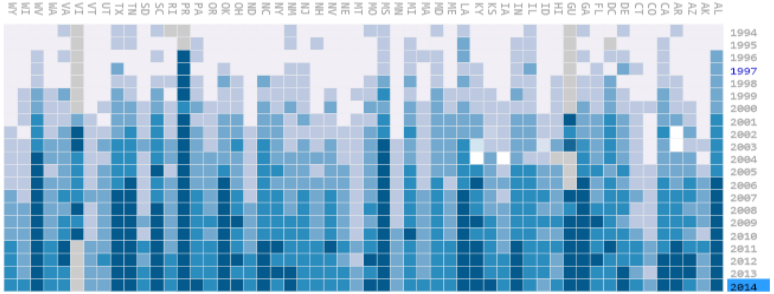
Health Data Correlation Analysis

Zeyang Gong, Stella Zheng

Background

Cancer is among the leading causes of death worldwide. In 2012, there were 14 million new cases and 8.2 million cancer-related deaths worldwide.

Results from the National Health and Nutrition Examination Survey (NHANES) showed that in 2011–2014, nearly 70% of U.S. adults age 20 years or older were overweight or obese and more than one-third (36.5%) were obese.



Project Goal

By conducting regression analysis, we want to answer the questions:

What variables in the dataset has a correlation with the cancer, obesity and diabetes rate?

Is there any regional differences? If there is, by how much?

Dataset Description

- The dataset is from Centers for Diseases Control and Presention.

- It contains data for 500 cities in the USA.

State	City	Region	NoInsurance	Arithritis	Drinking	HighBP	MedCol_Hig	Cancer	Asthma	Heart	Reg_Doc_V	Chol_Scree	Tests
AL	Birmingham	South	22.6	32.6	11.5	45.6	81.4	6.1	11.4	7.7	77.3	73.3	61.1
AL	Hoover	South	10.6	26.3	15.4	32.9	78.1	7	8.2	5.5	74.1	78.1	72.4
AL	Huntsville	South	17.4	30	12.2	37.3	79.1	6.8	9.7	7	71.2	75.5	65
AL	Mobile	South	20	33.1	12.5	44.1	81.1	6.9	10.7	8.2	73	73.2	61.6
AL	Montgomery	South	19.7	31	12.5	40.1	80.6	6.2	10.8	7.1	75	74	62.1
AL	Tuscaloosa	South	20.3	24.8	14.4	31.7	74.6	5	11.2	5.6	71.2	63.4	63.1
AK	Anchorage	West	15.5	19.3	21.6	27.8	62.1	5.3	8.7	4.2	58.9	68	57.5
AZ	Avondale	West	23.4	18.5	16.6	23.3	66.5	4.1	10	4.3	61.8	66.1	56.6
AZ	Chandler	West	13.1	19.4	17	24	69.1	5.2	9.2	4.2	64.6	71.8	64.8
AZ	Gilbert	West	10.5	18.2	18.5	22.2	67.2	5	9.1	3.7	64.2	72.3	66.5
AZ	Glendale	West	21.5	22.5	15.3	26.9	71.2	5.4	10.3	5.5	63.4	68.7	58.4
AZ	Mesa	West	18.2	24.5	15.3	28.6	74	6.5	10	6.3	65.1	71.3	62.5
AZ	Peoria	West	13.6	24.6	15.6	28.4	74.7	6.9	9.6	6	66.8	74.9	64.9
AZ	Phoenix	West	23.7	21.6	15.3	26.7	70.1	5.1	10.3	5.3	62.6	67.5	57.9
AZ	Scottsdale	West	9.2	25.6	14.8	30.9	77.1	8	8.9	6.2	69.3	77.9	70.3
AZ	Surprise	West	14	26.1	14.8	30.5	76.8	7.6	9.4	6.6	68	76.2	69.3
AZ	Tempe	West	16.9	17	17.9	22.1	66	4.4	10.1	4	61.4	60.6	62.9

Hypotheses

1. Cancer has a positive correlation with smoking and drinking.

2. Obesity has a positive relationship with lack of physical activities.

3. Diabetes has a correlation with

Variable Description

- Total of 31 variables in the whole dataset
- Reduced to 22 variables

Identifier

State, City

Explanatory Variable:

Three different groups:

Region

Health Condition: HighBP, HighChol, PoorPhyHealth, PoorMentalHealth, etc.

Other Condition: Drinking, Smoking, Regu_doc_visit, LackWorkOut, etc.

Repsonse Variable:

Cancer, Diabetes, Obesity

Regression Analytics - Cancer

1st Model: All variables

(Delete “Dental Visit”, “Lack of Workout” and “TeethLost” due to high VIF value)

```
Call:
lm(formula = Cancer ~ Region + Drinking + HighBP + MedCol_HighBP +
    Smoking + HighChol + PoorPhyHealth + SleepLess + NoInsurance +
    Reg_Doc_Visit + Tests + Preventive_M + Preventive_W + Mammo +
    PapaniTest + Chol_Screening, data = y)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00627 -0.23564 -0.00644  0.24779  1.23757

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.324373   0.970497  -7.547 2.25e-13 ***
RegionNortheast  0.202397   0.073524   2.753 0.006132 **
RegionSouth      0.029868   0.068575   0.436 0.663362
RegionWest       0.547419   0.083839   6.529 1.68e-10 ***
Drinking         0.058063   0.010044   5.781 1.34e-08 ***
HighBP           0.019961   0.009166   2.178 0.029903 *
MedCol_HighBP    0.074857   0.009480   7.896 1.97e-14 ***
Smoking          0.005458   0.009738   0.560 0.575408
HighChol         0.090447   0.009513   9.508  < 2e-16 ***
PoorPhyHealth    0.147831   0.019844   7.450 4.37e-13 ***
SleepLess       -0.068786   0.008085  -8.508 2.26e-16 ***
NoInsurance     -0.023332   0.006149  -3.794 0.000167 ***
Reg_Doc_Visit   -0.008727   0.008421  -1.036 0.300615
Tests            0.096454   0.007843  12.298  < 2e-16 ***
Preventive_M    -0.001130   0.007719  -0.146 0.883700
Preventive_W     0.004177   0.007796   0.536 0.592374
Mammo           -0.039706   0.008833  -4.495 8.73e-06 ***
PapaniTest      -0.049624   0.009130  -5.435 8.71e-08 ***
Chol_Screening   0.067608   0.007096   9.527  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3667 on 481 degrees of freedom
Multiple R-squared:  0.8762,    Adjusted R-squared:  0.8715
F-statistic: 189.1 on 18 and 481 DF,  p-value: < 2.2e-16
```

Regression Analytics - Cancer

2nd Model:

Delete insignificant variables

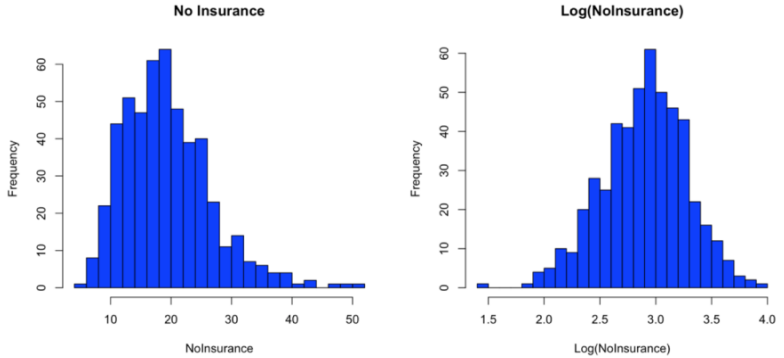
```
Call:
lm(formula = Cancer ~ Region + Drinking + HighBP + MedCol_HighBP +
    HighChol + PoorPhyHealth + SleepLess + NoInsurance + Tests +
    Mammo + PapaniTest + Chol_Screening, data = y)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01915 -0.24132 -0.00834  0.24618  1.24226

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.822573   0.899138  -7.588 1.68e-13 ***
RegionNortheast 0.184000   0.071757   2.564   0.0106 *
RegionSouth     0.024713   0.064107   0.385   0.7000
RegionWest      0.550099   0.075925   7.245 1.71e-12 ***
Drinking        0.059233   0.008971   6.602 1.06e-10 ***
HighBP          0.022285   0.008780   2.538   0.0115 *
MedCol_HighBP   0.072159   0.008901   8.107 4.27e-15 ***
HighChol        0.088826   0.009273   9.579  < 2e-16 ***
PoorPhyHealth   0.149767   0.016388   9.139  < 2e-16 ***
SleepLess      -0.071735   0.007058 -10.164  < 2e-16 ***
NoInsurance    -0.025003   0.005763  -4.339 1.74e-05 ***
Tests           0.096351   0.007251  13.289  < 2e-16 ***
Mammo          -0.044196   0.006902  -6.404 3.59e-10 ***
PapaniTest     -0.050493   0.008657  -5.833 9.97e-09 ***
Chol_Screening  0.064776   0.006381  10.151  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3661 on 485 degrees of freedom
Multiple R-squared:  0.8756,    Adjusted R-squared:  0.872
F-statistic: 243.8 on 14 and 485 DF,  p-value: < 2.2e-16
```

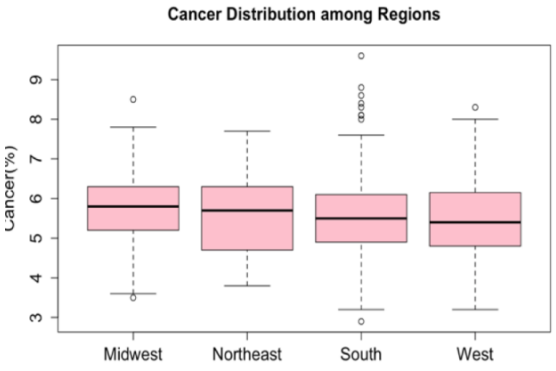
Regression Analytics - Cancer

Final Model:

-Delete not linear variables

-Log transform “NoInsurance”






```
One-way analysis of means

data:  y$Cancer and y$Region
F = 1.9318, num df = 3, denom df = 496, p-value = 0.1235
```

H0: all means are equal
H1: at least one mean is different
p-value = 0.12>0.05
Do not have enough evidence to reject H0.

Regression Analytics - Cancer

Final Model:

-Delete not linear variables

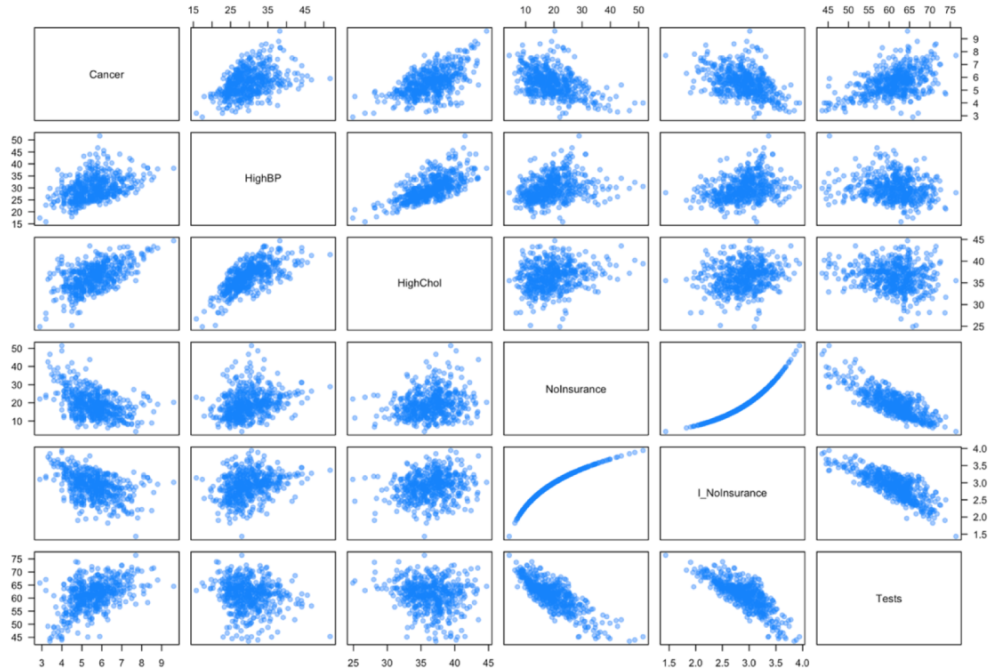
-Log transform “NoInsurance”

-Exclude “Region”

```
Call:
lm(formula = Cancer ~ HighBP + HighChol + log(NoInsurance) +
    Tests, data = y)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19108 -0.40183 -0.03559  0.35175  2.04816

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.329059   0.805935  -5.371 1.20e-07 ***
HighBP            0.034047   0.007122   4.780 2.31e-06 ***
HighChol          0.195939   0.011347  17.267  < 2e-16 ***
log(NoInsurance) -0.818300   0.113708  -7.196 2.30e-12 ***
Tests             0.067964   0.007326   9.277  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5539 on 495 degrees of freedom
Multiple R-squared:  0.7092,    Adjusted R-squared:  0.7069
F-statistic: 301.8 on 4 and 495 DF,  p-value: < 2.2e-16
```

Regression Analytics - Cancer

Check Assumptions - 1. Linearity ✔

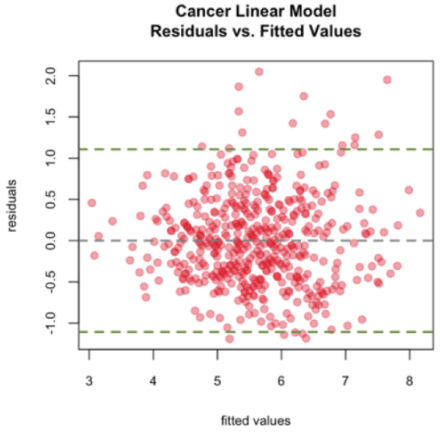


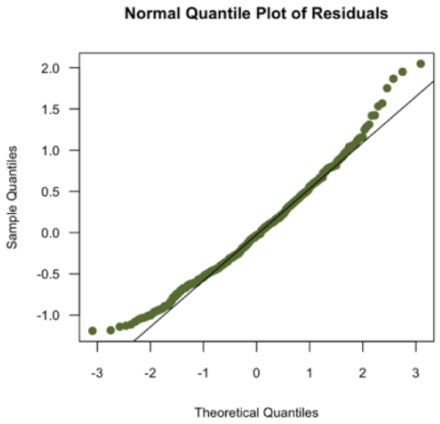
Regression Analytics - Cancer

Check Assumptions

2. Multicollinearity ✔

	Variables	VIF
1	HighBP	1.960838
2	HighChol	1.881112
3	NoInsurance	3.056786
4	Tests	2.924986





3. Constant Variance ✔ :
   no curve, no trend, white noise
4. Normality: right-skew
5. Outliers: No, influence.measures()
6. Independence: not time series data

Regression Analytics - Cancer

Interpretation:

Cancer = -5.37+ 0.03*HighBP + 0.2*HighChol-0.81 log(NoInsurance) + 0.06 Tests

1. High Blood Pressure, High Cholesterol and Medical Tests are positively correlated with Cancer, No Insurance is negatively correlated with Cancer.
2. The more people who has high blood pressure and high cholesterol in a place, the higher prevalence of cancer will be in the place.
3. If there are a lot of people who has no insurance in a place, the prevalence of cancer will also increase in the place.

Regression Analytics - Obesity

1st Model: All Variables

```
Call:
lm(formula = Obesity ~ Region + Drinking + HighBP + MedCol_HighBP +
    Smoking + HighChol + PoorPhyHealth + SleepLess + NoInsurance +
    Reg_Doc_Visit + Tests + Preventive_M + Preventive_W + Mammo +
    PapaniTest + Chol_Screening, data = y)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1666 -1.1880 -0.0245  1.1098  6.9654

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.000681   4.529560   2.649  0.00833 **
RegionNortheast -1.440787   0.343153  -4.199 3.20e-05 ***
RegionSouth     -2.458535   0.320059  -7.682 8.88e-14 ***
RegionWest      -4.070533   0.391300 -10.403  < 2e-16 ***
Drinking         0.017402   0.046880   0.371  0.71066
HighBP           0.485844   0.042779  11.357  < 2e-16 ***
MedCol_HighBP    0.040153   0.044248   0.907  0.36462
Smoking          0.101510   0.045452   2.233  0.02598 *
HighChol        -0.111373   0.044400  -2.508  0.01246 *
PoorPhyHealth    0.070427   0.092617   0.760  0.44738
SleepLess        0.187721   0.037733   4.975 9.10e-07 ***
NoInsurance      0.155027   0.028701   5.401 1.04e-07 ***
Reg_Doc_Visit   -0.057437   0.039305  -1.461  0.14458
Tests            0.083170   0.036607   2.272  0.02353 *
Preventive_M    -0.006457   0.036026  -0.179  0.85783
Preventive_W     0.056129   0.036387   1.543  0.12360
Mammo           -0.054510   0.041228  -1.322  0.18674
PapaniTest       0.270547   0.042612   6.349 5.02e-10 ***
Chol_Screening  -0.446069   0.033120 -13.468  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.711 on 481 degrees of freedom
Multiple R-squared:  0.863,     Adjusted R-squared:  0.8579
F-statistic: 168.4 on 18 and 481 DF,  p-value: < 2.2e-16
```

Regression Analytics - Obesity

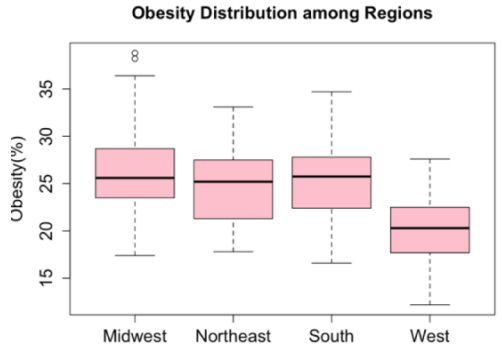
2nd Model: Delete insignificant variables

```
Call:
lm(formula = Obesity ~ Region + HighBP + Smoking + HighChol +
    SleepLess + NoInsurance + PapaniTest + Chol_Screening, data = y)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2842 -1.2621 -0.0577  1.1083  6.9496

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.03059    2.78958   7.181 2.60e-12 ***
RegionNortheast -1.41777    0.32502  -4.362 1.57e-05 ***
RegionSouth     -2.20027    0.26711  -8.237 1.62e-15 ***
RegionWest      -4.05809    0.25369 -15.996  < 2e-16 ***
HighBP           0.51588    0.03467  14.881  < 2e-16 ***
Smoking          0.12546    0.03345   3.751 0.000197 ***
HighChol        -0.10141    0.04164  -2.435 0.015229 *
SleepLess        0.10611    0.02762   3.843 0.000138 ***
NoInsurance      0.10472    0.01745   6.001 3.82e-09 ***
PapaniTest       0.24190    0.03480   6.951 1.16e-11 ***
Chol_Screening  -0.45968    0.02964 -15.510  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.725 on 489 degrees of freedom
Multiple R-squared:  0.8585,    Adjusted R-squared:  0.8556
F-statistic: 296.6 on 10 and 489 DF,  p-value: < 2.2e-16
```

Regression Analytics - Obesity

Final Model: Delete not linear variables

Is “Region” significant?

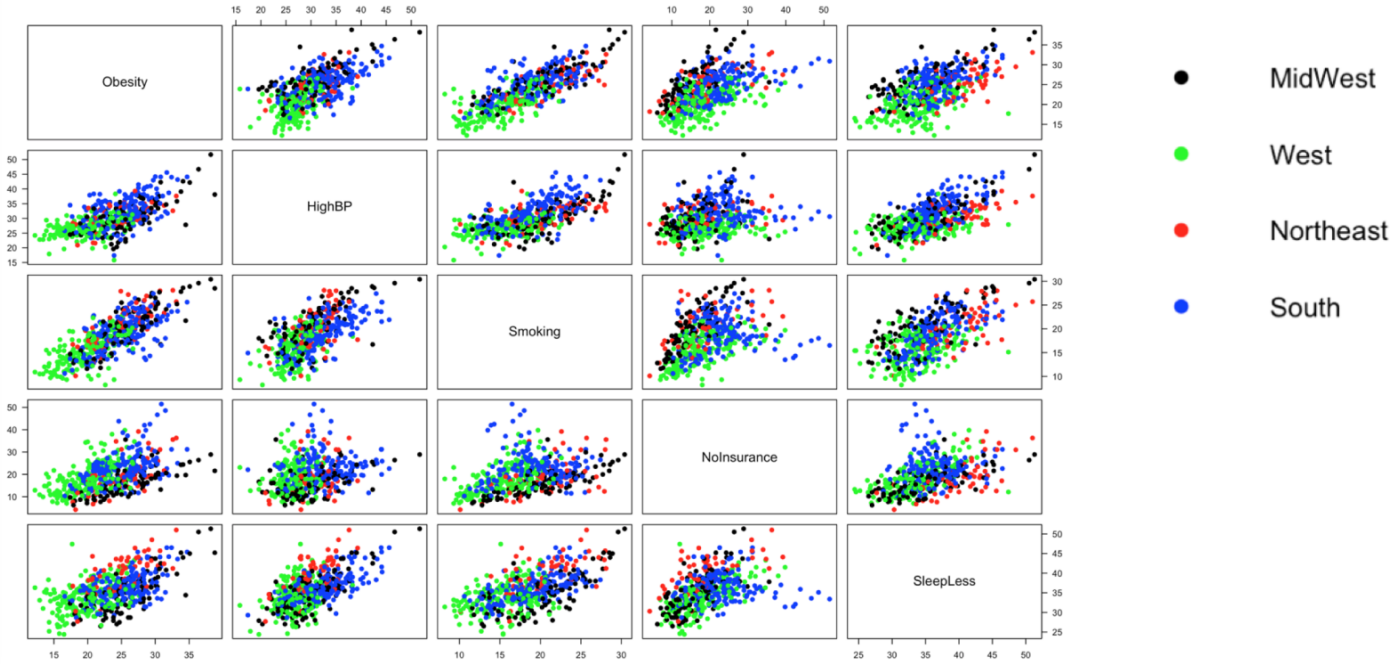



```
        One-way analysis of means

data:  y$Obesity and y$Region
F = 89.303, num df = 3, denom df = 496, p-value < 2.2e-16
```

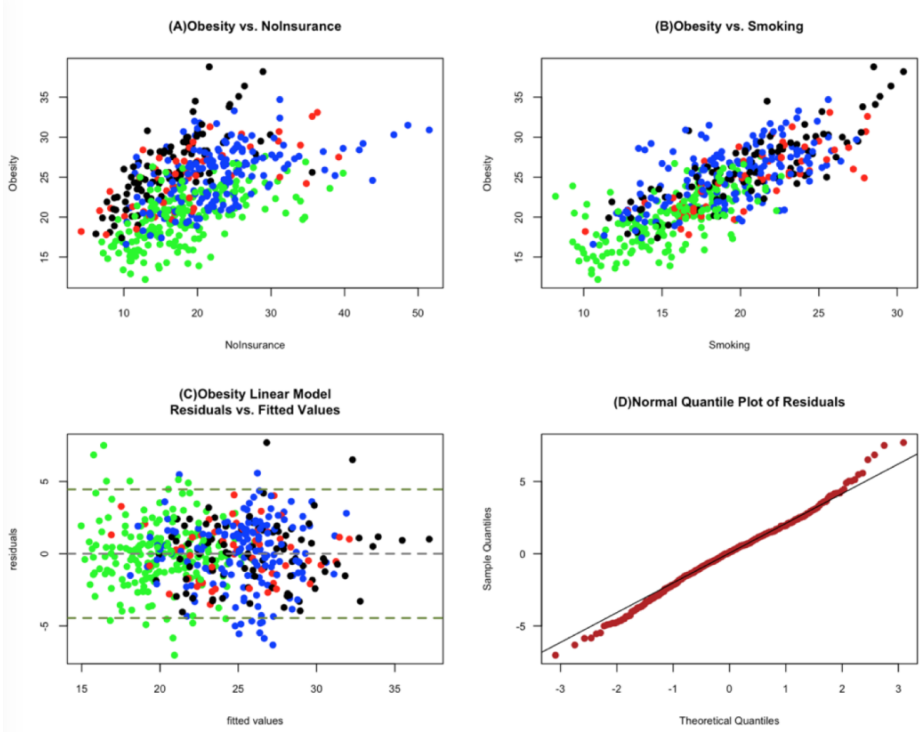
```
Call:
lm(formula = Obesity ~ Region + HighBP + Smoking + SleepLess +
    NoInsurance, data = y)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0253 -1.3476  0.0993  1.4281  7.6842

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.59617    0.93178   8.152 2.98e-15 ***
RegionNortheast -2.16699    0.41218  -5.257 2.18e-07 ***
RegionSouth     -2.15892    0.33120  -6.518 1.76e-10 ***
RegionWest      -3.90140    0.32508 -12.001  < 2e-16 ***
HighBP           0.17584    0.03153   5.577 4.04e-08 ***
Smoking          0.39472    0.03742  10.548  < 2e-16 ***
SleepLess        0.05783    0.03438   1.682   0.0932 .
NoInsurance      0.19169    0.01717  11.162  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.225 on 492 degrees of freedom
Multiple R-squared:  0.7632,    Adjusted R-squared:  0.7598
F-statistic: 226.5 on 7 and 492 DF,  p-value: < 2.2e-16
```

Regression Analytics - Obesity

Check Assumptions - 1. Linearity ✔



Regression Analytics - Obesity



Check Assumptions

2. Multicollinearity ✔

	Variables	VIF
1	HighBP	2.017949
2	Smoking	2.031903
3	NoInsurance	1.273966
4	SleepLess	1.945178

3. Constant Variance ✔ :
   no curve, no trend, white noise
4. Normality:
   S-shape, heavy tail
5. Outliers: No, influence.measures()
6. Independence: not time series data

Regression Analytics - Obesity

Interpretations

Obesity(Midwest) = 7.59 + 0.17*HighBP+0.39*Smoking + 0.05*Sleepless + 0.19*NoInsurance
Obesity(Northeast) = 5.43+ 0.17*HighBP+0.39*Smoking + 0.05*Sleepless + 0.19*NoInsurance
Obesity(South)= 5.44+ 0.17*HighBP+0.39*Smoking + 0.05*Sleepless + 0.19*NoInsurance
Obesity(West) = 3.69+ 0.17*HighBP+0.39*Smoking + 0.05*Sleepless + 0.19*NoInsurance

1. The intercept is representing the Region. West has the lowest predicted Obesity value, and Midwest has the highest.
2. High Blood Pressure is positively correlated with Obesity.
3. Sleepless has little positive correlation with Obesity.
4. Again, people who is lack of insurance will have higher chance to have obesity issue.

Regression Analysis - Diabetes

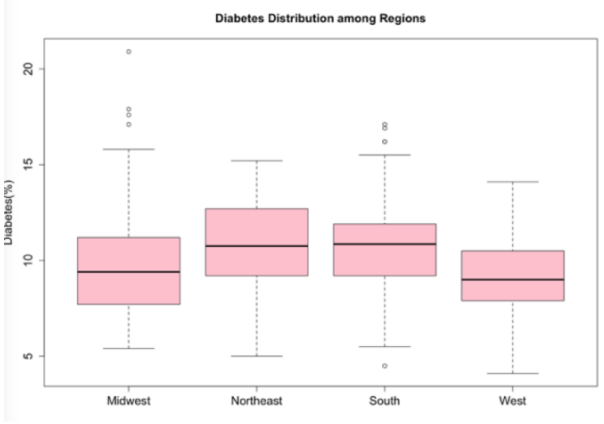
1st Model: All Variables

```
Call:
lm(formula = Diabetes ~ Region + Drinking + HighBP + MedCol_HighBP +
    Smoking + HighChol + PoorPhyHealth + SleepLess + NoInsurance +
    Reg_Doc_Visit + Tests + Preventive_M + Preventive_W + Mammo +
    PapaniTest + Chol_Screening, data = y)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4747 -0.3621 -0.0241  0.3141  1.8496

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.515931   1.410479  -5.329 1.52e-07 ***
RegionNortheast -0.637921   0.106856  -5.970 4.61e-09 ***
RegionSouth     -0.577476   0.099664  -5.794 1.24e-08 ***
RegionWest      -0.279505   0.121849  -2.294  0.02223 *
Drinking        -0.038078   0.014598  -2.608  0.00938 **
HighBP           0.151447   0.013321  11.369  < 2e-16 ***
MedCol_HighBP    0.122625   0.013779   8.900  < 2e-16 ***
Smoking         -0.102430   0.014153  -7.237 1.82e-12 ***
HighChol        -0.033991   0.013826  -2.458  0.01430 *
PoorPhyHealth    0.323692   0.028840  11.224  < 2e-16 ***
SleepLess        0.058055   0.011750   4.941 1.08e-06 ***
NoInsurance      0.075975   0.008937   8.501 2.38e-16 ***
Reg_Doc_Visit    0.021675   0.012239   1.771  0.07720 .
Tests           -0.067014   0.011399  -5.879 7.73e-09 ***
Preventive_M    -0.031234   0.011218  -2.784  0.00558 **
Preventive_W    -0.024226   0.011331  -2.138  0.03302 *
Mammo            0.059651   0.012838   4.646 4.37e-06 ***
PapaniTest      -0.011668   0.013269  -0.879  0.37968
Chol_Screening   0.018738   0.010313   1.817  0.06986 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5329 on 481 degrees of freedom
Multiple R-squared:  0.952,     Adjusted R-squared:  0.9502
F-statistic: 530.5 on 18 and 481 DF,  p-value: < 2.2e-16
```

Regression Analysis - Diabetes

2nd Model: Delete insignificant variables

Is “Region” significant?

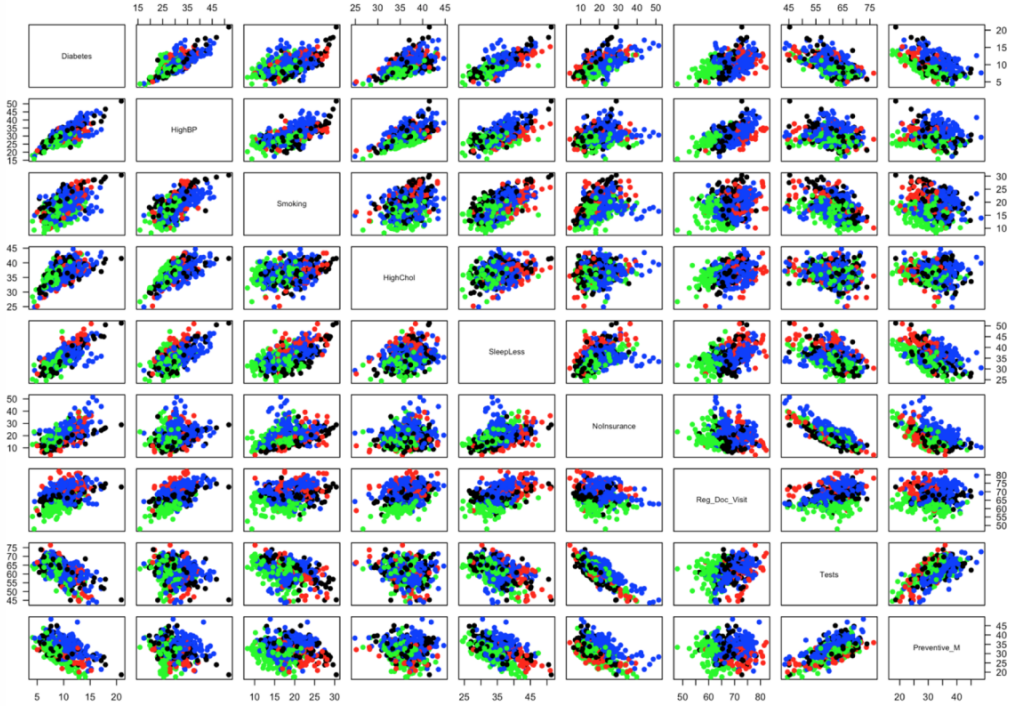



```
        One-way analysis of means

data:  y$Diabetes and y$Region
F = 18.67, num df = 3, denom df = 496, p-value = 1.729e-11
```

```
Call:
lm(formula = Diabetes ~ Region + HighBP + Smoking + HighChol +
    SleepLess + NoInsurance + Reg_Doc_Visit + Tests + Preventive_M,
    data = y)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17142 -0.37270 -0.04408  0.40288  2.24177

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.03893    1.13175  -1.802 0.072230 .
RegionNortheast -0.49343    0.12705  -3.884 0.000117 ***
RegionSouth     -0.82711    0.11696  -7.071 5.34e-12 ***
RegionWest      -0.06177    0.11044  -0.559 0.576177
HighBP           0.25879    0.01337  19.360  < 2e-16 ***
Smoking         -0.04785    0.01296  -3.693 0.000247 ***
HighChol         0.04602    0.01496   3.076 0.002218 **
SleepLess        0.06065    0.01293   4.691 3.53e-06 ***
NoInsurance      0.11105    0.00930  11.941  < 2e-16 ***
Reg_Doc_Visit    0.09009    0.01110   8.118 3.89e-15 ***
Tests           -0.08053    0.01270  -6.339 5.26e-10 ***
Preventive_M    -0.05352    0.01086  -4.926 1.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6625 on 488 degrees of freedom
Multiple R-squared:  0.9248,    Adjusted R-squared:  0.9231
F-statistic: 545.6 on 11 and 488 DF,  p-value: < 2.2e-16
```

Regression Analysis - Diabetes

Check Assumptions - 1. Linearity ✔



MidWest

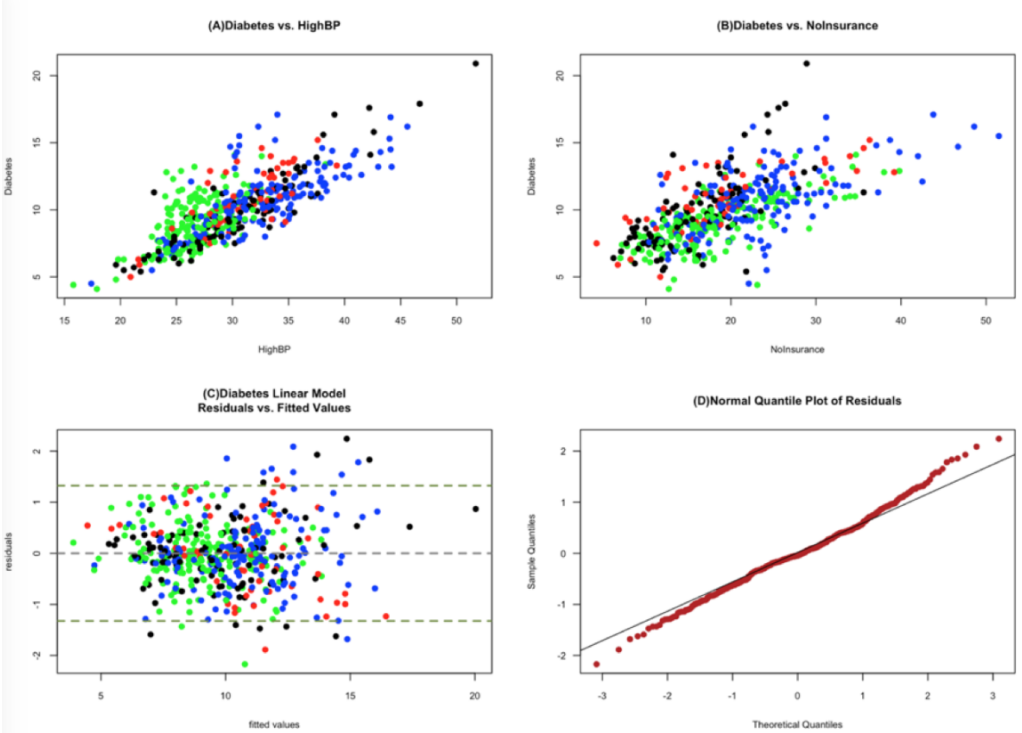
West

Northeast

South

Regression Analysis - Diabetes

Check Assumptions
2. Multicollinearity ✔



	Variables	VIF
1	HighBP	4.606761
2	Smoking	2.936591
3	HighChol	2.137107
4	SleepLess	3.601196
5	NoInsurance	3.183254
6	Reg_Doc_Visit	2.546921
7	Tests	5.576193
8	Preventive_M	2.621399

3. Constant Variance ✔ :
   no curve, no trend, white noise
4. Normality:
   S-shape, heavy tail
5. Outliers: No, influence.measures()
6. Independence: not time series data

Regression Analysis - Diabetes

Interpretations

Diabetes (Midwest) = -2.04+ 0.26*HighBP-0.04*Smoking + 0.06*Sleepless +0.04*HighChol+ 0.11*NoInsurance+0.09Reg_Doc_Visit-0.08*Tests-0.05Preventive_M

Diabetes (Northeast) = -2.53+ 0.26*HighBP-0.04*Smoking + 0.06*Sleepless +0.04*HighChol+ 0.11*NoInsurance+0.09Reg_Doc_Visit-0.08*Tests-0.05Preventive_M

Diabetes (South) = -2.86+ 0.26*HighBP-0.04*Smoking + 0.06*Sleepless +0.04*HighChol+ 0.11*NoInsurance+0.09Reg_Doc_Visit-0.08*Tests-0.05Preventive_M

Diabetes (West) = -2.1+ 0.26*HighBP-0.04*Smoking + 0.06*Sleepless +0.04*HighChol+ 0.11*NoInsurance+0.09Reg_Doc_Visit-0.08*Tests-0.05Preventive_M

1. The intercept is representing the Region. South has the lowest predicted Diabetes value, and Midwest has the highest.
2. High Blood Pressure, Sleepless, High Cholesterol, No Insurance, Regular Doctor Visit have positive correlations with Diabetes.
3. Tests, Preventive Test have negative correlations with Diabetes.

Conclusion

1. Cancer crude rate in a region is correlated with high blood pressure, high cholesterol, the percentage of people without health insurance, and medical tests.

2. Obesity crude rate is correlated with high blood pressure, frequently sleep than seven hours, smoking, and number of people without insurance.

3. Diabetes crude rate is also correlated with regular doctor visit and preventive tests for men.

4. High blood pressure, insurance access rate is correlated with the crude rate of all three diseases.

5. Midwest region has the highest intercept for obesity and diabetes.

Reference

[1]https://chronicdata.cdc.gov/500-Cities/500-Cities-City-level-Data-GIS-Friendly-Format-/dxpw-cm5u?category=500-Cities&view_name=500-Cities-City-level-Data-GIS-Friendly-Format-

[2]https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States

[3]https://www.cancer.gov/about-cancer/causes-prevention/risk/obesity/obesity-fact-sheet