BIS620 Final Project

Name: Jiayi Wang NetID: jw2775

1. Background and motivation

1. Definition: Social Vulnerability refers to the potential negative effects on communities caused by external stresses on human health.
2. Significance:
    1. Every community must prepare for and respond to hazardous events, whether a natural disaster like a tornado or disease outbreak, or a human-made event such as a harmful chemical spill. A community's capacity to stop human suffering and monetary loss in the event of a disaster may be weakened by a number of variables, including poverty, lack of access to transportation, and congested housing. Social vulnerability refers to these elements.
    2. Additionally, it is helpful in assisting local authorities in identifying communities that might require assistance prior to, during, or after disasters.
3. Research interest:
    1. I want to explore and analyze the crucial relationship between key social factors and the social vulnerability index in the US. For example:
        1. Does each social factor has an impact/correlation on every SVI? If any, is it positive or negative?
        2. What's the most influential social factor of every SVI? What social factors have the greatest impact on every SVI?
        3. Are there any regional differences in different states in the US?
    2. In addition, after doing the statistical analysis using R, I can further raise key social factors that deserve the attention of local officials and make some policy recommendations for regional governance, in order to increase the community's resistance to external stresses, making the community less vulnerable and better prepared.

2. Research Question

1. The research about the qualitative and quantitative relationship between key social factors and the social vulnerability index in the US.

3. Data Description

1. Dataset: ATSDR's Social Vulnerability Index Database.
2. Method: ATSDR's Geospatial Research, Analysis & Services Program (GRASP) uses 16 U.S. census variables, including poverty, lack of vehicle access, and crowded housing, to construct the SVI index into the following four related themes:
    1. Socioeconomic theme summary
    2. Household Composition theme summary

3. Minority Status/Language theme
4. Housing Type/Transportation theme

3. In my research:
    1. Dependent variable: SVI index. Specifically, we use the data from California, Florida, Michigan, and New York State, to investigate the relationship and regional difference between key social factors and SVI.
    2. Independent variable: Key social factors, including poverty, vehicle access, income, education level, and so on.

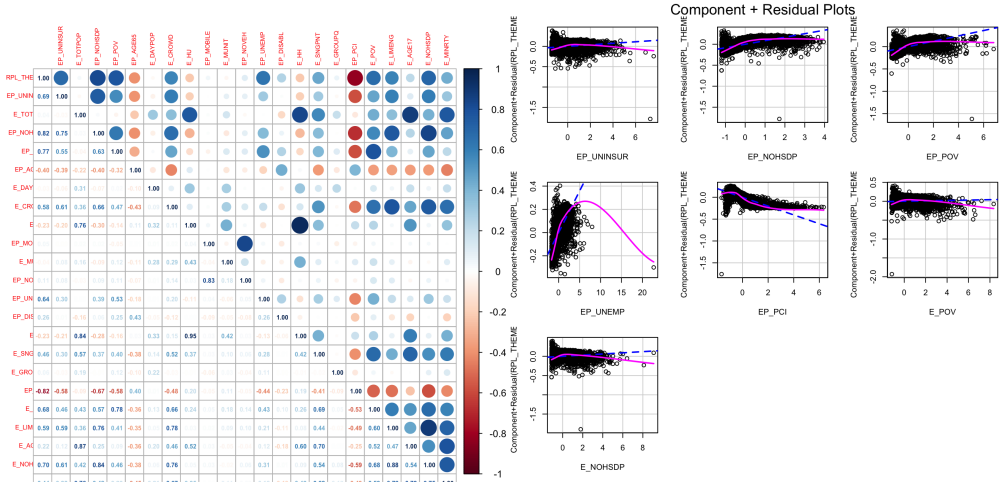
4. R Function Construction

1. Data cleaning and exploration
    1. Read data: Read data in a specific state
    2. Data cleaning function: I write a data cleaning function to be used in the dataset of different states, including the functions:
        1. Data standardization
        2. Delete missing and useless sample, which contains '-999'
    3. EDA: I write an EDA function to be used in the dataset of different states, including the functions:
        1. Explore the correlation between variables using the Pearson method
        2. Visualize the correlation using a 'heat map'
    4. Regression analysis: I write a regression function to be used in the dataset of different states, including the functions:
        1. Use a linear regression model to fit the data
        2. Visualize the component and residuals plots
    5. Test: I write a Durbin Watson test function to be used in the dataset of different states, including the functions:
        1. Linear test
        2. Independent test

5. Analysis

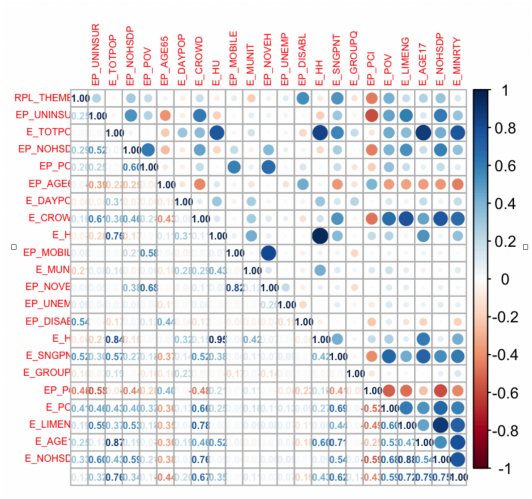
1. Analysis for CALIFORNIA
    1. SVI in Socioeconomic theme
        1. Run the data cleaning and EDA function I wrote before
            1. As the correlation matrix between the SVI1 'RPL_THEME1' (Percentile ranking for Socioeconomic theme summary) and census variables, we can select the most correlated census variables (absolute Pearson coefficient>0.7) including 'EP_UNINSUR', 'EP_NOHSDP', 'EP_POV', 'EP_UNEMP', 'EP_PCI', 'E_POV', 'E_NOHSDP'
            2. In the selected variables, the RPL_THEME1 is negatively correlated to EP_PCI, but is positively correlated to the others.

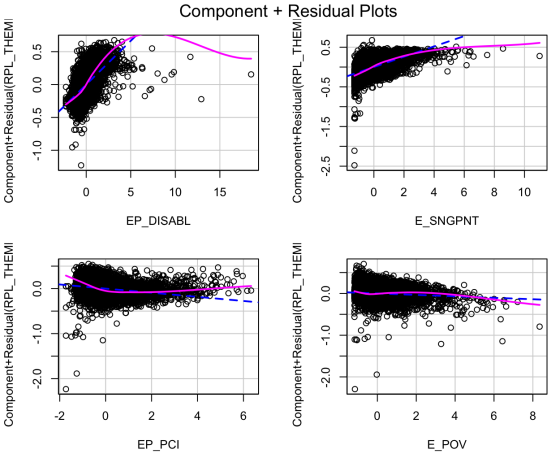
2. Run the regression and test function I wrote before
    1. After multivariate regression on the data, according to the estimate of coefficients of each variable, we know that SVI1 RPL_THEME1 are most influenced by EP_NOHSDP (poor education estimate), EP_POV (poverty estimate), EP_UNEMP (Unemployment Rate estimate), EP_PCI (income estimate).
    2. As shown in the linear test figure, Linear regression can be a good fit for all of the variables.
    3. The results of Durbin Watson Test D-W Statistic is 1.29, which means positive autocorrelation in the residuals and it is acceptable.




```
> summary(fit)

Call:
lm(formula = expression, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92372 -0.06160  0.00640  0.06731  0.41430

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.528006   0.001031 512.219  < 2e-16 ***
EP_UNINSUR   0.018915   0.001589  11.900  < 2e-16 ***
EP_NOHSDP    0.083689   0.002939  28.474  < 2e-16 ***
EP_POV       0.059997   0.002532  23.699  < 2e-16 ***
EP_UNEMP     0.071603   0.001246  57.488  < 2e-16 ***
EP_PCI      -0.100986   0.001495 -67.533  < 2e-16 ***
E_POV        0.005218   0.002531   2.062   0.0393 *
E_NOHSDP     0.015179   0.002905   5.225 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09208 on 7972 degrees of freedom
Multiple R-squared:  0.9057,    Adjusted R-squared:  0.9057
F-statistic: 1.094e+04 on 7 and 7972 DF,  p-value: < 2.2e-16

> hyppthsis_test(fit,processed_data)
 lag Autocorrelation D-W Statistic p-value
   1         0.35353       1.292417       0
 Alternative hypothesis: rho != 0
```

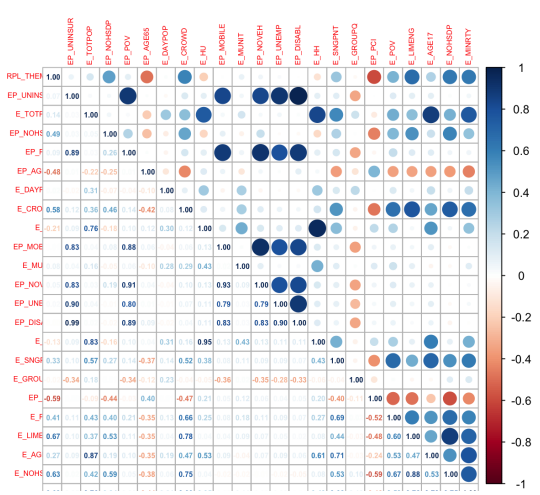
2. SVI in Household Composition theme
    1. Run the data cleaning and EDA function I wrote before
        1. As the correlation matrix between the SVI2 ‘RPL_THEME2' (Percentile ranking for Household Composition theme summary) and census variables, we can select the most correlated census variables (absolute Pearson coefficient>0.45) including: ‘EP_DISABL’, ‘E_SNGPNT', ‘EP_PCI', 'E_POV'.
        2. In the selected variables, the RPL_THEME2 is negatively correlated to EP_PCI, but is positively correlated to the others.
    2. Run the regression and test function I wrote before
        1. After multivariate regression on the data, according to the estimate of coefficients of each variable, we know that SVI2 RPL_THEME2 are most influenced by EP_DISABL (Disability estimate), E_SNGPNT (Single parent household with children under 18 estimate).
        2. As shown in the linear test figure, Linear regression can be a good fit for all of the variables.
        3. The results of Durbin Watson Test D-W Statistic is 1.42, which means positive autocorrelation in the residuals and it is acceptable.

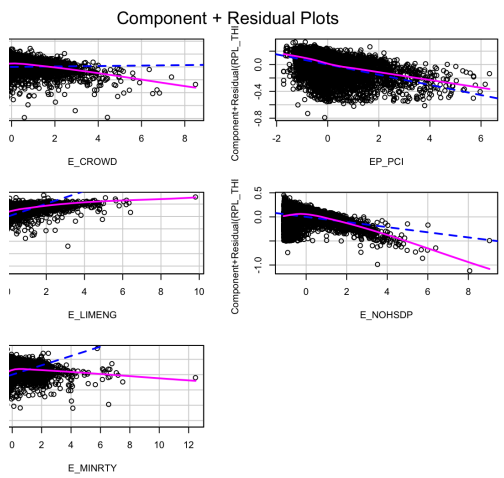





```
> summary(fit)

Call:
lm(formula = expression, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-2.31184 -0.11263 -0.00478  0.11301  0.69425

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.399807   0.001909 209.455  < 2e-16 ***
EP_DISABL    0.133292   0.001969  67.696  < 2e-16 ***
E_SNGPNT     0.130095   0.002646  49.160  < 2e-16 ***
EP_PCI      -0.045831   0.002299 -19.939  < 2e-16 ***
E_POV       -0.017414   0.002823  -6.168 7.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1706 on 7985 degrees of freedom
Multiple R-squared:  0.5818,    Adjusted R-squared:  0.5816
F-statistic:  2778 on 4 and 7985 DF,  p-value: < 2.2e-16

> hyppthsis_test(fit,processed_data)
 lag Autocorrelation D-W Statistic p-value
   1       0.2857073      1.428248       0
 Alternative hypothesis: rho != 0
```

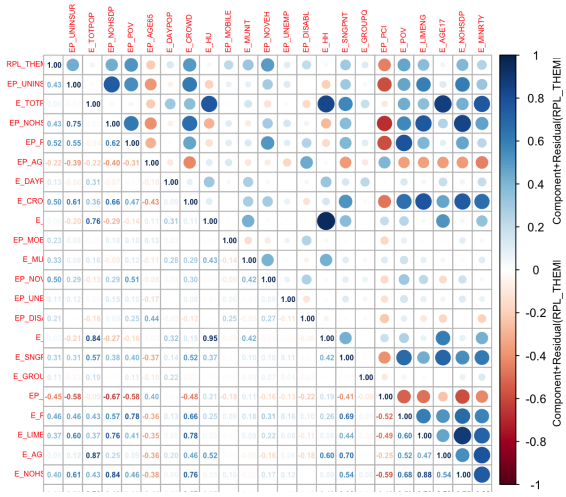
3. SVI in Minority Status/Language theme
    1. Run the data cleaning and EDA function I wrote before
        1. As the correlation matrix between the SVI3 ‘RPL_THEME3' (Percentile ranking for Minority Status/Language theme) and census variables, we can select the most correlated census variables (absolute Pearson coefficient>0.50) including: ‘E_CROWD', ‘EP_PCI', ‘E_LIMENG', ‘E_NOHSDP', ‘E_MINRTY'.
        2. In the selected variables, the RPL_THEME3 is negatively correlated to EP_PCI, but is positively correlated to the others.
    2. Run the regression and test function I wrote before
        1. After multivariate regression on the data, according to the estimate of coefficients of each variable, we know that SVI3 RPL_THEME3 are most influenced by E_LIMENG (Persons who speak English "less than well" estimate).
        2. As shown in the linear test figure, Linear regression can be a good fit for all of the variables.
        3. The results of Durbin Watson Test D-W Statistic is 1.1, which means positive autocorrelation in the residuals and it is acceptable.

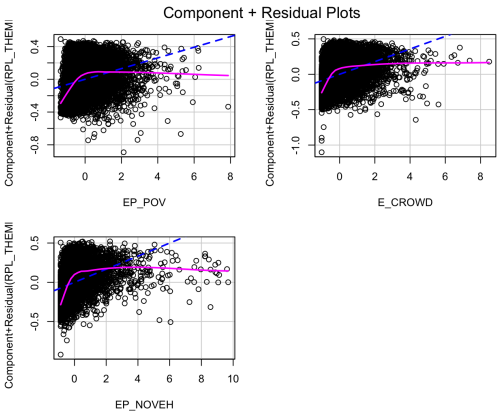





```
> summary(fit)

Call:
lm(formula = expression, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.80000 -0.05735  0.02714  0.08665  0.38997

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.758498   0.001479 512.942   <2e-16 ***
E_CROWD      0.003544   0.002469   1.435    0.151
EP_PCI      -0.075870   0.001841 -41.221   <2e-16 ***
E_LIMENG     0.102453   0.003396  30.165   <2e-16 ***
E_NOHSDP    -0.053301   0.003588 -14.856   <2e-16 ***
E_MINRTY     0.061001   0.002317  26.330   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1324 on 8006 degrees of freedom
Multiple R-squared:  0.5865,    Adjusted R-squared:  0.5863
F-statistic:  2271 on 5 and 8006 DF,  p-value: < 2.2e-16

> hyppthsis_test(fit,processed_data)
 lag Autocorrelation D-W Statistic p-value
   1       0.4419662      1.115941       0
 Alternative hypothesis: rho != 0
```

4. SVI in Housing Type/Transportation theme
    1. Run the data cleaning and EDA function I wrote before
        1. As the correlation matrix between the SVI4 ‘RPL_THEME4' (Percentile ranking for Housing Type/Transportation theme) and census variables, we can select the most correlated census variables (absolute Pearson coefficient>0.45) including: ‘EP_POV',E_CROWD','EP_NOVEH'.
        2. In the selected variables, the RPL_THEME4 is positively correlated to all selected variables.
    2. Run the regression and test function I wrote before
        1. After multivariate regression on the data, according to the estimate of coefficients of each variables, we know that SVI4 RPL_THEME4 are equally influenced by EP_POV (Poverty estimate), E_CROWD (More people than rooms estimate), EP_NOVEH (No vehicle available estimate).
        2. As shown in the linear test figure, Linear regression can be a good fit for all of the variables.
        3. The results of Durbin Watson Test D-W Statistic is 1.50, which means positive autocorrelation in the residuals and it is acceptable.






```
> summary(fit)

Call:
lm(formula = expression, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.0295 -0.1788  0.0099  0.1724  0.5781

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.573070   0.002454  233.48   <2e-16 ***
EP_POV      0.065221   0.003095   21.07   <2e-16 ***
E_CROWD     0.089335   0.002789   32.03   <2e-16 ***
EP_NOVEH    0.083278   0.002863   29.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2193 on 7977 degrees of freedom
Multiple R-squared:  0.4188,    Adjusted R-squared:  0.4186
F-statistic:  1916 on 3 and 7977 DF,  p-value: < 2.2e-16

> hyppthsis_test(fit,processed_data)
 lag Autocorrelation D-W Statistic p-value
   1       0.2503523      1.498932       0
 Alternative hypothesis: rho != 0
```

2. Analysis for Florida, Michigan, and New York
    1. Repeat the analysis process above using each state’s data. In order to save space in this report, the details are not repeated here. The code, plots, and results can be found in the R script

6. Interpretation and conclusions

1. The most influential factor of different SVI index are listed below:
    1. SVI index in Socioeconomic theme:
        1. Social factors that have a positive impact:
            1. Poor education: Percentage of persons with no high school diploma (age 25+)

    2. Poverty: Percentage of persons below poverty
    3. Unemployment rate
    4. People uninsured: Uninsured in the total civilian non-institutionalized population
  2. Social factors that have a negative impact:
    1. Income: Per capita income
  3. Intuitive explanation:
    1. In terms of Socioeconomic SVI, intuitively, low education level, poverty, unemployment and the number of uninsured people will increase the socioeconomic vulnerability of one community.
    2. However, if the income per capita increases, people can be much more prepared for hazardous events, thus decreasing social vulnerability.
2. SVI index in Household Composition theme:
  1. Social factors that have a positive impact:
    1. Disability: Percentage of civilian non-institutionalized population with a disability
    2. Single-parent household with children under 18
  2. Social factors that have a negative impact:
    1. Income: Per capita income
  3. Intuitive explanation:
    1. In terms of Household Composition SVI, intuitively, a community with more disabled people and single-parent families is much more vulnerable.
    2. However, if the income per capita increases, people and families can be much more prepared for hazardous events, thus decreasing social vulnerability.
3. SVI index in Minority Status/Language theme
  1. Social factors that have a positive impact:
    1. Persons who speak English ‘less than well’
    2. Minority: Minority (all persons except white, non-Hispanic)
  2. Social factors that have a negative impact:
    1. Income: Per capita income
    2. Poor education: Percentage of persons with no high school diploma (age 25+)
  3. Intuitive explanation:
    1. In terms of Minority Status SVI, intuitively, the more people who are not proficient in English and the more minorities there are, the more vulnerable the community will be.
    2. However, if the income per capita increases, people and families can be much more prepared for hazardous events, thus decreasing social vulnerability.
    3. Surprisingly, poor education also hurts the Minority Status SVI. One possible explanation is that highly educated people are prone to a kind of elitist arrogance, which can make them have a stronger sense of race and identity, and thus less

tolerance for minorities. So, if the level of education is lower, the community will be more inclusive and friendly, thus decreasing social vulnerability.

4. SVI index in Housing Type/Transportation theme
    1. Social factors that have a positive impact:
        1. Poverty: Percentage of persons below poverty
        2. More people than rooms: At household level (occupied housing units), more people than rooms estimate
        3. No vehicle available: Percentage of households with no vehicle available
    2. Intuitive explanation:
        1. In terms of Housing Type/Transportation SVI, intuitively, poverty, lack of vehicle access, and crowded housing will increase the housing and transportation vulnerability of one community.

2. Regional Difference

	California	**Florida**	**Michigan**	**New York**
Socioeconomic theme	Poor education Poverty Unemployment rate Income People uninsured	Poor education Poverty Unemployment rate Income People uninsured	Poverty Unemployment rate Income People uninsured	Poor education Poverty Unemployment rate Income People uninsured
Household Composition theme	Disability Single parent household with children under 18 Income	Disability Single parent household with children under 18 Income	Disability Single parent household with children under 18	Disability Single parent household with children under 18
Minority Status/ Language theme	Income Persons who speak English "less than well" Poor education Minority	Income Persons who speak English "less than well" Poor education Minority	Income Persons who speak English "less than well" Poor education Minority	Income Persons who speak English "less than well" Poor education Minority
Housing Type/ Transportation theme	Poverty More people than rooms No vehicle available	Poverty More people than rooms No vehicle available	Poverty More people than rooms No vehicle available	Poverty More people than rooms No vehicle available

1. The most influential social factors of SVI in different themes in different states are listed above. From the regression results in four different states in the US , I can conclude that the relationship between key social factors and the SVI in the US doesn't have the regional difference, because the key social factors of SVI are almost the same in different states.

3. Suggestions and policy recommendations for regional governance
    1. The income per capita is the most important social factor in almost every SVI theme. If we want to decrease the vulnerability of one community, the best option for local officials is to try to increase people's income, whether through direct subsidies or by creating more high-paying jobs.

2. Education level has different effects on the Socioeconomic SVI and Minority Status SVI. With the improvement of education level, the local officials should pay attention to the protection of minority groups, to reduce SVI, making the community less vulnerable and better prepared.
3. In addition to income, poverty, and education, the degree of room congestion, access to the vehicle, English proficiency, and the number of single-parent families are also key social factors worthy of local officials' attention.
4. Similar improvement strategies can be adopted for different states in the United States without over-considering intercontinental regional differences.

7. **Reference**: <https://www.atsdr.cdc.gov/placeandhealth/svi/index.html>