Gentrification and Health Disparities: An Integrated Analysis of Inequality in Philadelphia

Jenea Adams Annan Timon

april 30

Contents

Packages

Executive Summary

Abstract & Goal of the Study

Gentrification is a process of affluent residents and businesses displacing existing low-income residents and businesses. Beyond prospects for so-called "urban renewal", gentrification has real, tangible effects on the landscape and trajectories of existing communities who often don't benefit from the changes of a neighborhood and are disenfranchised from participating in the growth of their area. Gentrification also has documented

health effects on communities, such as shortened life expectancy, higher cancer rates, higher infant mortality, and cardiovascular diseases. Income inequality can be used to estimate gentrification rates. It can be quantified by a Gini index which is a value from 0 to 1 indicating inequality in the dispersion of income in a given unit. **This study begins to to investigate a statistical framework for capturing the relationship between income inequality and health effects in Philadelphia, especially the historic Black Bottom (Figure 1), from integrated datasets.**

Data

In this project, we use the following three nearly cleaned data:

final_data.RDS: Census-Tract-level socialeconomic information that combines the following datqsets:

- **Food access**: The Neighborhood Food Retail dataset includes GEOID level assessments of food access relevant to distance and types of high produce grocery stores, as well as if that area is a high poverty area.
- **Hospital locations**: locations of hospitals by type of care provided. We chose this to understand the distribution and access to healthcare in each census block.
- **Heat Vulnerability**: scores and indicators for heat vulnerability by census block and prevalence of heat-related illnesses. This may or may not be related to access to green space and tree canopy. This data gives us an idea of environmental variables which contribute to health outcomes and quality of life for Philadelphia residents.
- **Affordable Housing**: locations of affordable and accessible housing projects recorded by the city. This data will provide information on the distribution of affordable housing options mapped to census blocks.
- **Philadelphia population metrics**: Demographic information of Philadelphia census blocks by race and ethnicity
- **Socioeconomic data**: income inequality calculated as a census tract-level Gini index.
- **Health data**: measures of prevalence of health measures such as cancer prevalence, access to health insurance, blood pressure, heart disease, and more by census tract.
- **Transitscore/Walkscore/Bikescore**: measures how walkable, well served by public transportation, and bikeable a location is

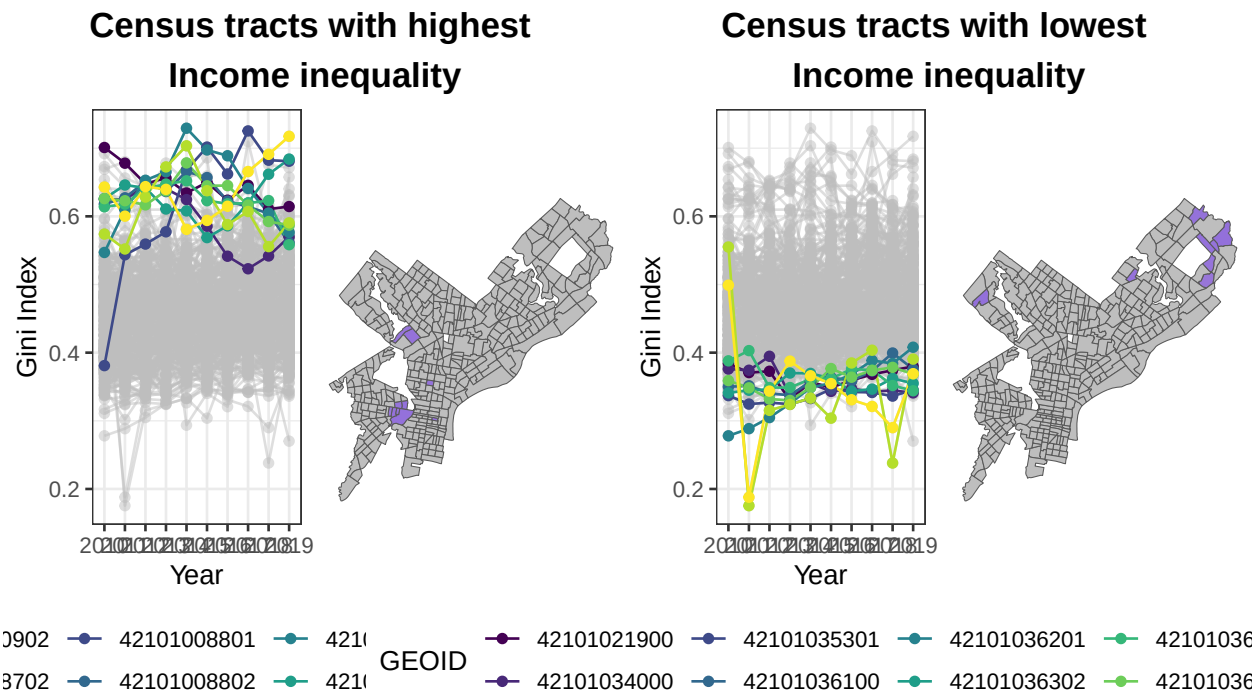
Philadelphia's Black Bottom

Figure 1: Philadelphia's Black Bottom from 1930 to 1973

Gini: measuring income inequality

Gini indexes have changed over time. Here we select the top and bottom 5 census tracts based on theer mean Gini coefficient from 2011 to 2019.

```
##   |                                                                      |
```



Brief Summary of Findings

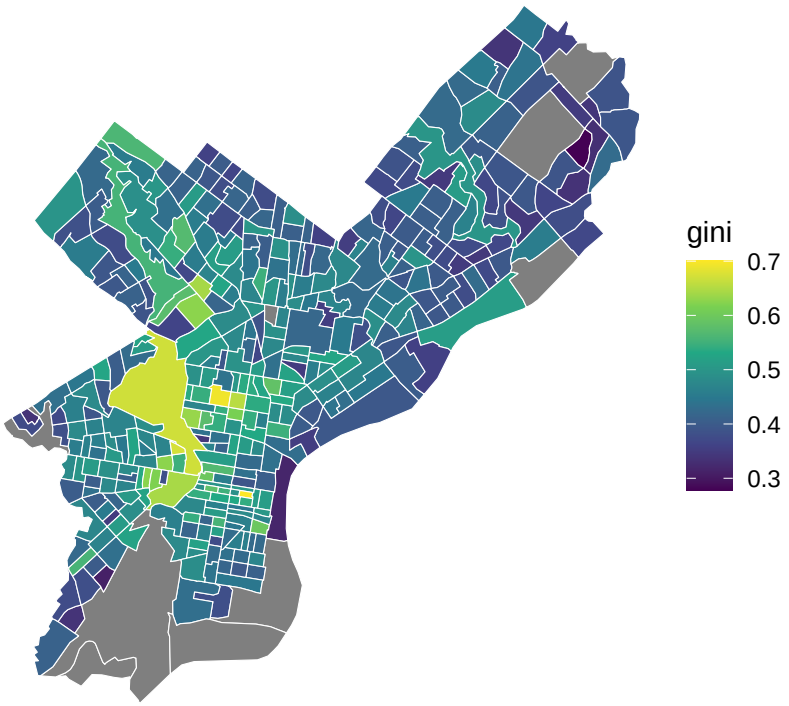
We found that there were strong associations between indicators of income inequality and adverse health outcomes in and closely bordering the footprint of Philadelphia's Black bottom. This association is particularly strong in in models used to predict COPD prevalence and Diabeates prevalence, diseases for which we see strong evidence of geospatial disparities in and around Philadelphia's Historic Black Bottom.

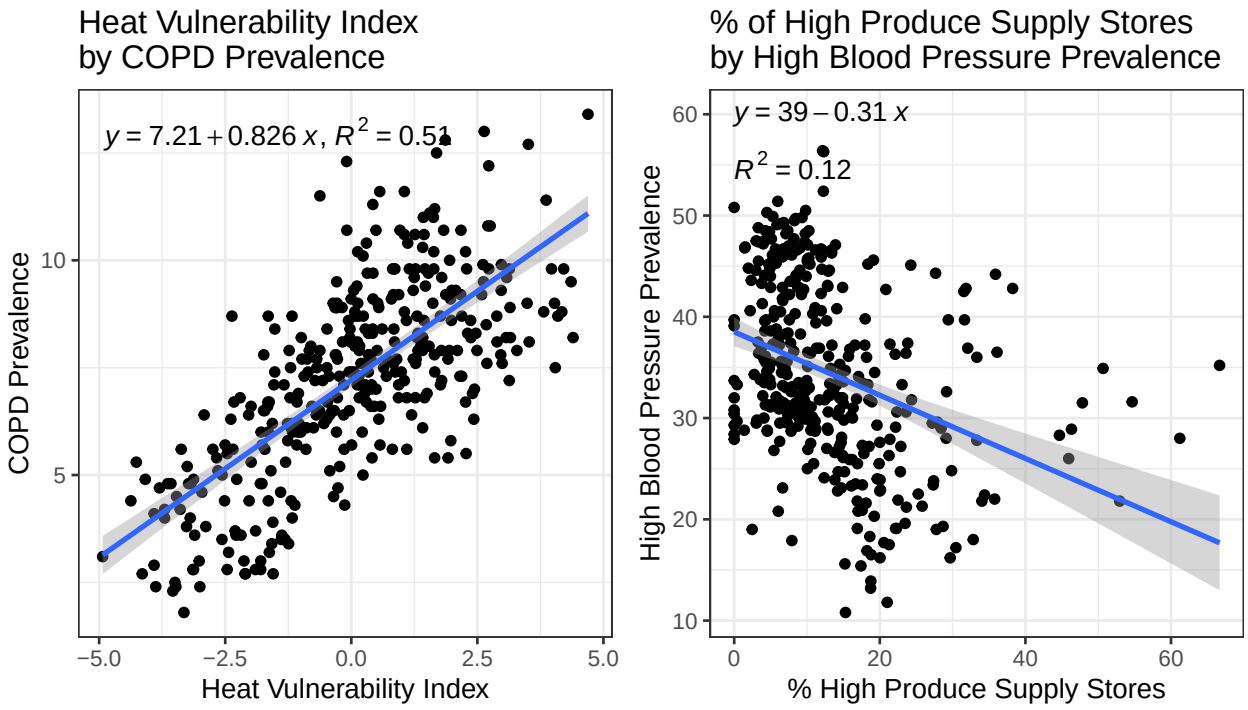
EDA

looking specifically at 2010

Even though the last census concluded in 2020, we are using 2010 data because that was the year for the most robust and informative data available from our many sources.

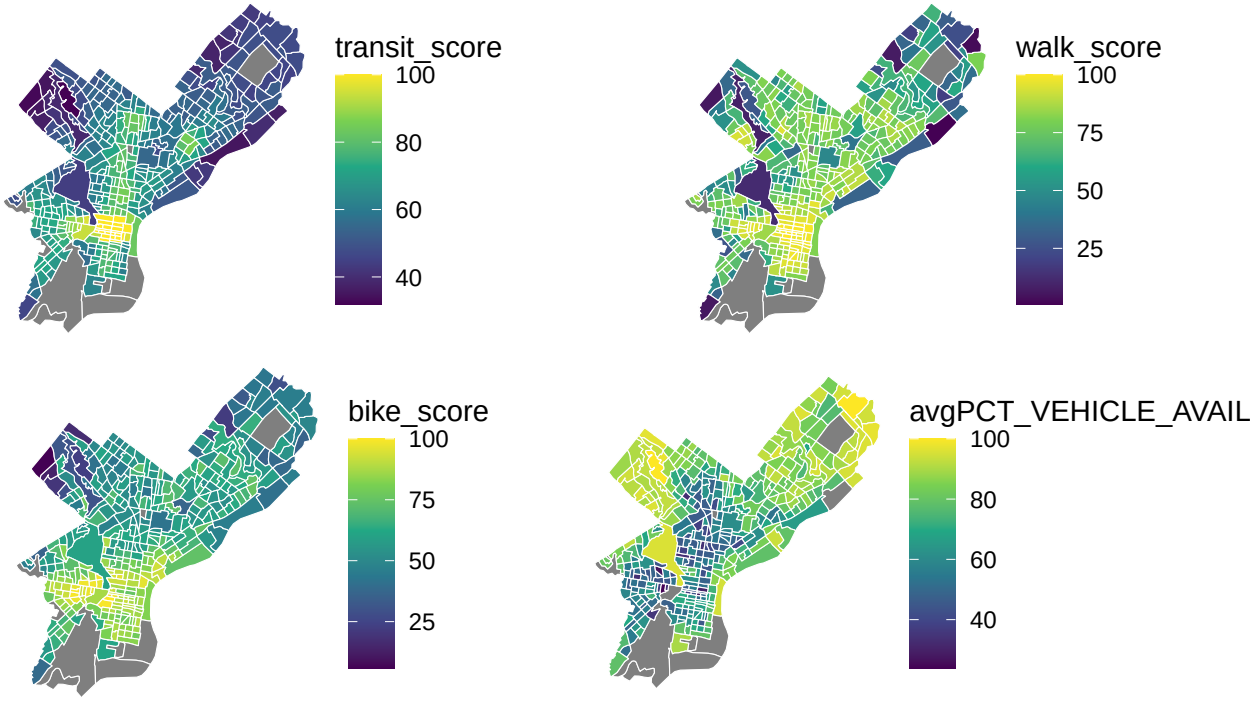
```
##   |                                                                      |
```





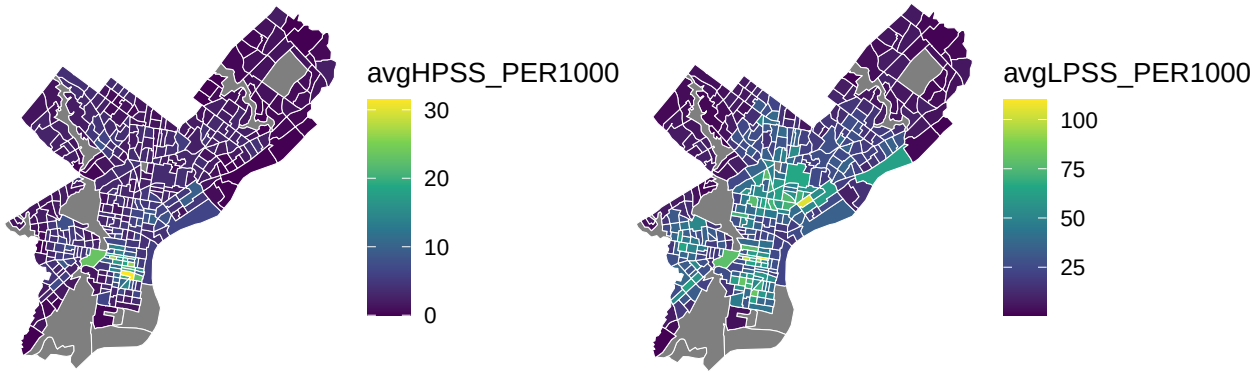
[talk about how transit and walk score is a little lower in our tracts with high inequality]

Utility of Public Transport, Walkability, Bikebility, and Car availability by Census Tracts



[Talk about we see both high and low produce supermarkets and stores in our tracts with high inequality] we also have number of restaurants but don't include the anlysis here for brevity

High vs. Low Produce Supemarkets and Stores by Census Tracts



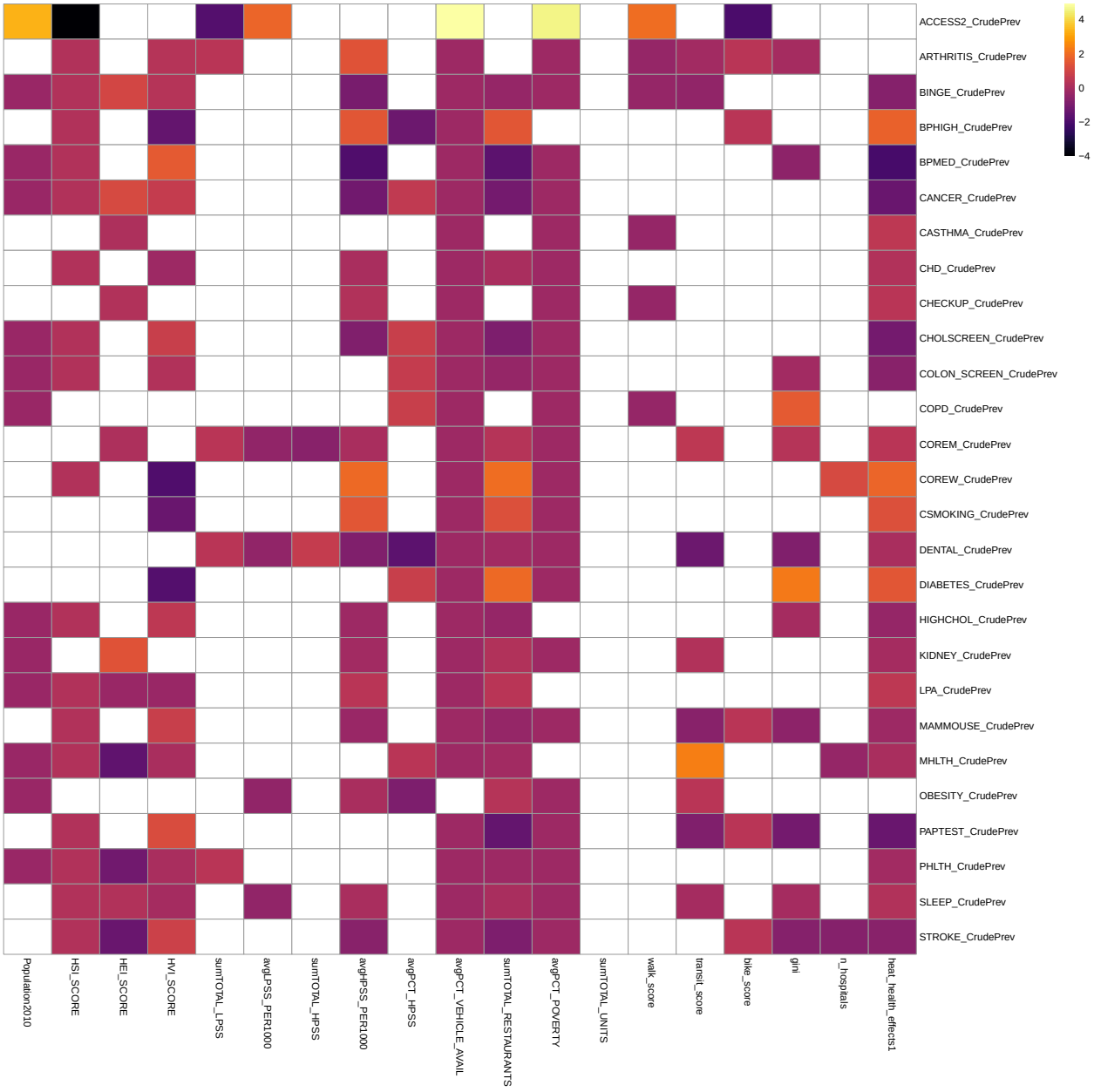
Analysis

Linear Model

Model data prep

```
## [1] 258  56
```

```
## [1] 111  56
```



pos gini beta: "COPD_CrudePrev", "COREM_CrudePrev", "DIABETES_CrudePrev" negative gini beta: "ARTHRITIS_CrudePrev", "BPMED_CrudePrev", "COLON_SCREEN_CrudePrev", "DENTAL_CrudePrev", "HIGHCHOL_CrudePrev", "MAMMOUSE_CrudePrev", "PAPTEST_CrudePrev","SLEEP_CrudePrev", "STROKE_CrudePrev"

Model Selection

Next we perform model selection on models to predict diabetes and COPD prevelance since they appear to have a strong positive correlation to gini index.

Pepare the data subsetting for model prediction of disease variables vs socioeconomic indicators

Forward Selection

Diabetes

```
##  [1] 9.153170e+03 1.046238e-27 6.566840e-28 5.313684e-28 4.621224e-28
##  [6] 4.278931e-28 4.071975e-28 3.881340e-28 3.770668e-28 3.643891e-28
## [11] 3.541627e-28 3.455380e-28 3.416352e-28 3.395181e-28 3.382157e-28
## [16] 3.375172e-28 3.372498e-28 3.369850e-28 3.367666e-28 3.365448e-28
## [21] 3.363454e-28
```

COPD

```
## [1] 21
```

Backward Selection

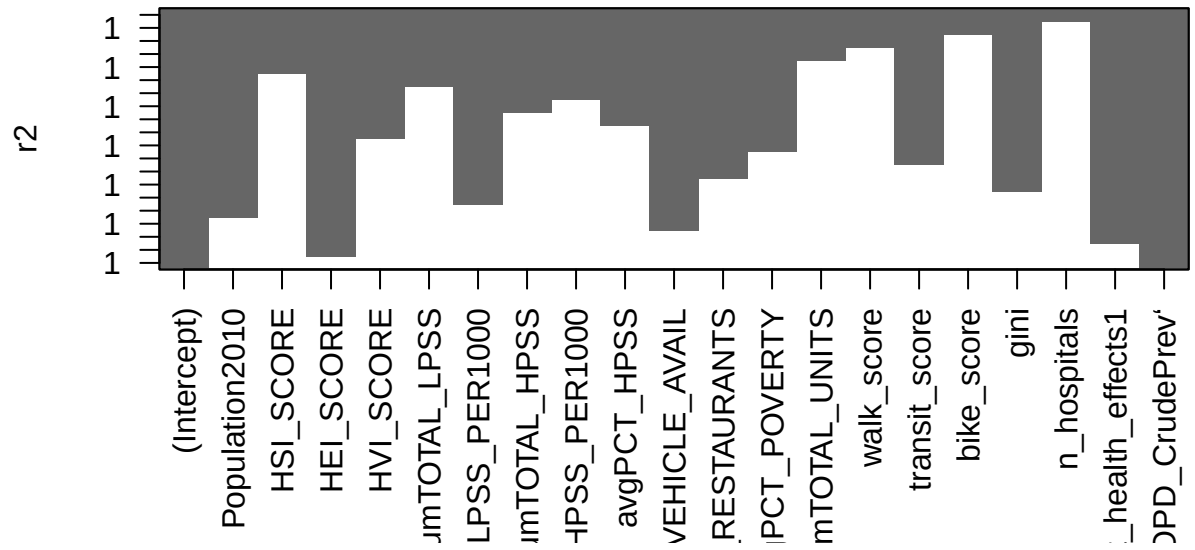
Diabetes

```
## [1] 21
```

COPD

```
##                 (Intercept)                   HEI_SCORE
##               2.455555e-15                1.858652e-16
##        heat_health_effects1 `final_data$COPD_CrudePrev`
##              -2.292235e-16                1.000000e+00
```

```
plot(fit.backward.copd, scale = "r2")
```



Evaluation of Forward and Backward selection

Diabetes

```
which.min(summary(fit.forward.diab)$rss)
```

```
## [1] 20
```

```
which.min(summary(fit.forward.diab)$rss)
```

```
## [1] 20
```

```
coef(fit.forward.diab, 20)
```

```
##                       (Intercept)                    Population2010
##                     -1.569968e-14                      1.986646e-19
##                         HSI_SCORE                         HEI_SCORE
##                      5.605775e-16                     -5.781967e-16
##                         HVI_SCORE                     sumTOTAL_LPSS
##                     -7.199359e-16                      3.778789e-18
##                   avgLPSS_PER1000                     sumTOTAL_HPSS
##                     -2.491557e-17                     -2.082416e-17
##                   avgHPSS_PER1000                       avgPCT_HPSS
##                      1.929901e-16                     -4.339753e-18
##              avgPCT_VEHICLE_AVAIL              sumTOTAL_RESTAURANTS
##                      6.488506e-17                      1.649184e-17
##                    avgPCT_POVERTY                    sumTOTAL_UNITS
##                     -3.217352e-18                     -4.200607e-19
##                        walk_score                     transit_score
##                      1.940334e-18                      2.792690e-18
##                        bike_score                              gini
##                      4.247765e-18                      1.407938e-15
##                       n_hospitals             heat_health_effects1
##                      1.347626e-16                      6.313680e-16
## `final_data$DIABETES_CrudePrev`
##                      1.000000e+00
```

```
which.min(summary(fit.backward.diab)$rss)
```

```
## [1] 20
```

```
which.min(summary(fit.backward.diab)$rss)
```

```
## [1] 20
```

```
coef(fit.backward.diab, 20)
```

```
##                       (Intercept)                    Population2010
##                     -4.459538e-15                      8.905011e-20
##                         HSI_SCORE                         HEI_SCORE
##                      3.712970e-16                     -1.080441e-16
##                         HVI_SCORE                     sumTOTAL_LPSS
##                     -3.453036e-16                      1.606824e-19
##                   avgLPSS_PER1000                     sumTOTAL_HPSS
##                     -2.945169e-18                     -5.262045e-18
##                   avgHPSS_PER1000                       avgPCT_HPSS
##                      6.168921e-17                     -5.572651e-18
##              avgPCT_VEHICLE_AVAIL              sumTOTAL_RESTAURANTS
##                      3.968924e-17                      1.026054e-17
##                    avgPCT_POVERTY                    sumTOTAL_UNITS
##                     -2.484314e-18                     -3.615268e-19
##                        walk_score                     transit_score
##                      1.504105e-18                      2.343544e-18
##                        bike_score                              gini
##                      3.765364e-18                      1.277819e-15
##                       n_hospitals             heat_health_effects1
##                      1.232242e-16                      7.573147e-16
## `final_data$DIABETES_CrudePrev`
```

```
##                    1.000000e+00
```

The RSS is the smallest for the 20-variable model for Diabetes in both forward and backward selection

COPD

```
which.min(summary(fit.forward.copd)$rss)
```

```
## [1] 20
```

```
which.min(summary(fit.forward.copd)$rss)
```

```
## [1] 20
```

```
coef(fit.forward.copd, 20)
```

```
##                  (Intercept)              Population2010
##                3.397340e-15                3.191191e-20
##                   HSI_SCORE                   HEI_SCORE
##               -1.241787e-17                1.580660e-16
##                   HVI_SCORE               sumTOTAL_LPSS
##                3.095724e-17               -4.975750e-19
##             avgLPSS_PER1000               sumTOTAL_HPSS
##               -5.322894e-19                8.596414e-18
##             avgHPSS_PER1000                 avgPCT_HPSS
##               -3.001471e-17               -1.170297e-18
##        avgPCT_VEHICLE_AVAIL        sumTOTAL_RESTAURANTS
##               -8.572170e-18               -1.127614e-18
##              avgPCT_POVERTY              sumTOTAL_UNITS
##                2.193157e-18                6.101153e-20
##                  walk_score               transit_score
##               -2.947037e-19               -2.388980e-18
##                  bike_score                        gini
##                3.005885e-19               -6.206644e-16
##                 n_hospitals       heat_health_effects1
##               -8.900563e-19               -2.411462e-16
## `final_data$COPD_CrudePrev`
##                1.000000e+00
```

```
which.min(summary(fit.backward.copd)$rss)
```

```
## [1] 20
```

```
which.min(summary(fit.backward.copd)$rss)
```

```
## [1] 20
```

```
coef(fit.backward.copd, 20)
```

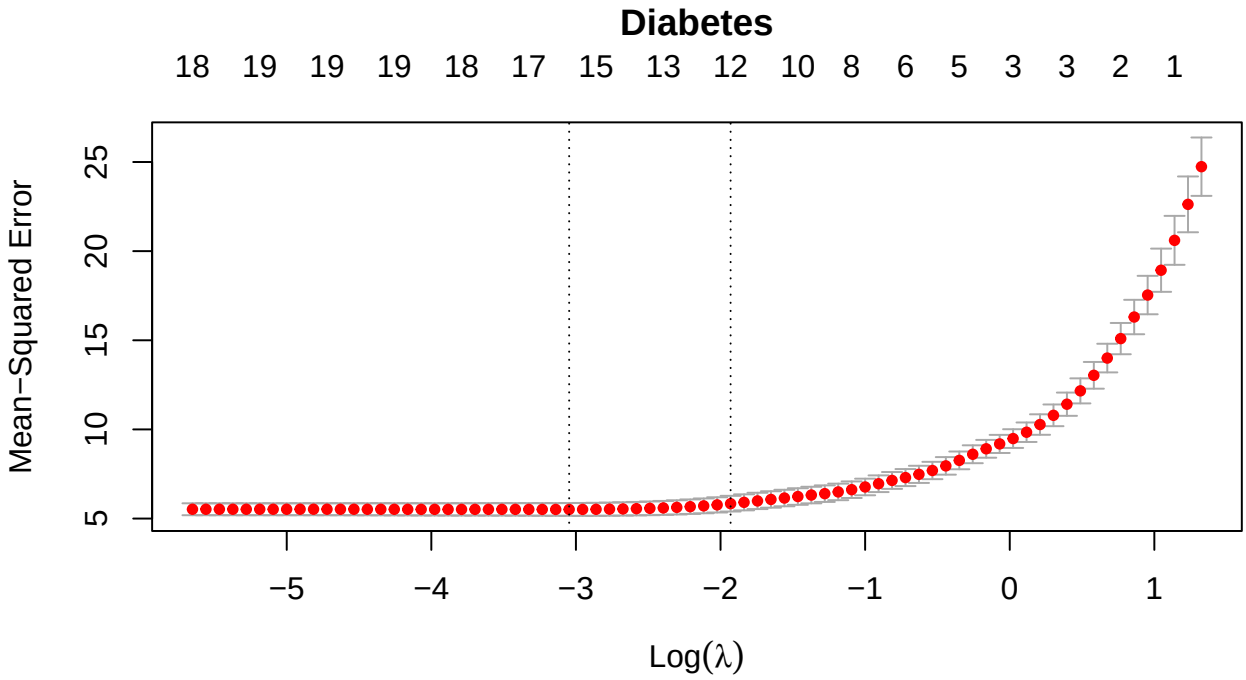
```
##                  (Intercept)              Population2010
##                3.397340e-15                3.191191e-20
##                   HSI_SCORE                   HEI_SCORE
##               -1.241787e-17                1.580660e-16
##                   HVI_SCORE               sumTOTAL_LPSS
##                3.095724e-17               -4.975750e-19
##             avgLPSS_PER1000               sumTOTAL_HPSS
##               -5.322894e-19                8.596414e-18
##             avgHPSS_PER1000                 avgPCT_HPSS
##               -3.001471e-17               -1.170297e-18
```

```
##        avgPCT_VEHICLE_AVAIL        sumTOTAL_RESTAURANTS
##               -8.572170e-18               -1.127614e-18
##               avgPCT_POVERTY               sumTOTAL_UNITS
##                2.193157e-18                6.101153e-20
##                  walk_score               transit_score
##               -2.947037e-19               -2.388980e-18
##                  bike_score                        gini
##                3.005885e-19               -6.206644e-16
##                 n_hospitals       heat_health_effects1
##               -8.900563e-19               -2.411462e-16
## `final_data$COPD_CrudePrev`
##                1.000000e+00
```

The RSS is the smallest for the 20-variable model for COPD in both forward and backward selection

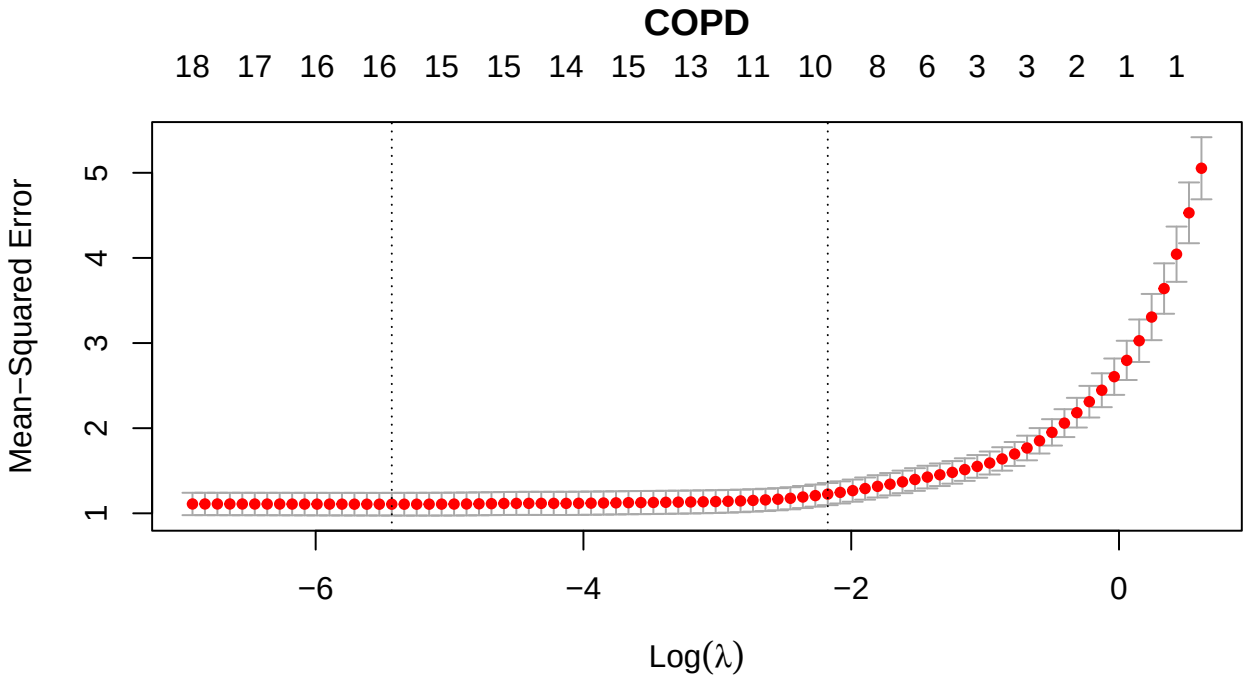
LASSO - predicting all diseases

LASSO loop




```
##
## Call:
## lm(formula = f_new, data = final_data_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1819 -1.4017 -0.1343  1.4373  6.3745
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.874e+01  1.920e+00  14.968  < 2e-16 ***
## Population2010      -3.486e-04  8.561e-05  -4.072 5.74e-05 ***
## HSI_SCORE           -7.054e-01  1.749e-01  -4.034 6.71e-05 ***
## HEI_SCORE            6.594e-01  3.240e-01   2.035 0.042588 *
```

```
## HVI_SCORE             1.353e+00  2.911e-01   4.647 4.75e-06 ***
## avgHPSS_PER1000      -1.357e-01  3.764e-02  -3.606 0.000356 ***
## avgPCT_VEHICLE_AVAIL -1.311e-01  1.405e-02  -9.331  < 2e-16 ***
## sumTOTAL_RESTAURANTS -3.423e-02  6.405e-03  -5.344 1.63e-07 ***
## avgPCT_POVERTY        9.262e-03  1.368e-02   0.677 0.498865
## bike_score           -2.118e-02  1.013e-02  -2.090 0.037359 *
## gini                 -5.069e+00  2.229e+00  -2.274 0.023564 *
## n_hospitals          -5.984e-01  3.860e-01  -1.550 0.121940
## heat_health_effects1 -2.821e+00  4.238e-01  -6.656 1.07e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.281 on 356 degrees of freedom
## Multiple R-squared:  0.7976, Adjusted R-squared:  0.7908
## F-statistic: 116.9 on 12 and 356 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = f_new, data = final_data_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6771 -0.5763 -0.0015  0.6450  3.6077
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          13.449200   0.893747  15.048  < 2e-16 ***
## HEI_SCORE             0.878069   0.054300  16.171  < 2e-16 ***
## avgHPSS_PER1000      -0.036264   0.018637  -1.946 0.052457 .
## avgPCT_HPSS          -0.025461   0.006810  -3.739 0.000215 ***
## avgPCT_VEHICLE_AVAIL -0.042243   0.006187  -6.827 3.71e-11 ***
## sumTOTAL_RESTAURANTS -0.007578   0.002935  -2.582 0.010223 *
```

```
## avgPCT_POVERTY        0.012661   0.006041   2.096 0.036787 *
## walk_score           -0.001614   0.003903  -0.414 0.679468
## transit_score        -0.020714   0.006127  -3.381 0.000802 ***
## gini                 -3.174678   0.990450  -3.205 0.001471 **
## heat_health_effects1 -0.663080   0.184491  -3.594 0.000371 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 358 degrees of freedom
## Multiple R-squared:  0.7971, Adjusted R-squared:  0.7914
## F-statistic: 140.6 on 10 and 358 DF,  p-value: < 2.2e-16
```

Reduced Linear Model

Diabetes

Appendix I: Data Cleaning

```
## [1] "3352b422b0ef5aaabae7a20651522b5a7688e0bf"
```

Data landing and inspection

In order to directly download data from the Census API, you need a key. You can sign up for a free key here. Type your key in quotes using the census_api_key() command.

376 tracts here

384 census tracts here (2010 data)

384 census tracts here (2010 data)

```
##   OBJECTID_1  TractFIPS NAME10 OBJECTID HSI_SCORE HEI_SCORE HVI_SCORE
## 1        276 4.2101e+10      1      188 0.9768257  -3.31447 -3.477334
##   N_VERYHIGH Shape__Area Shape__Length
## 1          0     1202257      4528.079
```

philly_census_health x population_philly (by race)

```
## [1] 376  77
```

```
## [1] 376  86
```

Next we just need to separate out the last digit in the food_access GEOID10 to transform the census block group number into a census tract number (https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html#:~:text=Census%20Tract,482012231001)

The new census tract column is now `TractFIPS` like the other datasets.

Then we can average the estimates for each census tract to get one row of data for each census tract.

```
## # A tibble: 6 x 9
##   TractFIPS   sumTOTAL_LPSS avgLPSS_PER1000 sumTOTAL_HPSS avgHPSS_PER1000
##   <chr>               <int>           <dbl>         <dbl>           <dbl>
## 1 42101000100           106            25.2          19.5            4.60
## 2 42101000200           120            44.7          17.8            6.62
## 3 42101000300           182            77.7          35.2           14.1
## 4 42101000401            74            28.1          20              7.60
## 5 42101000402           199            59.5          45.8           13.8
## 6 42101000500           148            58.9          23.8            9.45
## # i 4 more variables: avgPCT_HPSS <dbl>, avgPCT_VEHICLE_AVAIL <dbl>,
## #   sumTOTAL_RESTAURANTS <int>, avgPCT_POVERTY <dbl>
```

So now that we are down to about 380 census tracts, we can merge with the other data for `merege3`

```
## [1] 376  94
```

Geolocating addresses (for the housing and hospital data) to census tracts

Getting census tract-level gini indices

```
## [1] "GEOID"  "state"  "county" "tract"  "gini"
```

Merge these with the rest of the data matrix

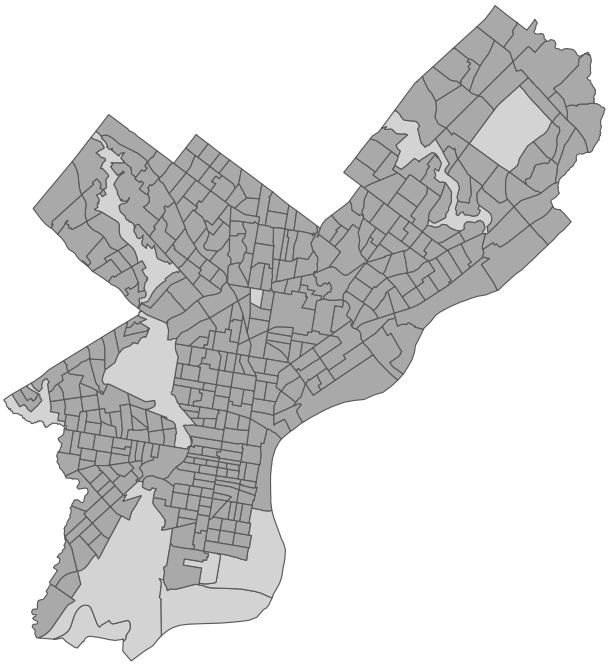
Preparing hospital data

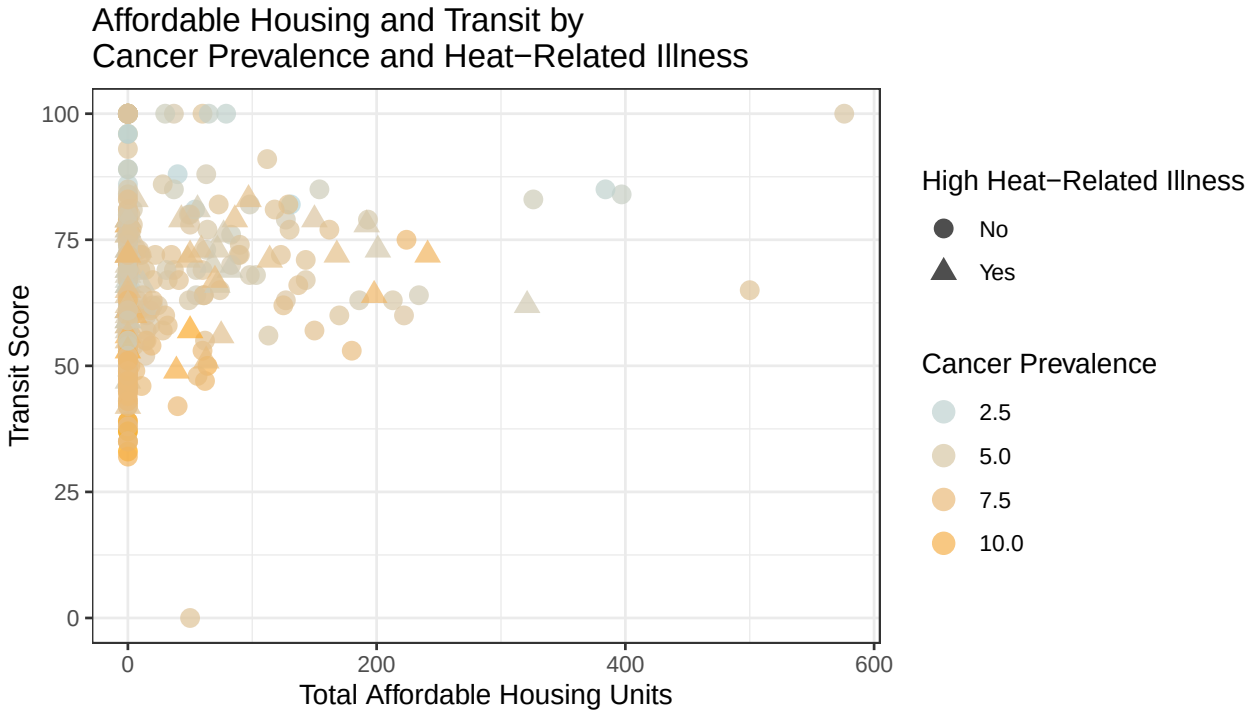
```
##    OBJECTID                              HOSPITAL_NAME         STREET_ADDRESS
## 25        1              Aria Health- Frankford Campus  4900 Frankford Avenue
## 22        2             Aria Health- Torresdale Campus       3998 Red Lion Rd
## 23        3 Belmont Center for Comprehensive Treatment     4200 Monument Road
## 36        4                      Chestnut Hill Hospital 8835 Germantown Avenue
## 17        5    The Children's Hospital of Philadelphia 3401 Civic Center Blvd
## 6         6     Einstein Medical Center - Philadelphia   1200 West Tabor Road
##            CITY STATE ZIP_CODE PHONE_NUMBER     HOSPITAL_TYPE   cxy_lon
## 25 Philadelphia    PA    19124 215-831-2000   General medical -75.08039
## 22 Philadelphia    PA    19114 215-612-4000   General medical -74.98026
## 23 Philadelphia    PA    19131 215-877-2000 Behavioral health -75.21645
## 36 Philadelphia    PA    19118 215-248-8200   General medical -75.21210
## 17 Philadelphia    PA    19104 215-590-1000   General medical -75.19318
## 6  Philadelphia    PA    19141 215-456-7890   General medical -75.14383
##     cxy_lat
## 25 40.02035
## 22 40.06777
## 23 39.99804
## 36 40.07856
## 17 39.94810
## 6  40.03806
```

Final merge

Appendix II: Extended EDA

We used only the following residential census tracts for the analysis





EDA looking at relationship of our variables

[distributions of hospitals]

[distribution of affordable housing]

Appendix III: Data Dictionary

Here is the break down of variable names:

Variable	Description
TractFIPS	Census tract identifier
Population2010	Population in 2010
ACCESS2_CrudePrev	Access to health insurance prevalence
ARTHRITIS_CrudePrev	Arthritis prevalence
BINGE_CrudePrev	Binge drinking
BPHIGH_CrudePrev	High blood pressure
BPMED_CrudePrev	On blood pressure medication
CANCER_CrudePrev	cancer
CASTHMA_CrudePrev	asthma
CHD_CrudePrev	heart disease
CHECKUP_CrudePrev	up-to-date on checkups
CHOLSCREEN_CrudePrev	cholesterol screen
COLON_SCREEN_CrudePrev	colon screen
COPD_CrudePrev	COPD
COREM_CrudePrev	male core checkups
COREW_CrudePrev	female core checkups
CSMOKING_CrudePrev	smoking
DENTAL_CrudePrev	dental visits
DIABETES_CrudePrev	diabietes
HIGHCHOL_CrudePrev	high cholesterol
KIDNEY_CrudePrev	kidney disease
LPA_CrudePrev	leisure-time physical activity
MAMMOUSE_CrudePrev	mamogram use
MHLTH_CrudePrev	mental health poor for > 14 days
OBESITY_CrudePrev	obesity
PAPTEST_CrudePrev	pap test
PHLTH_CrudePrev	poor physical health
SLEEP_CrudePrev	sleep issues
STROKE_CrudePrev	stroke
TEETHLOST_CrudePrev	teeth lost
COUNT_WHITE_NH	Count of White residents
COUNT_BLACK_NH	Count of Black residents
COUNT_ASIAN_NH	Count of Asian residents
COUNT_HISPANIC	Count of Hispanic Residents
PERCENT_WHITE_NH	Percent White
PERCENT_BLACK_NH	Percent Black
PERCENT_ASIAN_NH	Percent Asian
PERCENT_HISPANIC	Percent Hispanic

Variable	Description
PERCENT_ASIAN_NH	Percent Asian
PERCENT_HISPANIC	Percent Hispanic
HSI_SCORE	Displays sensitivity to heat by census tract, incorporating demographic, health and disability indicators (2019)
HEI_SCORE	Displays heat exposure by census tract incorporating daytime and nighttime land surface temperature, surface reflectivity, building density, and vegetation (2017-2019).
HVI_SCORE	Displays heat vulnerability by census tract, incorporating heat exposure and sensitivity indicators (2017 - 2019).
sumTOTAL_LPSS	Total low produce supply stores within a mile of internal census block groups
avgLPSS_PER1000	Average number of low-produce supply stores per 1k people
sumTOTAL_HPSS	Total high produce supply stores within a mile of internal census block groups
avgHPSS_PER1000	Average number of high-produce supply stores per 1k people
avgPCT_HPSS	Average percentang of all stores within half mile walking distance of the block group that are high produce supply
sumTOTAL_RESTAURANTS	Total restaurants in the tract
avgPCT_POVERTY	Average percent of people in poverty across census block groups
sumTOTAL_UNITS	Total affordable housing unit within the census tract
walk_score	Redfin walk score
transit_score	REdfin transit score
bike_score	Redfin bike score
gini	Gini index
n_hospitals	Number of hospitals int he census tract
heat_health_effects	Indicator variable caputring if this census tract is among the topp 75 most vulnerable to experience health impacts of heat vulnerability

Figure 2: Data Dictionary