IEOR 242 - Homework 1

Predicting Life Expectancy

Helen Yifang Liu
Joel Varghese
Kevin Danser
Mihir Tamhankar
Xilin Sun

TEAM 6

Context

- **Background:** Predicting Life Expectancy in the 50 states of the US.
- **Given:** Dataset containing 7 variables of the 50 states.
- **Objective**: Perform a Linear Regression to predict the life expectancy.

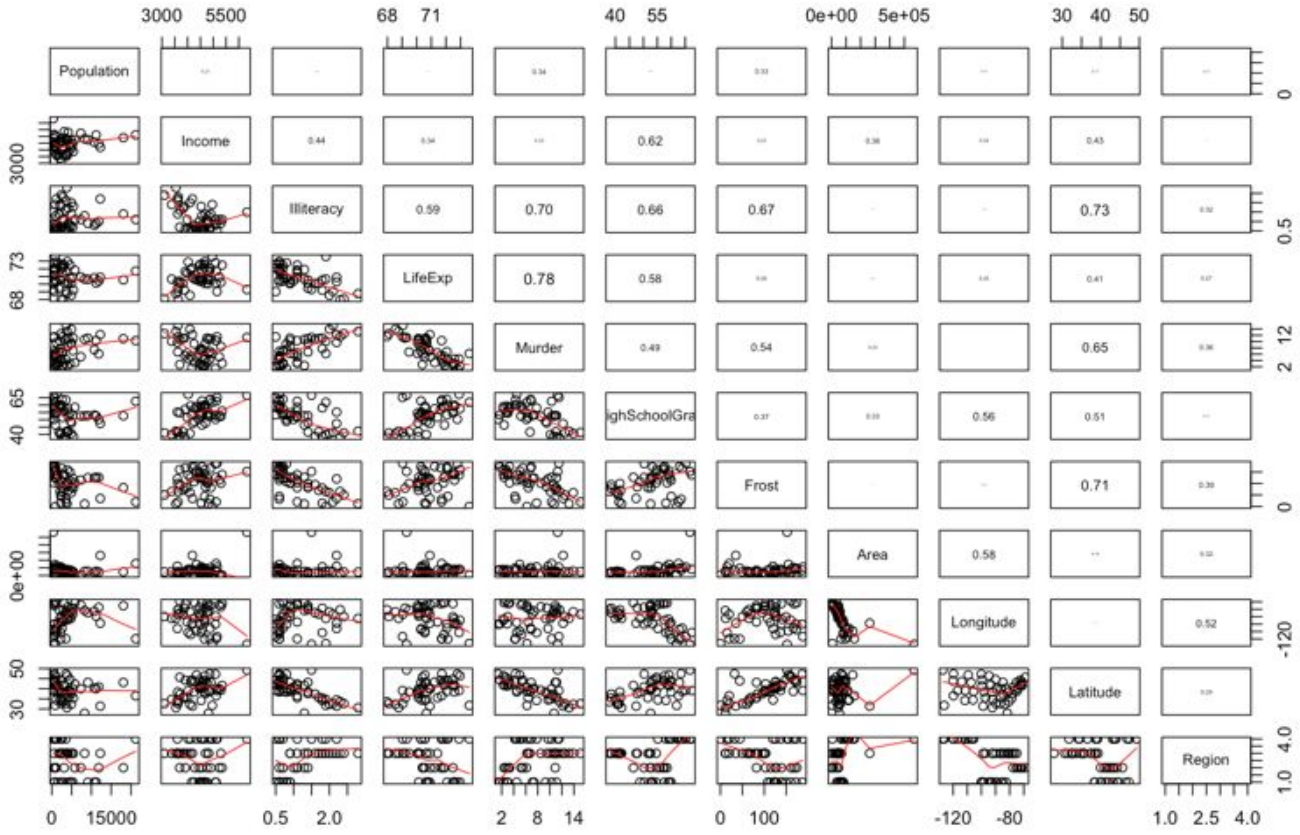
Dataset Summary - Descriptive Statistics

Population	Income	Illiteracy	LifeExp
Min. : 365	Min. :3098	Min. :0.500	Min. :67.96
1st Qu.: 1080	1st Qu.:3993	1st Qu.:0.625	1st Qu.:70.12
Median : 2838	Median :4519	Median :0.950	Median :70.67
Mean : 4246	Mean :4436	Mean :1.170	Mean :70.88
3rd Qu.: 4968	3rd Qu.:4814	3rd Qu.:1.575	3rd Qu.:71.89
Max. :21198	Max. :6315	Max. :2.800	Max. :73.60

Murder	HighSchoolGrad	Frost	Area
Min. : 1.400	Min. :37.80	Min. : 0.00	Min. : 1049
1st Qu.: 4.350	1st Qu.:48.05	1st Qu.: 66.25	1st Qu.: 36985
Median : 6.850	Median :53.25	Median :114.50	Median : 54277
Mean : 7.378	Mean :53.11	Mean :104.46	Mean : 70736
3rd Qu.:10.675	3rd Qu.:59.15	3rd Qu.:139.75	3rd Qu.: 81163
Max. :15.100	Max. :67.30	Max. :188.00	Max. :566432

Longitude	Latitude	Region
Min. :-127.25	Min. :27.87	North Central:12
1st Qu.:-104.16	1st Qu.:35.55	Northeast : 9
Median : -89.90	Median :39.62	South :16
Mean : -92.46	Mean :39.41	West :13
3rd Qu.: -78.98	3rd Qu.:43.14	
Max. : -68.98	Max. :49.25	

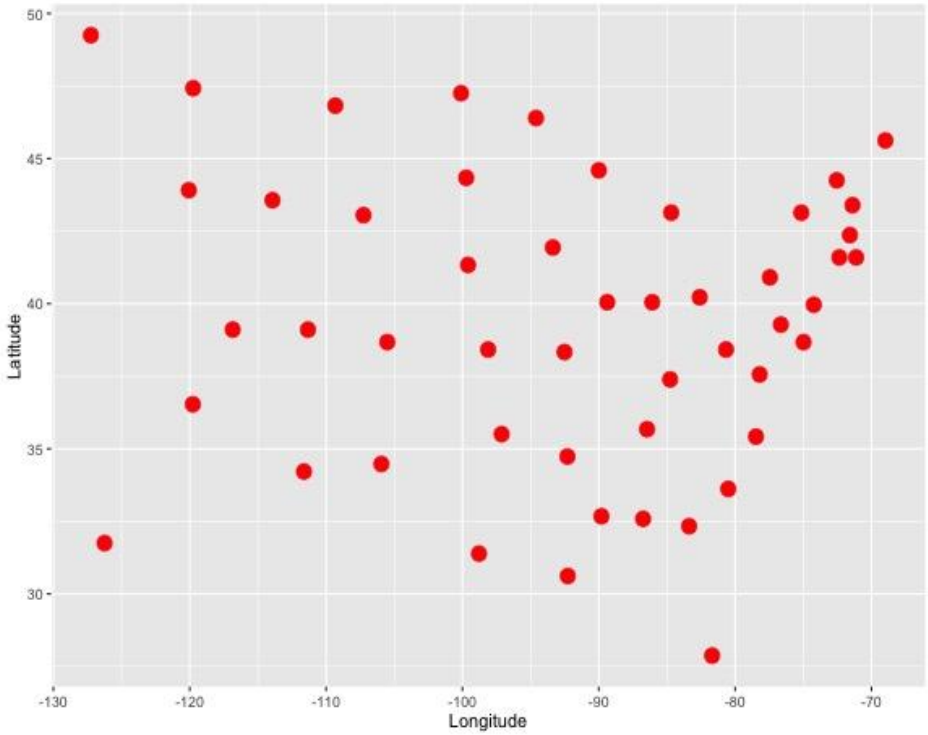
Dataset Summary - Correlation Plot



Scatterplot of the 50 States (Longitude vs Latitude)

- The plot mirrors the outline of the United States

- Plot was generated by ggplot



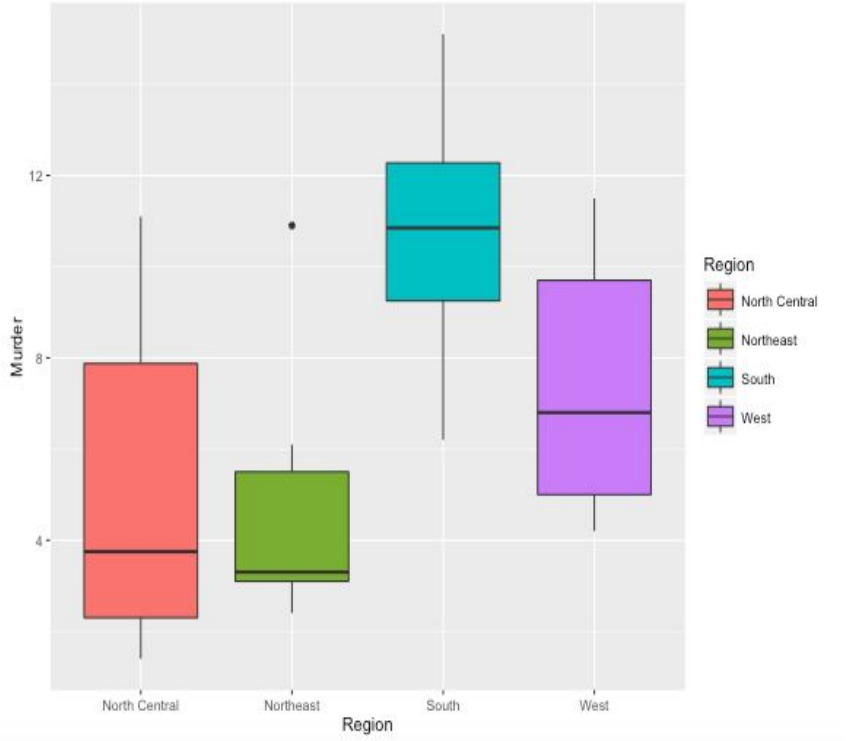
High School Graduation Rate per Region

- States are categorized into 4 regions
- Then an average High School Grad rate is calculated for each Region

	Region	mean(HighSchoolGrad)
	(fctr)	(dbl)
1	North Central	54.51667
2	Northeast	53.96667
3	South	44.34375
4	West	62.00000

Box Plot of Murder in each Region

	Minumum	Lower Quatile	Median	Upper Quartile	Maximum	Range
North Central	1.4	2.3	3.75	8.35	11.1	9.7
Northeast	2.4	3.1	3.30	5.50	6.1	3.7
South	6.2	9.0	10.85	12.35	15.1	8.9
West	4.2	5.0	6.80	9.70	11.5	7.3



Regression Model

- **LifeExp ~ Population + Income + Illiteracy + Murder + HighSchoolGrad + Frost + Area +Region**

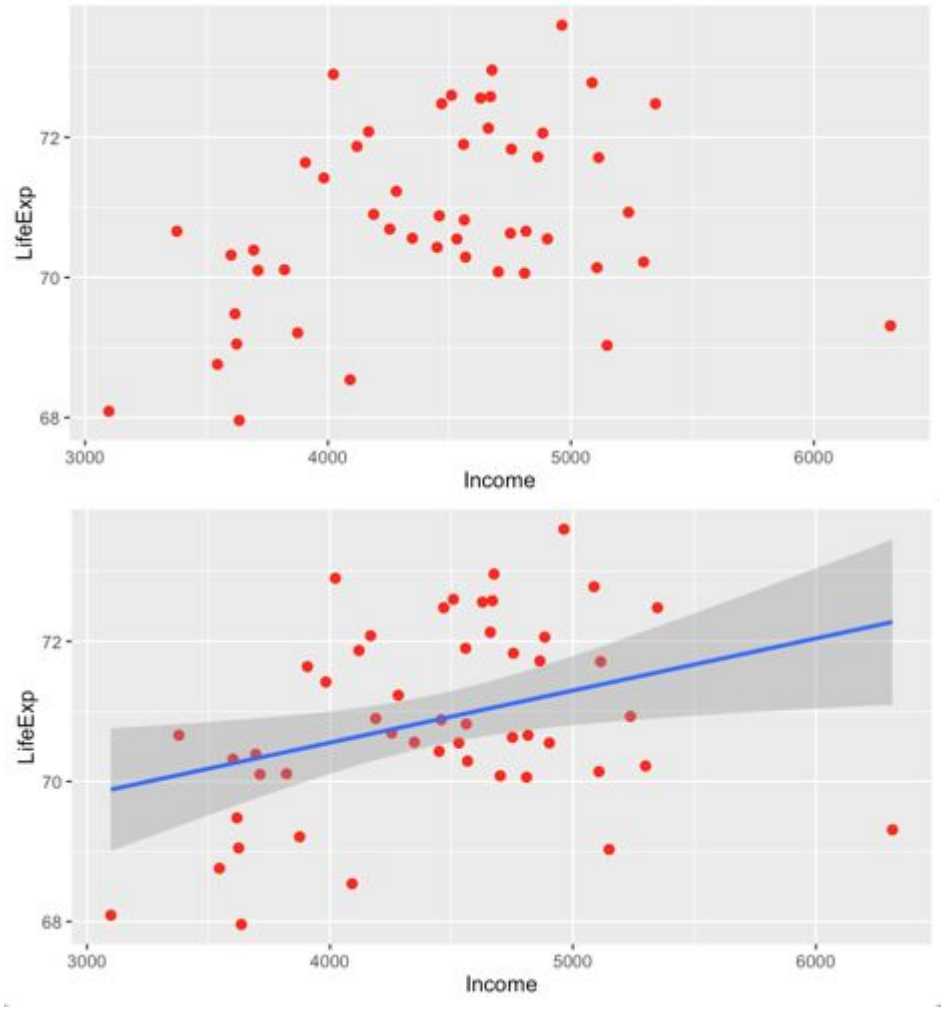
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.898e+01  3.902e+00  17.679  < 2e-16 ***
Population      5.867e-05  3.217e-05   1.824   0.0763 .
Income          1.182e-04  2.309e-04   0.512   0.6118
Illiteracy      6.112e-02  4.210e-01   0.145   0.8854
Murder         -3.036e-01  5.292e-02  -5.737 1.42e-06 ***
HighSchoolGrad  4.000e-02  3.506e-02   1.141   0.2611
Frost          -3.565e-04  3.770e-03  -0.095   0.9252
Area           -1.983e-06  2.130e-06  -0.931   0.3578
RegionNortheast 1.556e-01  5.009e-01   0.311   0.7578
RegionSouth    -2.002e-01  4.748e-01  -0.422   0.6757
RegionWest     -1.183e+00  5.309e-01  -2.228   0.0320 *
Longitude      -4.665e-02  2.004e-02  -2.328   0.0255 *
Latitude       -6.647e-02  4.179e-02  -1.591   0.1202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$-1.182*10^{-4}$ -> irrelevance to the dependent var

Residual standard error: 0.6613
Multiple R^2: 0.8618
Adjusted R^2: 0.7573
F-statistic: 13.74
p-value: $3.384*10^{-10}$

Regression Model

- LifeExp ~ Income
  - coefficient: $-1.182*10^{-4}$
- Plot:
  - Shows a positive correlation
  - Negative coefficient due to multicollinearity



Income vs High School Graduates

Model Rebuilt

- “Step” function:
  - try different combinations
  - get lowest AIC

```
step(model.1,direction="both")
#Lower the AIC, better is the model
```

- Rebuilt model:
  - LifeExp ~ Population + Murder + HighSchoolGrad + Longitude + Latitude + Region
  -

```
Coefficients:
    (Intercept)       Population           Murder    HighSchoolGrad  RegionNortheast
      7.120e+01        6.487e-05       -3.278e-01        3.702e-02        8.345e-02
    RegionSouth       RegionWest        Longitude         Latitude
     -2.650e-01       -1.045e+00       -3.839e-02       -8.500e-02
```

Comparison

LifeExp ~ Population + Income + Illiteracy + Murder + HighSchoolGrad + Frost + Area + Longitude + Latitude + Region

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.898e+01  3.902e+00  17.679  < 2e-16 ***
Population      5.867e-05  3.217e-05   1.824   0.0763 .
Income          1.182e-04  2.309e-04   0.512   0.6118
Illiteracy      6.112e-02  4.210e-01   0.145   0.8854
Murder         -3.036e-01  5.292e-02  -5.737 1.42e-06 ***
HighSchoolGrad  4.000e-02  3.506e-02   1.141   0.2611
Frost          -3.565e-04  3.770e-03  -0.095   0.9252
Area           -1.983e-06  2.130e-06  -0.931   0.3578
RegionNortheast 1.556e-01  5.009e-01   0.311   0.7578
RegionSouth    -2.002e-01  4.748e-01  -0.422   0.6757
RegionWest     -1.183e+00  5.309e-01  -2.228   0.0320 *
Longitude      -4.665e-02  2.004e-02  -2.328   0.0255 *
Latitude       -6.647e-02  4.179e-02  -1.591   0.1202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.6613
Multiple R^2: 0.8618
Adjusted R^2: 0.7573
F-statistic: 13.74
p-value: $3.384*10^{-10}$

LifeExp ~ Population + Murder + HighSchoolGrad + Longitude + Latitude + Region

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.120e+01  2.226e+00  31.984  < 2e-16 ***
Population      6.487e-05  2.516e-05   2.578   0.0136 *
Murder         -3.278e-01  4.247e-02  -7.719 1.62e-09 ***
HighSchoolGrad  3.702e-02  2.269e-02   1.632   0.1104
RegionNortheast 8.345e-02  4.092e-01   0.204   0.8394
RegionSouth    -2.650e-01  3.953e-01  -0.670   0.5064
RegionWest     -1.045e+00  4.759e-01  -2.195   0.0339 *
Longitude      -3.839e-02  1.479e-02  -2.596   0.0130 *
Latitude       -8.500e-02  2.752e-02  -3.088   0.0036 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.6375
Multiple R^2: 0.8103
Adjusted R^2: **0.7733**
F-statistic: 21.89
p-value: $1.654*10^{-12}$

Predictions

$\text{LifeExp}\|_{max}$	73.6	HI	$\text{LifeExp}\|_{min}$	67.96	SC
$\text{Prediction}\|_{max}$	72.6	NE	$\text{Prediction}\|_{min}$	69.05	AL