Introduction to Spatial Data Science: Final Project

March 20, 2020

Jesus Cantu Jr.

"…by studying death we gain capacity to understand life as well as to protect it." (Noguchi & DiMona, 1985)

Introduction

In July of 1995, the midwestern United States experienced a severe heat wave, for which, over the course of just a single week, the city of Chicago, Illinois, documented about 514 heat-related deaths and 696 excess deaths [2], a baseline mortality rate of about 72 individuals per day—among the highest in the history of the nation. Following the legacy of Durkheim and Mauss, Eric Klinenberg (1999) carried what he called a "social autopsy" of this disaster by combining elements of urban sociology, which highlights the importance of political and economic power in socio-spatial organization of the city, with those of the Chicago School, whose approach tried to account for the geography of urban vulnerability. In doing so, his analysis of the heat wave helps to explain the mortality patterns observed across the city beyond what can be officially attributed to the weather. Patterns that clearly reflect the inequalities of Chicago's built infrastructure.

Throughout the 1995 heat wave "geography was linked to destiny" [3]. As *Figure 1* illustrates, the Chicago community areas with the highest death rates lie on the south and southwest sides of the city. Of the fifteen community areas with the highest mortality rates during the heat wave, eleven contained high proportions of people living below half of the federal poverty line[1] and ten are home to populations between 94-99% African-American (see *Table 1*); city regions that have historically been marginalized. Klinenberg (1999) notes a strong positive relationship between death rates among community areas and homicide rates. High mortality rates were also strongly associated with the number of seniors living alone, as deaths were more prevalent among low-income, elderly, African-Americans. Judith Helfand's (2019) film, *Cooked: Survival by Zip Code,* captures the life and death stories of these individuals, many of whom spent their last moments alone, confined to small, single-bedroom apartments, too afraid of neighborhood crime to even open a window and far-removed from the safety networks allocated to the rest of [WHITE] America in times of disaster to receive any immediate assistance.

The purpose of this research project is to survey the city's socio-economic environment, 25 years after the tragedy, for changes in structure against the backdrop of current population health statistics. Do current mortality rates follow the same patterns as observed by Klinenberg (1999)? If so, what can modern data tell us about people's living conditions and their capacity to deal with future disasters in the face of an ever-changing climate? In the wake of the 2020 COVID-19 (coronavirus) pandemic, and the spread of the disease in the U.S., it is of outmost importance to identify areas and individuals who are at increased risk of suffering a severe infection, particularly, as a result of complications with other chronic illnesses or auto-immune disorders. These susceptible individuals should be prioritized when allocating medical/financial assistance by state and local governments. Ultimately, I hope the results of this preliminary spatial analysis can inspire further research into the transmission dynamics and control of the virus across different geographic regions.

---

[1] About $15,150 for a four-person family in 1995 [4].

Data

The City of Chicago Data Portal (https://data.cityofchicago.org/) holds a rich repository of data regarding the city's infrastructure. Datasets can be downloaded for free and are available, at different organizational levels, in a variety of categories ranging from administration and finance information to sanitation and public safety requests. Here, we will be focusing on a dataset that contains data on selected underlying causes of death in Chicago.[2] It includes the cumulative number of deaths, average number of deaths annually, average annual crude and adjusted rates with corresponding 95% confidence intervals, and average annual years of potential life lost per 100,00 residents aged 75 and younger due to selected causes of death[3], by Chicago community area, for the years 2006 to 2010. A ranking for each measure is also provided, with the highest value indicated with a ranking of 1.

Age-adjustment was done using a slight modification of the direct method to the 2000 US standard million population (specifically, using standard distribution 1, the less than 1 year-old and 1 to 4 year-old groups were combined). Rates are expressed per 100,000 population. Age-specific population estimates by community area were obtained through a linear interpolation of counts from the 2000 and 2010 Census. Because the age-adjusted death rate is a mortality rate that controls for the effects of differences in population age distributions, it is useful for comparing data over time and by geographic area. For this study, we will be using this measure as our main dependent variable while paying particular attention to the following causes of death: coronary heart disease, diabetes-related, and stroke as they have been found to increase the risk of severe complications from coronavirus infection among individuals [6-9]. Data from deaths resulting from unintentional injury was also included as a possible control; however, this assumes that most of these types of deaths are a result of chance events. Spatially, its distribution should be random (i.e., null hypothesis). We should also expect there to be little association between this mortality rate and our independent variables.

To explore the city's socio-economic environment, I am using a dataset that contains a selection of six socioeconomic indicators of public health significance and a "hardship index", by Chicago community area, for the years 2008 to 2012. The variables include: the percent of occupied housing units with more than one person per room (i.e., crowded housing), the percent of households living below the federal poverty line, the percent of persons aged 16 years and older in the labor force that are unemployed, the percent of people aged 25 years or older without a high school diploma, the percent of the population under 18 or over 64 years of age (i.e., dependency), and per capita income. The Chicago Department of Public Health (CDPH) calculated the indicators using census tract-level estimates obtained from the 2008-2012 American Community Survey 5-year estimates.[4] A community area's per capita income was estimated by dividing aggregate

[2] An "underlying cause of death" is defined as the main disease, accident, or injury that caused the death.

[3] The years of potential life lost rate (YPLL) is a summary of years lost due to premature death per 100,000 population at or below 75. In contrast to mortality rates, YPLL emphasizes the effect of premature mortality on a population. YPLL is the sum of the differences between a predetermined end point (average life expectancy) and the ages of death for those who died before that end point, divided by the total population at or below that end point, and multiplied by 100,000.

[4] According to U.S. Census Bureau 2008-2012 American Community Survey 5-year estimates, 3.2% of occupied housing units in the U.S. had more than one person per room; 10.9% of households were living below the federal poverty level; 9.3% of persons aged 16 years or older in the labor force were unemployed; 14.2% of persons aged 25 years or older did not have a high school diploma; 37.2% of the population was under 18 or over 64 years of age; and per capita income was $28,051.

income of the census tracts within community area by the number of residents. The hardship index is a score that incorporates each of the six socio-economic variables according to the methods described in [10]. Scores on the index can range from 1 to 100, with a higher index number representing a greater level of hardship; these scores are standardized for the 77 community areas, and cannot be compared to scores generated from data of other jurisdictions.

As a proxy to survey individual's access to healthcare, I calculated the relative distance (in kilometers) from each community area's centroid to the nearest hospital (see *Figure 2*). A list of hospitals with their corresponding addresses was obtained from CPH; the list was last updated as of August 28, 2011. Precaution was taken to remove institutions that would not likely treat individuals suffering from COVID-19 infections (e.g., rehabilitation centers). This process generated a list of 39 institutions, for a complete list see the *Appendix*. Addresses were geocoded using the 'ggmap' package in R. Distances to the nearest hospital were calculated using the Haversine formula, which determines the great-circle-distance between two points on a sphere given their longitudes and latitudes. It should be noted that this method, as implemented using the 'geosphere' package, assumes a spherical earth, ignoring ellipsoidal effects. Unless specified, all other data wrangling and analysis was carried out using R (version 3.6.1).

Exploratory Data Analysis

Descriptive statistics for our selected underlying causes of death are shown in *Table 2*. Coronary heart disease appears to have the highest average annual death rate[5] (age-adjusted) amongst community areas in Chicago, 2006-2010. It is then followed by stroke, unintentional injury, and diabetes-related deaths. For the time period, no outliers were identified among diabetes-related mortality rates, but were present for all other underlying causes (see *Figure 3*). *Figure 4* shows a scatter plot of mortality rates vs. YPLL. In the plot, every dot represents a community area while each color represents a different underlying cause of death. Because of the way in which YPLL is calculated, this measure gives more weight to a death the earlier it occurs. It appears, then, that unintentional injury resulting in death is occurring at a higher proportion within younger individuals compared to the other causes, which tend to afflict older populations. From this plot, we can also observe more of a cloud-like pattern for deaths from coronary heart disease; there is more variability in the average annual mortality rates for this underlying cause of death between community areas compared to that of the others.

Summary statistics for our independent variables are shown in *Table 3*. In order to reduce the number of independent variables, and to minimize issues of multicollinearity later in regression analysis, correlation between independent variables was tested using the Pearson correlation method. The results are visualized in a correlogram (see *Figure 5*). Most variables, with the exception of per capita income, are positively correlated with each other. Of significance, is the hardship index which appears to share a moderate to strong uphill linear relationship with most socioeconomic variables, and per capita income which shares the opposite. From here on out, the analysis focuses on these two variables and our proxy for access to healthcare services (i.e., kilometer distance to nearest hospital).

[5] From here on out, reference to any mortality rates will always be average annual, age-adjusted, rates.

Geo-visualization

Mapping was used to describe the spatial aspects of the data and to assess potential spatial heterogeneity in our variables. Box maps are shown in *Figures 6-7*, a box map is an augmented quartile map with an additional lower and upper category. When there are lower outliers, for example, the starting point for the breaks will be the minimum value for that variable, and the second its lower fence. In contrast, when there are no outliers, then the starting point will be lower fence, and the second break its minimum value. The same method is applied at the upper end of the distribution.

Visually, we can detect some heterogeneity in mortality rates across community areas in Chicago, for each of our underlying causes of death (see *Figure 6*). Highest mortality rates are concentrated within the south and southwest regions of the city, particularly for stroke and diabetes-related deaths. Surprisingly, these regions also have high indices of death as a result of unintentional, injury which we expected to have a more random distribution across the city. The appearance of spatial heterogeneity is more pronounced among our main independent variables, where there is also more evidence of clustering of like values (see *Figure 7*). For per capita income, there is almost an incline seen across the city going from the north to the south sides— rich to poor. The opposite occurs for the hardship index which increases for community areas as you travel southwards. For these regions, the larger presence of hospitals in the upper side of Chicago makes most distances to the nearest hospital seem almost like outliers.

Results from multivariable regression analysis are shown in *Tables 4-5*. Because of issues of collinearity, per capita income and the hardship index had to be fitted separately. Both of these variables are significantly associated with mortality rates ($p < 0.01$), across each of the underlying causes of death. As they are, however, the models can at best explain about 37% of the variance in mortality rates for some of the datasets, with those that include the hardship index showing more explanatory power. Further analysis will require more extensive variables at the community level.

Spatial Autocorrelation Analysis

Tests for spatial autocorrelation were carried out in GeoDa (version 1.14.0), as this program provided an easier way to implement Moran's I. For each underlying cause of death, global spatial autocorrelation was examined across our variables by creating a univariate, Moran scatter plot (e.g., *Figure 8*) with significance being assessed by means of a permutation test. Queen-based contiguity was chosen as a weights measure between the polygon figures (i.e., community areas) after sensitivity analysis. This measure resulted in a minimum and maximum of 7 and 9 neighbors, respectively (a mean of 5.12). Among our global search, results show the appearance of significant clustering for all our dependent and independent measures (pseudo p-value < 0.001).

Identification of clusters and the assessment of their significance was carried out by computing a local, univariate Moran statistic and analyzing the associated significance map and cluster map (e.g., *Figure 9 & 10*), with queen-based contiguity being kept as the choice of weights. Results for the analysis of our three socioeconomic variables show positive spatial autocorrelation or clusters of like values. In their Moran scatter plots (not shown), clusters were found to lie within the upper right and lower left quadrants, indicating low-low (L-L) and high-high (H-H) values relative to the mean. Brushing of the scatter plot helped identify clusters shared between our socioeconomic variables, with the following community areas sharing the same qualities among the hardship index and the measure of per capita income (pseudo p-values < 0.001): Lincoln Park (H-H), Near North Side (H-H), New City (L-L), Englewood (L-L). For each underlying cause of

death, this process was repeated to identify patterns of mortality rates across the city. The following clusters were observed among all causes (pseudo p-values < 0.01): Greater Grand Crossing (H-H), Pullman (H-H), Fuller Park (H-H), Englewood (H-H), and Lincoln Square (L-L). Of particular significance are Fuller Park, Grand Crossing, and Englewood which were noted by Klinenberg (1999) as community areas to have had experienced some of the highest heat-related deaths during the 1995 Chicago heat wave. It appears that these areas continue to be afflicted by poor socioeconomic conditions (i.e., lower per capita income and a higher hardship index). This study reveals that these areas may also lack proper healthcare infrastructure, in terms of access to nearby hospitals.

Conclusion

The indication of clustering does not provide an explanation for why the clustering occurs. Different processes can result in the same spatial patterns (e.g., true contagion and apparent contagion). However, the implementation of regression analysis alongside novel geo-visualization techniques allows us to start building a better picture of the city's environment, its resources, and their effect on individuals health and well-being. Of relevance to our epidemiological question is *Figure 11* which shows a linear association between the percent of individuals (aged 25 and older) without a high-school diploma and the percent of households living in crowded conditions across community areas in Chicago— a clear sign of hard living conditions (as enumerated by the hardship index). Evaluating *Figure 12* in light of this information tells us that the community areas facing the hardest living conditions as a result of lower educational attainment, particularly in terms of income and mortality rates, also tend to be the most crowded and with the highest number of dependents (either children or elderly). These living conditions could result in catastrophic events as COVID-19 spreads across Chicago. Disaster preparedness requires us to identify vulnerable populations early and to plan methods for which we can allocate essential resources to them expeditiously. It is unwise to do so while fighting a global pandemic— it is almost impossible; thus, only time will tell if we are able to avoid yet another tragedy, much like the 1995 heath wave.

Bibliography

1. Noguchi, Thomas T., and Joseph DiMona. *Coroner at Large*. Pochet Books, 1985.
2. Whitman, S., Good, G., Donoghue, E. R., Benbow, N., Shou, W., & Mou, S. (1997). Mortality in Chicago attributed to the July 1995 heat wave. *American journal of public health*, *87*(9), 1515–1518. https://doi.org/10.2105/ajph.87.9.1515
3. Klinenberg, E. (1999). Denaturalizing Disaster: A Social Autopsy of the 1995 Chicago Heat Wave. *Theory and Society*, *28*(2), 239-295. Retrieved March 18, 2020, from www.jstor.org/stable/3108472
4. Prior HHS Poverty Guidelines and Federal Register References. (2020, January 21). Retrieved from https://aspe.hhs.gov/prior-hhs-poverty-guidelines-and-federal-register-references
5. *COOKED: Survival by Zip Code*. 2019. [film] Directed by J. Helfand. Chicago, Illinois, USA: ITVS, Kartemquin Films.
6. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., … Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), 507–513. doi: 10.1016/s0140-6736(20)30211-7
7. Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., … Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, *395*(10223), 514–523. doi: 10.1016/s0140-6736(20)30154-9
8. Guan, W.-jie, Doremalen, N. van, Cao, B., Lu, X., & China Medical Treatment Expert Group. (2020).Clinical Characteristics of Coronavirus Disease 2019 in China. *NEJM*, doi:10.1056/NEJMoa2002032
9. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., … Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, *395*(10229), 1054–1062. doi: 10.1016/s0140-6736(20)30566-3
10. Montiel, M. Lisa, Nathan, R.P., Wright, D.J. (2004). An Update on Urban Hardship. The *Nelson A. Rockefeller Institute of Government*. http://www.phasocal.org/wp-content/uploads/2014/03/EH-urban-hardship-Rockefeller-Institute-2004.pdf

Figures & Tables

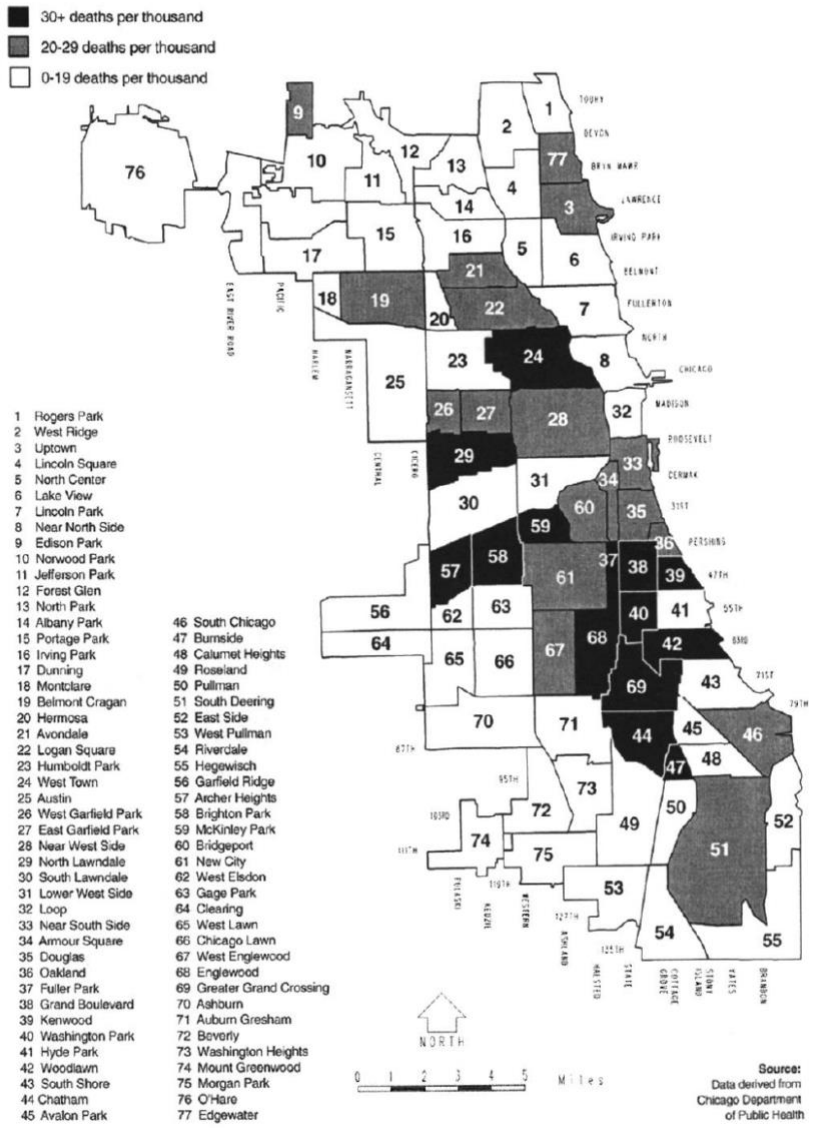

Figure 3. Chicago community areas with highest heat-related death rates.

Figure 1. Chicago community areas with highest heat-related death rates. Obtained from "Denaturalizing Disaster: A Social Autopsy of the 1995 Chicago Heat Wave," by Klinenberg, E., 1999, *Theory and Society, 28*(2), p. 251.

254

Table 1. Chicago community areas with highest heat-related death rates

Community area	Heat-related deaths/100,000	Percent of population black	Median family income ($)
Fuller Park	92	99	8,371
Woodlawn	73	96	17,714
Archer Heights	54	0	37,744
Grand Crossing	52	99	22,931
Washington Park	51	99	9,657
Grand Boulevard	47	99	8,371
McKinley Park	45	0	31,597
North Lawndale	40	96	14,209
Chatham	35	99	29,258
Kenwood	33	77	31,954
Englewood	33	99	22,931
West Town	32	10	20,532
Brighton Park	31	0	30,677
Burnside	30	98	30,179
Near South Side	29	94	7,576
Chicago	7	39	30,707

Data based on 514 heat-related deaths located by the Illinois Department of Public Health.

Table 1. Chicago community areas with highest heat-related death rates. Obtained from "Denaturalizing Disaster: A Social Autopsy of the 1995 Chicago Heat Wave," by Klinenberg, E., 1999, *Theory and Society, 28*(2), p. 254.

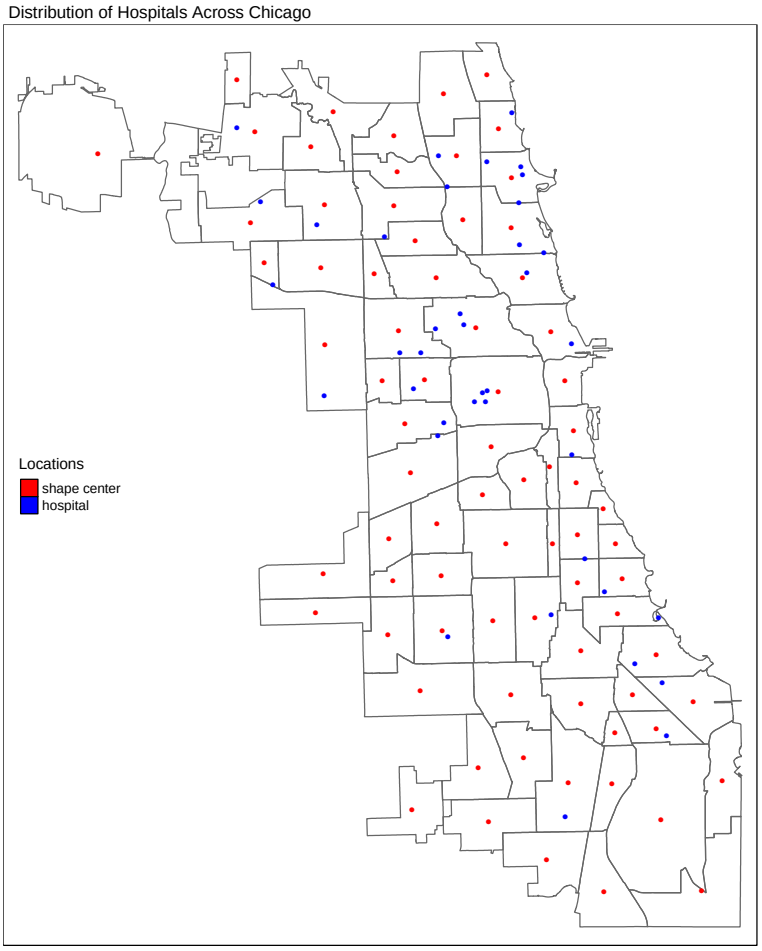

Figure 2. Map showing Chicago community areas with centroid (red) and the locations of hospitals throughout the city (blue); hospital data as of August 28, 2011.

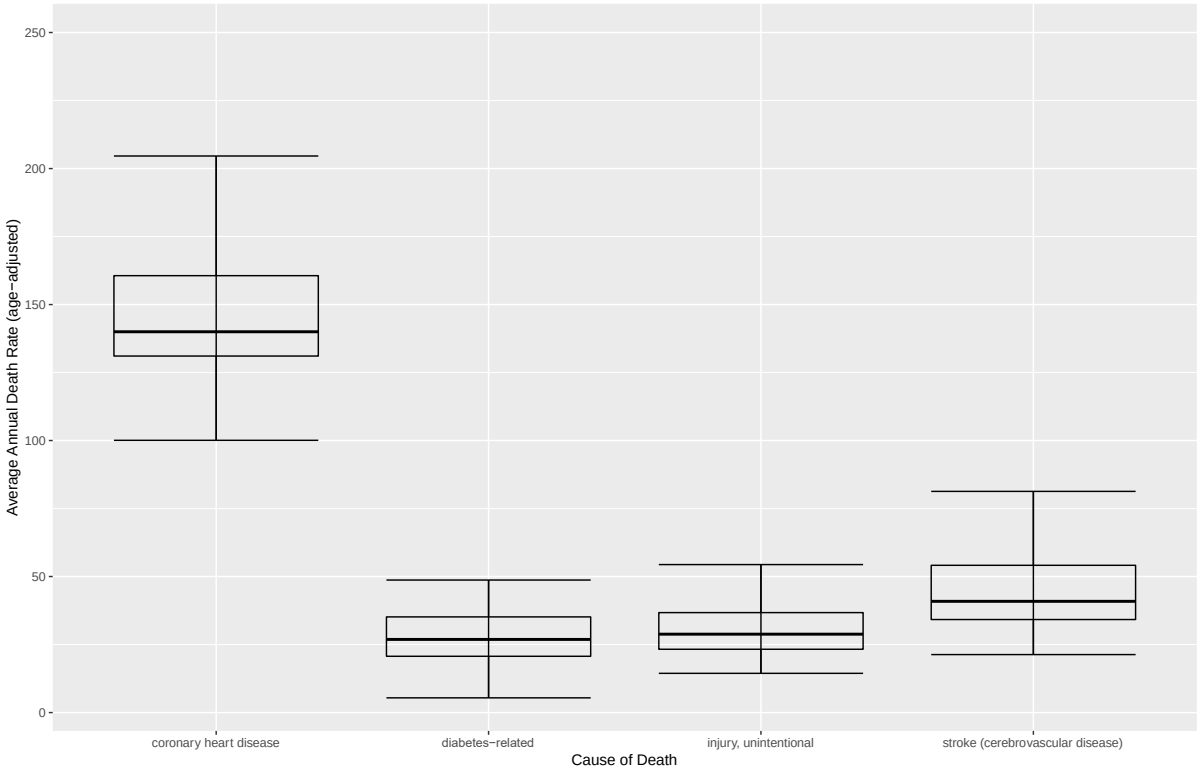

Figure 3. Box and whisker plot of dependent variables. For each underlying cause of death, outliers (red) and observations (black) are shown as points; community area data for Chicago (2006-2010).

Descriptive statistics for dependent variables

Cause.of.Death	*mean*	*SD*	*median*	*min*	*max*
coronary heart disease	144	30	140	83	248
diabetes-related	28	10	27	5	49
injury, unintentional	32	12	29	14	78
stroke (cerebrovascular disease)	45	14	41	21	88

Table 2. Summary statistics of dependent variables; community area data for Chicago (2006-2010).

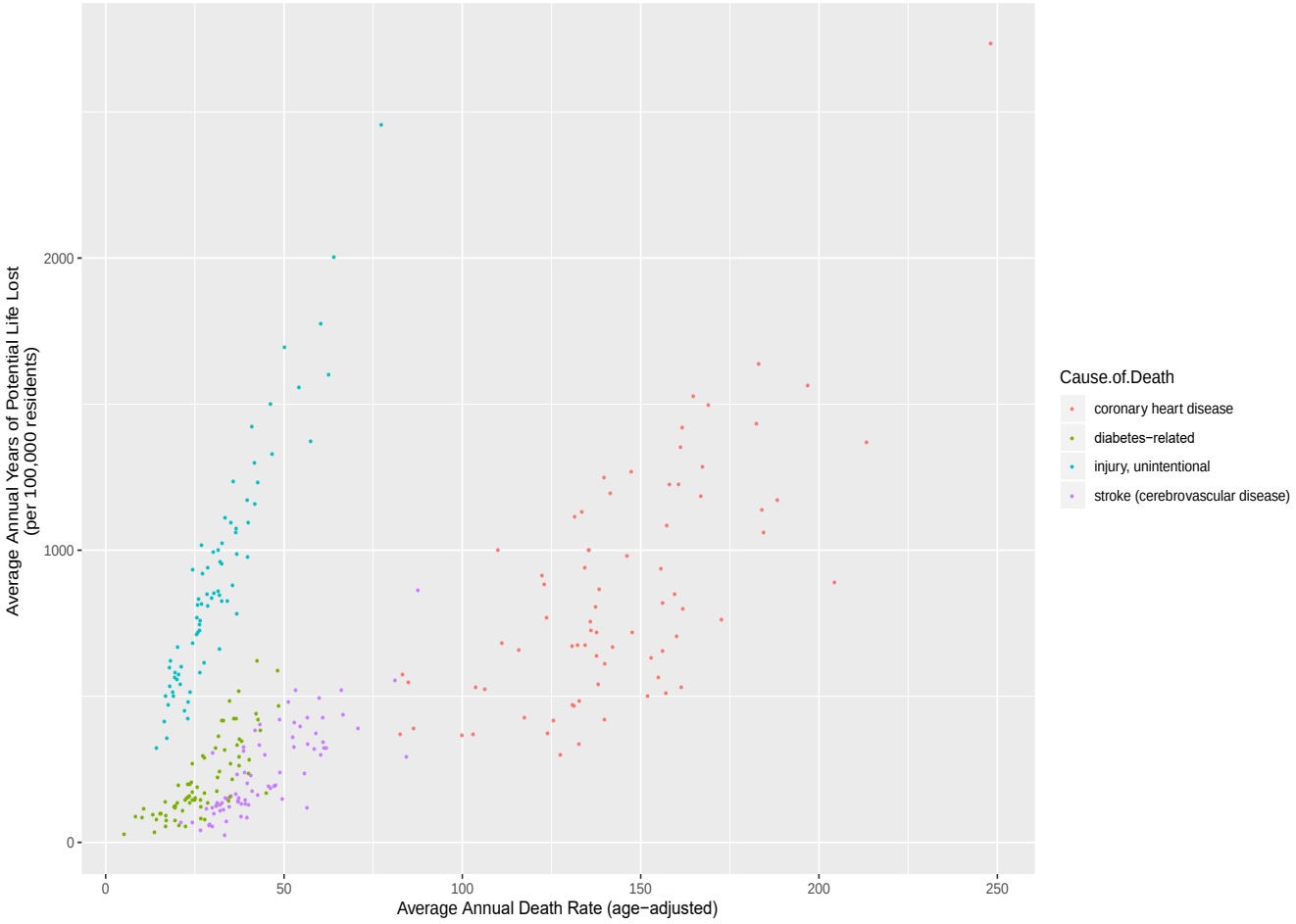

Figure 4. Bivariate analysis of average annual death rate vs. average annual years of potential life lost (YPLL) colored by underlying cause of death; community area data for Chicago (2006-2010).

Descriptive statistics for independent variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
PERCENT.OF.HOUSING.CROWDED	77	5	4	0	2	7	16
PERCENT.HOUSEHOLDS.BELOW.POVERTY	77	22	12	3	13	29	56
PERCENT.AGED.16..UNEMPLOYED	77	15	8	5	9	20	36
PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA	77	20	12	2	12	27	55
PERCENT.AGED.UNDER.18.OR.OVER.64	77	36	7	14	32	40	52
PER.CAPITA.INCOME	77	25,563	15,293	8,201	15,754	28,887	88,669
HARDSHIP.INDEX	77	50	29	1	25	74	98
DISTANCE.NEAREST.HOSPITAL	74	2	2	0	1	3	8

Table 3. Summary statistics of independent variables; community area data for Chicago (2008-2012).

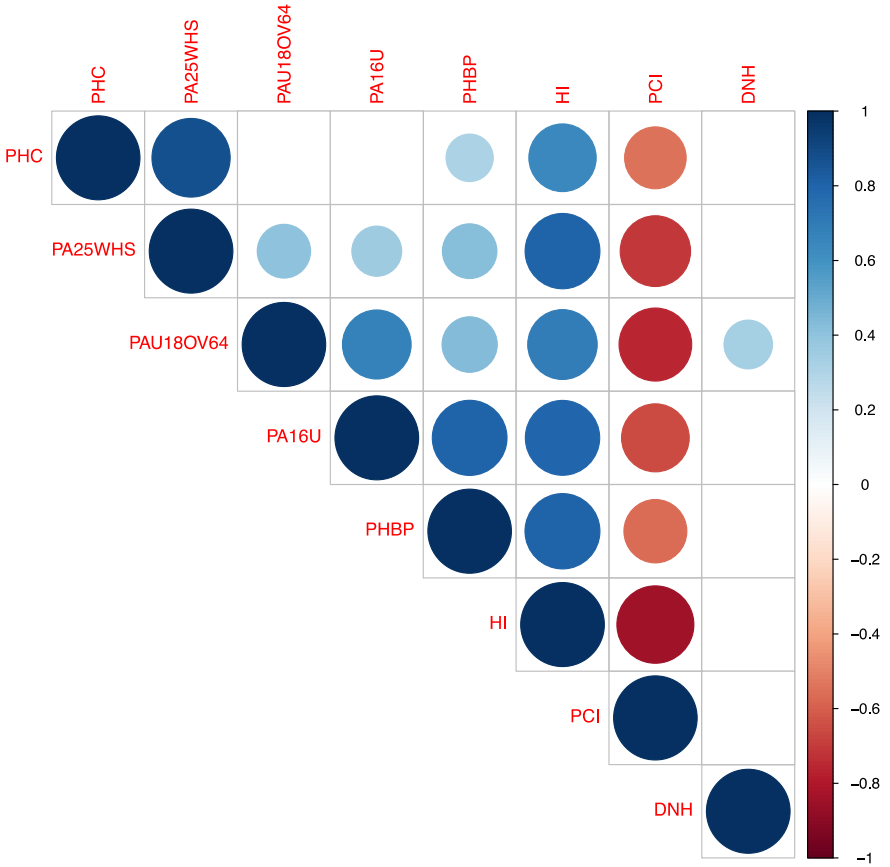

Figure 5. Correlogram for independent variables; community area data for Chicago (2008-2012). Variable names were shortened to fit plot and are as follows: percent of occupied housing units with more than one person per room (PHC), percent of individuals aged 25 years or older without a high school diploma (PA25WHS), percent of the population under 18 or over 64 years of age (PAU18OV64), percent of persons aged 16 years and older in the labor force that are unemployed (PA16U), percent of households living below the federal poverty line (PHBP), hardship index (HI), per capita income (PCI), kilometer distance to the nearest hospital (DNH). Correlation coefficients with corresponding significance levels larger than 0.01 were ignored.

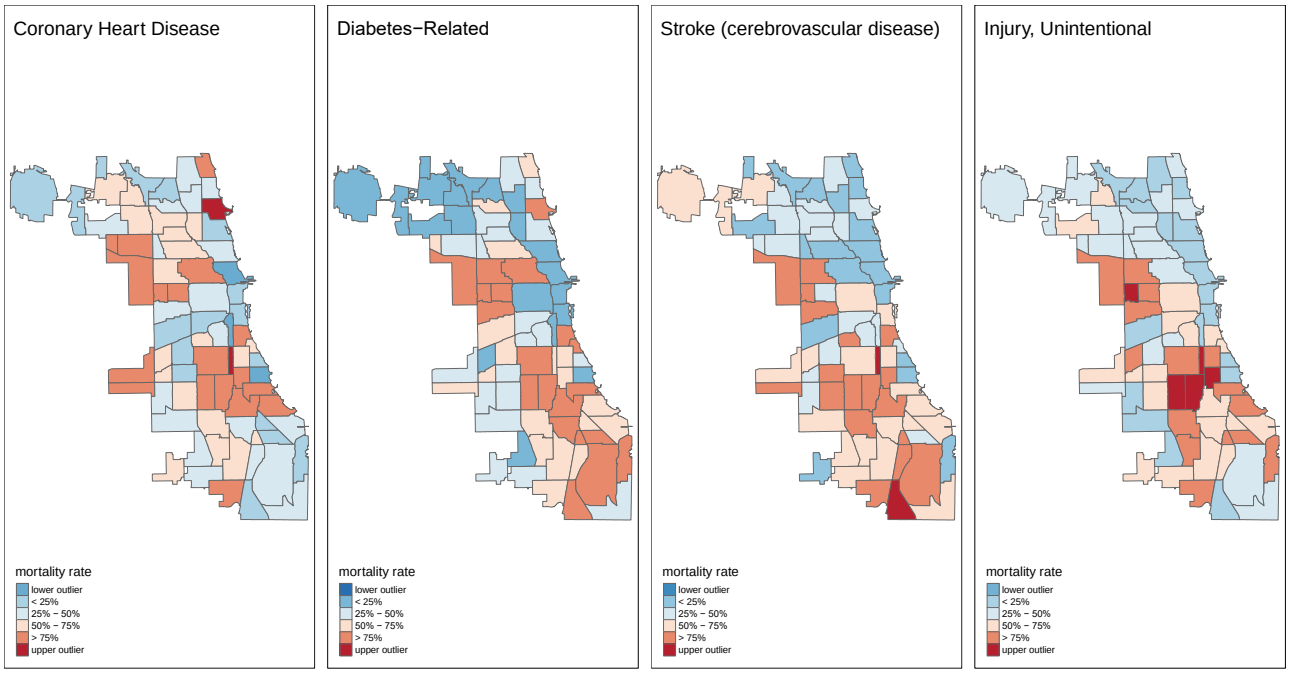

Figure 6. Box map showing average, annual mortality rates (age-adjusted) across the city of Chicago for each underlying cause of death; community area data (2006-2010).

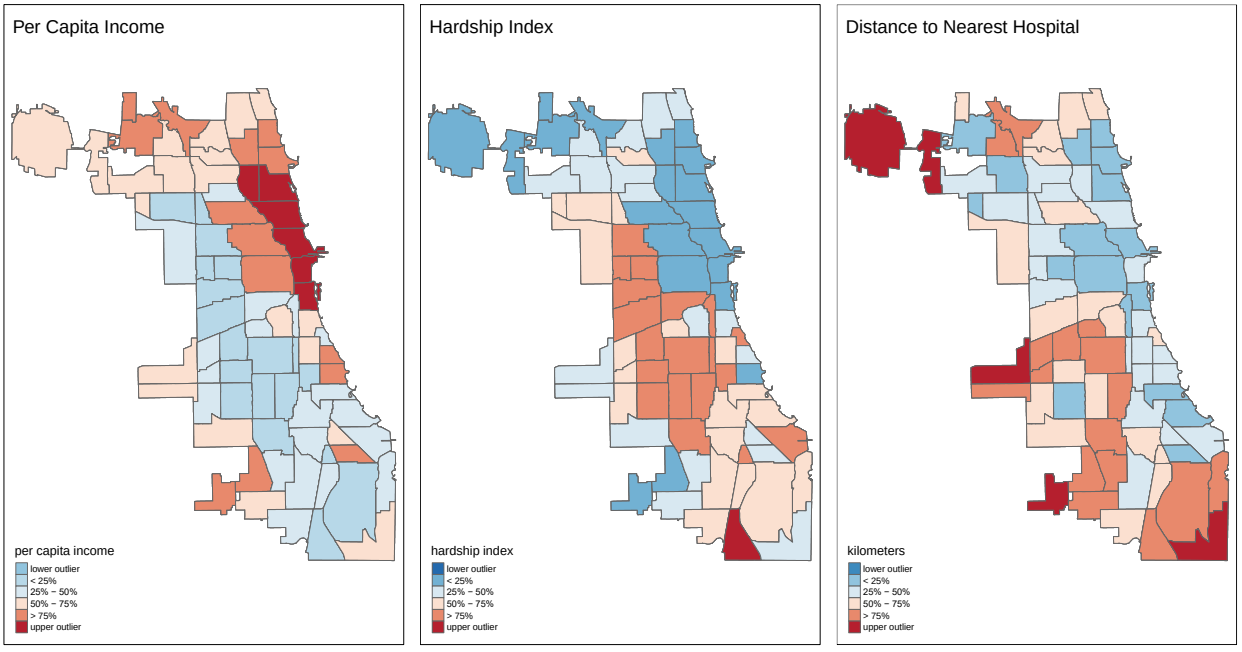

Figure 7. Box map showing data for each socioeconomic variable; community area data for Chicago (2008-2012).

Dependent Variable	Mortality Rate			
	(Model 1)	(Model 2)	(Model 3)	(Model 4)
Independent Variable				
Hardship Index	0.331***	0.199***	0.298***	0.259***
	(0.114)	(0.032)	(0.044)	(0.040)
Distance from Nearest Hospital (km)	-0.786	-0.384	0.732	-0.548
	(1.986)	(0.556)	(0.767)	(0.690)
Constant	129.631***	18.740***	28.419***	20.019***
	(7.988)	(2.239)	(3.084)	(2.775)
Observations	77	77	77	77
R_2	0.104	0.345	0.390	0.369
Adjusted R_2	0.079	0.327	0.374	0.352
Residual Std. Error (df = 74)	28.446	7.972	10.983	9.883
F Statistic (df = 2; 74)	4.275**	19.483***	23.669***	21.658***
Note:				*p**p***p<0.01

Table 4. Results from multivariable regression analysis. Models 1-4 vary in underling cause of death; data is as follows: for model (1) coronary heart disease, (2) diabetes-related, (3) stroke (cerebrovascular disease), and (4) injury, unintentional.

Dependent Variable	Mortality Rates			
Independent Variable	(Model 1)	(Model 2)	(Model 3)	(Model 4)
Per Capita Income	-0.001***	-0.0004***	-0.0005***	-0.0004***
	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Distance from Nearest Hospital (km)	-1.908	-0.882	0.112	-1.107
	(1.942)	(0.566)	(0.839)	(0.744)
Constant	169.343***	39.416***	57.147***	45.381***
	(8.349)	(2.433)	(3.606)	(3.196)
Observations	77	77	77	77
R_2	0.168	0.342	0.292	0.289
Adjusted R_2	0.146	0.325	0.273	0.270
Residual Std. Error (df = 74)	27.403	7.988	11.835	10.491
F Statistic (df = 2; 74)	7.477***	19.265***	15.250***	15.052***
Note:				*p**p***p<0.01

Table 5. Results from multivariable regression analysis. Models 1-4 vary in underling cause of death; data is as follows: model (1) coronary heart disease, (2) diabetes-related, (3) stroke (cerebrovascular disease), and (4) injury, unintentional.

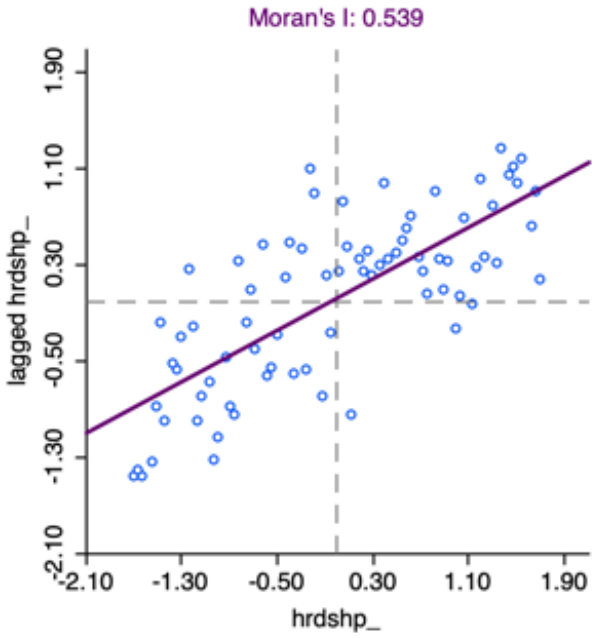

Figure 8. Moran scatter plot for the description of univariate global spatial autocorrelation. Significance was examined by means of a permutation test (~999 permutations). Here, we can observed that values for the hardship index most generally fall within the low-low (L-L) and high-high (H-H) categories, the lower left and upper right quadrants respectively; community area data for Chicago (2008-2012).

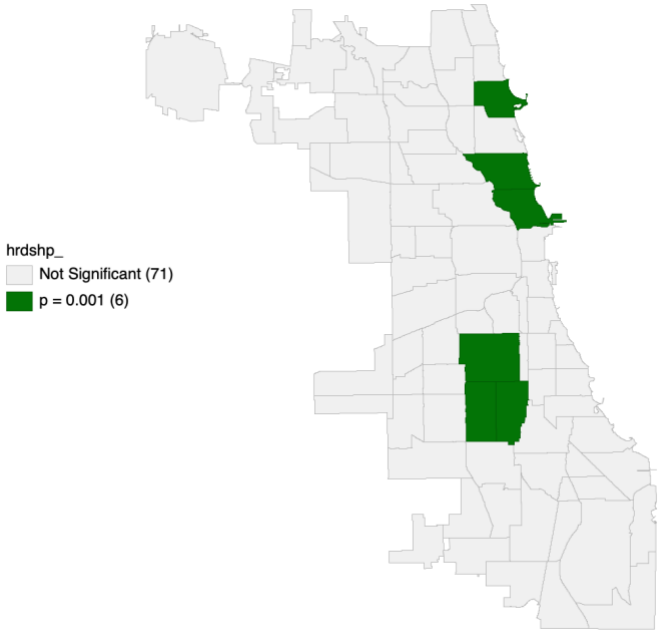

Figure 9. Associated significance map for local, univariate Moran statistic. Significance was examined by means of a permutation test (~999 permutations). Here, we can detect specific clusters of like values for our hardship index (pseudo p value < 0.001); community area data for Chicago (2008-2012).

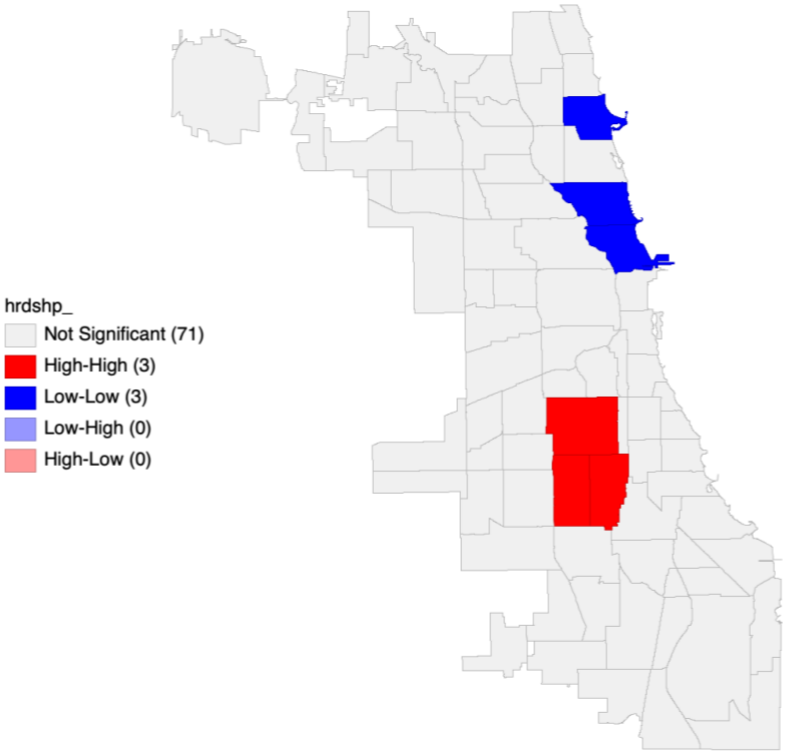

Figure 10. Associated cluster map for local, univariate Moran statistic. Significance was examined by means of a permutation test (~999 permutations). Here, we can detect specific clusters of like values, H-H and L-L, for our hardship index; community area data for Chicago (2008-2012).

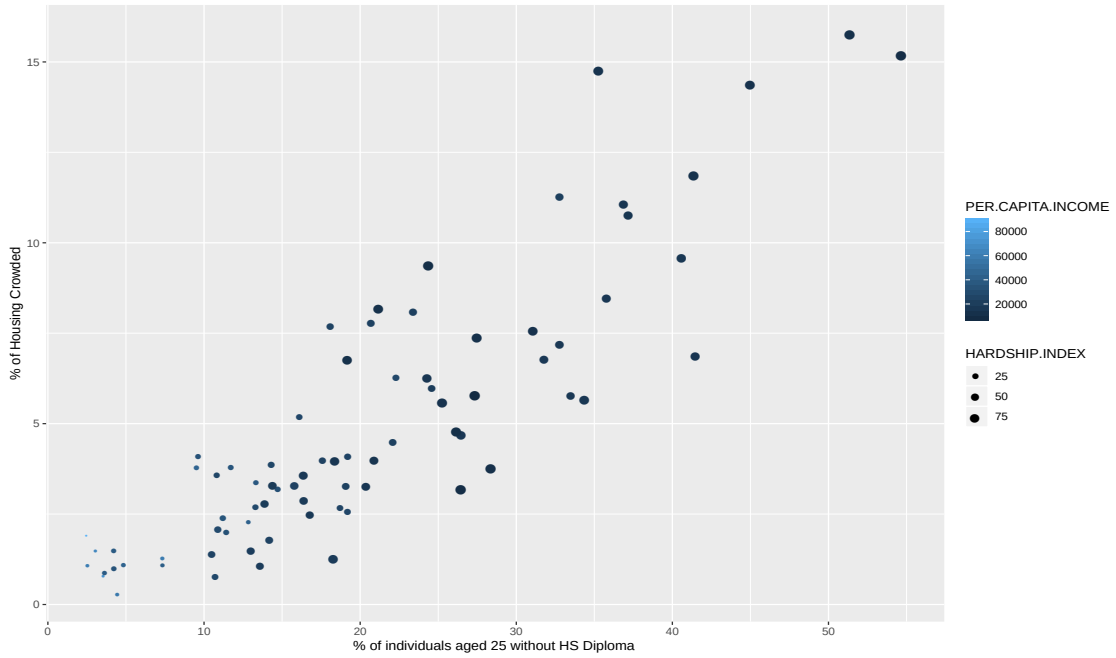

Figure 11. Bubble chart showing the association between some of our socioeconomic variables; community area data for Chicago (2008-2012).

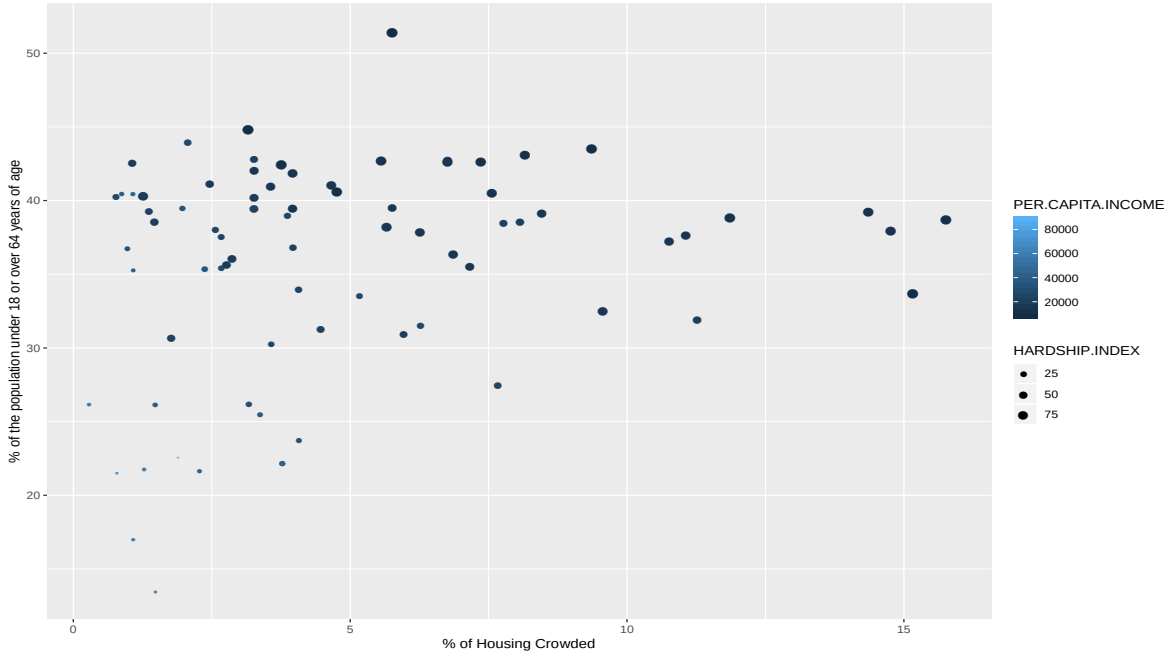

Figure 12. Bubble chart showing the association between some of our socioeconomic variables; community area data for Chicago (2008-2012).

Appendix

List of Chicago Hospitals

	LABEL	FACILITY	ADDRESS	COMMUNITY
1	Kindred- Lakeshore	Kindred Hospital - Chicago Lakeshore	6130 N. Sheridan	EDGEWATER
2	Children's	Children's Memorial Hospital	2300 Children's Plaza	LINCOLN PARK
3	St Bernard	St Bernard Hospital	326 W. 64th Street	ENGLEWOOD
4	Holy Cross	Holy Cross Hospital	2701 W. 68th Street	CHICAGO LAWN
5	IL Masonic	Advocate Illinois Masonic Medical Center	836 W. Wellington	LAKE VIEW
6	St Joseph	Saint Joseph Hospital	2900 N. Lake Shore Drive	LAKE VIEW
7	Roseland	Roseland Community Hospital	45 W. 111th Street	ROSELAND
8	Trinity	Advocate Trinity Hospital	2320 E. 93rd Street	CALUMET HEIGHTS
9	South Shore	South Shore Hospital	8012 S. Crandon	SOUTH CHICAGO
10	La Rabida	La Rabida Children's Hospital	East 65th St. at Lake Michigan	WOODLAWN
11	Jackson Pk	Jackson Park Hospital	Medical Center	7531 S. Stony Island
12	Swedish	Swedish Covenant Hospital	5145 N. California	LINCOLN SQUARE
13	Kindred-North	Kindred Hospital - Chicago North	2544 W. Montrose	LINCOLN SQUARE
14	U of C	University of Chicago Hospital	5841 S. Maryland	HYDE PARK
15	Provident	Provident Hospital of Cook County	500 E. 51st Street	GRAND BOULEVARD
16	Mercy	Mercy Hospital and Medical Center	2525 S. Michigan	NEAR SOUTH SIDE
17	St Anthony	Saint Anthony Hospital	2875 W. 19th Street	SOUTH LAWNDALE
18	Methodist	Methodist Hospital of Chicago	5025 N. Paulina	UPTOWN
19	Lakeshore	Chicago Lakeshore Hospital	4840 N. Marine Drive	UPTOWN

20	Weiss	Louis A. Weiss Memorial Hospital	4646 N. Marine Drive	UPTOWN
21	Thorek	Thorek Hospital and Medical Center	850 W. Irving Park Road	UPTOWN
22	Bethany	Advocate Bethany Hospital	3435 W. Van Buren	EAST GARFIELD PARK
23	Rush	Rush University Medical Center	1653 W. Congress	NEAR WEST SIDE
24	VA-Brown	Jesse Brown Department of Veteran's Affairs Medical Center	820 S. Damen	NEAR WEST SIDE
25	Stroger	John H. Stroger, Jr. Hospital of Cook County	1901 W. Harrison	NEAR WEST SIDE
26	U of I at Chicago	University of Illinois at Chicago Hospital	1740 W. Taylor	NEAR WEST SIDE
27	Mount Sinai	Mount Sinai Hospital	Medical Center	1500 S California
28	Schwab	Schwab Rehabilitation Hospital	1401 S. California	NORTH LAWNDALE
29	Norwegian American	Norwegian American Hospital	1044 N. Francisco	WEST TOWN
30	Sts. Mary and Eliz MC-St. Eliz	Sts Mary and Elizabeth Medical Center-St. Elizabeth	1431 N. Claremont	WEST TOWN
31	Sts. Mary and Eliz MC-St. Mary	Sts Mary and Elizabeth Medical Center-St. Mary	2233 W. Division	WEST TOWN
32	Loretto	Loretto Hospital	645 S. Central Avenue	AUSTIN
33	Hartgrove	Hartgrove Hospital	520 N.Ridgeway	HUMBOLDT PARK
34	Sacred Heart	Sacred Heart Hospital	3240 W. Franklin	HUMBOLDT PARK
35	Our Lady of Resurrection	Our Lady of Resurrection Medical Center	5645 W. Addison	PORTAGE PARK
36	Kindred-Central	Kindred Hospital - Chicago Central	4058 W. Melrose	IRVING PARK
37	Read Mental HC	Chicago Read Mental Health Center	4200 N. Oak Park Avenue	DUNNING
38	Shriners	Shriners Hospital for Children	2211 N. Oak Park Avenue	MONTCLARE
39	Resurrection	Resurrection Medical Center	7435 W. Talcott	NORWOOD PARK

40	Northwestern	Northwestern Memorial Hospital	251 E. Huron	NEAR NORTH SIDE
41	Rehabilitation Institute	Rehabilitation Institute of Chicago	345 E. Superior	NEAR NORTH SIDE
42	Orthopedic Institute of Chicago	Chicago Institute of Neurosurgery and Neuroresearch	4501 WINCHESTER AVENUE	LINCOLN SQUARE

RCODE

```
##############################################
# The purpose of this script is to
# explore a dataset containing location (addresses) of hospitals
# in Chicago, the goal is to calculate distance to the nearest medical center
# for each community area (centroid)
# this measure will be used as a proxy for access to healthcare facilities
##############################################

# load packages
library(sp) #spatial data wrangling & analysis
library(sf) #spatial data wrangling & analysis

library(rgdal) #spatial data wrangling
library(rgeos) #spatial data wrangling & analytics
library(tidyverse) #data wrangling

library(tmap) #modern data visualizations
library(leaflet) #modern data visualizations
library(ggmap) #to obtain lat/lon coords from addresses
```

```
library(geosphere) #to calculate distances
library(tidyr) #data wrangling
library(sjPlot) #create nice data tables
library(stargazer) #create nice data tables

# read data of Chicago hospital locations, last updated August 28, 2011
chicago.hospitals <-
st_read("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Raw_Data/Hospitals/Hospitals.shp")
class(chicago.hospitals)
# check the projection information using st_crs
st_crs(chicago.hospitals)
# list the content of the data frame
dat.for.table <-
read.csv("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Generated_Data/Chicago_hospital_list_address.csv")
stargazer(dat.for.table, summary = FALSE, type = "html", title = "List of Chicago Hospitals", digits = 1, out = "table2.htm")

# remove rehabilitation centers from the list, as well as a neurology center
chicago.hospitals.sub <- subset(chicago.hospitals, chicago.hospitals$FACILITY !=
"Schwab Rehabilitation Hospital" & chicago.hospitals$FACILITY != "Rehabilitation Institute of Chicago" &
                                chicago.hospitals$FACILITY != "Chicago Institute of Neurosurgery and Neuroresearch")
# this should have remove 3 institutions, bringing the number of obs down to 39
```

```
# select variables of interest
chicago.hospitals.sub <- chicago.hospitals.sub[, c("FACILITY", "ADDRESS",
"AREA_NUMBE", "COMMUNITY")]

# to convert addresses to points (lat & long) we will use a function of ggmap
# this function now requires a registered API key
# to obtain an API key and enable services, go to https://cloud.google.com/maps-
platform/

# this sets your google map permanently
# register_google(key = "AIzaSyA05W_1xRb01Wf7jALMGzbjveGvSL7oM1Q", write =
TRUE)
has_google_key()
google_key()

# obtain lat and long for each address, create a new dataset
locs <- as.character(chicago.hospitals.sub$ADDRESS) # list of addresses
hospital.coords <- geocode(locs) # ggmap function
hospital.coords$address <- locs # adds addresses to dataset

# the next step is to obtain shape centers for each of our community areas
# read data of boundaries for current community areas in Chicago
chicago.comm <-
read_sf("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Raw_Dat
a/Boundaries_Community_Areas_current/current_boundaries_community_areas.geojso
n")
class(chicago.comm)
```

```
# check the projection information using st_crs
st_crs(chicago.comm)
# the layer is unprojected in decimal degrees. Also, a quick plot
# note that, by default, sf draws a choropleth map for each variable included in the data
frame
# since we won't be using sf for mapping, we ignore that aspect for now
plot(chicago.comm)

# ensure area numbers are read as integer
class(chicago.comm$area_num_1)
chicago.comm$area_num_1 <- as.integer(chicago.comm$area_num_1)

# obtain shape centers
# the st_centroid function is part of sf (there is no obvious counterpart to the mean
center functionality)
# it creates a point simple features layer and contains all the variables of the original
layer
chicago.centroid <- st_centroid(chicago.comm)

# explore results visually
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
pdf("Chicago_comm_area_centroids_and_hospital_locations.pdf", height = 8, width =
15)
# plot centroid locations (red)
plot1 <- tm_shape(chicago.comm) +
  tm_borders() +
  tm_shape(chicago.centroid) +
```

```
  tm_dots(size = 0.2, col = "red")
# plot hospital locations (blue)
plot2 <- plot1 +
  tm_shape(chicago.hospitals.sub) +
  tm_dots(size = 0.2, col = "blue") +
  tm_layout(main.title = "Distribution of Hospitals Across Chicago",
            main.title.position = "left", main.title.size = 0.85)
# add legend
plot3 <- plot2 +
  tm_add_legend(
    type = c("fill", "symbol", "text", "line"),
    labels = c("shape center", "hospital"),
    col = c("red", "blue"),
    size = 0.60,
    shape = NULL,
    lwd = NULL,
    lty = 0.45,
    text = 0.45,
    alpha = NA,
    border.col = "black",
    border.lwd = 1,
    border.alpha = NA,
    title = "Locations",
    is.portrait = TRUE,
    legend.format = list(),
    reverse = FALSE,
    z = NA,
```

```
    group = NULL
  ) +
  # adjust legend
  tm_layout(legend.position = c("left", "center"), legend.title.size = 0.95)

print(plot3)
dev.off()

# save new formatted dataset as a shapefile
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Generated_Data")
st_write(chicago.centroid, "chicago_comm_area_centroids", driver = "ESRI Shapefile")

# obtain lat and long coordinates for the different centroids
centroid_coords <- do.call(rbind, st_geometry(chicago.centroid)) %>%
  as_tibble() %>% setNames(c("lon","lat"))
# add comunity area information (name and area number)
centroid_coords$community <- chicago.centroid$community
centroid_coords$area_number <- chicago.centroid$area_num_1

# calculate distance to nearest hospital using 'Haversine' Great Circle Distance
# the shortest distance between two points (i.e., the 'great-circle-distance' or 'as the crow flies'), according to the 'haversine method'
# this method assumes a spherical earth, ignoring ellipsoidal effects. As default, estimates are in meters.

# obtains list of names to run loop
```

```
community.area.list <- unique(centroid_coords$community)
hospital.list <- unique(hospital.coords$address)
# create new matrix to save values
to_save <- as.data.frame(matrix(NA, nrow = length(community.area.list), ncol = 3))
colnames(to_save) <-
c("community","closest.hospital.address","distance.to.hospital.km")

for (i in 1:nrow(centroid_coords)) {
  # select a community area
  current_comm_area <- community.area.list[i]
  # select the data for that area
  current_comm_area_dat <- filter(centroid_coords, centroid_coords$community ==
current_comm_area)
  # save the lat and long for that community area
  comm_area_lat <- current_comm_area_dat$lat
  comm_area_lon <- current_comm_area_dat$lon

  # create empty vector to store calculated distances
  distance_vect <- rep(NA, nrow(geocoded))

  for (j in 1:nrow(hospital.coords)) {
    # for every hospital on the list
    current_hospital <- hospital.list[j]
    # obtains it data
    current_hospital_dat <- filter(hospital.coords, hospital.coords$address ==
current_hospital)
    # save its coordinates
```

```
    current_hospital_lat <- current_hospital_dat$lat
    current_hospital_lon <- current_hospital_dat$lon

    # calculate distance to current hospital from community area centroid
    distance_to_current_hospital <- distm(c(comm_area_lon, comm_area_lat),
                                          c(current_hospital_lon, current_hospital_lat), fun =
distHaversine)
    # save it in vector, results are in meters
    distance_vect[j]  <- distance_to_current_hospital
  }
  # obtain the shortest distance to hospital
  shortest_dist_to_hospital <- distance_vect[which(distance_vect ==
min(distance_vect))]
  shortest_dist_to_hospital_km <- shortest_dist_to_hospital/1000
  # save values to dataset
  to_save$community[i] <- current_comm_area
  to_save$closest.hospital.address[i] <- hospital.list[which(distance_vect ==
min(distance_vect))]
  to_save$distance.to.hospital.km[i] <- round(shortest_dist_to_hospital_km, digits = 2)
}

# save data in a new csv
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Generated
_Data")
write.csv(to_save,
"km_distance_to_nearest_hospital_chicago_comm_area_dat_2011.csv", row.names =
FALSE)
```

```
##############################################
# The purpose of this script is to
# consolidate data from three diff datasets
# two which were obtained from the Chicago Data Portal
# in order to create a simple features spatial object
# The hope for this is to create a simple pipeline
# to analyze future spatial data
# for tutorial see
https://spatialanalysis.github.io/lab_tutorials/1_R_Spatial_Data_Handling.html
##############################################


# load packages
```

```
library(sp) #spatial data wrangling & analysis
library(sf) #spatial data wrangling & analysis

library(rgdal) #spatial data wrangling
library(rgeos) #spatial data wrangling & analytics
library(tidyverse) #data wrangling

library(tmap) #modern data visualizations
library(leaflet) #modern data visualizations

# set working directory
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Raw_Data
")

# read and inspect mortality data by underlying cause of death, Chicago 2006-2010 by
community area
mort.dat <-
read.csv("Public_Health_Statistics_selected_underlying_causes_of_death_in_Chicago_2
006_2010.csv")
head(mort.dat)
# subset data to include only variables of interest
mort.dat.sub <- mort.dat %>% select(Community.Area.Name, area_num_1 =
Community.Area, cause_of_death = Cause.of.Death,
                                    average_adjusted_rate = Average.Adjusted.Rate.2006...2010,
Adjusted.Rate.Rank)
# set comm names as uppercase
mort.dat.sub$Community.Area.Name <- toupper(mort.dat.sub$Community.Area.Name)
```

```
# and cause of death as lowercase
mort.dat.sub$cause_of_death <- tolower(mort.dat.sub$cause_of_death)
# ensure area numbers are read as integer
class(mort.dat.sub$area_num_1)

# subset data to include the underlying causes of death for which we are more
interested in
mort.dat.sub <- filter(mort.dat.sub, cause_of_death == "injury, unintentional" |
cause_of_death == "coronary heart disease" |
                cause_of_death == "diabetes-related" | cause_of_death == "stroke
(cerebrovascular disease)")

# remove data for Chicago
mort.dat.sub <- subset(mort.dat.sub, mort.dat.sub$Community.Area.Name !=
"CHICAGO")
comm.list <- unique(mort.dat.sub$Community.Area.Name)
length(comm.list) # should be 77

# set column names to uppercase to match socieconomic data
names(mort.dat.sub) <- toupper(names(mort.dat.sub))

# read and inspect dataset for selected socioeconomic indicators in Chicago
# this dataset contains a selection of six socioeconomic indicators of public health
# significance and a "hardship index" by Chicago community area, for the years 2008-
2012
```

```
socioeconomic.dat <-
read.csv("Census_Data__selected_socioeconomic_indicators_Chicago_2008_2012.csv"
)
head(socioeconomic.dat)
# remove data for all of Chicago
socioeconomic.dat  <- na.omit(socioeconomic.dat )
socioeconomic.dat$COMMUNITY.AREA.NAME <-
toupper(socioeconomic.dat$COMMUNITY.AREA.NAME) #change names to uppercase
to match the other dataset
# select only socioeconomic variables of interes
socioeconomic.dat <- socioeconomic.dat[, c("COMMUNITY.AREA.NAME",
"HARDSHIP.INDEX", "PER.CAPITA.INCOME")]

# add data of distance to nearest (km)
hospit.dis.dat <-
read.csv("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Genera
ted_Data/km_distance_to_nearest_hospital_chicago_comm_area_dat_2011.csv")
colnames(hospit.dis.dat) <- c("COMMUNITY.AREA.NAME", "HOSPITAL.ADDRESS",
"DISTANCE.NEAREST.HOSPITAL")
hospit.dis.dat$COMMUNITY.AREA.NAME <-
toupper(hospit.dis.dat$COMMUNITY.AREA.NAME) #change names to uppercase to
match the other dataset
hospit.dis.dat <- hospit.dis.dat[, c(1,3)] #select variables of interest
# recode mispelled community area name
hospit.dis.dat$COMMUNITY.AREA.NAME[hospit.dis.dat$COMMUNITY.AREA.NAME
== "MONTCLARE"] <- "MONTCLAIRE"
```

```
# merge socioeconomic data
soc.eco.dat <- left_join(socioeconomic.dat, hospit.dis.dat, by =
"COMMUNITY.AREA.NAME")
# recode mispelled community area name
soc.eco.dat$COMMUNITY.AREA.NAME[soc.eco.dat$COMMUNITY.AREA.NAME ==
"WASHINGTON HEIGHT"] <- "WASHINGTON HEIGHTS"

# merge dependent and independent variable data
dat <- left_join(mort.dat.sub, soc.eco.dat , by = "COMMUNITY.AREA.NAME")
summary(dat) #there should not be any NA's

# recode community area name to match those in boundary data file
dat$COMMUNITY.AREA.NAME[dat$COMMUNITY.AREA.NAME == "O'HARE"] <-
"OHARE"
dat$COMMUNITY.AREA.NAME[dat$COMMUNITY.AREA.NAME == "MONTCLAIRE"]
<- "MONTCLARE"

# read data of boundaries for current community areas in Chicago
chicago.comm <-
read_sf("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Raw_Dat
a/Boundaries_Community_Areas_current/current_boundaries_community_areas.geojso
n")
class(chicago.comm)
# check the projection information using st_crs
st_crs(chicago.comm)
# the layer is unprojected in decimal degrees. Also, a quick plot
```

```
# note that, by default, sf draws a choropleth map for each variable included in the data
frame
# since we won't be using sf for mapping, we ignore that aspect for now
plot(chicago.comm)

# changing projections, we will assign the Universal Tranverse Mercator zone 16N,
which is
# the proper one for Chicago, with an ESPG code of 32616
chicago.comm <- st_transform(chicago.comm, 32616)
st_crs(chicago.comm) #check the result

# ensure area numbers are read as integer
class(chicago.comm$area_num_1)
chicago.comm$area_num_1 <- as.integer(chicago.comm$area_num_1)

# spatial join data
to.save <- left_join(chicago.comm, dat, by = c("community" =
"COMMUNITY.AREA.NAME", "area_num_1" = "AREA_NUM_1"))
# change all column names to lowercase
names(to.save) <- tolower(names(to.save))

# basic chloropleth map
tm_shape(to.save) +
  tm_polygons("per.capita.income") #there should not be any NA's

# explore data for diabetes-related deaths
diabetes.dat <- filter(to.save, cause_of_death == "diabetes-related")
```

```
head(diabetes.dat)


# create plot
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
pdf("diabetes_related_deaths_age_adjusted_Chicago_2006_2010.pdf", height = 8, width
= 15)
tm_shape(diabetes.dat) +
  tm_polygons("average_adjusted_rate", title = "average adjusted rate", legend.hist =
FALSE, legend.hist.z = 0) +
  tm_layout(inner.margins = 0,
            legend.text.size = .95,
            legend.title.size = .97,
            legend.position = c("left","bottom"),
            legend.hist.height = .2, legend.hist.width = .3)
dev.off()
# for futher information on any of the following aspects:
# reading and loading a shapefile
# creating choropleth maps for different classifications
# customizing choropleth maps
# selecting appropriate color schemes
# calculating and plotting polygon centroids
# omposing conditional maps
# creating a cartogram
# see the following link:
# https://spatialanalysis.github.io/lab_tutorials/4_R_Mapping.html


# save new formatted dataset as a shapefile
```

```
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Generated_Data")
st_write(to.save,"selected_underlying_cause_of_death_dat_plus_socioeconomic_variables_Chicago_2006_2010", driver = "ESRI Shapefile")

# create shapefiles for each underlying cause of death
to.save.hd <- filter(to.save, cause_of_death == "coronary heart disease")
st_write(to.save.hd,"coronary_heart_disease", driver = "ESRI Shapefile")

to.save.dr <- filter(to.save, cause_of_death == "diabetes-related")
st_write(to.save.dr,"diabetes_related", driver = "ESRI Shapefile")

to.save.st <- filter(to.save, cause_of_death == "stroke (cerebrovascular disease)")
st_write(to.save.st,"stroke_cerebrovascular_disease", driver = "ESRI Shapefile")

to.save.injury <- filter(to.save, cause_of_death == "injury, unintentional")
st_write(to.save.injury,"injury_unintentional", driver = "ESRI Shapefile")
```

```
##############################################
# The purpose of this script is to
# carry out preliminary data exploration
# on our dependent variables, using some of the methods employed in
# the Exploratory Data Analysis I tutorial
# https://spatialanalysis.github.io/lab_tutorials/2_R_EDA_1.html
##############################################

# load packages
library(sp) #spatial data wrangling & analysis
library(sf) #spatial data wrangling & analysis

library(rgdal) #spatial data wrangling
library(rgeos) #spatial data wrangling & analytics
library(tidyverse) #data wrangling

library(tmap) #modern data visualizations
library(leaflet) #modern data visualizations
library(Hmisc) #to utilize the LOWESS smoother
library(sjPlot) #to create nice tables
library(stargazer) #to create nice tables
library(Hmisc) #for correlation matrix with p-values
library(corrplot) #to visualize correlation matrix
library(PerformanceAnalytics) #for correlation chart

# set working directory
```

```
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Generated_Data/selected_underlying_cause_of_death_dat_Chicago_2006_2010")
# read and inspect mortality data by underlying cause of death, Chicago 2006-2010 by community area
dat <- read_sf("selected_underlying_cause_of_death_dat_Chicago_2006_2010.shp")
head(dat)

# select data for specific causes of death:
# injury, unintentional; coronary heart disease; diabetes-related; stroke (cerebrovascular disease)
dat.sub <- filter(dat, cs_f_dt == "injury, unintentional" | cs_f_dt == "coronary heart disease" |
                  cs_f_dt == "diabetes-related" | cs_f_dt == "stroke (cerebrovascular disease)")
head(dat.sub)
causes.list <- unique(dat.sub$cs_f_dt)
print(causes.list) # there should only be four causes of death!!

# remove data for Chicago
dat.sub <- filter(dat.sub, comm != "CHICAGO")
comm.list <- unique(dat.sub$comm)
length(comm.list) # there should be 77 community areas

# analyze the distribution of the average number of deaths for each different cause
# interleaved histograms
ggplot(dat.sub, aes(x = avrg_d_ ,color = cs_f_dt )) +
  geom_histogram(fill = "white", position = "dodge", bins = 55) +
```

```
  theme(legend.position = "top")
# ddd mean lines
library(plyr)
dat.mean <- ddply(dat.sub, "cs_f_dt", summarise, grp.mean = mean(avrg_d_))
head(dat.mean)

p1 <- ggplot(dat.sub, aes(x = avrg_d_ , color = cs_f_dt)) +
  geom_histogram(fill = "white", position = "dodge", bins = 55) +
  geom_vline(data = dat.mean, aes(xintercept = grp.mean, color = cs_f_dt),
             linetype = "dashed") +
  theme(legend.position = "top") +
  xlab("Average Annual Death Rate (age-adjusted)")

p2 <- p1 + labs(fill = "Cause of Death")
p2
# distributions differ greatly by cause of death!!!

# create a boxplot
base.plt <- ggplot(data = dat.sub, aes(x = cs_f_dt, y = avrg_d_))
base.plt + geom_boxplot() +
  xlab("") +
  ggtitle("") +
  xlab("Cause of Death") + ylab("Average Annual Death Rate (age-adjusted)") +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(face = "bold",
color = "black",
                                                          size = 10, angle = 0))
# another version
```

```
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
pdf("boxplot_underlying_cause_of_death_Chicago_2006_2010.pdf", height = 8, width =
12)
base.plt +
  geom_point(color = "black", alpha = 0.5) +
  geom_boxplot(color = "black", fill = "grey", outlier.color = "red", alpha = 0.5) +
  stat_boxplot(geom = "errorbar") +
  xlab("") +
  ggtitle("") +
  xlab("Cause of Death") + ylab("Average Annual Death Rate (age-adjusted)")
dev.off()

# create descriptive statistics table for dependent variables
# subset data for table
table.dat <- dat.sub[, c("commnty", "cs_f_dt", "avrg_d_")]
# summarize data
table.dat.sum <- to.save %>%
  group_by(cs_f_dt) %>%
  summarize(mean = round(mean(avrg_d_), digits = 0), SD = round(sd(avrg_d_), digits =
0), median = round(median(avrg_d_), digits = 0),
            min = round(min(avrg_d_), digits = 0), max = round(max(avrg_d_), digits = 0))
# rename column
table.dat.sum <- table.dat.sum %>%
  rename(Cause.of.Death = cs_f_dt)
# create table
tab_df(table.dat.sum,
       title = "Descriptive statistics for dependent variables",
```

```
      file = "table2_sum_stats_dpv.doc")

# bivariate analysis: the scatter plot
dat.sub <- dat.sub %>%
  rename(Cause.of.Death = cs_f_dt) #rename column

setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
pdf("scatter_plot_underlying_cause_of_death_YPLL_Chicago_2006_2010.pdf", height =
8, width = 12)
p3 <- ggplot(data = dat.sub, aes(x = avrg_d_, y = YPLL, color = Cause.of.Death)) +
  geom_point() +
  xlab("Average Annual Death Rate (age-adjusted)") +
  ylab("Average Annual Years of Potential Life Lost \n(per 100,000 residents) ") +
  ggtitle("")
#coord_fixed(ratio = 55.0/25.0)
print(p3)
dev.off()
# comparison of smoothing methods
ggplot(data = dat.sub, aes(x = avrg_d_, y = YPLL_rn)) +
  stat_plsmo(aes(color = "lowess")) +
  geom_point() +
  geom_smooth(aes(color = "lm"), method = lm, se = FALSE) +
  geom_smooth(aes(color = "loess"), method = loess,se = FALSE) +
  xlab("Average Number of Deaths lost to Diabetes-related illnesses") +
  ylab("Average Annual Years of Possible Life Lost (rank)") +
  ggtitle("Comparison of Smoothing Methods") +
  theme(plot.title = element_text(hjust = 0.5)) +
```

```
  labs(color = "Method")

# Note, I need to find a way to differentiate among community areas (maybe North vs South?)
# in order to carry out tests of breaks in spatial heterogeneity









################################################
# The purpose of this script is to
# carry out preliminary data exploration
# on some of the independent variables from our dataset using some of the methods employed in
# the Exploratory Data Analysis II tutorial
# https://spatialanalysis.github.io/lab_tutorials/3_R_EDA_2.html
```

```
##############################################

# load packages
library(tidyverse) #data wrangling
library(GGally) #add-on package to create a scatterplot matrix and parallel coordinate plot
library(scatterplot3d) #to create a static 3d scatter plot
library(plotly) #to construct interactive 3d scatter and parallel coordinate plots
library(Hmisc) #for correlation matrix with p-values
library(corrplot) #to visualize correlation matrix
library(PerformanceAnalytics) #for correlation chart

# set working directory
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Raw_Data")
# read and inspect dataset for selected socioeconomic indicators in Chicago
# this dataset contains a selection of six socioeconomic indicators of public health
# significance and a "hardship index" by Chicago community area, for the years 2008-2012
socioeconomic.dat <-
read.csv("Census_Data__selected_socioeconomic_indicators_Chicago_2008_2012.csv")
head(socioeconomic.dat)
# remove data for all of Chicago
socioeconomic.dat  <- na.omit(socioeconomic.dat )
```

```
socioeconomic.dat$COMMUNITY.AREA.NAME <-
toupper(socioeconomic.dat$COMMUNITY.AREA.NAME) #change names to uppercase
to match the other dataset

# add data of distance to nearest (km)
hospit.dis.dat <-
read.csv("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Genera
ted_Data/km_distance_to_nearest_hospital_chicago_comm_area_dat_2011.csv")
colnames(hospit.dis.dat) <- c("COMMUNITY.AREA.NAME", "HOSPITAL.ADDRESS",
"DISTANCE.NEAREST.HOSPITAL")
hospit.dis.dat$COMMUNITY.AREA.NAME <-
toupper(hospit.dis.dat$COMMUNITY.AREA.NAME) #change names to uppercase to
match the other dataset

# merge datasets
dat <- left_join(socioeconomic.dat, hospit.dis.dat, by = "COMMUNITY.AREA.NAME")
# shorten column names
names(dat)
colnames(dat) <- c("AREA.N", "COMMUNITY","PHC", "PHBP", "PA16U", "PA25WHS",
"PAU18OV64", "PCI", "HI", "HOSPA", "DNH")

# arrange data for correlation matrix
table.dat <- dat[, -c(1,2, 10)]

# basic scatter plot matrix
ggscatmat(table.dat)
# another version
```

```
ggpairs(table.dat)

# correlation tests for independent variables
res <- cor(table.dat, method = "pearson", use = "complete.obs")
round(res, 2)

# correlation with significance values
res2 <- rcorr(as.matrix(table.dat), type = "pearson")
res2

# extract the correlation coefficients
res2$r
# extract p-values
res2$P

# visualize correlation
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)

setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
pdf("corr_plot_independent_variables_.pdf", height = 8, width = 12)
# insignificant correlation are crossed
corrplot(res2$r, type = "upper", order = "hclust",
         p.mat = res2$P, sig.level = 0.01, insig = "blank")
# insignificant correlations are leaved blank
corrplot(res2$r, type = "upper", order = "hclust",
         p.mat = res2$P, sig.level = 0.01, insig = "blank")
```

```
dev.off()

?corrplot()

# use heat map
col <- colorRampPalette(c("blue", "white", "red"))(20) # get some colors
heatmap(x = res, col = col, symm = TRUE)

# use correlation chart
pdf("corr_chart_independent_variables_.pdf", height = 8, width = 12)
chart.Correlation(table.dat, histogram = TRUE, pch = 19)
dev.off()

# to interpret its value, see which of the following values your correlation r is closest to:
# exactly –1. A perfect downhill (negative) linear relationship
# –0.70. A strong downhill (negative) linear relationship
# –0.50. A moderate downhill (negative) relationship
# –0.30. A weak downhill (negative) linear relationship
# 0. No linear relationship
# 0.30. A weak uphill (positive) linear relationship
# 0.50. A moderate uphill (positive) relationship
# 0.70. A strong uphill (positive) linear relationship
# exactly +1. A perfect uphill (positive) linear relationship

# pairwise scatter plots
```

```
ggpairs(dat, columns =
c("PERCENT.OF.HOUSING.CROWDED","PERCENT.HOUSEHOLDS.BELOW.POVERTY",
                  "PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA",
"PER.CAPITA.INCOME", "HARDSHIP.INDEX"),
      upper = list(continuous = "points"), diag = list(continuous = "barDiag"))

# scatter plot matrix with loess smoother
ggpairs(dat, columns =
c("PERCENT.OF.HOUSING.CROWDED","PERCENT.HOUSEHOLDS.BELOW.POVERTY",
                  "PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA",
"PER.CAPITA.INCOME", "HARDSHIP.INDEX"),
      upper = list(continuous = "smooth_loess"), lower = list(continuous =
"smooth_loess"))


# bubble chart
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
pdf("bubble_chart.pdf", height = 8, width = 12)
ggplot(data = dat, aes(x = PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA,
y = PERCENT.OF.HOUSING.CROWDED, size = HARDSHIP.INDEX, col =
PER.CAPITA.INCOME)) +
  geom_point() +
  xlab("% of individuals aged 25 without HS Diploma") +
  ylab("% of Housing Crowded") +
  ggtitle("") +
```

```
  theme(plot.title = element_text(hjust = 0.5))
dev.off()

# bubble chart V2
pdf("bubble_chart2.pdf", height = 8, width = 12)
ggplot(data = dat, aes(x = PERCENT.OF.HOUSING.CROWDED, y =
PERCENT.AGED.UNDER.18.OR.OVER.64  , size = HARDSHIP.INDEX, col =
PER.CAPITA.INCOME)) +
  geom_point() +
  xlab("% of Housing Crowded") +
  ylab("% of the population under 18 or over 64 years of age") +
  ggtitle("") +
  theme(plot.title = element_text(hjust = 0.5))
dev.off()

# parallel coordinate plot (PCP)
vars <-
c("PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA","PERCENT.OF.HOUSI
NG.CROWDED",
        "PER.CAPITA.INCOME","HARDSHIP.INDEX")
pcp.vars <- select(dat,vars)
ggparcoord(data = pcp.vars)
# consider reordering the columns to minimize the number of cross between series
# scaling the variables might also make comparisons easier


# create conditional plot with LOESS Smoother
```

```
# the facetting formula does not evaluate functions, so the conditioning categories need to be computed beforehand
# there are three so-called helper functions to make this easy: cut_interval, cut_width, and cut_number
# the closest to the median (2 quantiles) conditioning illustrated in the GeoDa Workbook is the cut_number function
# variable will be split in two groups on the median value
dat$xcut <- cut_number(dat$PERCENT.OF.HOUSING.CROWDED, n = 2) # condition variable for the x-axis
dat$ycut <- cut_number(dat$PER.CAPITA.INCOME, n = 2) # condition variable for the y-axis


ggplot(data = dat, aes(x = PERCENT.AGED.UNDER.18.OR.OVER.64 , y = HARDSHIP.INDEX )) +
  geom_point() +
  geom_smooth(method = "loess") +
  facet_grid(ycut ~ xcut, as.table = FALSE) # make sure to have the right order!!
```

```
##############################################
# The purpose of this script is to
# create geo-visualizations for our data
# for tutorial see https://spatialanalysis.github.io/lab_tutorials/4_R_Mapping.html
##############################################

# load packages
library(sp) #spatial data wrangling & analysis
library(sf) #spatial data wrangling & analysis

library(rgdal) #spatial data wrangling
library(rgeos) #spatial data wrangling & analytics
library(tidyverse) #data wrangling

library(tmap) #modern data visualizations
library(leaflet) #modern data visualizations

library(tmap) #for mapping
library(RColorBrewer) #to create colors
library(cartogram) #for mapping
```

```
################################
# LOAD DATA
################################
dat <-
st_read("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Generated_Data/selected_underlying_cause_of_death_dat_plus_socioeconomic_variables_Chicago_2006_2010/selected_underlying_cause_of_death_dat_plus_socioeconomic_variables_Chicago_2006_2010.shp")

# inspect column names
names(dat)

# basic chloropleth map example
tm_shape(dat) +
  tm_polygons("pr_cpt_") #there should not be any NA's

################################
# BOX MAP
################################
```

```r
# function to get variables from shapefile
get.var <- function(vname, df) {
  # function to extract a variable as a vector out of an sf data frame
  # arguments:
  #    vname: variable name (as character, in quotes)
  #    df: name of sf data frame
  # returns:
  #    v: vector with values (without a column name)
  v <- df[vname] %>% st_set_geometry(NULL)
  v <- unname(v[,1])
  return(v)
}

# function to computete the box map break points
boxbreaks <- function(v, mult = 1.5) {
  # break points for box map
  # arguments:
  #   v: vector with observations
  #   mult: multiplier for IQR (default 1.5)
  # returns:
  #   bb: vector with 7 break points
  # compute quartile and fences
  qv <- unname(quantile(v))
  iqr <- qv[4] - qv[2]
  upfence <- qv[4] + mult * iqr
  lofence <- qv[2] - mult * iqr
```

```
  # initialize break points vector
  bb <- vector(mode = "numeric", length = 7)
  # logic for lower and upper fences
  if (lofence < qv[1]) {  # no lower outliers
    bb[1] <- lofence
    bb[2] <- floor(qv[1])
  } else {
    bb[2] <- lofence
    bb[1] <- qv[1]
  }
  if (upfence > qv[5]) { # no upper outliers
    bb[7] <- upfence
    bb[6] <- ceiling(qv[5])
  } else {
    bb[6] <- upfence
    bb[7] <- qv[5]
  }
  bb[3:5] <- qv[2:4]
  return(bb)
}

# fucntion to plot boxmap
boxmap <- function(vnam, df, legtitle = NA, mtitle = "Box Map", mult = 1.5) {
  # box map
  # arguments:
  #   vnam: variable name (as character, in quotes)
  #   df: simple features polygon layer
```

```
  #   legtitle: legend title
  #   mtitle: map title
  #   mult: multiplier for IQR
  # returns:
  #   a tmap-element (plots a map)
  var <- get.var(vnam, df)
  bb <- boxbreaks(var)
  tm_shape(df) +
    tm_fill(vnam,title = legtitle,breaks = bb,palette = "-RdBu",
            labels = c("lower outlier", "< 25%", "25% - 50%", "50% - 75%","> 75%", "upper
outlier"))  +
    tm_borders() +
    tm_layout(title = mtitle, title.position = c("right","bottom"))
}

#explore box plot for main independent variables
plot_1 <- boxmap("pr_cpt_", dat , mtitle = "", legtitle = "per capita income", mult = 1.5) +
  tm_layout(title = "Per Capita Income", title.position = c("left","top"))
plot_2 <- boxmap("hrdshp_", dat , mtitle = "", legtitle = "hardship index", mult = 1.5) +
  tm_layout(title = "Hardship Index", title.position = c("left","top"))
plot_3 <- boxmap("dstnc__", dat , mtitle = "", legtitle = "kilometers", mult = 1.5) +
  tm_layout(title = "Distance to Nearest Hospital", title.position = c("left","top"))

# explore data for coronary heart disease
hdisease.dat <- filter(dat, cs_f_dt == "coronary heart disease")
# box map for mortality rates
setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")
```

```
pdf("box_map_heart_disease.pdf", height = 8, width = 15)

plot1 <- boxmap("avrg_d_", hdisease.dat, mtitle = "", legtitle = "mortality rate", mult =
1.5) +
tm_layout(title = "Coronary Heart Disease", title.position = c("left","top"))
print(plot1)
dev.off()

# explore data for diabetes-related deaths
diabetes.dat <- filter(dat, cs_f_dt == "diabetes-related")
# box map for mortality rates
pdf("box_map_diabetes_related.pdf", height = 8, width = 15)
plot2 <- boxmap("avrg_d_", diabetes.dat, mtitle = "", legtitle = "mortality rate", mult = 1.5)
+
tm_layout(title = "Diabetes-Related", title.position = c("left","top"))
print(plot2)
dev.off()

# explore data for stroke (cerebrovascular disease)
stroke.dat <- filter(dat, cs_f_dt == "stroke (cerebrovascular disease)")
# box map for mortality rates
pdf("box_map_stroke.pdf", height = 8, width = 15)
plot3 <- boxmap("avrg_d_", stroke.dat, mtitle = "", legtitle = "mortality rate", mult = 1.5) +
tm_layout(title = "Stroke (cerebrovascular disease)", title.position = c("left","top"))
print(plot3)
dev.off()
```

```
# explore data for injury, unintentional
injury.dat <- filter(dat, cs_f_dt == "injury, unintentional")
# box map for mortality rates
pdf("box_map_heart_injury.pdf", height = 8, width = 15)
plot4 <- boxmap("avrg_d_", injury.dat, mtitle = "", legtitle = "mortality rate", mult = 1.5) +
tm_layout(title = "Injury, Unintentional", title.position = c("left","top"))
print(plot4)
dev.off()
```

```
################################
# PLACE PLOTS ON THE SAME PANEL
################################
pdf("mortality_rates_box_maps_panel_view.pdf", height = 8, width = 15)
current.mode <- tmap_mode("plot")
tmap_arrange(plot1, plot2, plot3, plot4)
tmap_mode(current.mode)
dev.off()

pdf("socioeconomic_variables_box_maps_panel_view.pdf", height = 8, width = 15)
current.mode <- tmap_mode("plot")
tmap_arrange(plot_1, plot_2, plot_3)
tmap_mode(current.mode)
dev.off()
```

```
################################
# SD MAP
################################
tm_shape(hdisease.dat) +
  tm_fill("avrg_d_", title = "", style = "sd",palette = "-RdBu")  +
  tm_borders() +
  tm_layout(title = "Standard Deviation Map", title.position = c("right","bottom"))

################################
# REGRESSION ANAYLIS
################################
# fit regression model for each underlying cause of death
m1 <- lm(avrg_d_ ~ hrdshp_ + dstnc__ , data = hdisease.dat)
summary(m1)
m2 <- lm(avrg_d_ ~ hrdshp_ + dstnc__ , data = diabetes.dat)
summary(m2)
m3 <- lm(avrg_d_ ~ hrdshp_ + dstnc__ , data = stroke.dat)
summary(m3)
m4 <- lm(avrg_d_ ~ hrdshp_ + dstnc__ , data = injury.dat)
summary(m4)

setwd("/Users/jesuscantu/Desktop/U_Chicago/Spatial_Data_Science_Project/Figures")

library(stargazer)
stargazer(m1, m2, m3, m4, type = "html",
          dep.var.labels = c("(1)", "(2)", "(3)", "(4)"),
```

```
        covariate.labels = c("Hardship Index", "Distance from Nearest Hospital (km)"),
        dep.var.caption = "Mortality Rates",
        out = "reg_models1.doc")

# fit regression model for each underlying cause of death
m1 <- lm(avrg_d_ ~ pr_cpt_ + dstnc__ , data = hdisease.dat)
summary(m1)
m2 <- lm(avrg_d_ ~ pr_cpt_ + dstnc__ , data = diabetes.dat)
summary(m2)
m3 <- lm(avrg_d_ ~ pr_cpt_ + dstnc__ , data = stroke.dat)
summary(m3)
m4 <- lm(avrg_d_ ~ pr_cpt_ + dstnc__ , data = injury.dat)
summary(m4)

stargazer(m1, m2, m3, m4, type = "html",
        dep.var.labels = c("(1)", "(2)", "(3)", "(4)"),
        covariate.labels = c("Per Capita Income", "Distance from Nearest Hospital (km)"),
        dep.var.caption = "Mortality Rates",
        out = "reg_models2.doc")


################################
# EXTRA TABLES
################################
table.dat <- stroke.dat[, c("commnty", "hrdshp_", "avrg_d_", "pr_cpt_")]

# sort data by per capita income (descending)
```

```
table.dat.st <- table.dat[order(-table.dat$hrdshp_),]
table.dat.st.sub <- table.dat.st[c(1:15), ] #select first 20 rows
```