COVID-19 Relevant Social Determinants and Risk Factors at the County Level

Yue Chen

5/5/2020

Overview

As of May 5, 2020, New York State still has the highest number of confirmed cases than any other states in the U.S. There have been 1,029,000 tests, of which there were 321,200 confirmed cases, and 19,645 people have died. We know that the spread of coronavirus is more likely when people are in close contact withon another. Older people who have severe underlying medical conditions like lung disease or diabetes seem to be at higher risk for developing more serious symptoms. One may wonder if the fatality rate at county level is also related to these risk factors. This notebook thus analyzes and maps some of the county level social determinants and risk factors like percentage smokers, percentage adults with diabetes, median household income etc. to see if they are related to the fatality rate.

Preparation: .

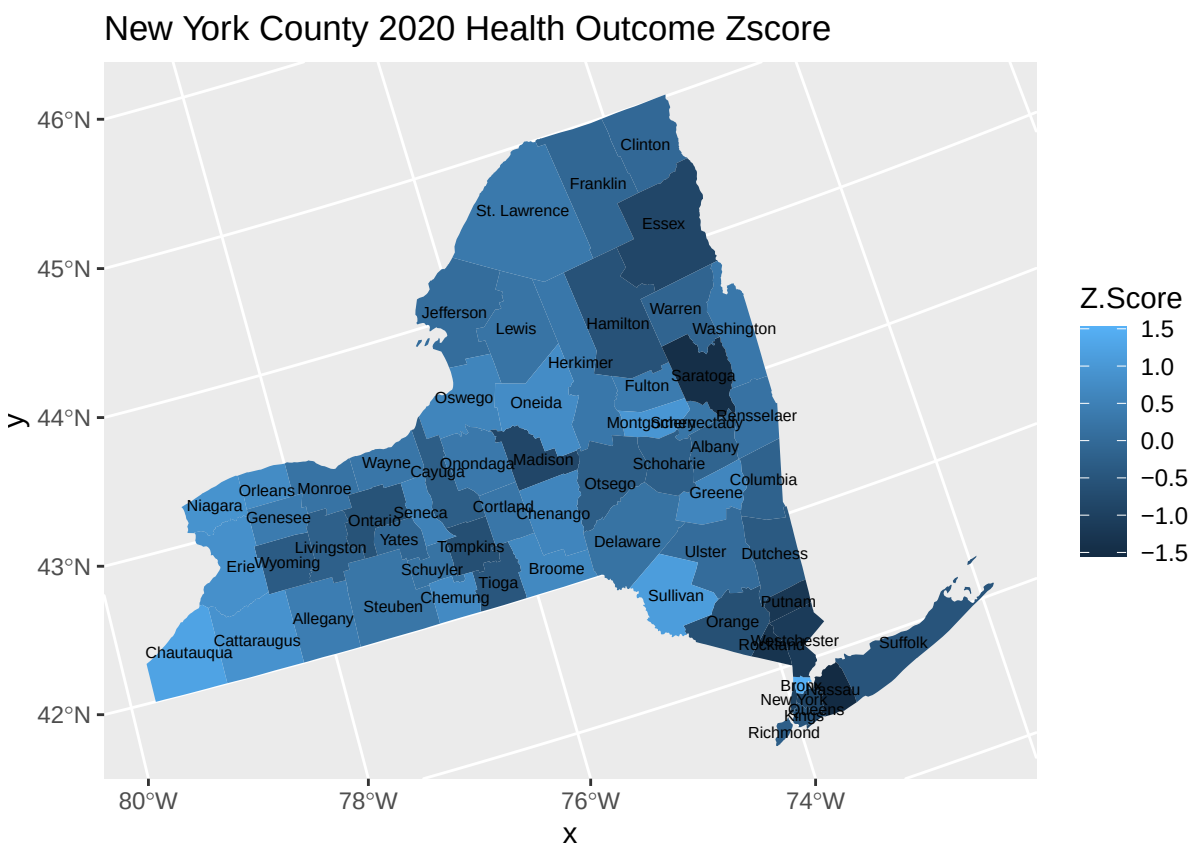
Table of Content

Mapping NY County Health Outcomes & Factors Rankings

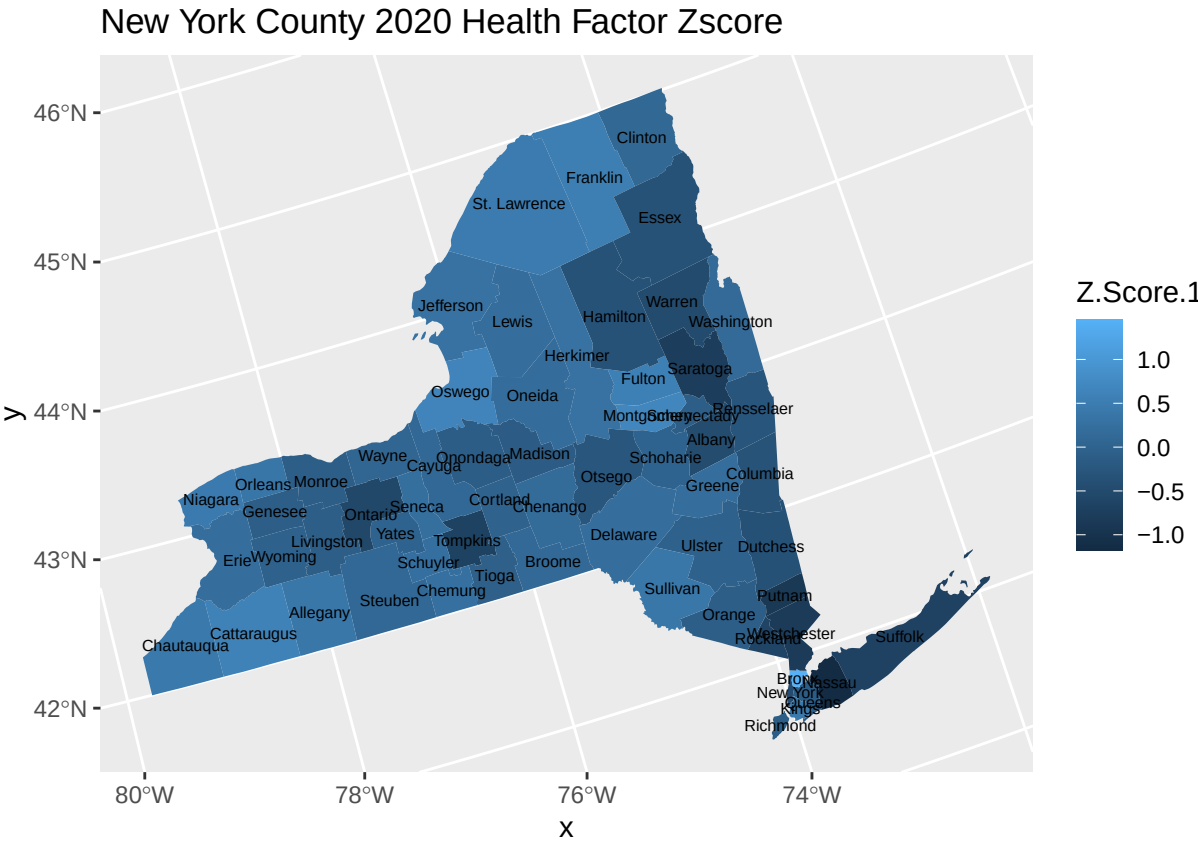
The 2020 county Health Rankings State Reports is a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute. This section maps the health outcomes and health factors z-score of the NY counties. The health outcomes are calculated based on length of life, and quality of life data collected from year 2012-2018. The health factors are calculated based upon health behaviors, clinical care, social & economic factors, physical environment, and demographics data from year 2010-2018. Data is downloaded from https://www.countyhealthrankings.org/.

```
# Read in the NY county factor data
NYfactor <- read.csv('NY2020County.csv')
# Convert fips column from numeric type to character
NYfactor <- transform(NYfactor, county_fips = as.character(county_fips))
# Combine geometry features to work with map
counties_sf <- get_urbn_map(map = "counties", sf = TRUE)
counties_sf <- counties_sf[counties_sf$state_abbv=="NY",]
healthFactor <- left_join(counties_sf, NYfactor, by = "county_fips")
# Map NY Health outcome Zscore
```

```
healthFactor %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Z.Score),
                color = NA, size = 0.05)+
    geom_sf_text(data = healthFactor,
                aes(label = County), size = 2)+
  ggtitle("New York County 2020 Health Outcome Zscore")
```




```
# Map NY Health factor Zscore
healthFactor %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Z.Score.1),
                color = NA, size = 0.05)+
    geom_sf_text(data = healthFactor,
                aes(label = County), size = 2)+
  ggtitle("New York County 2020 Health Factor Zscore")
```



Darker color indicates negative Z-score and higher rank. One can conclude that counties like Nassau, Saratoga, and NYC have better health outcomes and factors than counties like Bronx, Chautauqua, or Sullivan.

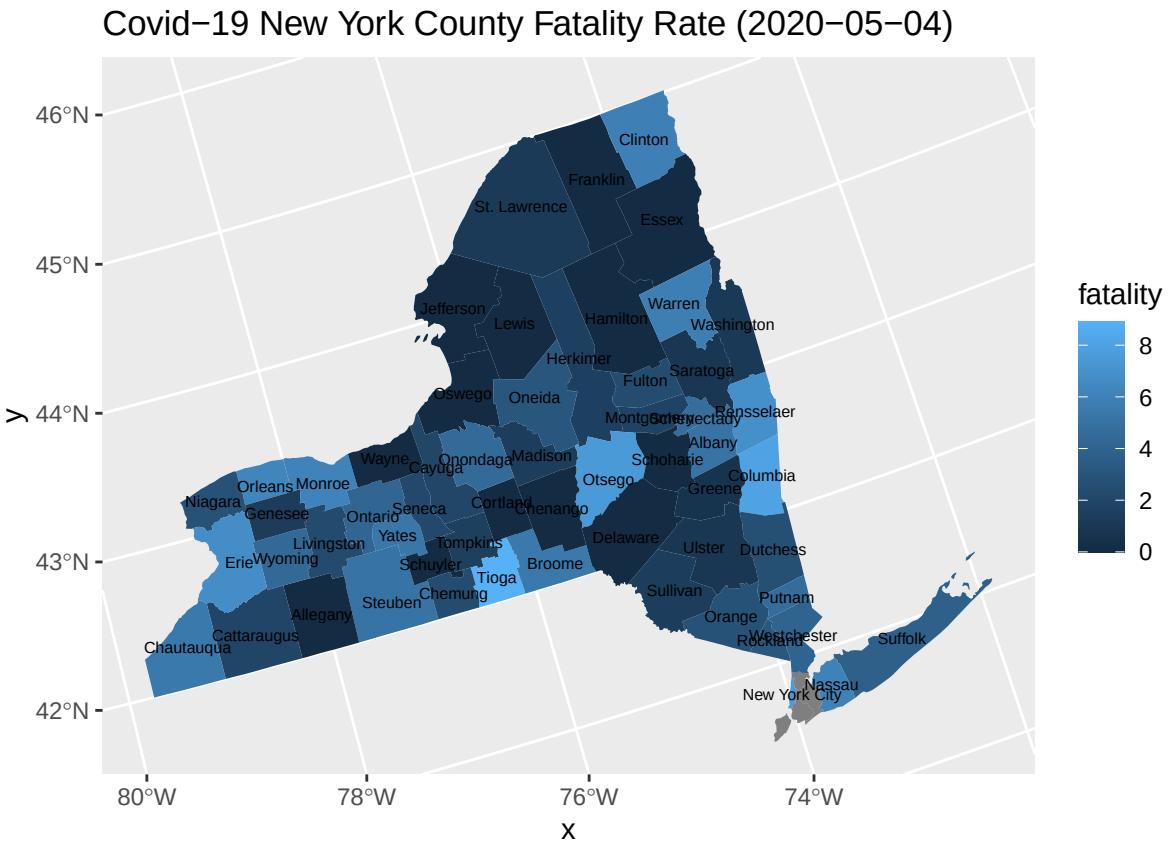
Mapping NY Covid-19 Fatality Rate

This section calculates and maps the fatality rate of each county based on cases and deaths of 2020-05-04. Data is downloaded from https://github.com/nytimes/covid-19-data

```
# Read in COVID-19 county case and death data
county <- read.csv('us-counties.csv')
# Work with NY COVID-19 county data
NYcounty <- county[county$state=="New York",]
# Fill fips for New York City entries
NYcounty[is.na(NYcounty)] <- 36061
colnames(NYcounty)[4] <- "county_fips"
# Calculate fatality rate as to May 4th.
NYcounty <- NYcounty[NYcounty$date=="2020-05-04",]
NYcounty <- transform(NYcounty, fatality=deaths*100/cases)
# Combine geometry features to work with map
NYcounty <- transform(NYcounty, county_fips = as.character(county_fips))
NYcovid <- left_join(counties_sf, NYcounty, by = "county_fips")
# Map NY Covid-19 Fatality Rate
NYcovid %>%
  ggplot() +
    geom_sf(mapping = aes(fill = fatality),
                color = NA, size = 0.05)+
    geom_sf_text(data = NYcovid,
                aes(label = county), size = 2)+
```

```
  ggtitle("Covid-19 New York County Fatality Rate (2020-05-04)")
```

```
## Warning: Removed 4 rows containing missing values (geom_text).
```



Light color indicates higher fatality rate. Counties like Tioga, Columbia, and NYC have higher fatality rate compare to counties like Allegany, Chenango, and Cortland. One can not spot direct correlations between NY counties' fatality rate and overall health outcomes and factors from the graphs.

Building Linear Regression Model for Fatality Rate

We want to test if health outcomes and factors data can explain changes in fatality rate. The dependent varables are continuous, and not binary varables. Thus we try building linear regression model.

```
# Create dataset of fatality rate, health outcome and health factor
data <- as.data.frame(NYcounty[c("county","fatality")])
data1 <- as.data.frame(NYfactor[c("Z.Score","Z.Score.1","County")])
colnames(data1) <- cbind("outcome","factor","county")
data <- merge(x=data,y=data1,by="county")
# Build Linear Regression Model
fit <- lm(fatality ~ outcome + factor, data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = fatality ~ outcome + factor, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

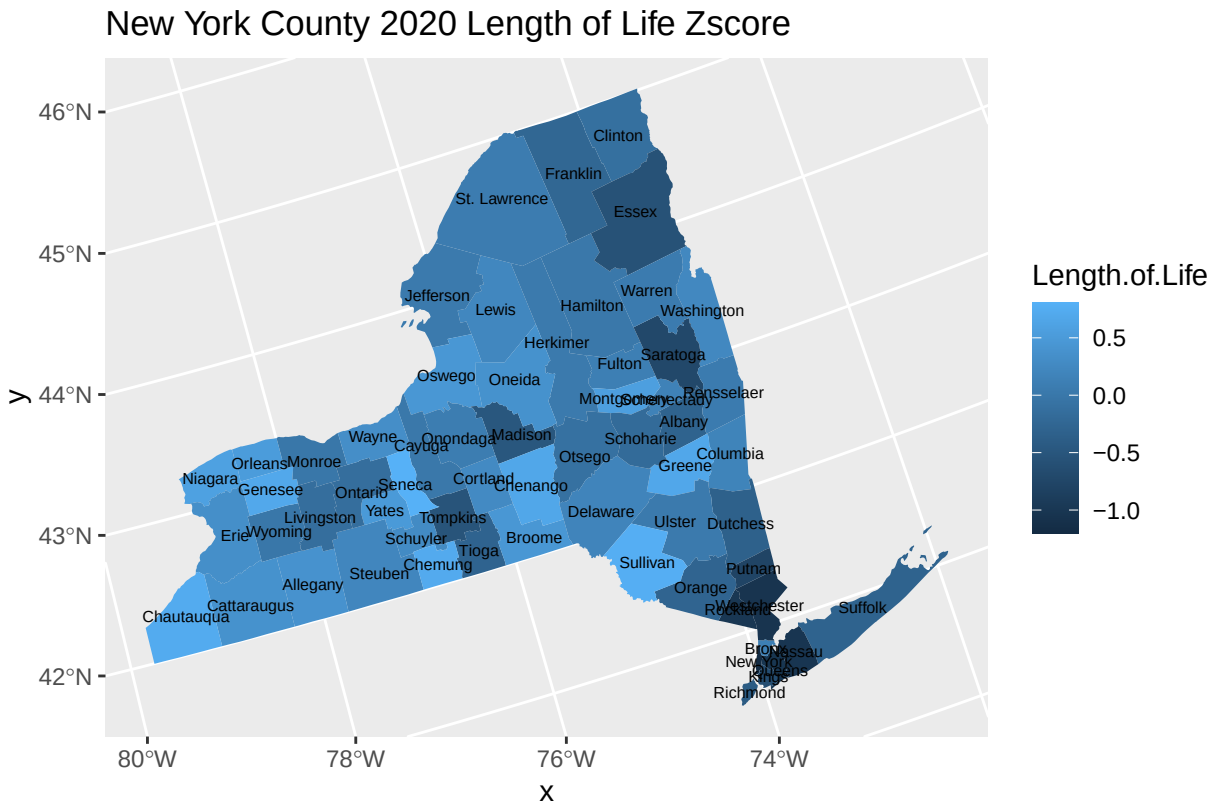
```
## -3.3603 -1.9800 -0.0714  1.0399  6.9260
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7465     0.3029   9.067 1.95e-12 ***
## outcome       2.2418     0.8597   2.607 0.011769 *
## factor       -4.8750     1.3542  -3.600 0.000692 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.263 on 54 degrees of freedom
## Multiple R-squared:  0.2009, Adjusted R-squared:  0.1713
## F-statistic: 6.788 on 2 and 54 DF,  p-value: 0.002345
```

R-squared is 0.2 which is below 0.6. One can conclude that NY county health outcome and factor did poorly on explaining fatality rate in the linear regression model.

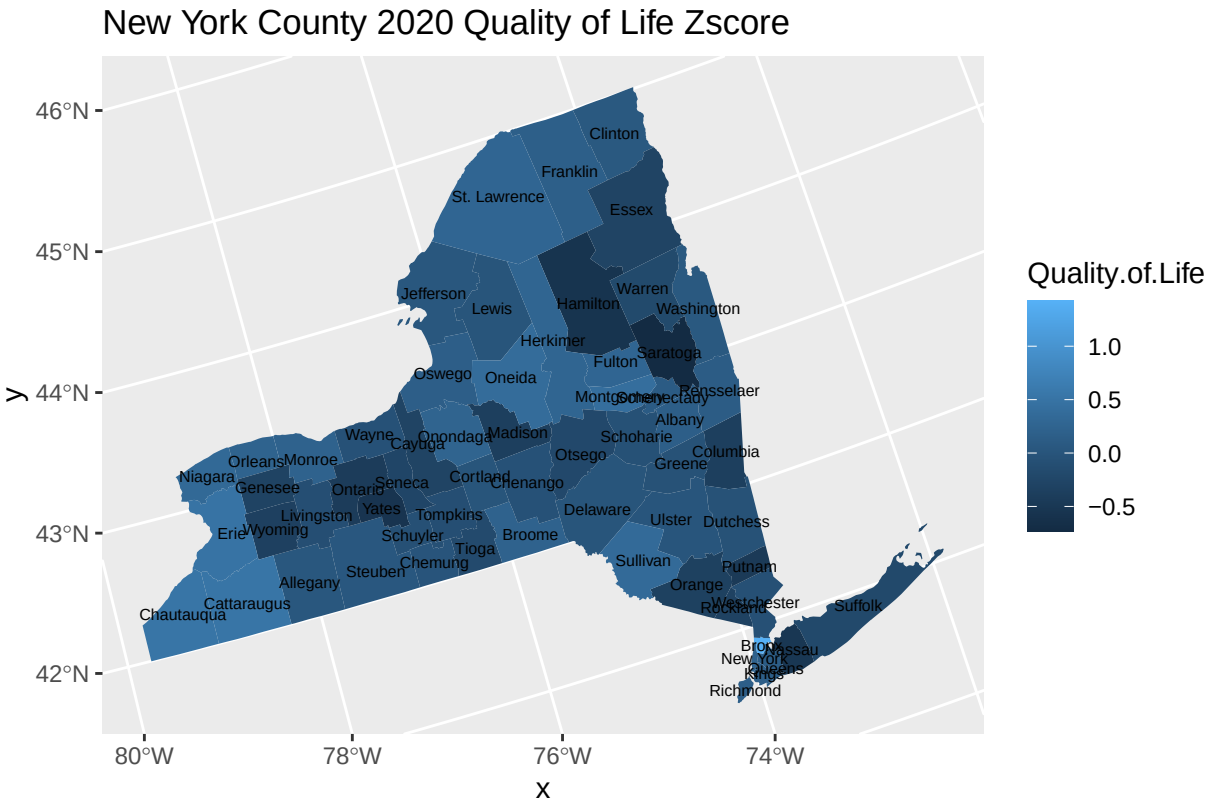
Mapping NY County Health Outcomes & Factors Subrankings

This section will map health outcomes and factors subrankings to see if there's any can serve as indicators for fatality rate. Subrankings include: length of life, quality of life, health behaviors, and clinical care, social & economic factors, and physical environment. Data is downloaded from https://www.countyhealthrankings.org/.

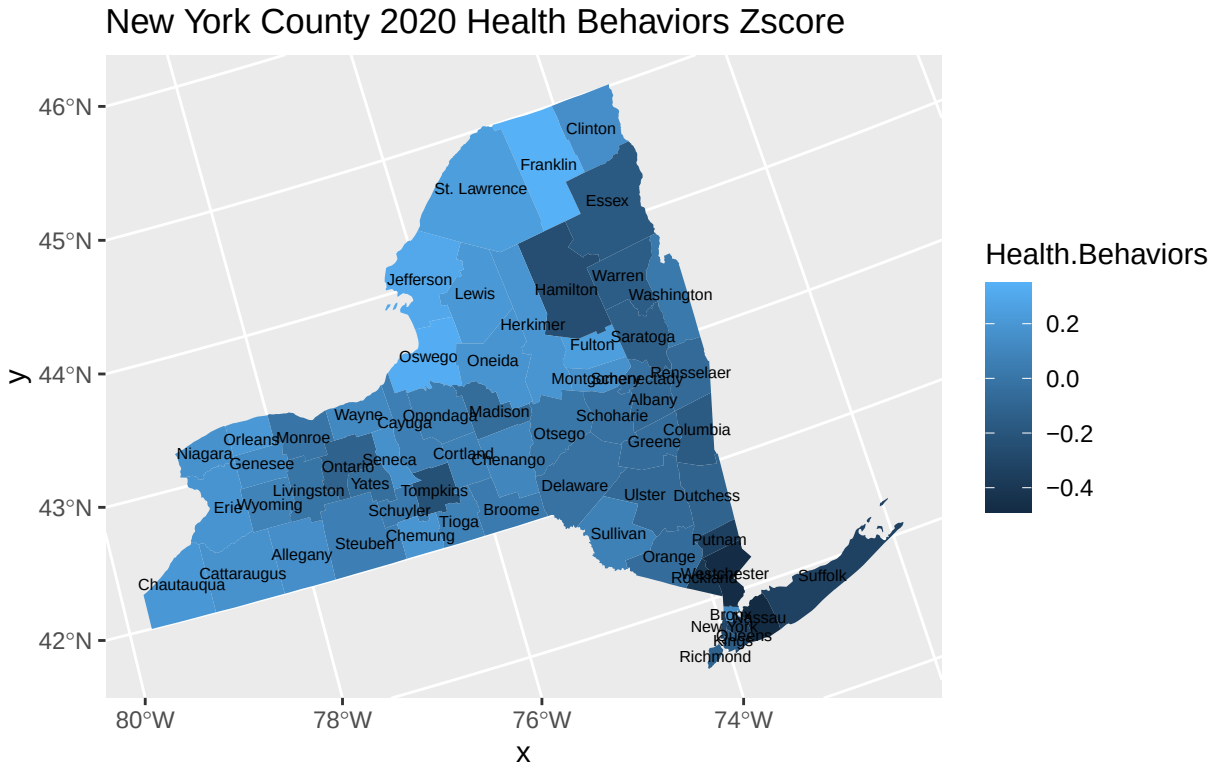
```
# Read in the NY county subrankings data
NYsubrank <- read.csv('NY2020CountySubrankings.csv')
# Convert fips column from numeric type to character
NYsubrank <- transform(NYsubrank, county_fips = as.character(county_fips))
# Combine geometry features to work with map
subrank <- left_join(counties_sf, NYsubrank, by = "county_fips")
# Map NY county length of life
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Length.of.Life),
                color = NA, size = 0.05)+
    geom_sf_text(data = subrank,
                 aes(label = County), size = 2)+
  ggtitle("New York County 2020 Length of Life Zscore")
```



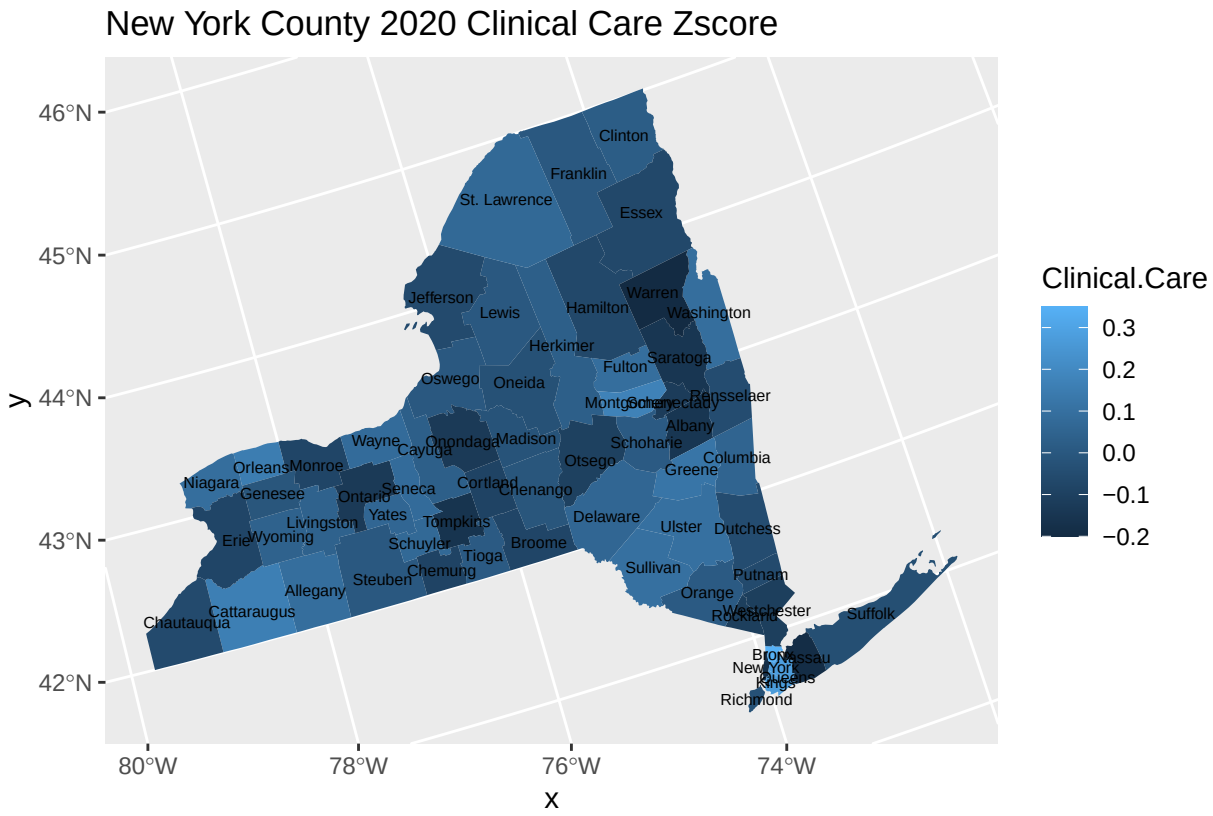

```
# Map NY county quality of life
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Quality.of.Life),
                  color = NA, size = 0.05)+
    geom_sf_text(data = subrank,
                 aes(label = County), size = 2)+
  ggtitle("New York County 2020 Quality of Life Zscore")
```



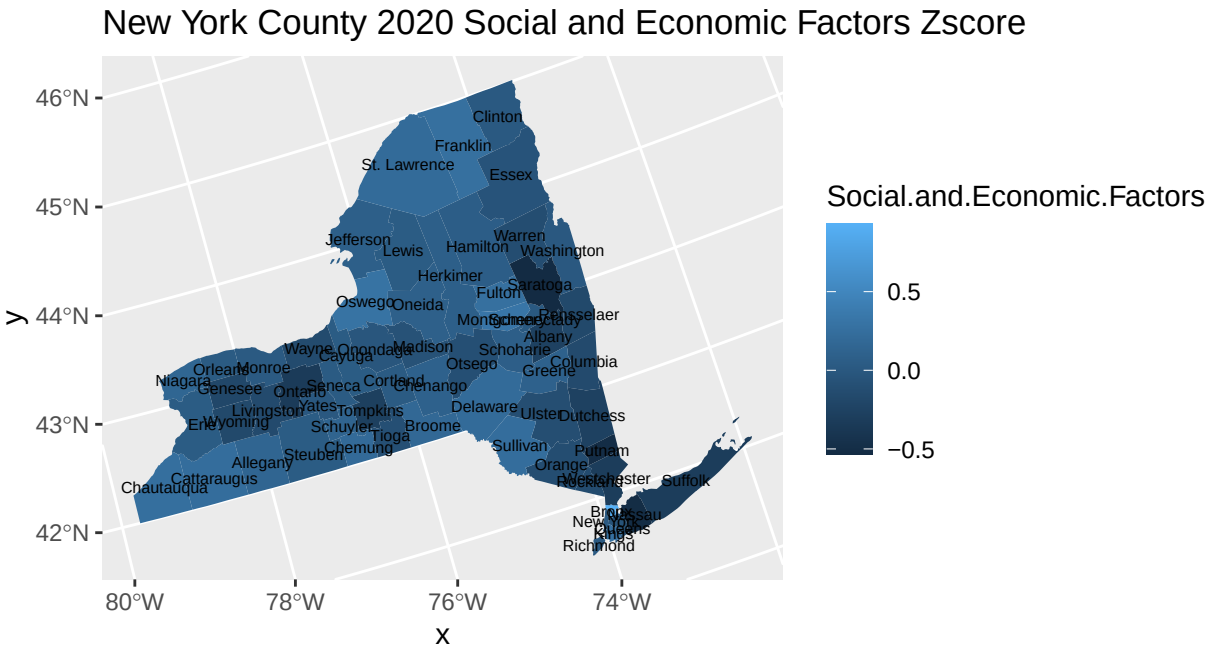

```r
# Map NY county Health Behaviors
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill =Health.Behaviors),
                  color = NA, size = 0.05)+
    geom_sf_text(data = subrank,
                 aes(label = County), size = 2)+
  ggtitle("New York County 2020 Health Behaviors Zscore")
```



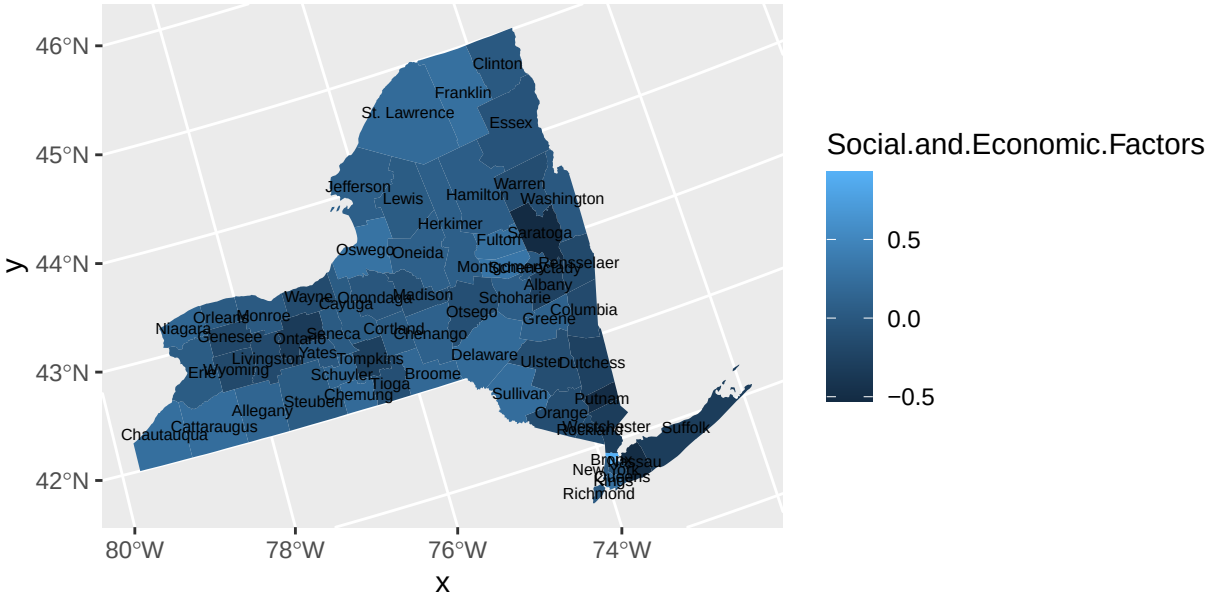

```
# Map NY county Clinical Care
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Clinical.Care ),
                  color = NA, size = 0.05)+
    geom_sf_text(data = subrank,
                 aes(label = County), size = 2)+
  ggtitle("New York County 2020 Clinical Care Zscore")
```




```
# Map NY county Social and Economic Factors
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Social.and.Economic.Factors ),
                color = NA, size = 0.05)+
    geom_sf_text(data = subrank,
                 aes(label = County), size = 2)+
  ggtitle("New York County 2020 Social and Economic Factors Zscore")
```

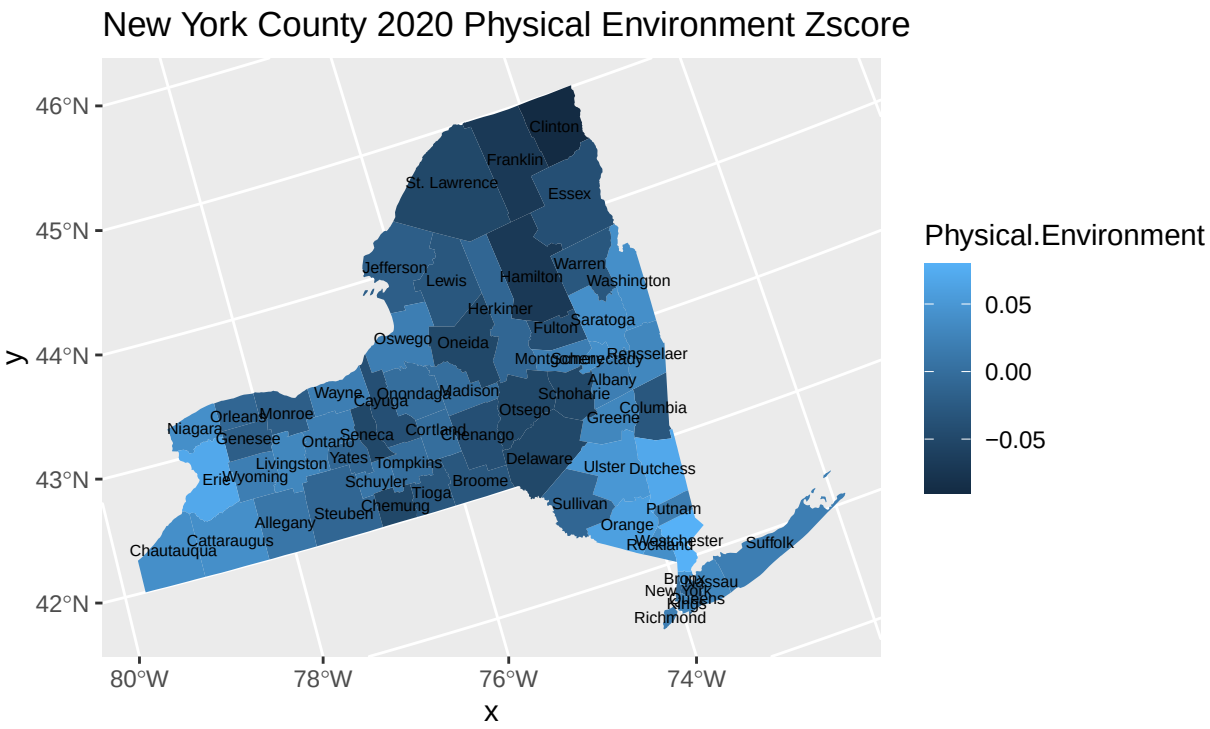



```
# Map NY county Social and Economic Factors
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Social.and.Economic.Factors ),
                 color = NA, size = 0.05)+
    geom_sf_text(data = subrank,
                 aes(label = County), size = 2)+
  ggtitle("New York County 2020 Social and Economic Factors Zscore")
```




```
# Map NY county Physical Environment
subrank %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Physical.Environment ),
```

```
                color = NA, size = 0.05)+
  geom_sf_text(data = subrank,
               aes(label = County), size = 2)+
  ggtitle("New York County 2020 Physical Environment Zscore")
```

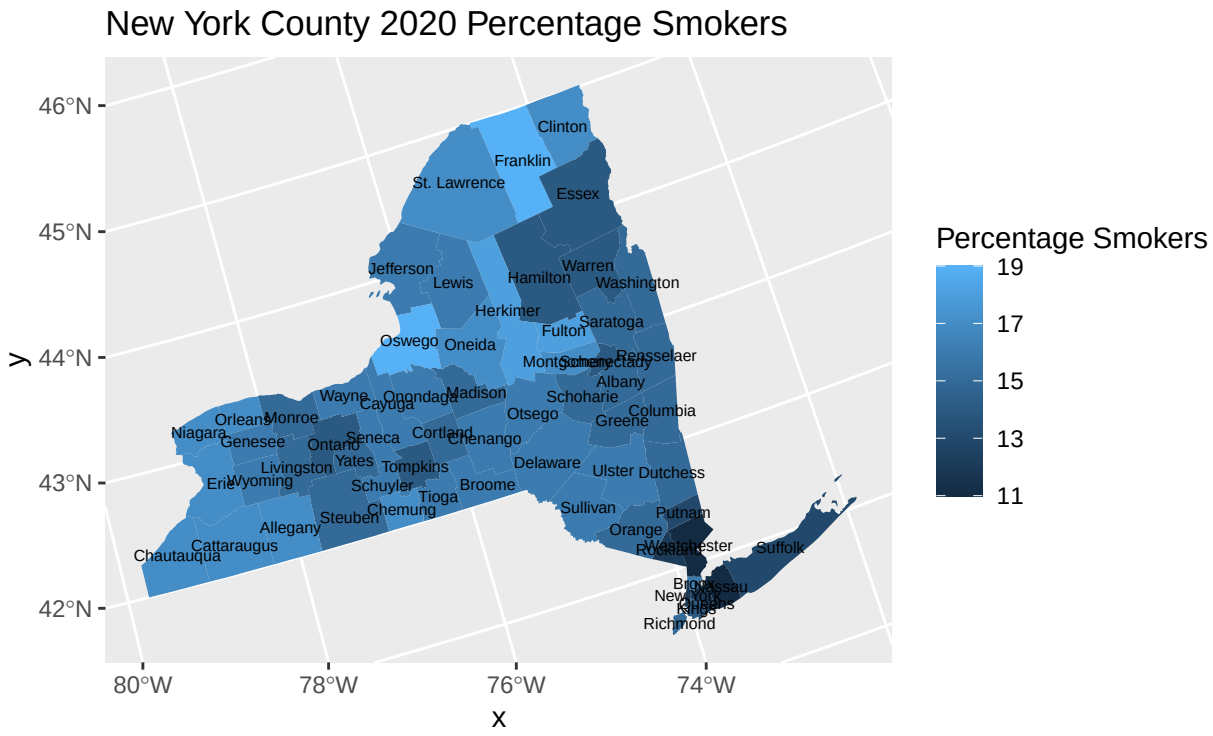


One can not spot direct correlations between NY counties' fatality rate and subrankings of health outcomes and factors from the graphs.

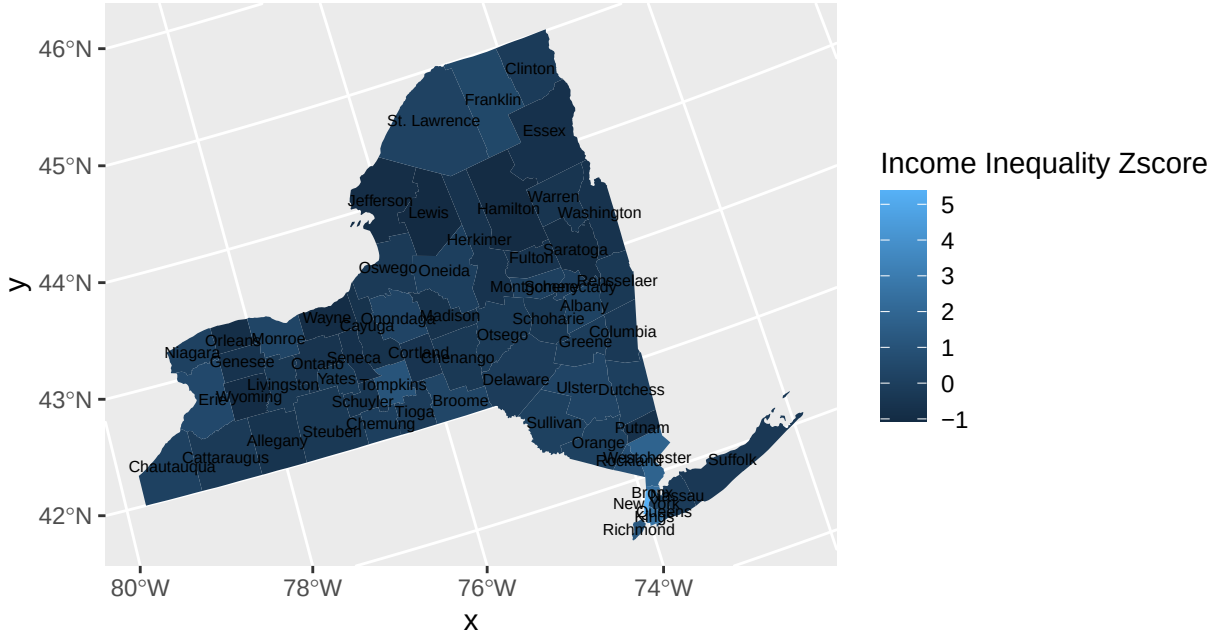
Mapping NY County Measures

This section maps some measures that may be related to Covid-19 Fatality rate including percentage smokers, income inequality, and social association.

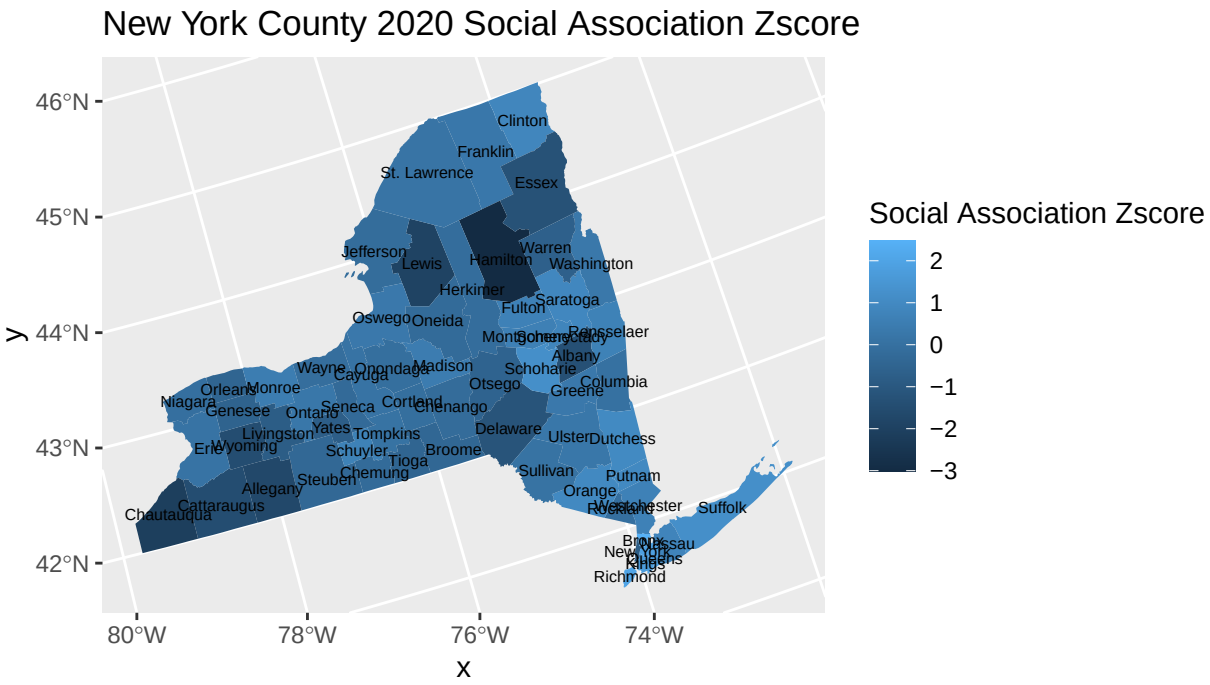
```
# Read in the NY county measure data
NYmeasure <- read.csv('NY2020CountyMeasure.csv')
# Convert fips column from numeric type to character
NYmeasure <- transform(NYmeasure, county_fips = as.character(county_fips))
# Combine geometry features to work with map
measure <- left_join(counties_sf, NYmeasure, by = "county_fips")
# Map NY county percentage smokers
measure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = X..Smokers),
                  color = NA, size = 0.05)+
    geom_sf_text(data = measure,
                 aes(label = County), size = 2)+
    labs(fill = "Percentage Smokers") +
  ggtitle("New York County 2020 Percentage Smokers")
```




```r
# Map NY county Income Inequality
measure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Z.Score.25),
                color = NA, size = 0.05)+
    geom_sf_text(data = measure,
                aes(label = County), size = 2)+
    labs(fill = "Income Inequality Zscore") +
  ggtitle("New York County 2020 Income Inequality Zscore")
```



```
# Map NY county Social Association
measure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Z.Score.27),
                 color = NA, size = 0.05)+
    geom_sf_text(data = measure,
                aes(label = County), size = 2)+
    labs(fill = "Social Association Zscore") +
  ggtitle("New York County 2020 Social Association Zscore")
```

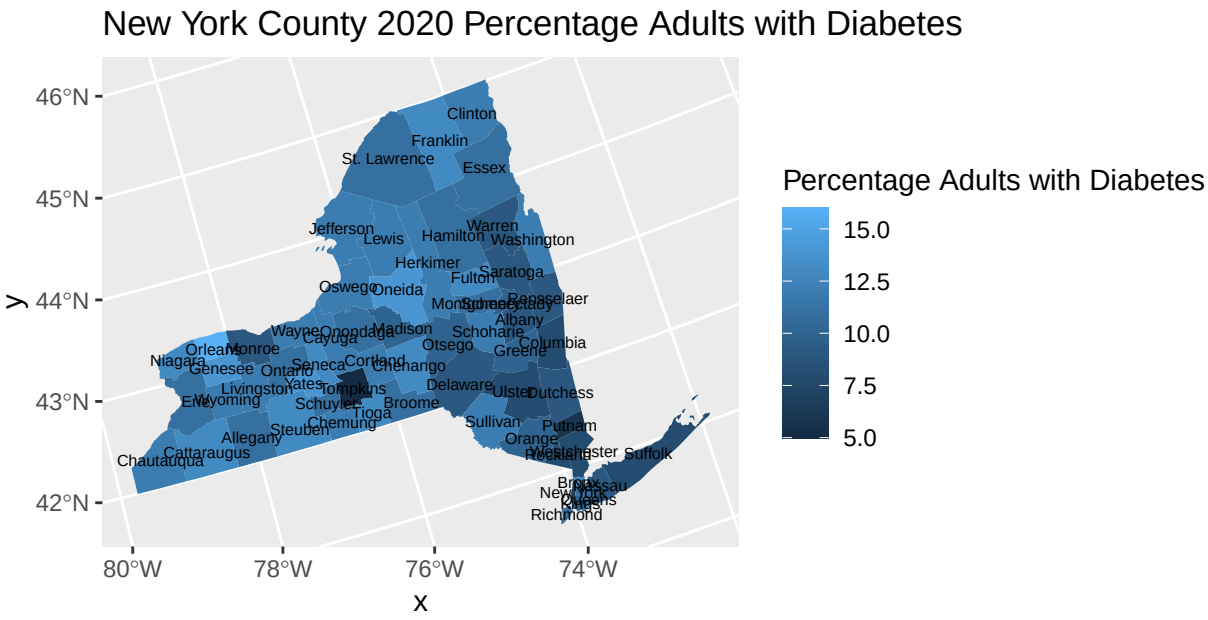


One can not spot direct correlations between NY counties' fatality rate and these measures.

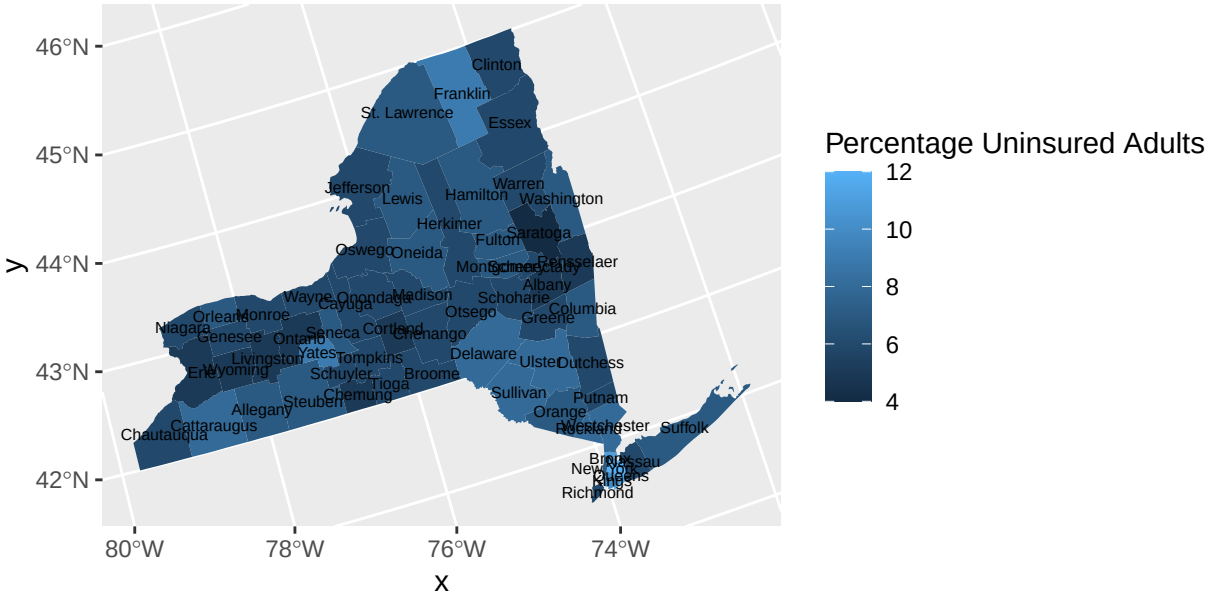
Mapping NY County Additional Measures

This section maps some additional measures that may be related to Covid-19 Fatality rate including percentage adults with diabetes, percentage uninsured adults, median household income, and average traffic volume per meter of major roadways.

```
# Read in the NY county Additional measure data
NYAddmeasure <- read.csv('NY2020CountyAddMeasure.csv')
# Convert fips column from numeric type to character
NYAddmeasure <- transform(NYAddmeasure, county_fips = as.character(county_fips))
# Combine geometry features to work with map
Addmeasure <- left_join(counties_sf, NYAddmeasure, by = "county_fips")
# Map NY county Percentage Adults with Diabetes
Addmeasure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = X..Adults.with.Diabetes),
                 color = NA, size = 0.05)+
    geom_sf_text(data = measure,
                aes(label = County), size = 2)+
    labs(fill = "Percentage Adults with Diabetes") +
  ggtitle("New York County 2020 Percentage Adults with Diabetes")
```

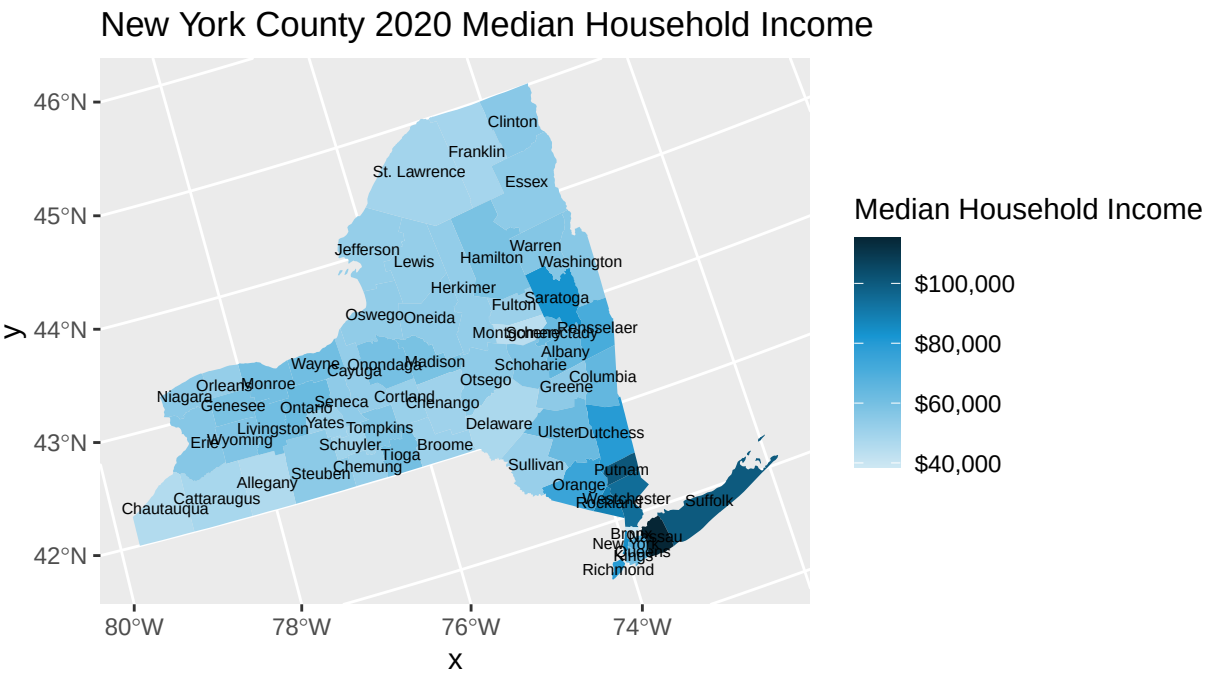



```
# Map NY county Percentage Uninsured Adults
Addmeasure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = X..Uninsured.1),
                color = NA, size = 0.05)+
    geom_sf_text(data = measure,
                 aes(label = County), size = 2)+
    labs(fill = "Percentage Uninsured Adults") +
  ggtitle("New York County 2020 Percentage Uninsured Adults")
```

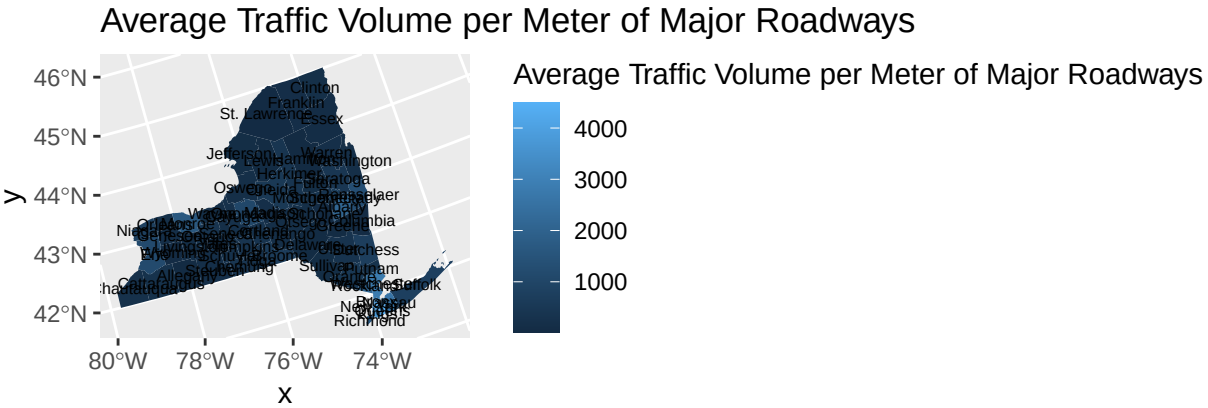



```
# Map NY county Median Household Income
Addmeasure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Median.Household.Income),
```

```
                color = NA, size = 0.05)+
  geom_sf_text(data = measure,
               aes(label = County), size = 2)+
  scale_fill_gradientn(labels = scales::dollar) +
  labs(fill = "Median Household Income") +
ggtitle("New York County 2020 Median Household Income")
```




```
# Map NY county traffic volume
Addmeasure %>%
  ggplot() +
    geom_sf(mapping = aes(fill = Average.Traffic.Volume.per.Meter.of.Major.Roadways),
                color = NA, size = 0.05)+
    geom_sf_text(data = measure,
                 aes(label = County), size = 2)+
    labs(fill = "Average Traffic Volume per Meter of Major Roadways") +
  ggtitle("Average Traffic Volume per Meter of Major Roadways")
```



One can not spot direct correlations between NY counties' fatality rate and these additional measures.

Building Linear Regression Model with Measures

We want to test if the above measures can explain changes in fatality rate by building linear regression model.

```
# Create dataset of fatality rate, health outcome and health factor
data2 <- as.data.frame(NYcounty[c("county","fatality")])
data3 <- as.data.frame(NYmeasure[c("X..Smokers","Z.Score.25","Z.Score.27","County")])
data4 <- as.data.frame(NYAddmeasure[c("X..Adults.with.Diabetes","X..Uninsured.1",
                                      "Median.Household.Income",
                                      "Average.Traffic.Volume.per.Meter.of.Major.Roadways","County")])

colnames(data3) <- cbind("smoker","income_inequality","social_association","county")
colnames(data4) <- cbind("diabete","uninsured","MHI","traffic","county")
data5 <- merge(x=data2,y=data3,by="county")
data6 <- merge(x=data5,y=data4,by="county")
# Build Linear Regression Model for percentage smoker
fit2 <- lm(fatality ~ smoker, data = data6)
summary(fit2)
```

```
##
## Call:
## lm(formula = fatality ~ smoker, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3907 -1.8614 -0.5174  1.6807  6.1412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.8919     3.1480   2.507   0.0152 *
## smoker       -0.3215     0.2014  -1.596   0.1162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.452 on 55 degrees of freedom
## Multiple R-squared:  0.04427,    Adjusted R-squared:  0.02689
## F-statistic: 2.547 on 1 and 55 DF,  p-value: 0.1162
```

```
# Build Linear Regression Model for income inequality
fit3 <- lm(fatality ~ income_inequality, data = data6)
summary(fit3)
```

```
##
## Call:
## lm(formula = fatality ~ income_inequality, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6257 -2.1044 -0.5462  2.1409  5.9308
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.1222     0.3471   8.995 2.15e-12 ***
## income_inequality   1.0944     0.6112   1.790   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.438 on 55 degrees of freedom
## Multiple R-squared:  0.05508,    Adjusted R-squared:  0.0379
## F-statistic: 3.206 on 1 and 55 DF,  p-value: 0.07888
```

```
# Build Linear Regression Model for social association
fit4 <- lm(fatality ~ social_association, data = data6)
summary(fit4)
```

```
##
## Call:
## lm(formula = fatality ~ social_association, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2023 -2.2220 -0.5661  2.1785  6.0750
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.9239     0.3345   8.740 5.52e-12 ***
## social_association   0.2340     0.3815   0.613    0.542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.499 on 55 degrees of freedom
## Multiple R-squared:  0.006791,   Adjusted R-squared:  -0.01127
## F-statistic: 0.3761 on 1 and 55 DF,  p-value: 0.5422
```

```
# Build Linear Regression Model for percentage diabete
fit5 <- lm(fatality ~ diabete, data = data6)
summary(fit5)
```

```
##
## Call:
## lm(formula = fatality ~ diabete, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1913 -2.3908 -0.3646  1.8593  6.3021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5512     1.7845   2.550   0.0136 *
## diabete      -0.1511     0.1599  -0.945   0.3489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 55 degrees of freedom
## Multiple R-squared:  0.01597,    Adjusted R-squared:  -0.001923
## F-statistic: 0.8925 on 1 and 55 DF,  p-value: 0.3489
```

```
# Build Linear Regression Model for percentage uninsured adults
fit6 <- lm(fatality ~ uninsured, data = data6)
summary(fit6)
```

```
##
```

```
## Call:
## lm(formula = fatality ~ uninsured, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3899 -1.9597 -0.4453  1.7363  5.8565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.1777     2.0793   2.490   0.0158 *
## uninsured    -0.3576     0.3215  -1.112   0.2709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.48 on 55 degrees of freedom
## Multiple R-squared:  0.02199,    Adjusted R-squared:  0.00421
## F-statistic: 1.237 on 1 and 55 DF,  p-value: 0.2709
```

```
# Build Linear Regression Model for median household income
fit7 <- lm(fatality ~ MHI, data = data6)
summary(fit7)
```

```
##
## Call:
## lm(formula = fatality ~ MHI, data = data6)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.127 -2.131 -0.570  1.645  5.999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.124e-01  1.363e+00   0.082   0.9346
## MHI         4.605e-05  2.193e-05   2.100   0.0403 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.413 on 55 degrees of freedom
## Multiple R-squared:  0.07424,    Adjusted R-squared:  0.0574
## F-statistic:  4.41 on 1 and 55 DF,  p-value: 0.04032
```

```
# Build Linear Regression Model for traffic volume
fit8 <- lm(fatality ~ traffic, data = data6)
summary(fit8)
```

```
##
## Call:
## lm(formula = fatality ~ traffic, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6696 -2.0421 -0.6064  1.4705  6.2817
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.1886536  0.3863813   5.664 5.57e-07 ***
```

```
## traffic     0.0014843  0.0004916   3.019  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.323 on 55 degrees of freedom
## Multiple R-squared:  0.1422, Adjusted R-squared:  0.1266
## F-statistic: 9.116 on 1 and 55 DF,  p-value: 0.003837
```

R-squred value for the tested seven measures are all below 0.6, indicating that these measures did poorly on explaining fatality rate.

Conclusion and Discussion

Based on the results, NY county level fatality rate can not be explained by health outcomes and factors like percentage smokers, percentage adults with diabetes, etc. These factors may have higher explanatory power in predicting individual cases. But when modeling the disease at a county level, it is possible that other factors like amount of necessary medical equipments, number of nurses avaliable, percentage of people practicing safe social distancing may have better explanatory power. However, these data can not be easily obtained. Future work can be carried into a different direction. One may be interested looking at correlations between social factors and the recover rates at county level.