Exploring Health and Food Influences of Walkability

Temi Kassim, Megan Klein, and Maddison Clancy

Spring 2024

Overview and Goal of Study

The goal of this study is to investigate the question, "Does accessibility to food and healthcare connect to an area's walkability index and what can be improved?"

We were particularly interested in this question as Chicago is a neighboring city to South Bend that many of us plan to move to in the following years. We were curious about how location affects a person's access to healthcare and food, directly impacting their lifestyle. We wanted to look further into the relationship between areas with food insecurity and their accessibility to healthcare to get a deeper understanding of how walkability may play a role.

Data Description

National Walkability Index

The main dataset is from the EPA Smart Location Database, version 3, dated January 2021, comprises a collection of variables aimed at assessing locations based on urban transportation factors. It covers a wide range of data points across 220,740 entries, each of which represent different geographic areas within the US. There are 117 columns, and these include both numerical and categorical data types. We narrowed down this data to relevant variables that relate to food and healthcare. We also combined three other datasets "US_FIPS_Codes.csv", "Chicago Health Atlas Data" and "Chicago_CensusTracts2010" data to help us answer our question. Finally, after data cleaning and manipulation, we were left with 38 columns and 2,185 rows.

The Data Dictionary can be referenced here: https://docs.google.com/document/d/1nbo9Qp6IDW9rJSgrV38W-hXUsyy8n0B_uqH1spD7PpA/edit?usp=sharing

Key Aspects: The EKW_2022 and NatWalkInd variables provide a summary score reflecting the walkability and pedestrian-friendliness of an area. Thus, these metrics are a crucial component of smart urban planning.

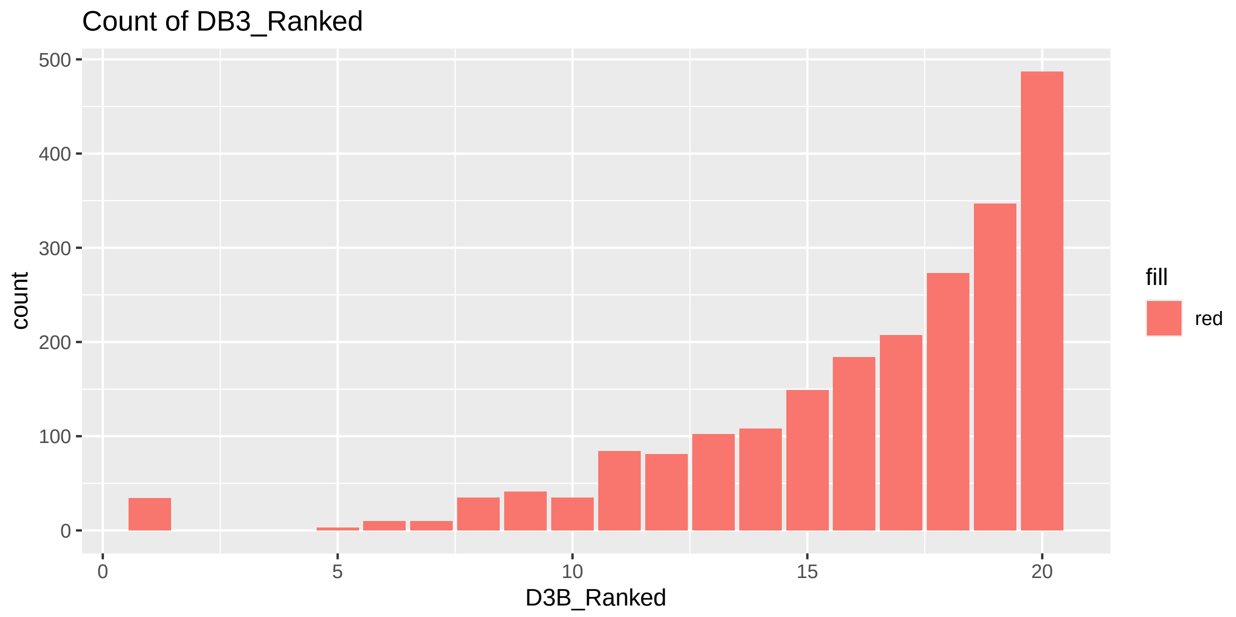
Quick Look Into Some Key Variables:

The dataset includes geographic identifiers like GEOID10, STATEFP, COUNTYFP, and TRACTCE, which are crucial for mapping and spatial analysis. Columns like CSA and CSA_Name (Combined Statistical Area) and CBSA (Core Based Statistical Area) provide insights into the urban or rural classification of regions, critical for understanding the spatial context of the data. Indicators like D2A_Ranked, D2B_Ranked, D3B_Ranked, and D4A_Ranked offer ranked assessments of the areas based on specific criteria, aiding in comparative analyses across different regions.

Exploratory Data Analysis (EDA)

First, we wanted to look at the D3B and D3B Ranking variables which represent Street Intersection Density. The D3B Ranked variable is essentially considering the D3B and ranking it on a scale from 0-20. We were curious about these two variables as we know that the Ranked variables contribute to NatWalkInd thanks to our data guide.

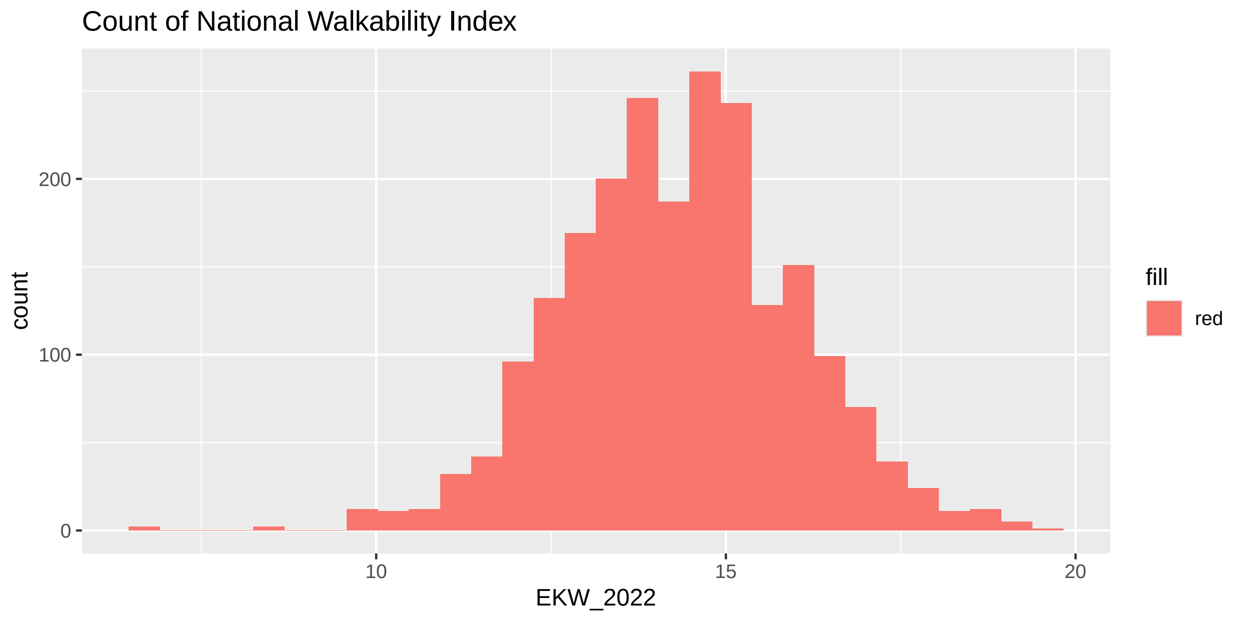
Let's take a look at the distribution of the D3B_Ranked variables to see how prevalent each ranking is.



The bar chart of ranked D3B shows that interconnected density rankings (D3B_Ranked) are most commonly 20. Thus, we have to be careful when interpreting results with this variable, as the distribution of this ranking is very left skewed.

Next we want to look at the distribution of the National Walkability Index:

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



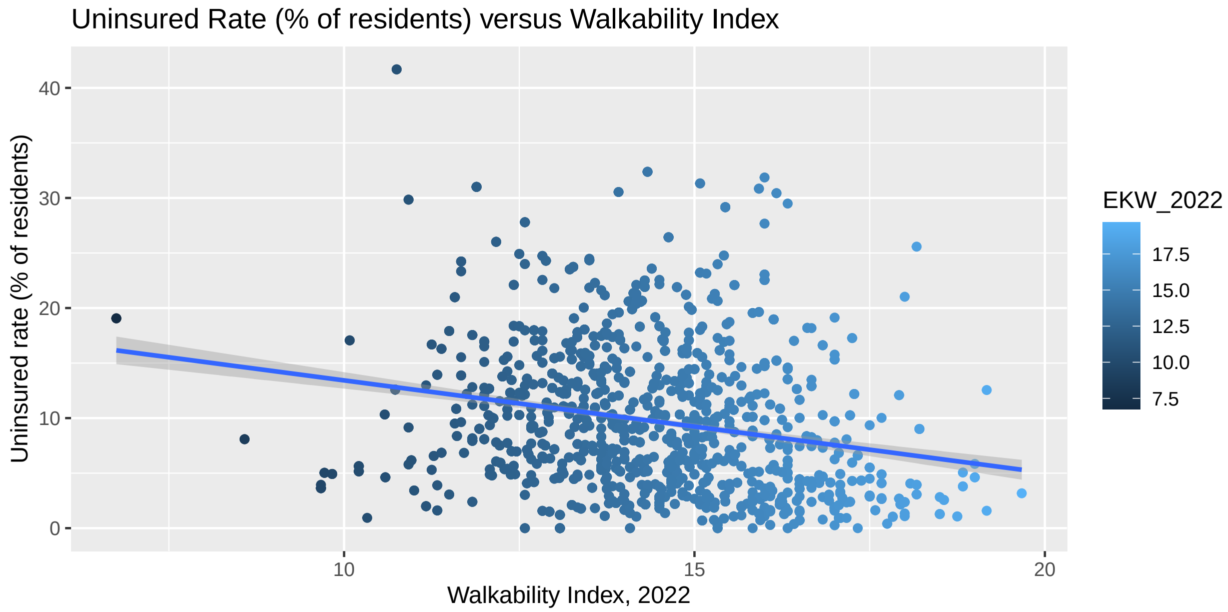
This histogram shows that the National Walkability Index from 2022 is normally distributed but not around the expected mean (rating of 10). Unexpectedly, the distribution lies mostly higher than 10, indicating decent to moderate walkability. This plot also shows that there may be a few outliers.

We took a quick look at the distribution of the National Walkability Index in 2022 and found the min to be 6.75 while the max was 19.67. The scale of the National Walkability Index ranges from 0 to 20, with 20 being most walkable.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    6.75   13.25   14.33   14.33   15.33   19.67       3 
```

Lets look at the relationship between Healthcare and the National Walkability Index (EKW_2022), specifically the rate of those who are uninsured (UNS_2018.2022).

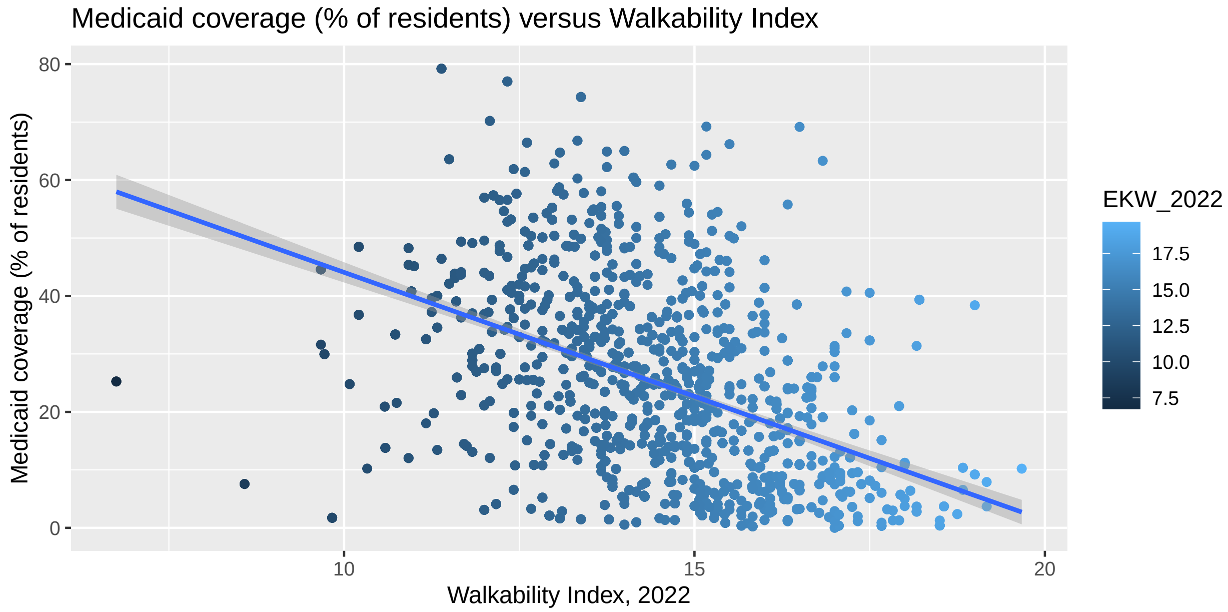
```
## `geom_smooth()` using formula = 'y ~ x'
```



We see that the graph tells us that there is a slight negative linear relationship between the two. This indicates that as the walkability index increases the uninsured rate of residents decreases, suggesting that there are less people uninsured in highly walkable areas.

Next lets look at the Medicaid Coverage Rate against the National Walkability Index:

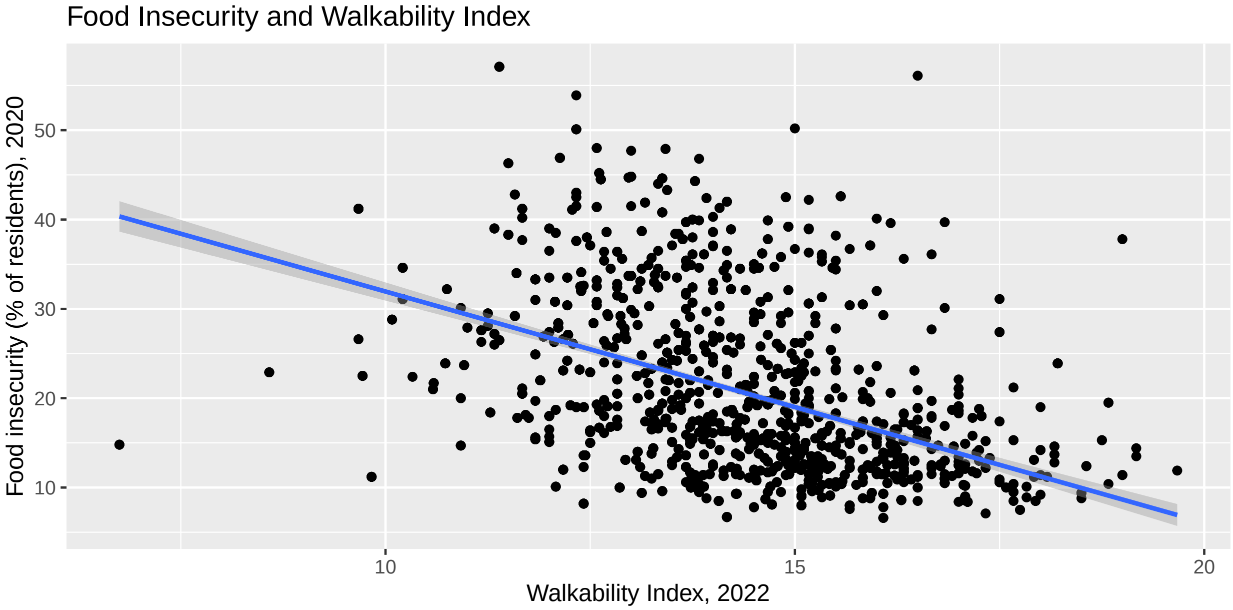
```
## `geom_smooth()` using formula = 'y ~ x'
```



Here there appears to be a clear negative linear relationship. This suggests that as the walkability of an area increases, the rate of individuals that are covered by medicaid decreases, suggesting that there are less people covered by medicaid in highly walkable areas.

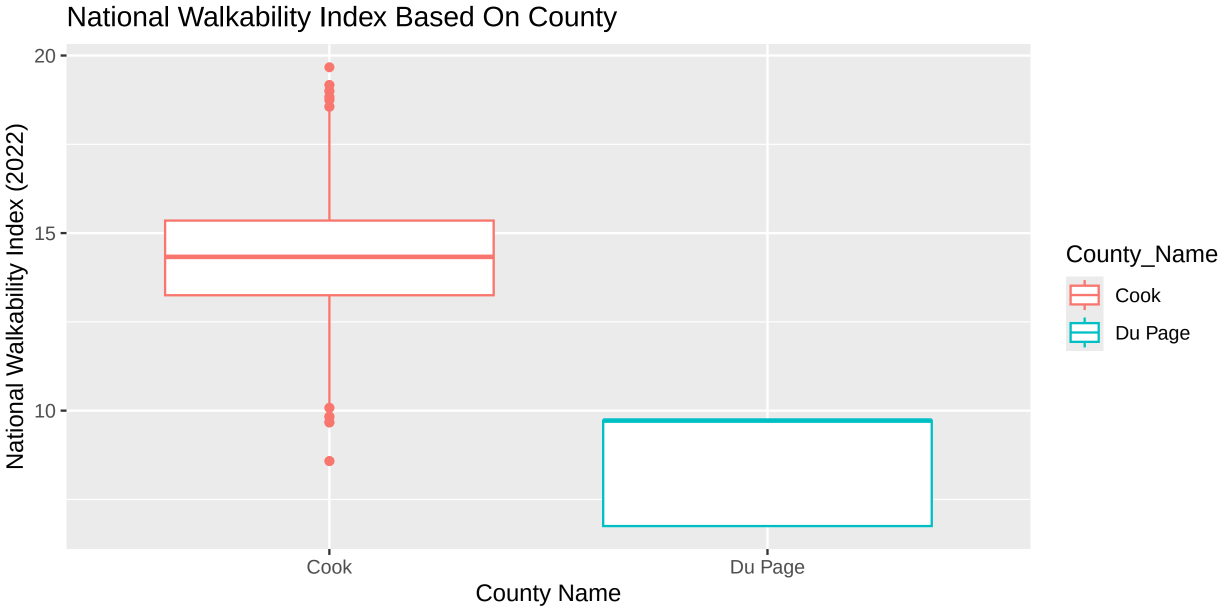
Next Food Insecurity Rate versus the National Walkability Index:

```
## 'geom_smooth()' using formula = 'y ~ x'
```



There also appears to be a negative linear relationship. Again suggesting that as areas become more walkable, the rate of food insecurity decreases.

Next, lets view the distribution of the National Walkability Index in 2022 based on counties, notice that there are only two counties (Cook and DuPage) available. This may be interesting to research as we see the counties have slightly different linear relationships to the percent of low wage workers.

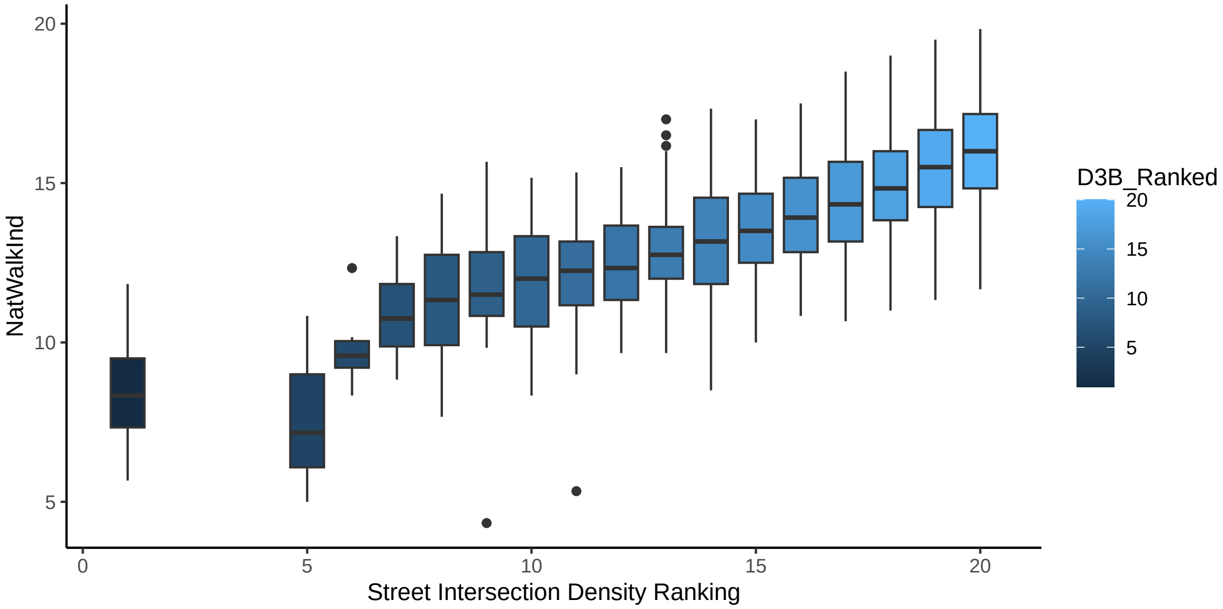


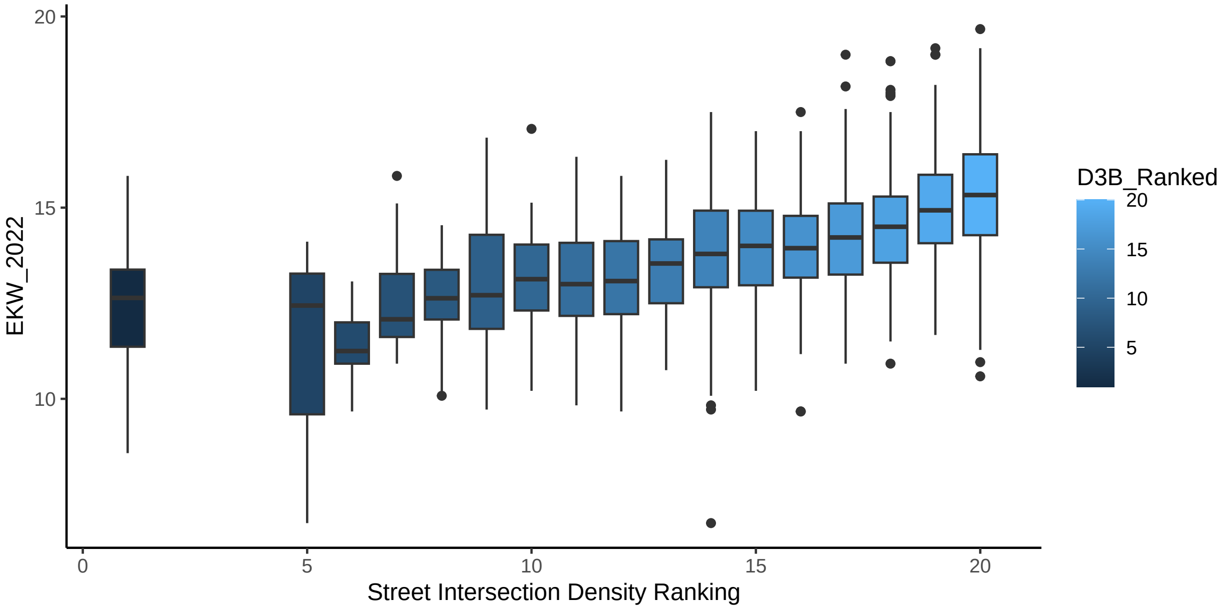
```
## # A tibble: 2 x 2
##   County_Name ‘n()‘
##   <chr>       <int>
## 1 Cook         2185
## 2 Du Page         5
```

Looking into the differences between counties could prove interesting in understanding factors and interactions that impact walkability. Based on the pivot table, it is clear that we have a biased amount of data favoring Cook county, so further analysis may not be possible. This may be a limitation later as DuPage may not be accurately represented by our findings.

Lets look at variables' relationship with the old National Walkability Index Variable (NatWalkInd):

Walkability Per Street Intersection Density Rankings (D3B_Ranked) Compared to National Walkability Index (NatWalkInd and EKW_2022) :





These boxplots visualize the relationship between street intersection density and our original NatWalkInd variable as well as our chosen one , EKW_2022. Each box represents the IQR of the NavWalkInd/EKW_2022 for a specific street intersection density rank. The color gradient from lighter to darker shades as the density ranking increases indicate higher values of rankings.

These plots show a pretty strong correlation between the complexity of street intersections and walkability, suggesting that as street intersections become more densely ranked, the variation and median values of the walkability index might change.

By choosing EKW_2022 vs NatWalkInd, it shows that these metrics are essentially the same, but there could be some colineariaty between the two variables, intuitively because they are measuring the same thing but come from different sources. This may pose issues in the future when modeling.

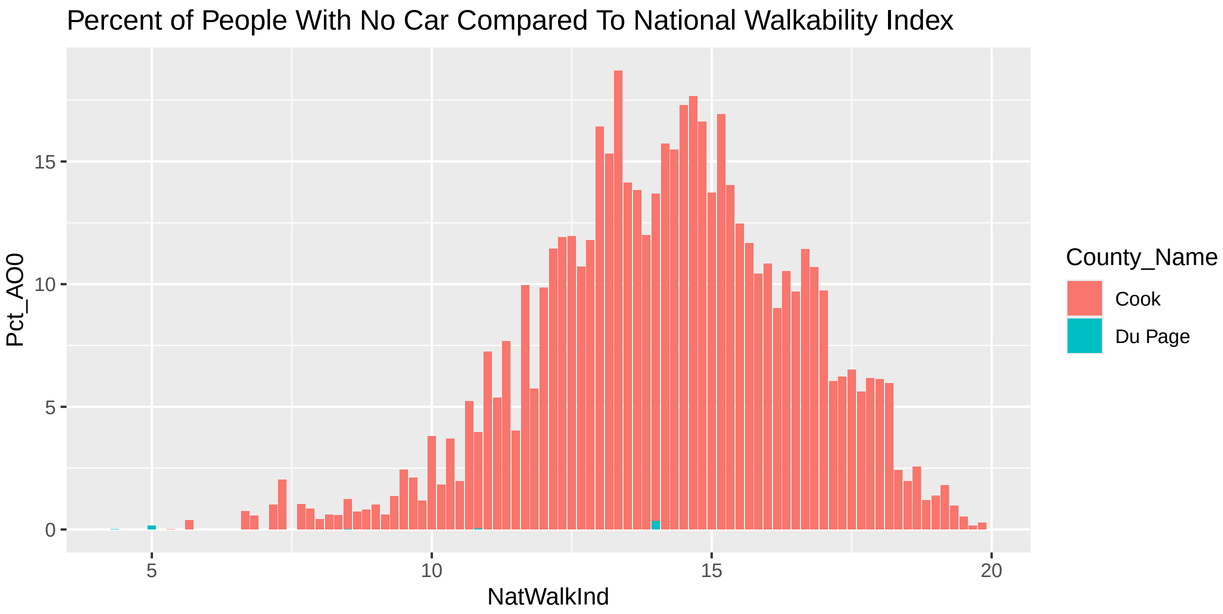
Further, Let's compare the spreads of the National Walkability Index (EKW_2022 and NatWalkInd):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.333  12.833  14.500  14.323  16.000  19.833
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    6.75   13.25   14.33   14.33   15.33   19.67       3
```

They have pretty similar spreads with the same mean! They are both ranked 0-20,but have sight differences throughout.

Lastly, when considering walkability, we also want to explore car ownership to consider other modes of possible transportation. We decided to explore the Pct_AO0 variable which is the percent of households that own zero automobiles. This will give us a quick understanding of those that do have a car versus those who do not.



This graph demonstrates an interesting topic, since with Pct_AO0, we see that areas with medium walkability have less cars than expected. Surprisingly areas with high walkability own more cars than those in areas with medium to moderately good walkability. Unsurprisingly, those with low walkability also own cars. This may be interesting to promote alternative sources of transportation in the medium walkability areas because they have less cars and also help show that areas with lower walkability may need better access to essential resources, like healthcare and food.

Testing Logisitic Regression with National Walkability (NatWalkInd)

Before determining if we should use EKW_2022 or NatWalkInd, we wanted to research if the variables suggested in the data guide, used to compose NatWalkInd, would show up as the only contributors to NatWalkInd.

We decided the best way to research this was running a simple logistic regression against NatWalkInd, dividing the binary categories into high walkability =1 (>mean) and low walkability = 0 (<mean).

We removed all variables that were not predictive and changed numeric variables into the numeric data type in order to run our logistic regression model.

We suspect that NatWalkInd is very reliant on the Ranked variables. Let’s take a closer look at what variables go into NatWalkInd to understand if we should remove these variables when testing EKW_2022 to avoid colinearity.

```
## Morgan-Tatar search since family is non-gaussian.
```

```
##     D3B TotPop CountHU    HH R_PCTLOWWAGE D2A_JPHH D3AMM D2A_Ranked D2B_Ranked
## 1 FALSE  FALSE   FALSE FALSE        FALSE    FALSE FALSE       TRUE       TRUE
##   D3B_Ranked D4A_Ranked MCD_2018.2022 UNS_2018.2022 EKW_2022 TRV_2018.2022
## 1       TRUE       TRUE         FALSE         FALSE    FALSE         FALSE
##   Criterion
## 1  8.000002
```

```
## 
## Call:
## glm(formula = HighWalkability ~ . - TotPop - HH - CountHU - D2A_JPHH - 
##     D3AMM - R_PCTLOWWAGE - MCD_2018.2022 - EKW_2022 - UNS_2018.2022 - 
##     TRV_2018.2022 - D3B - D2A_Ranked, family = binomial(), data = data_cities_log2)
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -29.52192    1.55154  -19.03   <2e-16 ***
## D2B_Ranked    0.43692    0.02150   20.32   <2e-16 ***
## D3B_Ranked    0.72586    0.03775   19.23   <2e-16 ***
## D4A_Ranked    0.73674    0.05936   12.41   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 3022.1  on 2184  degrees of freedom
## Residual deviance: 1224.6  on 2181  degrees of freedom
## AIC: 1232.6
## 
## Number of Fisher Scoring iterations: 7
```

```
## 
## Call:
## glm(formula = HighWalkability ~ . - TotPop - HH - CountHU - D2A_JPHH - 
##     D3AMM - R_PCTLOWWAGE - MCD_2018.2022 - EKW_2022 - UNS_2018.2022 - 
##     TRV_2018.2022 - D3B, family = binomial(), data = data_cities_log2)
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3311.12  114674.60  -0.029    0.977
## D2A_Ranked      37.81    1310.80   0.029    0.977
## D2B_Ranked      37.81    1308.83   0.029    0.977
## D3B_Ranked      75.66    2616.04   0.029    0.977
## D4A_Ranked      75.74    2665.01   0.028    0.977
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.0221e+03  on 2184  degrees of freedom
## Residual deviance: 1.7270e-06  on 2180  degrees of freedom
## AIC: 10
##
## Number of Fisher Scoring iterations: 25
```

Based on this model, if we utilize the model with the lowest Criterion/AIC (AIC = 8), the variables D2A_Ranked, D2B_Ranked, D3B_Ranked, and D4A_Ranked all play a role in NatWalkInd. We removed D2A_Ranked in order to see the actual p values without colinearity interference of these ranked variables. When including all ranked variables, the p-values shot up to .977 although the AIC was estimated at 10. When removing one of the ranked variables (D2A_Ranked), the p-values shot down to <.01 and had an AIC of 1232.6. This analysis helps suggest to not use NatWalkInd and these ranked variables when testing EKW_2022, as they are collinear and will not pass the independence assumption. Further, we think that using the ranked variables for analysis may be more insightful since they each are metrics for different things. Thus, we will continue this analysis with linear modeling and backwards selection for our EKW_2022 variable.

Linear Modeling And Backwards Selection

When performing backwards selection on EKW_2022, we wanted to try including NatWalkInd to see how it would perform and the effects there.

```
##
## Call:
## lm(formula = EKW_2022 ~ . - CBSA_EMP - D4A_Ranked - D3APO - D4E -
##     Shape_Area - Shape_Length - D4A - CountHU - HH - R_PCTLOWWAGE -
##     D1C8_HLTH - D3B_Ranked - D2B_Ranked - D2A_JPHH - Pct_AO0 -
##     D2A_Ranked, data = data_cities_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4740 -0.6284  0.0509  0.6838  3.4205
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.9923398  0.4388859  15.932  < 2e-16 ***
## NatWalkInd     0.4313904  0.0123772  34.854  < 2e-16 ***
## D3B           -0.0009016  0.0003715  -2.427 0.015301 *
## TotPop         0.0001493  0.0000377   3.960 7.74e-05 ***
## P_WrkAge       0.7031112  0.2122182   3.313 0.000938 ***
## D3A            0.0265551  0.0052647   5.044 4.94e-07 ***
## D3AAO         -0.0170343  0.0077535  -2.197 0.028127 *
## D3AMM         -0.0198583  0.0043943  -4.519 6.54e-06 ***
## EMP_2018.2022  0.0188269  0.0044333   4.247 2.26e-05 ***
## MCD_2018.2022  0.0125943  0.0044581   2.825 0.004771 **
## UNI_2018.2022 -0.0002953  0.0001082  -2.729 0.006409 **
## UNS_2018.2022  0.0185656  0.0074026   2.508 0.012215 *
## VHA_2018.2022 -0.0792966  0.0174792  -4.537 6.03e-06 ***
## FAI_2020      -0.0206716  0.0042410  -4.874 1.17e-06 ***
## TRV_2018.2022 -0.0215786  0.0044042  -4.900 1.03e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.02 on 2170 degrees of freedom
## Multiple R-squared:   0.62,  Adjusted R-squared:  0.6176
## F-statistic: 252.9 on 14 and 2170 DF,  p-value: < 2.2e-16
```

Here we can see the Ranked variables appeared not significant, again pointing to collinearity between the Ranked variables and NatWalkInd.

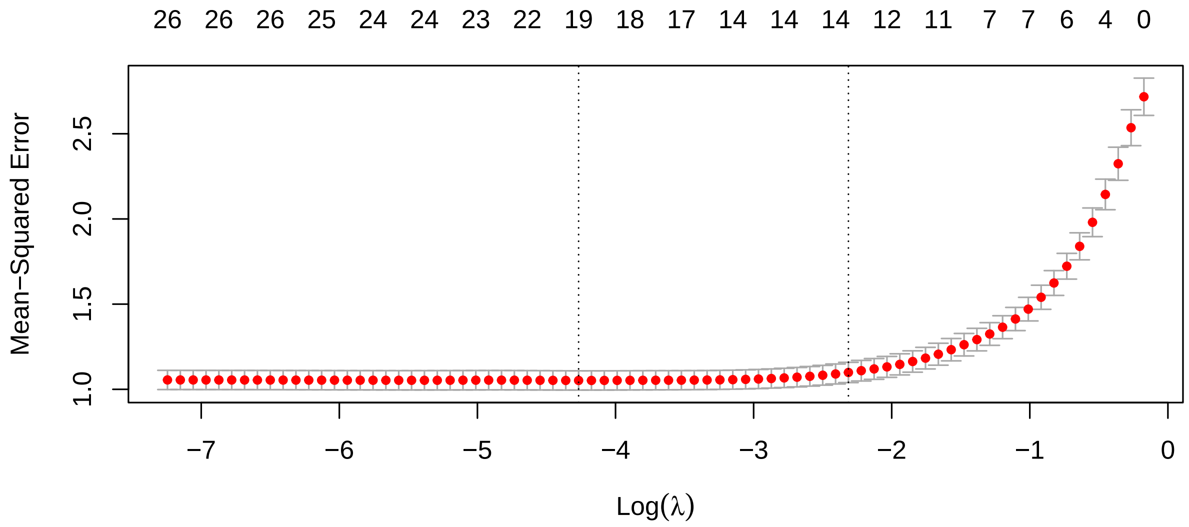
Instead of including NatWalkInd, we decided to allow room for the Ranked variables to see their effects on our response variable (EKW_2022) since each of these variables may differ slightly.

```
##
## Call:
## lm(formula = EKW_2022 ~ . - NatWalkInd - CBSA_EMP - D3APO - Shape_Area -
##     CountHU - HH - D1C8_HLTH - D4E - Pct_AO0 - D2A_JPHH - Shape_Length -
##     D3B - D3AAO, data = data_cities_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5685 -0.6278  0.0468  0.6835  3.4838
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.493e+00  5.095e-01  16.669  < 2e-16 ***
## TotPop         1.328e-04  3.793e-05   3.501 0.000473 ***
## P_WrkAge       6.579e-01  2.150e-01   3.060 0.002243 **
## R_PCTLOWWAGE  -1.699e+00  5.832e-01  -2.913 0.003617 **
## D3A            1.739e-02  3.654e-03   4.759 2.08e-06 ***
## D3AMM         -1.324e-02  3.812e-03  -3.473 0.000525 ***
## D4A            4.092e-05  6.341e-06   6.454 1.34e-10 ***
## D2A_Ranked     7.862e-02  4.480e-03  17.548  < 2e-16 ***
## D2B_Ranked     7.199e-02  4.590e-03  15.683  < 2e-16 ***
## D3B_Ranked     1.331e-01  7.564e-03  17.594  < 2e-16 ***
## D4A_Ranked     8.551e-02  1.533e-02   5.576 2.76e-08 ***
## EMP_2018.2022  1.806e-02  4.444e-03   4.064 4.99e-05 ***
## MCD_2018.2022  1.513e-02  4.468e-03   3.386 0.000721 ***
## UNI_2018.2022 -3.380e-04  1.075e-04  -3.144 0.001689 **
## UNS_2018.2022  2.401e-02  7.373e-03   3.256 0.001147 **
## VHA_2018.2022 -7.290e-02  1.750e-02  -4.166 3.22e-05 ***
## FAI_2020      -1.773e-02  4.373e-03  -4.053 5.23e-05 ***
## TRV_2018.2022 -2.130e-02  4.400e-03  -4.841 1.39e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 1.01 on 2167 degrees of freedom
## Multiple R-squared:  0.6276, Adjusted R-squared:  0.6246
## F-statistic: 214.8 on 17 and 2167 DF,  p-value: < 2.2e-16
```

Our Findings: Here we can see that the FAI_2020 variable, UNI_2018.2022 variable, UNS_2018.2022, MCD_2018.2022, and the VHA_2018.2022 are all significant to National Walkability Index, indicating that food insecurity and healthcare accessibility are likely related to walkability. The model suggests that food insecurity decreases at a rate of 1.773% as walkability increases. It also suggests that the uninsured rate increases at a rate of 2.4% while the rate of those who are covered by VA Health Care decreases at a rate of 7.29% as walkability increases. Assumptions of this model are below in the appendix.

LASSO



Shows that we need a very small lambda, which we then went back up to adjust.

Build the lasso model

```
##
## Call:
## lm(formula = lm.input, data = data_cities_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.4807 -0.6359  0.0465  0.6902  3.3055 
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      9.306e+00  4.144e-01  22.456  < 2e-16 ***
## TotPop           1.315e-04  3.772e-05   3.487 0.000498 ***
## P_WrkAge         6.545e-01  2.143e-01   3.054 0.002286 ** 
## R_PCTLOWWAGE  -1.454e+00  5.810e-01  -2.503 0.012381 *
## D3A            6.033e-03  3.636e-03   1.659 0.097186 .
## D3APO          1.511e-02  3.942e-03   3.834 0.000130 ***
## D4A            3.937e-05  6.350e-06   6.200 6.73e-10 ***
## D2A_Ranked     8.069e-02  4.447e-03  18.144  < 2e-16 ***
## D2B_Ranked     7.154e-02  4.601e-03  15.548  < 2e-16 ***
## D3B_Ranked     1.303e-01  8.079e-03  16.123  < 2e-16 ***
## D4A_Ranked     8.578e-02  1.536e-02   5.584 2.64e-08 ***
## EMP_2018.2022  6.521e-03  2.090e-03   3.120 0.001831 **
## VHA_2018.2022 -7.545e-02  1.713e-02  -4.404 1.12e-05 ***
## FAI_2020      -1.039e-02  3.834e-03  -2.712 0.006751 **
## TRV_2018.2022 -2.058e-02  4.363e-03  -4.718 2.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014 on 2170 degrees of freedom
## Multiple R-squared:  0.6248, Adjusted R-squared:  0.6223
## F-statistic: 258.1 on 14 and 2170 DF,  p-value: < 2.2e-16
```

The ranked variables are able to come back into the model due to NatWalkInd being gone. The adjusted R squared is also a bit higher in this model, signaling that by taking out the index, we can actually improve the model.

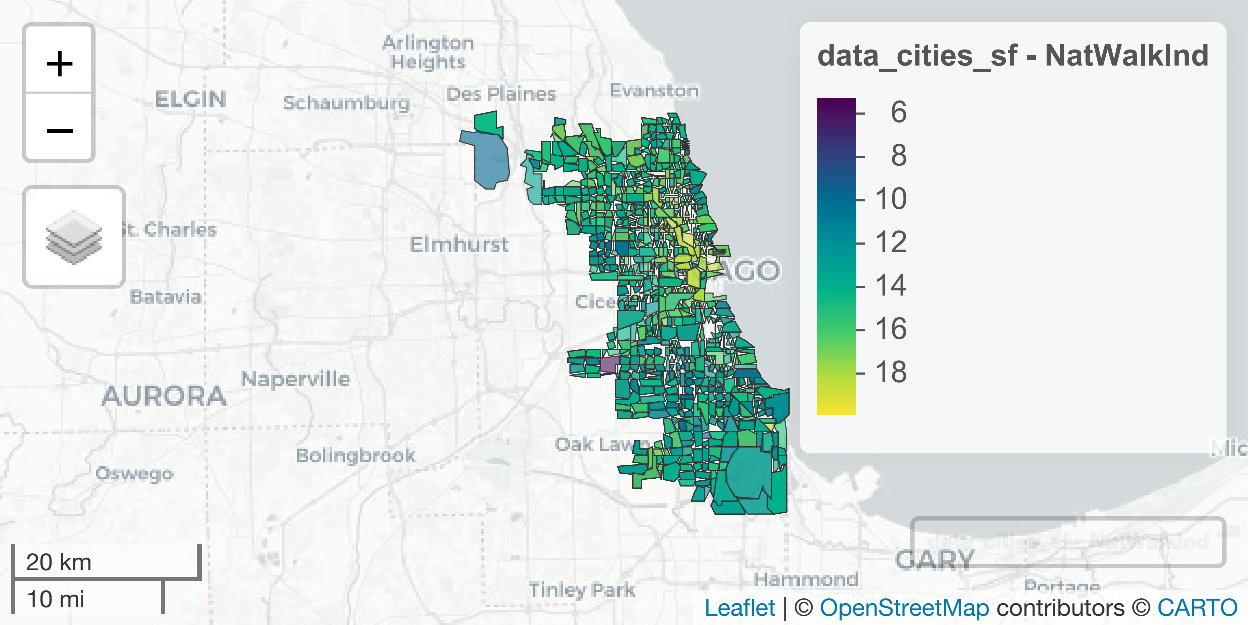
Another takeaway is that food insecurity decreases in significance and there is a removal of uninsured residents and uninsured residents rate from the model. VHA rate still remains in the model with a high negative coefficient suggesting that there may be a strong relationship between it and walkability. This is interesting as it brings insight on which factors may connect more closely with walkability.

Spatial Analysis

Let's do some spatial analysis to see how the variables look spatially:

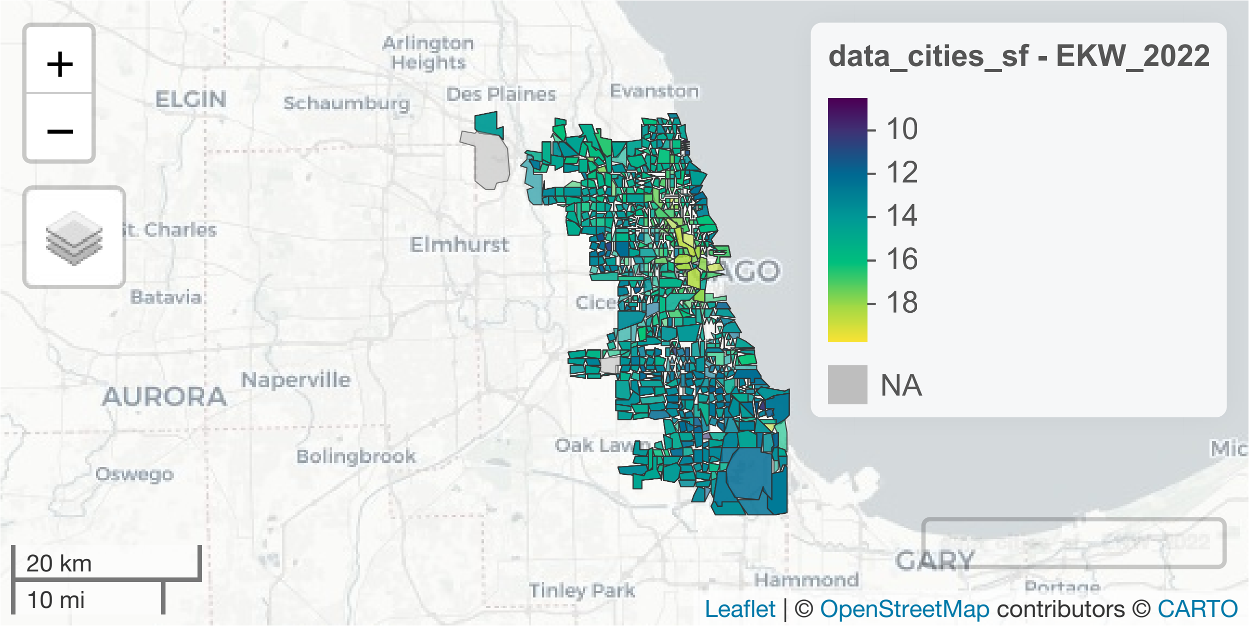
Spatial Maps

With NatWalkInd



Walkability varies across the city, with medium to low scores in darker green areas suggesting average pedestrian conditions, while lighter green and yellow areas indicate higher walkability. Higher walkability is noted in central and northern parts of Chicago, likely reflecting the better pedestrian infrastructure in urban and densely populated areas.

With EKW_2022

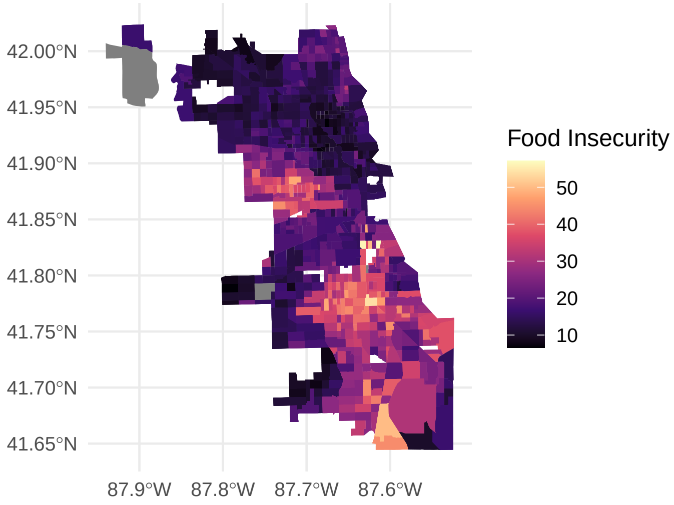


The maps with EKW_2022 highlights high walkability in central and northern regions, indicative of dense urban areas. The second map, EKW_2022, displays a wider distribution of walkability, possibly due to infrastructure enhancements or improved data collection from the year 2021 to 2022. From now on, we only use EKW_2022 because of the more recent data and because of aforementioned findings.

Food & Healthcare Access

Map of Food Insecurity in Chicago

Based on FAI_2020

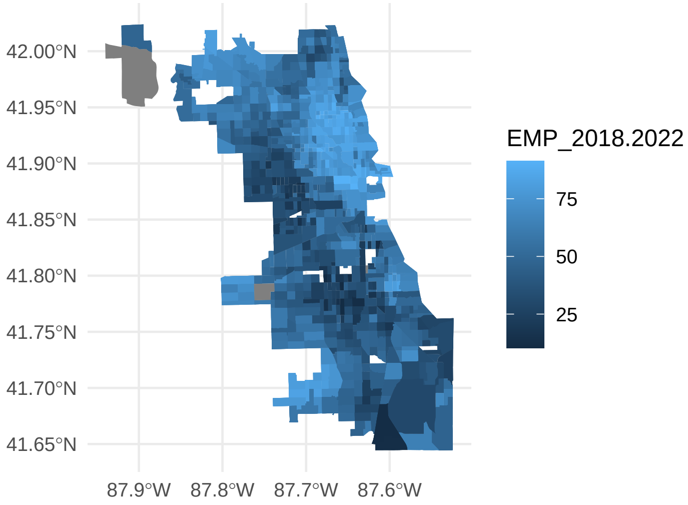


Source: Chicago Health Atlas Data

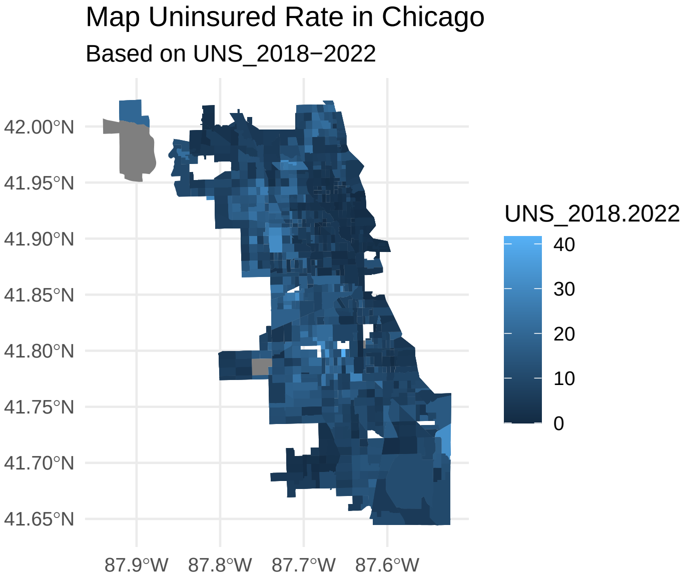
The map displays varying levels of food insecurity across Chicago, based on the Food Insecurity Index (FAI) for 2020. Areas shaded in darker purple, particularly concentrated in the central and northern regions of the city, indicate lower rates of food insecurity. Conversely, regions depicted in lighter colors, notably in the southern parts of the city, show relatively higher levels of food insecurity, with scores closer to 50. This geographic distribution suggests significant disparities in access to adequate food within the certain areas, likely reflecting underlying socio-economic differences across these areas.

Map Employment Insurance in Chicago

Based on EMP_2018.2022



Source: Chicago Health Atlas Data



Source: Chicago Health Atlas Data

The two maps of Chicago show different socioeconomic indicators: Employment Insurance rates and Uninsured Rates for the same period. The first map, illustrating employment insurance, uses a color gradient from light to dark blue to denote lower to higher rates of employment insurance across different neighborhoods. In contrast, the second map depicting uninsured rates also employs a blue gradient, but it indicates lower rates in darker areas, showing an inverse relationship compared to the employment insurance map. Areas with higher employment insurance might suggest a more stable economic environment which can correlate with better-maintained infrastructure and amenities, potentially enhancing walkability. Regions with higher uninsured rates might reflect economic strain, potentially diminishing walkability in these areas. Again we see an acceleration of the employment insurance rating in the north east region and a higher uninsured rating in areas away from the coast in the northwest and southern region.

Spatial Regression

Lag Model

Let’s try running a spatial lag model using the significant variables found earlier in our linear regression model (not including NatWalkInd).

```
##
## Call:lagsarlm(formula = EKW_2022 ~ TotPop + P_WrkAge + R_PCTLOWWAGE +
##     D3A + D3AMM + D4A + D2A_Ranked + D2B_Ranked + D3B_Ranked +
##     D4A_Ranked + EMP_2018.2022 + MCD_2018.2022 + UNI_2018.2022 +
##     UNS_2018.2022 + VHA_2018.2022 + FAI_2020 + TRV_2018.2022,
##     data = data_cities_sf, listw = weights)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -4.330804 -0.571945  0.013681  0.627912  3.658314
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    2.2478e+00  1.2356e+00  1.8192 0.0688863
## TotPop         8.8156e-05  3.4551e-05  2.5515 0.0107265
## P_WrkAge       1.5866e-01  1.9670e-01  0.8066 0.4198927
## R_PCTLOWWAGE  -1.3363e-02  5.3665e-01 -0.0249 0.9801334
## D3A            1.0108e-02  3.3104e-03  3.0535 0.0022617
## D3AMM         -8.8630e-03  3.4500e-03 -2.5689 0.0102008
## D4A           -1.1539e-04  5.9283e-04 -0.1947 0.8456668
## D2A_Ranked     6.7794e-02  4.0844e-03 16.5983 < 2.2e-16
## D2B_Ranked     6.0633e-02  4.1731e-03 14.5296 < 2.2e-16
## D3B_Ranked     1.1373e-01  6.8484e-03 16.6070 < 2.2e-16
## D4A_Ranked     6.7847e-02  5.3304e-02  1.2728 0.2030770
## EMP_2018.2022  6.8668e-03  4.0694e-03  1.6874 0.0915199
## MCD_2018.2022  9.5886e-03  4.0673e-03  2.3575 0.0183978
## UNI_2018.2022 -3.0515e-04  9.6763e-05 -3.1536 0.0016128
## UNS_2018.2022  1.6318e-02  6.6929e-03  2.4382 0.0147623
## VHA_2018.2022 -6.3993e-03  1.6006e-02 -0.3998 0.6892984
## FAI_2020      -1.3818e-02  3.9787e-03 -3.4731 0.0005146
## TRV_2018.2022 -2.7556e-03  4.0943e-03 -0.6730 0.5009162
##
## Rho: 0.50557, LR test value: 366.38, p-value: < 2.22e-16
## Asymptotic standard error: 0.026856
##     z-value: 18.825, p-value: < 2.22e-16
## Wald statistic: 354.39, p-value: < 2.22e-16
##
## Log likelihood: -2866.444 for lag model
## ML residual variance (sigma squared): 0.8215, (sigma: 0.90637)
## Number of observations: 2157
## Number of parameters estimated: 20
## AIC: 5772.9, (AIC for lm: 6137.3)
## LM test for residual autocorrelation
## test value: 95.438, p-value: < 2.22e-16
```

In this model including spatial data, we can see that the Ranked variables did show up when NatWalkInd was removed.

This model suggests that food insecurity has a negative relationship with the national walkability index in 2022, meaning that as the national walkability increases by one, the food insecurity rate decreases by 1.38e-2% on average, taking into account its neighbors. The medicaid rate and uninsured rate both have a positive linear relationships when taking into account their spatial data, on average increasing by .96e-3% and 1.63e-2% as the national walkability increases by one. This is an unexpected conclusion. Furthermore, the VHA insured rate has a negative relationship to EKW_2022, decreasing on average by .69e-3%.

Interestingly, we see that the uninsured resident amount has a negative linear relationship with the national walkability index. This could potentially be due to the spatial nature of this analysis, with the potential for future analysis to account more closely for things like population number. Additionally, it would be helpful for us to get a sense of all insurance measures in one metric, since there are clear differences in access to them.

Lastly, let's look at the variables that are significant to EKW_2022 using backwards selection from all the possible variables.

```
##
## Call:lagsarlm(formula = EKW_2022 ~ HH + D3APO + D2A_Ranked + D4A_Ranked +
##     D4A_Ranked + UNS_2018.2022 + D2B_Ranked + MCD_2018.2022 +
##     D3B_Ranked + UNI_2018.2022 + FAI_2020, data = data_cities_sf,
##     listw = weights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.28442 -0.57595  0.01830  0.62311  3.51188
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                   Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    2.4034e+00  4.2790e-01  5.6168 1.945e-08
## HH             2.1403e-04  8.1675e-05  2.6205  0.008780
## D3APO          1.0820e-02  3.2885e-03  3.2901  0.001001
## D2A_Ranked     6.8519e-02  4.0089e-03 17.0915 < 2.2e-16
## D4A_Ranked     7.6809e-02  1.3666e-02  5.6206 1.903e-08
## UNS_2018.2022  1.0817e-02  5.4114e-03  1.9989  0.045617
## D2B_Ranked     6.0262e-02  4.1773e-03 14.4262 < 2.2e-16
## MCD_2018.2022  4.5114e-03  2.2195e-03  2.0326  0.042090
## D3B_Ranked     1.1251e-01  7.0015e-03 16.0694 < 2.2e-16
## UNI_2018.2022 -2.9539e-04  9.6202e-05 -3.0706  0.002137
## FAI_2020      -1.5543e-02  3.8045e-03 -4.0855 4.398e-05
##
## Rho: 0.52318, LR test value: 471.31, p-value: < 2.22e-16
## Asymptotic standard error: 0.024458
##     z-value: 21.391, p-value: < 2.22e-16
## Wald statistic: 457.58, p-value: < 2.22e-16
##
## Log likelihood: -2870.598 for lag model
## ML residual variance (sigma squared): 0.8215, (sigma: 0.90637)
## Number of observations: 2159
## Number of parameters estimated: 13
## AIC: 5767.2, (AIC for lm: 6236.5)
## LM test for residual autocorrelation
## test value: 90.535, p-value: < 2.22e-16
```

This model narrows down the important related variables to EKW_20222, specifically not including NatWalkInd. The model finds a good amount of variables to be significant, but most importantly variables: UNS_2018.2022, MCD_2018.2022, UNI_2018.2022, and FAI_2020. This finally suggests that food insecurity and healthcare accesibility may be related to geo-spatial walkability. Similar to seen before, food insecurity had a negative relationship with wallability as well the number of uninsured people. The rate of uninsured people and the rate of medicaid users had a positive linear relationship with walkability, again seen previously. This model is likely the best to represent the relationship between walkability and these variables as it has the lowest AIC at 5767.2 and indicates that there is a spatial relationship between these variables and their neighbors.

Error Model

```
##
## Call:errorsarlm(formula = EKW_2022 ~ NatWalkInd + D3B + TotPop + P_WrkAge +
##     D3A + D3AAO + D3AMM + EMP_2018.2022 + MCD_2018.2022 + UNI_2018.2022 +
##     UNS_2018.2022 + VHA_2018.2022 + FAI_2020 + TRV_2018.2022,
##     data = data_cities_sf, listw = weights)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -4.5759869 -0.5757112  0.0098626  0.6584761  3.7889984
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    9.1597e+00  4.8298e-01 18.9648 < 2.2e-16
## NatWalkInd     3.4322e-01  1.1787e-02 29.1175 < 2.2e-16
## D3B           -4.8421e-04  3.4126e-04 -1.4189 0.1559314
## TotPop         1.0177e-04  3.5762e-05  2.8458 0.0044294
## P_WrkAge       1.5302e-01  2.1672e-01  0.7061 0.4801440
## D3A            1.5250e-02  4.9632e-03  3.0726 0.0021217
## D3AAO         -8.5217e-03  7.4727e-03 -1.1404 0.2541275
## D3AMM         -1.5273e-02  4.5030e-03 -3.3917 0.0006946
## EMP_2018.2022  6.2071e-03  4.7270e-03  1.3131 0.1891433
## MCD_2018.2022  5.8720e-03  4.6471e-03  1.2636 0.2063788
## UNI_2018.2022 -3.7130e-04  1.1618e-04 -3.1958 0.0013944
## UNS_2018.2022  1.2495e-02  7.7273e-03  1.6170 0.1058827
## VHA_2018.2022 -1.3195e-02  1.7824e-02 -0.7403 0.4591101
## FAI_2020      -2.4965e-02  5.0297e-03 -4.9635 6.923e-07
## TRV_2018.2022 -4.4178e-03  5.0647e-03 -0.8723 0.3830620
##
## Lambda: 0.71611, LR test value: 222.85, p-value: < 2.22e-16
## Asymptotic standard error: 0.030052
##     z-value: 23.829, p-value: < 2.22e-16
## Wald statistic: 567.81, p-value: < 2.22e-16
##
## Log likelihood: -2946.388 for error model
## ML residual variance (sigma squared): 0.86469, (sigma: 0.92989)
## Number of observations: 2157
## Number of parameters estimated: 17
## AIC: 5926.8, (AIC for lm: 6147.6)
```

```
##
## Call:errorsarlm(formula = EKW_2022 ~ TotPop + P_WrkAge + R_PCTLOWWAGE +
##     D3A + D3AMM + D4A + D2A_Ranked + D2B_Ranked + D3B_Ranked +
##     D4A_Ranked + EMP_2018.2022 + MCD_2018.2022 + UNI_2018.2022 +
##     UNS_2018.2022 + VHA_2018.2022 + FAI_2020 + TRV_2018.2022,
##     data = data_cities_sf, listw = weights)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -4.541888 -0.575404  0.016477  0.640578  3.816798
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    9.9155e+00  1.2155e+00  8.1575 4.441e-16
## TotPop         9.7459e-05  3.6419e-05  2.6760 0.0074499
## P_WrkAge       1.9528e-01  2.1620e-01  0.9032 0.3664035
## R_PCTLOWWAGE  -9.7415e-01  6.3994e-01 -1.5222 0.1279489
## D3A            1.2262e-02  3.5084e-03  3.4951 0.0004739
## D3AMM         -1.1240e-02  4.0171e-03 -2.7979 0.0051430
## D4A            1.1032e-04  6.0063e-04  0.1837 0.8542682
## D2A_Ranked     6.4415e-02  4.2768e-03 15.0617 < 2.2e-16
## D2B_Ranked     6.0377e-02  4.2736e-03 14.1281 < 2.2e-16
## D3B_Ranked     1.0574e-01  7.2662e-03 14.5525 < 2.2e-16
## D4A_Ranked     7.8101e-02  5.3948e-02  1.4477 0.1477009
## EMP_2018.2022  6.0861e-03  4.7362e-03  1.2850 0.1987878
## MCD_2018.2022  6.8677e-03  4.6514e-03  1.4765 0.1398148
## UNI_2018.2022 -3.6627e-04  1.1593e-04 -3.1593 0.0015816
## UNS_2018.2022  1.3240e-02  7.7117e-03  1.7168 0.0860087
## VHA_2018.2022 -1.0792e-02  1.7787e-02 -0.6067 0.5440479
## FAI_2020      -2.2802e-02  5.1407e-03 -4.4356 9.183e-06
## TRV_2018.2022 -4.0377e-03  5.0434e-03 -0.8006 0.4233655
##
## Lambda: 0.70541, LR test value: 218.89, p-value: < 2.22e-16
## Asymptotic standard error: 0.030801
##     z-value: 22.902, p-value: < 2.22e-16
## Wald statistic: 524.52, p-value: < 2.22e-16
##
## Log likelihood: -2940.189 for error model
## ML residual variance (sigma squared): 0.86109, (sigma: 0.92795)
## Number of observations: 2157
## Number of parameters estimated: 20
## AIC: 5920.4, (AIC for lm: 6137.3)
```

The results highlight several statistically significant predictors of walkability. Food insecurity specifically also shows a strong negative impact on walkability, with a significant negative coefficient. Other significant negative effects are explained by the variables Medicaid coverage, uninsured rates, and travel times. The model is highly significant overall, because of the likelihood ratio test statistic, and lambda.

Given the lower AIC and the higher log likelihood of the second fit, it is statistically superior to the first model.

Conclusion

Our study aimed to analyze and assess Chicago's walkability, particularly in relation to two significant urban challenges: food insecurity and healthcare coverage access. By examining the correlation between walkability and access to food and healthcare, we sought to contribute to the discourse on urban planning and its impact on health outcomes. Throughout the course of this analysis, we performed logistic regression, linear regression, lasso regularization, and both spatial error and lag models.

The logistic regression investigated NatWalkInd and its correlation to the ranked variables: D2A_Ranked, D2B_Ranked, D3B_Ranked, and D4A_Ranked, suggesting that NatWalkInd and the ranked variables should be used independently in future analysis. Through the linear model with backwards selection, we were able to find that there was a relationship of food insecurity and healthcare to the walkability index.

Our best model was spatial lag regression, with an AIC of 5767.2, which we utilized backwards selection without NatWalkInd, considering NatWalkInd's previously identified influence. From this model, our findings indicated that food insecurity and healthcare accessibility were spatially related with the walkability. We found that as walkability increases, on average the rate of uninsured residents would increase by 0.0108% and the rate of residents insured by medicaid by .00451%. On average, as the walkability index increases, the rate of food insecurity decreases by 0.0155 and uninsured individuals by .000295 while taking into account spatial location. Although food insecurity was similar to our EDA findings, we did not expect to see a positive relationship between our uninsured rate variable and our walkability variable.

Thus, our analysis suggests that although food insecurity is decreasing in areas with high walkability, it is important to decrease food insecurity in areas that are less walkable, like the south or inland Chicago. Food deserts (urban areas with limited access to affordable food) often overlap with regions of low walkability, exacerbating public health issues related to poor diet.

Our analysis also gave us insight on the relationship between healthcare accessibility and walkability, indicating that rate of people who are uninsured and rely on medicaid increases as walkability increases. This is interesting as walkability does not lead to higher independent healthcare accessibility unlike what we predicted. Instead, the rate of residents covered by medicaid coverage increases, indicating that there may be higher access to government based healthcare in more walkable areas. Similarly, areas with poor walkability can limit residents' ability to reach government healthcare services, which is crucial for preventative and emergency health care.

This analysis could inform policy recommendations to the Chicago mayor, the Department of Planning and Development (DPD) and the CDPH, that not only improve walkability but also enhance overall urban livability, thereby improving overall lifestyle. Such improvements are pivotal in transforming urban environments into more health-supportive spaces that encourage walking and other physical activities.

Limitations: Some limitations of our study include that we were only able to look at Cook and Du Page county. Further, Du Page only had a few data points (5), while Cook had many (2180), leading us to not have adequate comparison. There could also be other counties in Chicago that would provide good expansion for the project. Another limitation is that we had over 140 variables to begin with, so we had to cut down initially without statistical measure to narrow in our question. Thus, there is potential that we missed out on different contributions from other available data. There are also potential for further analysis with autocorrelation and Moran's I, that we decided not to focus on, but think would be good to continue understanding walkability in Chicago. Overall, the mass amounts of data provided useful for analysis, but also posed a limitation in needing to cut down on what we used and how we used it.

Appendix

Put all supplementary materials in Appendix so that we do not lose focus in the main text but at the same time provide all relevant information for your reader/colleague/boss as well as **the future you** for reference!

Data dictionary

Data Dictionary Link: https://docs.google.com/document/d/1nbo9Qp6IDW9rJSgrV38W-hXUsyy8n0B_uqH1spD7PpA/edit?usp=sharing

Data cleaning process:

###Read in the data

Switched this line to run from my computer: walkability <- fread(“EPA_SmartLocationDatabase_V3_Jan_2021_Final.csv”)

Same for the other: state_city <- fread(“US_FIPS_Codes.csv”, header = TRUE)

###Formatting State_City Data Set

###Clean Data: Column CSA Name provides context of the city statistical area which we will later need to make sure we select for specifically the Chicago city.

Selecting for rows that have Chicago in their CSA name:

```
## [1] "Chicago-Naperville-Elgin, IL-IN-WI"
```

```
## [1] "Chicago-Naperville, IL-IN-WI"
##       State County Name STATEFP COUNTYFP
##      <char>      <char>   <int>    <int>
## 1: Alabama     Autauga       1        1
## 2: Alabama     Baldwin       1        3
## 3: Alabama     Barbour       1        5
## 4: Alabama        Bibb       1        7
## 5: Alabama      Blount       1        9
## 6: Alabama     Bullock       1       11
```

Joining the data_cities and walkability data set so have state and county names.

```
## [1] "Cook"    "McHenry" "Lake"    "Will"    "Kendall" "Kane"    "Du Page"
```

```
## [1] "Illinois"
```

```
## [1] "Chicago-Naperville, IL-IN-WI"
```

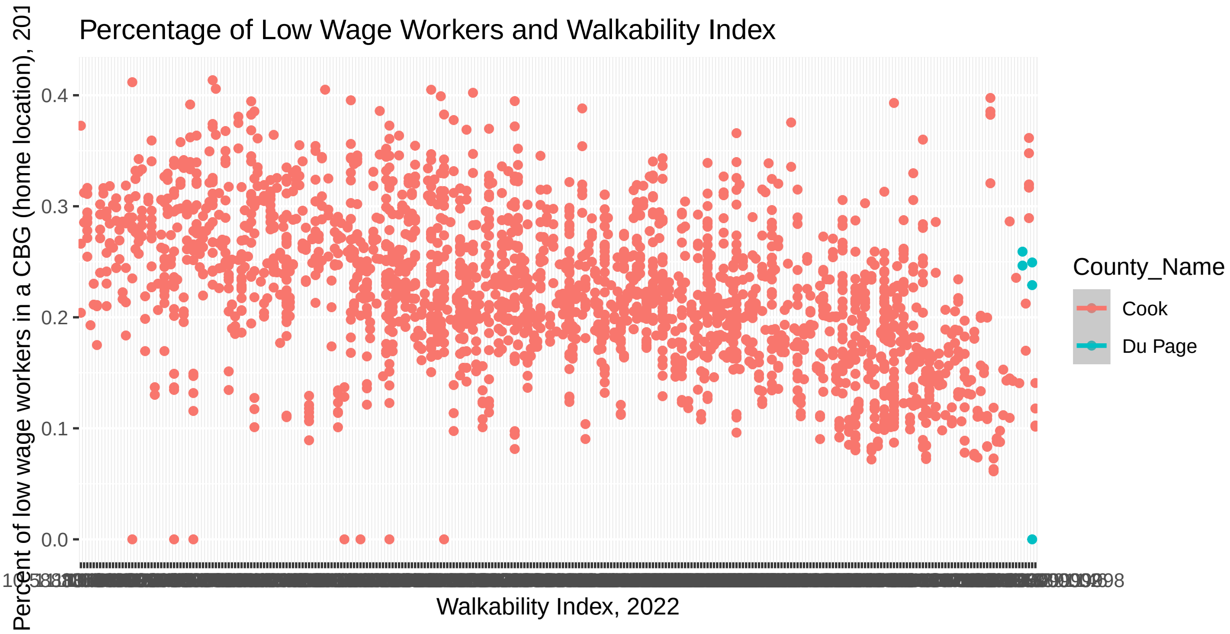
Getting Rid of Non-Useful Columns.

Do not run following code twice:

Further EDA Analysis

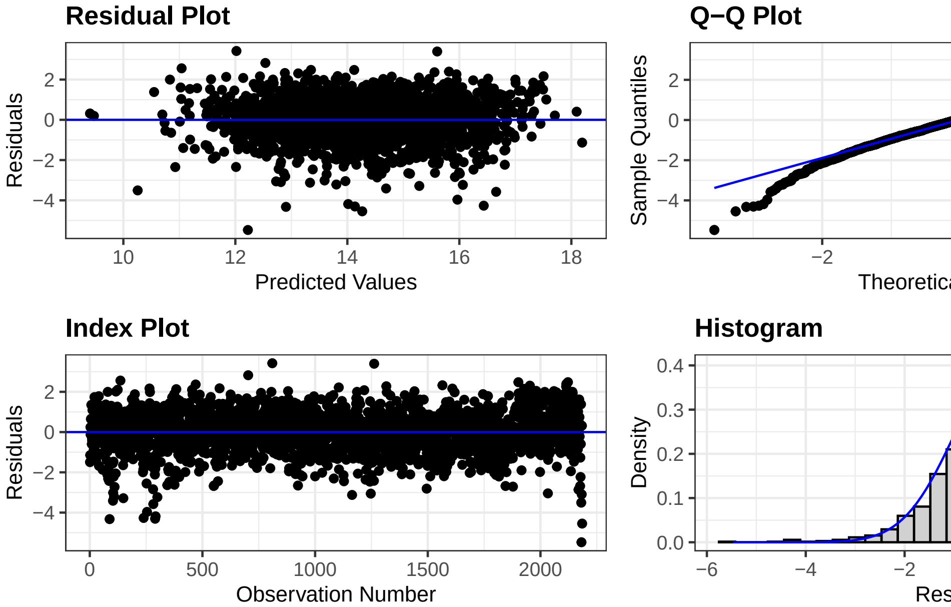
We also wanted to compare the percent of low wage workers against the National Walkability Index, including counties.

```
## `geom_smooth()` using formula = 'y ~ x'
```



Here we can see that there is a negative linear relationship between the two. This indicates that as the national walkability index increases, the percent of low wage workers in a CBG decreases. There is a slight difference based on counties, although there isn't much data for Du Page county compared to Cook county, and thus further analysis may not be possible between the two counties.

Assumptions



Checking the Assumptions of the model (:

The assumptions linearity, normality, independence, and homoscedasticity appear to be passed. There are a few outliers shown in the qq-plot but not enough to be significant.

Spatial Data

Further Spatial Analysis

```
##
## Call:lagsarlm(formula = EKW_2022 ~ NatWalkInd + D3B + TotPop + P_WrkAge +
##     D3A + D3AAO + D3AMM + EMP_2018.2022 + MCD_2018.2022 + UNI_2018.2022 +
##     UNS_2018.2022 + VHA_2018.2022 + FAI_2020 + TRV_2018.2022,
##     data = data_cities_sf, listw = weights)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.4266651 -0.5675582  0.0083143  0.6134090  3.6572672
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    1.3095e+00  5.2032e-01  2.5167 0.0118472
## NatWalkInd     3.6027e-01  1.1440e-02 31.4921 < 2.2e-16
## D3B           -4.5546e-04  3.3226e-04 -1.3708 0.1704423
## TotPop         9.2678e-05  3.4112e-05  2.7169 0.0065906
## P_WrkAge       1.0002e-01  1.9345e-01  0.5170 0.6051350
## D3A            1.3557e-02  4.7574e-03  2.8497 0.0043761
## D3AAO         -8.7445e-03  6.9795e-03 -1.2529 0.2102509
## D3AMM         -1.2859e-02  3.9592e-03 -3.2479 0.0011625
## EMP_2018.2022  6.8908e-03  4.0365e-03  1.7071 0.0878000
## MCD_2018.2022  9.2944e-03  4.0179e-03  2.3133 0.0207080
## UNI_2018.2022 -3.0375e-04  9.6817e-05 -3.1374 0.0017044
## UNS_2018.2022  1.5324e-02  6.6753e-03  2.2956 0.0216962
## VHA_2018.2022 -7.4991e-03  1.5953e-02 -0.4701 0.6383091
## FAI_2020      -1.4349e-02  3.8559e-03 -3.7212 0.0001983
## TRV_2018.2022 -3.1735e-03  4.0951e-03 -0.7749 0.4383712
##
## Rho: 0.50677, LR test value: 372.93, p-value: < 2.22e-16
## Asymptotic standard error: 0.0266
##     z-value: 19.051, p-value: < 2.22e-16
## Wald statistic: 362.94, p-value: < 2.22e-16
##
## Log likelihood: -2871.346 for lag model
## ML residual variance (sigma squared): 0.82516, (sigma: 0.90839)
## Number of observations: 2157
## Number of parameters estimated: 17
## AIC: 5776.7, (AIC for lm: 6147.6)
## LM test for residual autocorrelation
## test value: 93.933, p-value: < 2.22e-16
```

In Lag model 1 we took out NatWalkInd because of its collinearity to other variables to see if that will affect what variables appear as significant to EKW_2022.