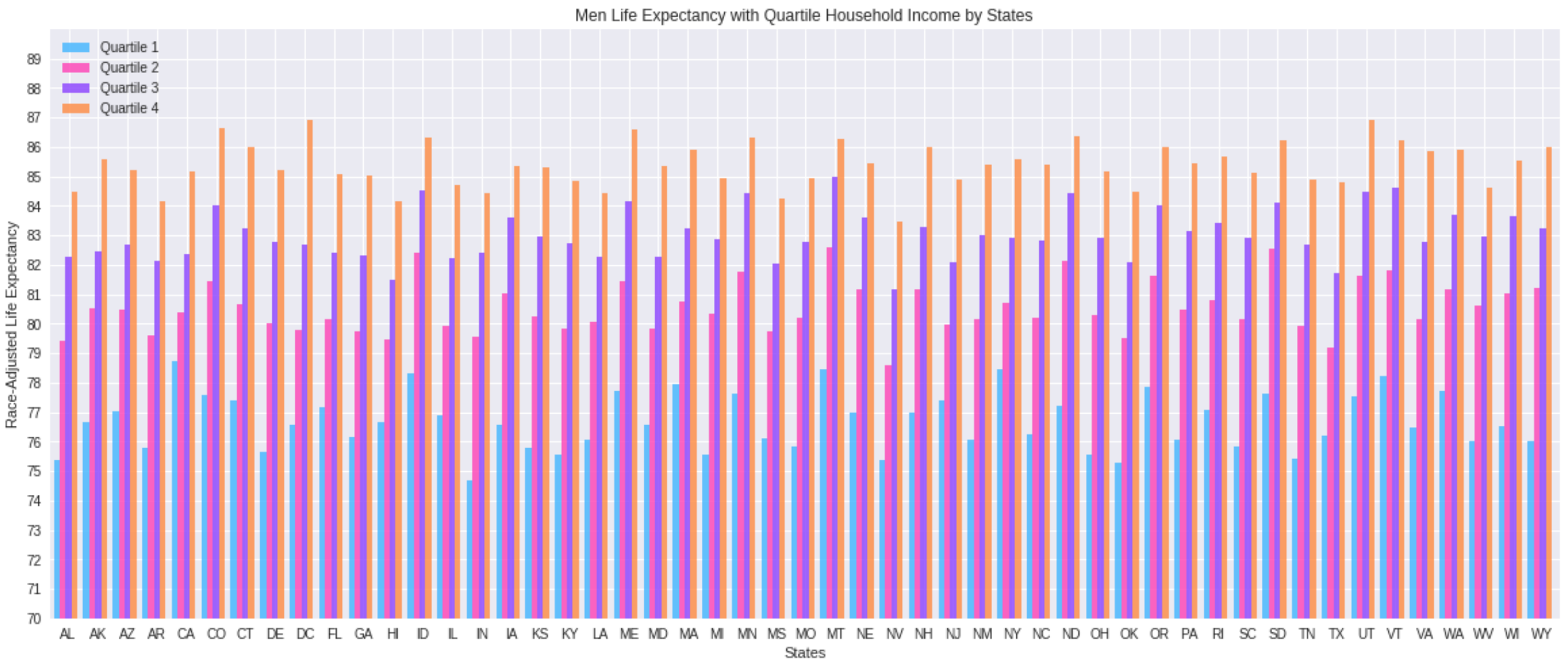


Local Area Variation in Life Expectancy Gap by Income

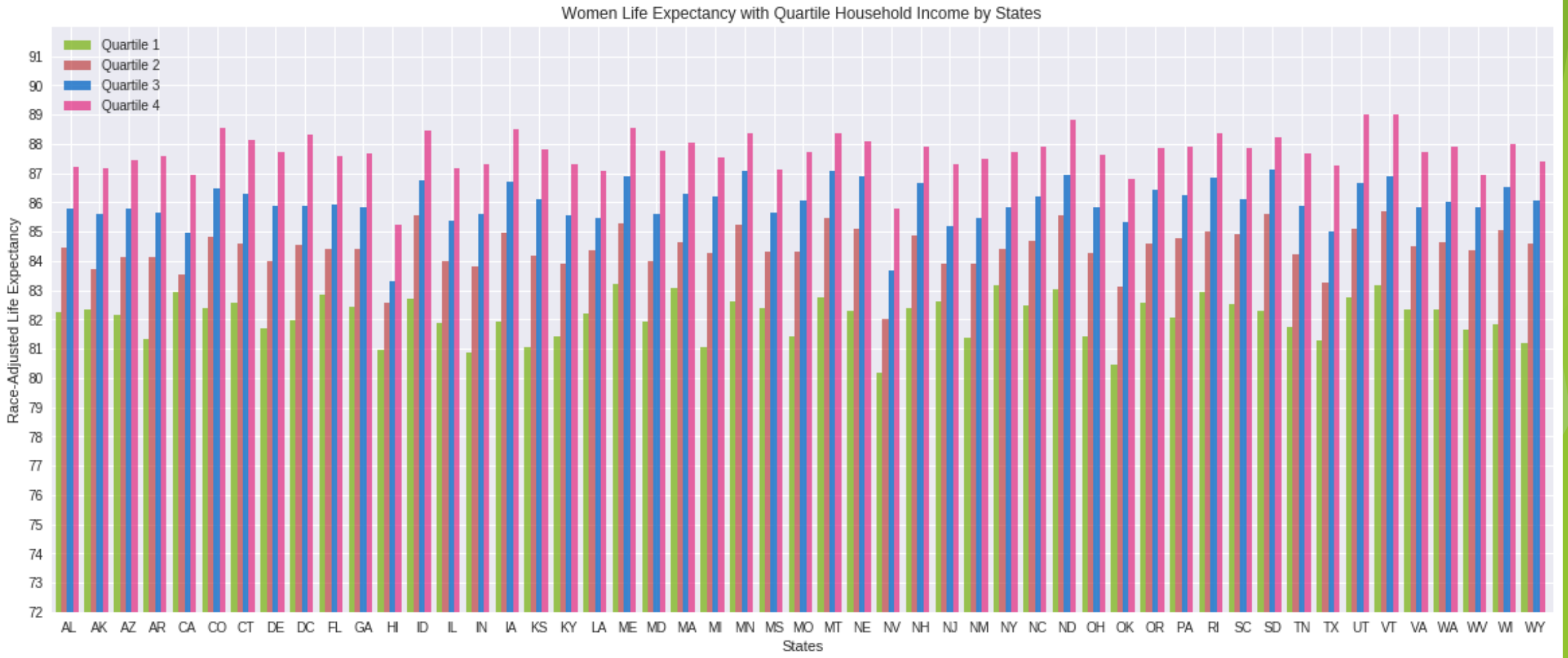
How do the gaps vary across local areas?

- Life Expectancy gap by State
- Life Expectancy gap by Commuting Zone

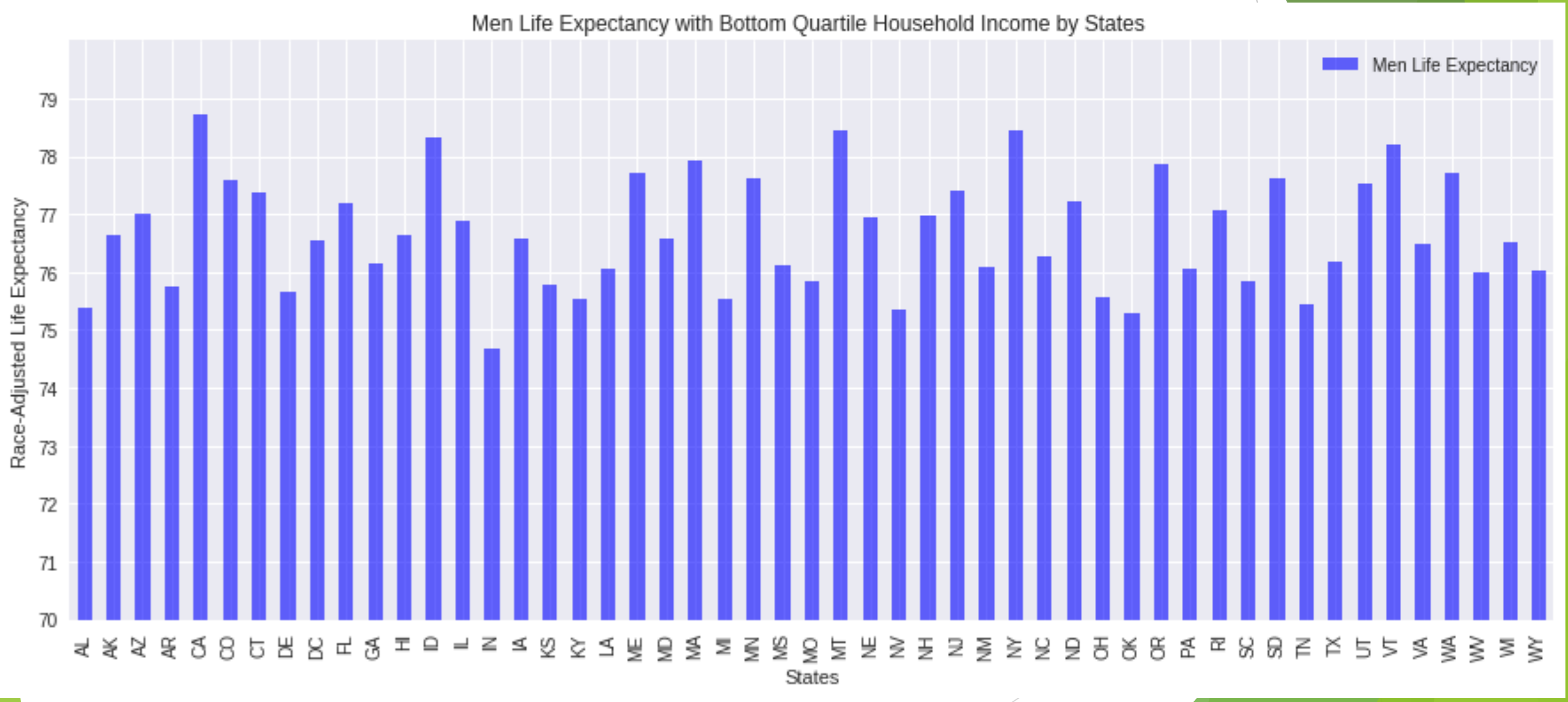
Life Expectancy gap by State



Life Expectancy gap by State



Life Expectancy of Q1 by State



Life Expectancy of Q1 by State

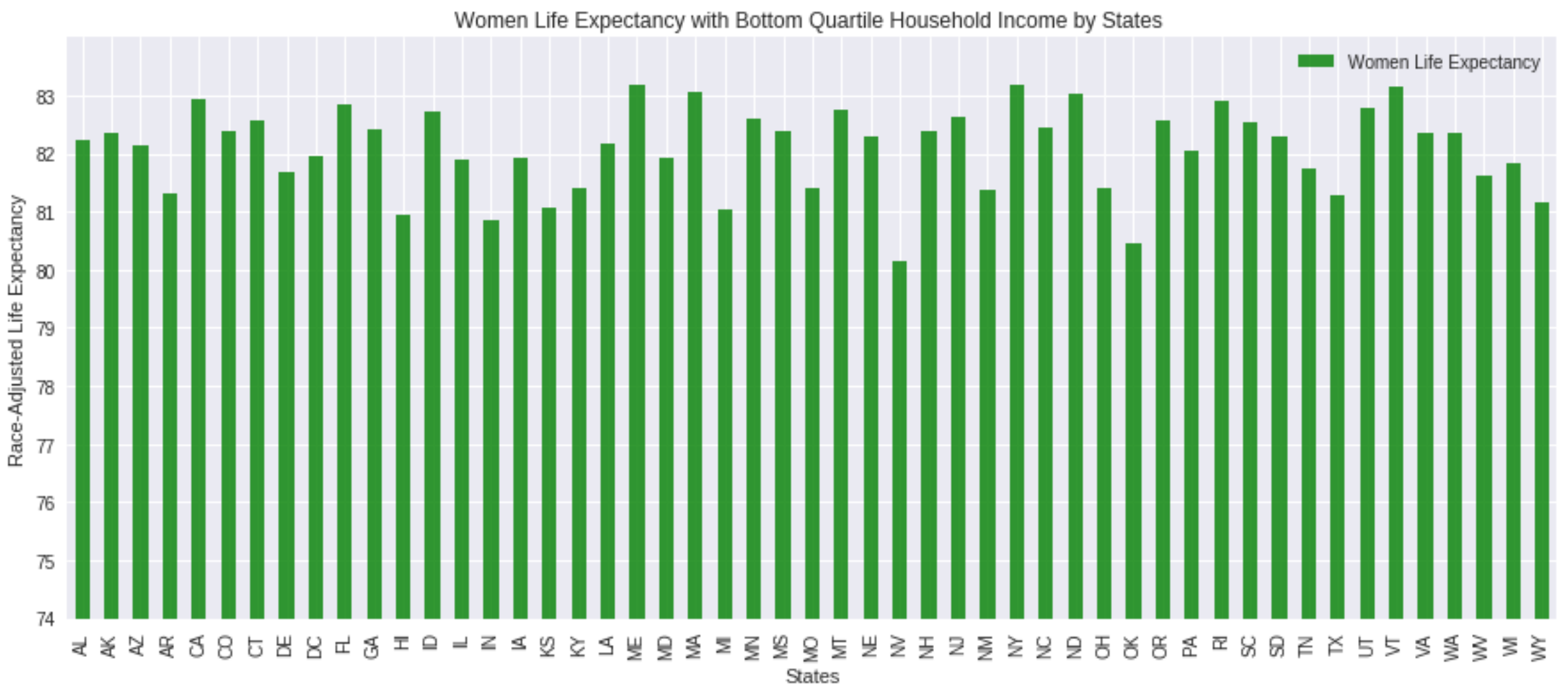
5 states with the highest men life expectancy of bottom quartile income

	statename	stateabbrv	le_raceadj_q1_M
0	California	CA	78.73162
1	New York	NY	78.45039
2	Montana	MT	78.44444
3	Idaho	ID	78.33078
4	Vermont	VT	78.20271

5 states with the lowest men life expectancy of bottom quartile income

	statename	stateabbrv	le_raceadj_q1_M
0	Indiana	IN	74.68581
1	Oklahoma	OK	75.28735
2	Nevada	NV	75.36532
3	Alabama	AL	75.37608
4	Tennessee	TN	75.43765

Life Expectancy of Q1 by State



Life Expectancy of Q1 by State

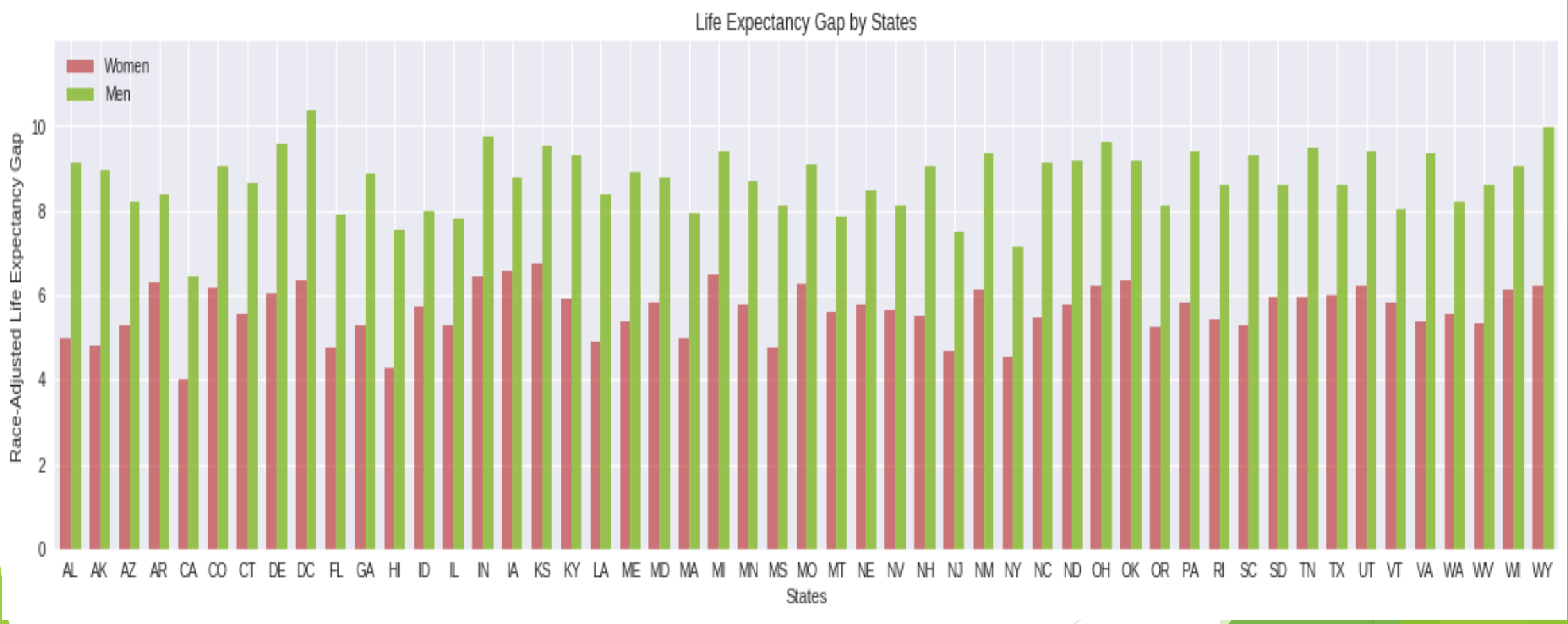
5 states with the highest women life expectancy of bottom quartile income

	statename	stateabbrv	le_raceadj_q1_F
0	Maine	ME	83.19597
1	New York	NY	83.17820
2	Vermont	VT	83.15334
3	Massachusetts	MA	83.06703
4	North Dakota	ND	83.01955

5 states with the lowest women life expectancy of bottom quartile income

	statename	stateabbrv	le_raceadj_q1_F
0	Nevada	NV	80.16355
1	Oklahoma	OK	80.44940
2	Indiana	IN	80.85909
3	Hawaii	HI	80.95252
4	Michigan	MI	81.03877

Life Expectancy gap by State



Life Expectancy gap by State

5 states with the highest life expectancy gap of men

	statename	stateabbrv	gap_M
0	District Of Columbia	DC	10.38028
1	Wyoming	WY	9.96908
2	Indiana	IN	9.74794
3	Ohio	OH	9.60021
4	Delaware	DE	9.56742

5 states with the lowest life expectancy gap of men

	statename	stateabbrv	gap_M
0	California	CA	6.43012
1	New York	NY	7.14489
2	New Jersey	NJ	7.50561
3	Hawaii	HI	7.51897
4	Illinois	IL	7.81921

Life Expectancy gap by State

5 states with the highest life expectancy gap of women

	statename	stateabbrv	gap_F
0	Kansas	KS	6.73606
1	Iowa	IA	6.58660
2	Michigan	MI	6.49275
3	Indiana	IN	6.42861
4	Oklahoma	OK	6.35006

5 states with the lowest life expectancy gap of women

	statename	stateabbrv	gap_F
0	California	CA	3.99338
1	Hawaii	HI	4.29524
2	New York	NY	4.53851
3	New Jersey	NJ	4.67141
4	Florida	FL	4.74321

Life Expectancy gap by Commuting Zone

CZs with the highest life expectancy gap of men

	statename	czname	gap_M
1	Texas	Lubbock	17.05468
2	South Carolina	Spartanburg	16.49910
3	West Virginia	Charleston	16.40359
4	Ohio	Mansfield	16.22799
5	Wisconsin	Appleton	16.11722
6	Michigan	Kalamazoo	16.10882
7	Nebraska	Lincoln	16.05373
8	Utah	Salt Lake City	15.87730
9	Indiana	Terre Haute	15.82364
10	New York	Union	15.66260

CZs with the highest life expectancy gap of women

	statename	czname	gap_F
1	Texas	Midland	12.78140
2	Illinois	Springfield	12.36650
3	Indiana	Lafayette	12.22876
4	Illinois	Davenport	12.01902
5	Nebraska	Lincoln	11.96011
6	Wisconsin	Green Bay	11.93846
7	Missouri	Columbia	11.88892
8	Pennsylvania	Hagerstown	11.87228
9	Wisconsin	Appleton	11.83465
10	Ohio	Mansfield	11.82840

Life Expectancy gap by Commuting Zone

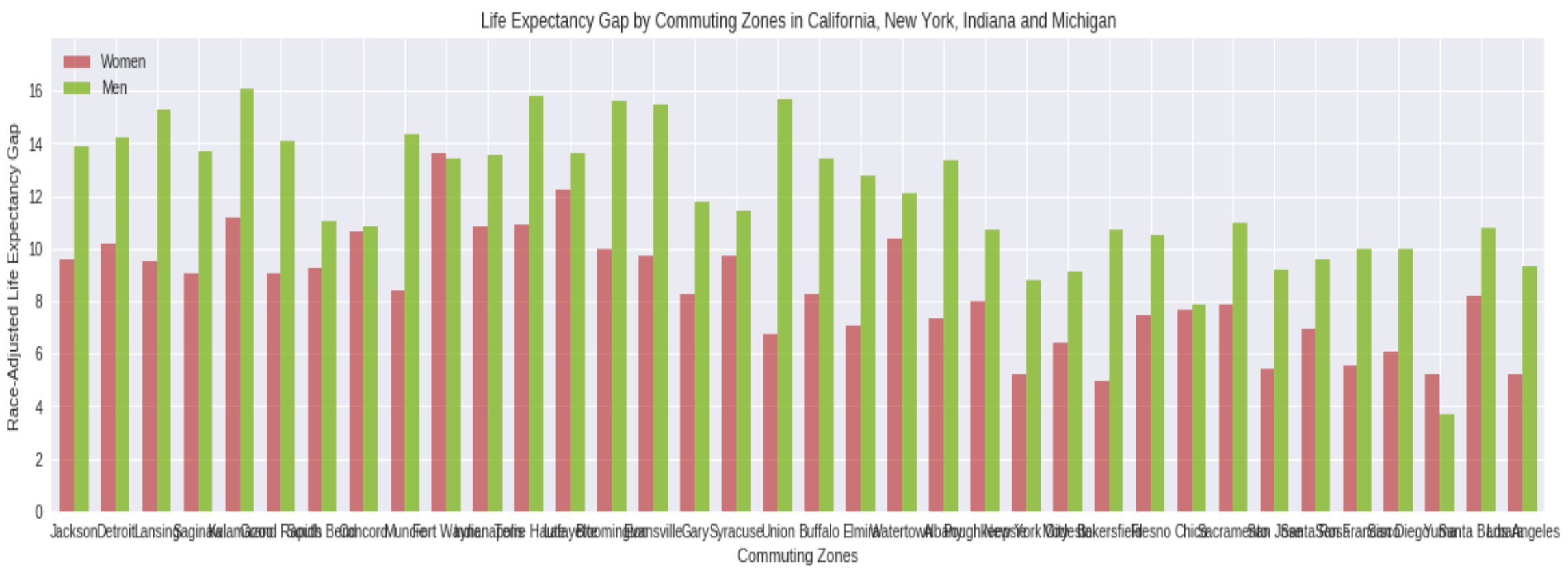
CZs with the lowest life expectancy gap of men

	statename	czname	gap_M
1	California	Chico	7.84758
2	Texas	Tyler	8.45538
3	New York	New York City	8.82618
4	Texas	El Paso	9.01521
5	North Carolina	Asheville	9.04417
6	Pennsylvania	State College	9.07298
7	California	Modesto	9.10118
8	Missouri	Springfield	9.13781
9	Maine	Bangor	9.15706
10	Texas	Longview	9.18776

CZs with the lowest life expectancy gap of women

	statename	czname	gap_F
1	Washington	Yakima	4.41043
2	Arizona	Flagstaff	4.59987
3	Wisconsin	Wausau	4.60628
4	Maine	Bangor	4.75503
5	Utah	Provo	4.92064
6	California	Bakersfield	4.96867
7	California	Los Angeles	5.19068
8	New York	New York City	5.20648
9	California	Yuma	5.21239
10	California	San Jose	5.41566

Life Expectancy gap by Commuting Zones in California, New York, Indiana and Michigan



Life Expectancy gap by Commuting Zones in California, New York, Indiana and Michigan

5 CZs with the highest life expectancy gap of men

	statename	czname	gap_M
1	Indiana	Terre Haute	15.82364
2	New York	Union	15.66260
3	Indiana	Bloomington	15.61690
4	Indiana	Evansville	15.46764
5	Michigan	Lansing	15.31108

5 CZs with the highest life expectancy gap of women

	statename	czname	gap_F
1	Indiana	Lafayette	12.22876
2	Michigan	Kalamazoo	11.20880
3	Indiana	Terre Haute	10.94084
4	Indiana	Indianapolis	10.82229
5	Indiana	Concord	10.62655

Life Expectancy gap by Commuting Zones in California, New York, Indiana and Michigan

5 CZs with the lowest life expectancy gap of men

	statename	czname	gap_M
1	California	Chico	7.84758
2	New York	New York City	8.82618
3	California	Modesto	9.10118
4	California	San Jose	9.22374
5	California	Los Angeles	9.32232

5 CZs with the lowest life expectancy gap of women

	statename	czname	gap_F
1	California	Los Angeles	5.19068
2	New York	New York City	5.20648
3	California	Yuma	5.21239
4	California	San Jose	5.41566
5	California	San Francisco	5.54166

Data Modeling

Find a model to predict average life expectancy of a county by factors associated with life expectancy:

- Machine Learning Models
- Feature Selection Methods
- Factors Affect Life Expectancy

Machine Learning Models

- Linear Regression
- Support Vector Regression
- Random Forest Regressor

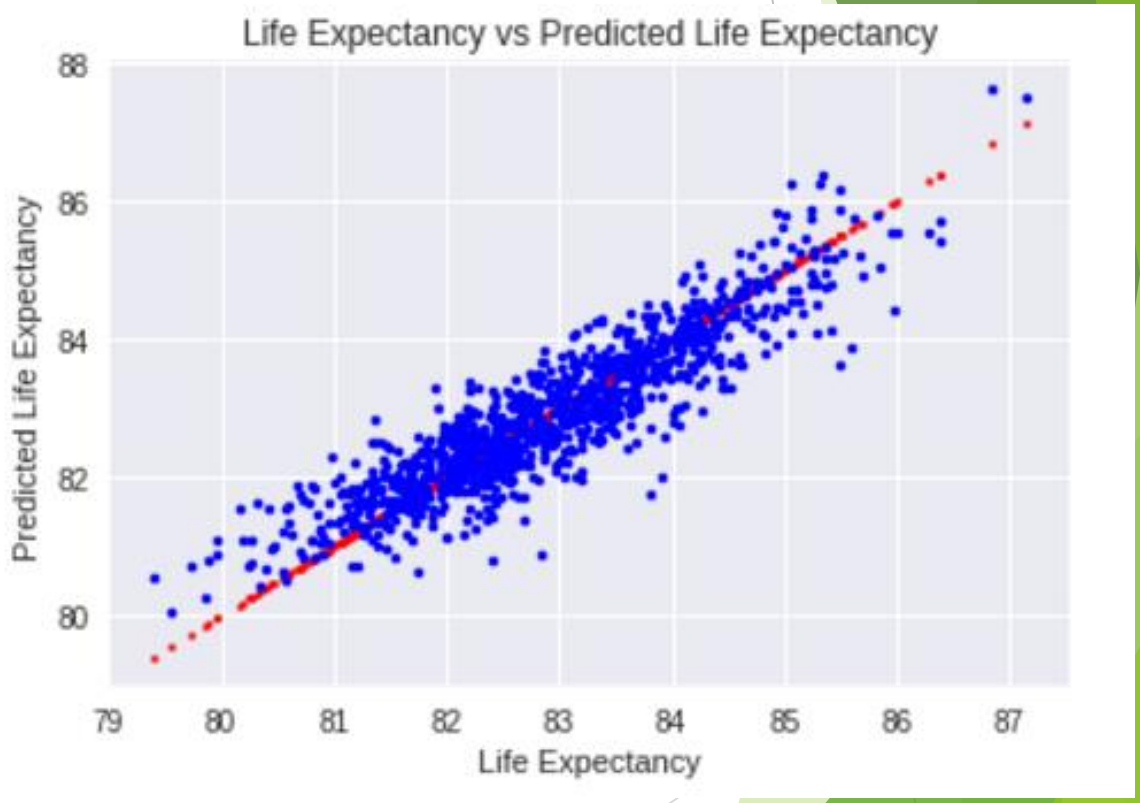
Linear Regression

Evaluation result of the model:

Model	Features	S^2	MSE
LinearRegression()	53	0.8249	0.2469

Predict test dataset with the model

Fit a model X_train, and calculate MSE with y_train: 0.2399
Fit a model X_train, and calculate MSE with X_test, y_test: 0.2964



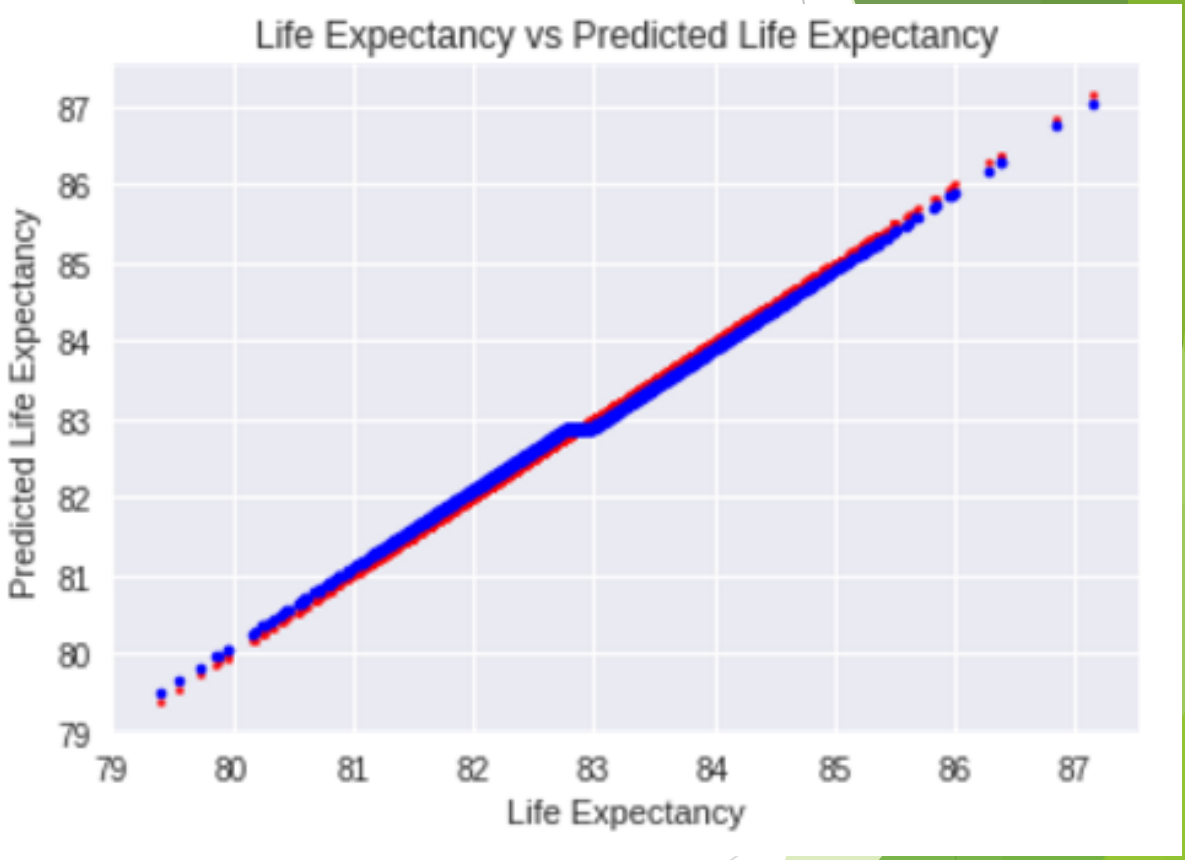
Support Vector Regression

Evaluation result of the model:

Model	C	S^2	MSE
SVR()	1.0	0.7835	0.3054
SVR(C=2)	2.0	0.9656	0.0486
SVR(C=5)	5.0	0.9932	0.0096

Predict test dataset with the model

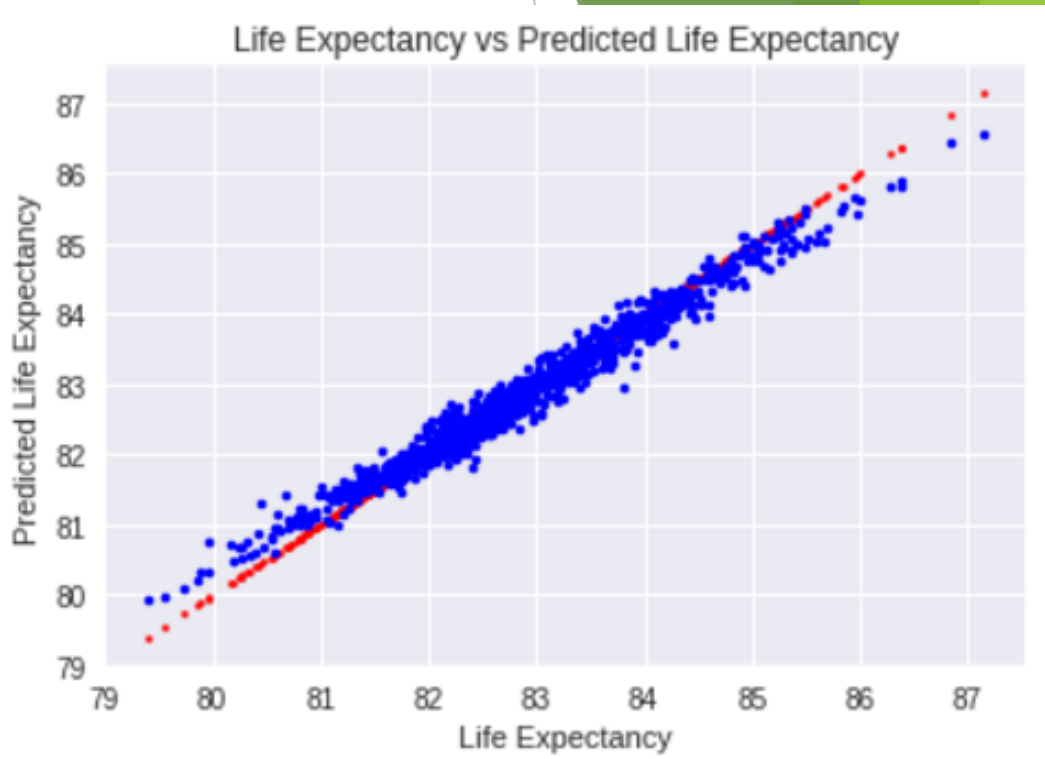
Fit a model X_train, and calculate MSE with y_train: 0.0096
Fit a model X_train, and calculate MSE with X_test, y_test: 1.4732



Random Forest Regressor

Evaluation result of the model:

Model	max_features	n_estimators	S^2	MSE
RandomForestRegressor()	53	10	0.9535	0.0656
RandomForestRegressor(max_features=20)	20	10	0.9506	0.0697
RandomForestRegressor(max_features=10)	10	10	0.9526	0.0668
RandomForestRegressor(max_features=5)	5	10	0.9497	0.0710
RandomForestRegressor(n_estimators=20)	53	20	0.9607	0.0555
RandomForestRegressor(n_estimators=100, oob_score=True)	53	100	0.9679	0.0452
RandomForestRegressor(n_estimators=200, oob_score=True, random_state=50)	53	200	0.9692	0.0434



Predict test dataset with the model

Fit a model X_train, and calculate MSE with y_train: 0.0441
Fit a model X_train, and calculate MSE with X_test, y_test: 0.3623

Feature Selection Methods

Feature selection is the process of selecting a subset of relevant features for use in model construction.

- Principal Component Analysis
- Regularization
- Random Forests

Principal Component Analysis

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize. The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations.

Evaluation result of the model:

Model	Number of Components	S^2	MSE
PCA(n_components='mle', svd_solver='full')	52	0.8249	0.2470
PCA(n_components=20)	20	0.7643	0.3324
PCA(n_components=10)	10	0.6944	0.4311
PCA(n_components=5)	5	0.5536	0.6296

Predict test dataset with the model

Model	MSE with training dataset	MSE with test dataset
PCA(n_components='mle', svd_solver='full')	0.2399	0.2955
PCA(n_components=20)	0.3230	0.3850
PCA(n_components=10)	0.4314	0.4424
PCA(n_components=5)	0.6392	0.6044

Regularization

Regularization is a technique used to avoid the overfitting problem. It is a process of introducing additional information in order to prevent overfitting. Lasso is a Linear Model trained with L1 prior as regularizer.

Evaluation result of the model:

Model	alpha	Number of non-zero coefficients	S^2	MSE
Lasso()	1.0	13	0.6733	0.4608
Lasso(alpha=0.1)	0.1	16	0.7632	0.3340
Lasso(alpha=0.01)	0.01	26	0.7847	0.3037
Lasso(alpha=0.001)	0.001	35	0.8127	0.2642

Predict test dataset with the model

Model	MSE with training dataset	MSE with test dataset
Lasso()	0.4554	0.4569
Lasso(alpha=0.1)	0.3248	0.3764
Lasso(alpha=0.01)	0.2971	0.3470
Lasso(alpha=0.001)	0.2594	0.2980

Random Forests

Random forests are among the most popular machine learning methods thanks to their relatively good accuracy, robustness and ease of use. They are often used for feature selection. The reason is because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node.

	features	Importance
41	median_house_value	0.244180
11	med_prev_qual_z	0.131244
2	cur_smoke	0.098932
46	cs_educ_ba	0.076523
5	puninsured2010	0.074873
47	e_rank_b	0.037186
6	reimb_penroll_adj10	0.035520
3	bmi_obese	0.035098
16	mammogram_10	0.024484
17	amb_disch_per1000_10	0.019673

Factors Affect Life Expectancy

- Result of Regularization with Lasso model
- Result of Random Forests Regressor model
- Factors affect life expectancy of people with bottom quartile income

Result of Regularization with Lasso model

No.	features	Feature Description	Coefficients
1	cs_fam_wkidsinglemom	Fraction of Children with Single Mother	-2.531824
2	cur_smoke	Fraction Current Smokers	-2.377746
3	poor_share	Poverty Rate	1.753844
4	cs_labforce	Labor Force Participation	-0.911347
5	frac_traveltime_lt15	Fraction with Commute < 15 Min	-0.586644
6	gini99	Gini Index Within Bottom 99%	0.462429
7	cs_elf_ind_man	Share Working in Manufacturing	0.405938
8	lf_d_2000_1980	Percent Change in Labor Force 1980-2000	0.275495
9	cs_race_theil_2000	Racial Segregation	0.208809
10	mort_30day_hosp_z	30-day Hospital Mortality Rate Index	-0.140977

Result of Random Forests Regressor model

No.	features	Feature Description	Importance
1	median_house_value	Median House Value	0.244180
2	med_prev_qual_z	Mean of Z-Scores for Dartmouth Atlas Ambulatory Care Measures	0.131244
3	cur_smoke	Fraction Current Smokers	0.098932
4	cs_educ_ba	Percent College Grads	0.076523
5	puninsured2010	Percent Uninsured	0.074873
6	e_rank_b	Absolute Mobility (Expected Rank at p25)	0.037186
7	reimb_penroll_adj10	Medicare $ Per Enrollee	0.035520
8	bmi_obese	Fraction Obese	0.035098
9	mammogram_10	Percent Female Aged 67-69 with Mammogram	0.024484
10	amb_disch_per1000_10	Discharges for Ambulatory Care Sensitive Conditions Among Medicare Enrollees	0.019673

Factors affect life expectancy of people with bottom quartile income

No.	features	Feature Description	Importance
1	median_house_value	Median House Value	0.140114
2	reimb_penroll_adj10	Medicare $ Per Enrollee	0.088326
3	cur_smoke_q1	Fraction Current Smokers in Q1	0.057331
4	cs_frac_black	Percent Black	0.044545
5	mammogram_10	Percent Female Aged 67-69 with Mammogram	0.033224
6	amb_disch_per1000_10	Discharges for Ambulatory Care Sensitive Conditions Among Medicare Enrollees	0.028584
7	med_prev_qual_z	Mean of Z-Scores for Dartmouth Atlas Ambulatory Care Measures	0.024302
8	adjmortmeas_pnall30day	30-day Mortality for Pneumonia	0.023432
9	frac_middleclass	Fraction Middle Class (p25-p75)	0.020439
10	lf_d_2000_1980	Percent Change in Labor Force 1980-2000	0.018807

Analysis Results

- Higher income was associated with longer life expectancy throughout the income distribution. The gap in life expectancy between the richest 1% and poorest 1% of individuals was 14.6 years for men and 10.1 years for women.
- Life expectancy of women is higher than life expectancy of men in the same income percentile. Gender gap of life expectancy decreased with higher income percentile. It's 6.0 for the poorest 1% and 1.5 for the richest 1% of individuals.
- Inequality in life expectancy increased over time. Between 2001 and 2014, life expectancy increased by 2.4 years for men and 2.7 years for women in the top 1% of the income distribution, but decreased by 0.1 years for men and women in the bottom 1%.
- Life expectancy for low-income individuals varied substantially across local areas. In the bottom income quartile, California and New York have the highest life expectancy while Indiana and Oklahoma have the lowest life expectancy. The difference is about 3~4 years.
- Geographic differences in life expectancy for individuals in the United States were significantly correlated with income inequality, health behaviors such as smoking and obese, access to medical care, education, and health status. Life expectancy for low-income individuals was correlated with Percent Black, Fraction Middle Class, and labor market conditions as well.

Future Work

- Do factors associated with life expectancy change by year?
- Collect data of 2015~2017 and do the research again.
- In areas with low life expectancy or high gap of life expectancy, improve the factors affect life expectancy most and check the result.

Thank You!