The Detection of Redlining in Chicago Insurance Data

STAT7026 Final Project Part B

Background

Redlining is a practice of denying services, either directly or through selectively raising prices, to residents of certain areas based on the racial or ethnic composition of those areas.[1] In this report, we are investigating data collected by the U.S. Commission on Civil Rights to examine charges that insurance companies were "redlining" certain neighbourhoods in Chicago in the 1970s.

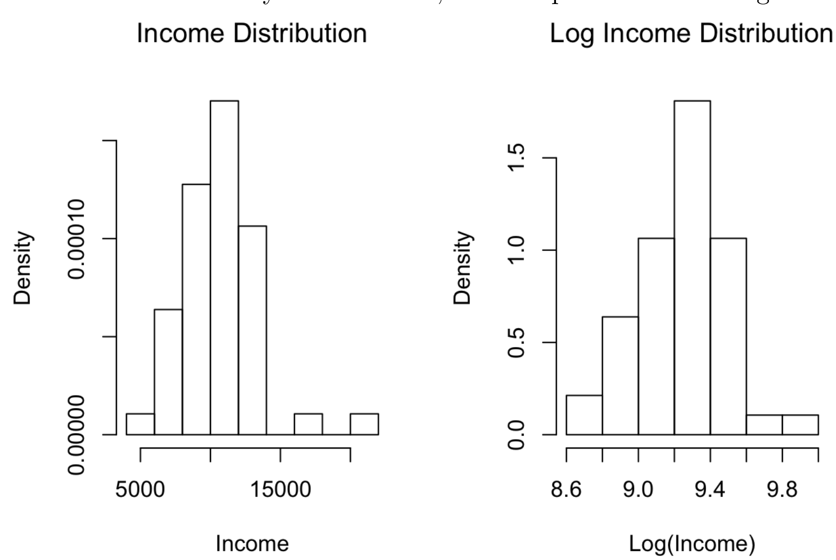
Data Cleaning and Manipulation

The data is stored by Zip codes and each zip code has a series of corresponding variables `Fire`,`Theft`,`Income`,`Race` and `Age`. The dependent variables we are interested are `Volun` and `Invol`. Using `Invol` only to represent insurability of insurance companies might not be appropriate, since the willingness of people to buy insurance differs. In other words, low value of `Invol` in a district (zip code area) does not guarantee it is redlining-free. It could be due to educational background that people simple don't want to buy insurance. Therefore, we construct a new variable called `rRej` (rate of rejection) by calculating the ratio between `Invol` and `Volun+Invol`. In this way, `rRej` is the proportion of people who are rejected by private insurance companies (hence have to buy from government), to **roughly** all people who wanted to buy insurance.

$$\text{rRej} = \frac{\text{Invol}}{\text{Invol+Volun}}$$

The only drawback of rewriting is that some people could give up on buying insurance from government after being rejected by insurance companies. The main obstruction here is low income. People can stop buying or renewing insurance simply because they were not able to afford it. So there can be loss from `Volun` to `Invol`, hence our `rRej` is overestimated. But generally we believe this is a better expression than `Invol` itself.

We transform the variable `Income` by taking logarithm and rename it as `LogIncome`. This is mainly a conventional action since the income distribution is usually skewed. Also, it is the predictor with largest scale.



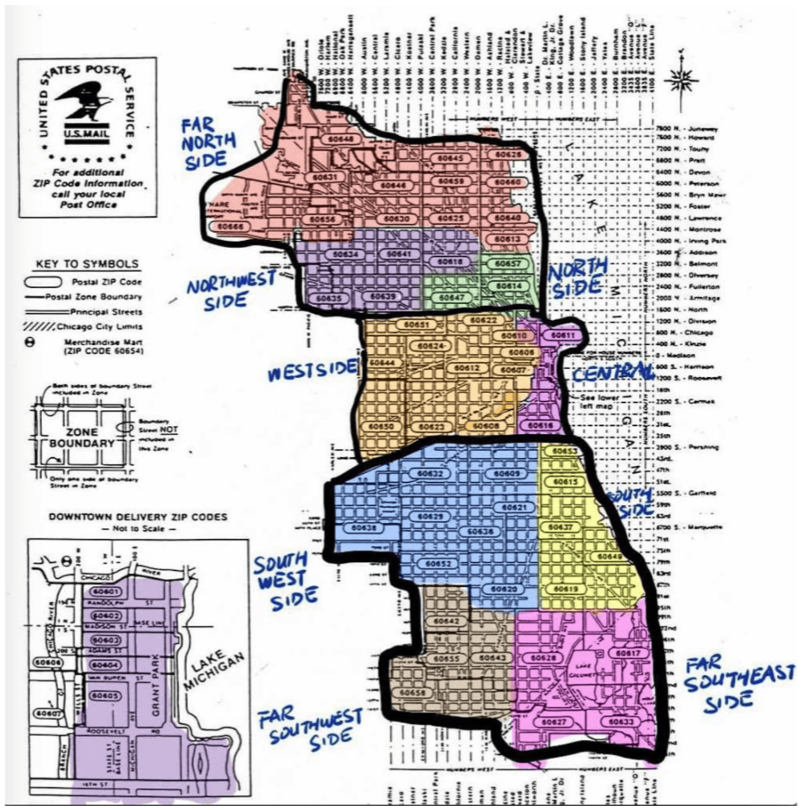
Moreover, we extract the longitude and latitude from Jeffrey Breen's R package `zipcode`. [2] One funny thing to notice is that Jeffrey's zipcode data was updated in 2011, but the zipcodes of Chicago we are dealing with can be traced

[1] Redlining, Wikipedia, https://en.wikipedia.org/wiki/Redlining.

[2] My first R package: zipcode, by Jeffrey Breen, https://jeffreybreen.wordpress.com/2011/01/05/cran-zipcode/.

back to 40 years ago. And in fact, by comparison, we find two of those zip codes 60627 (near Dolton) and 60635 (near Elmwood Park) are abandoned, so we input the missing geo-location information manually.

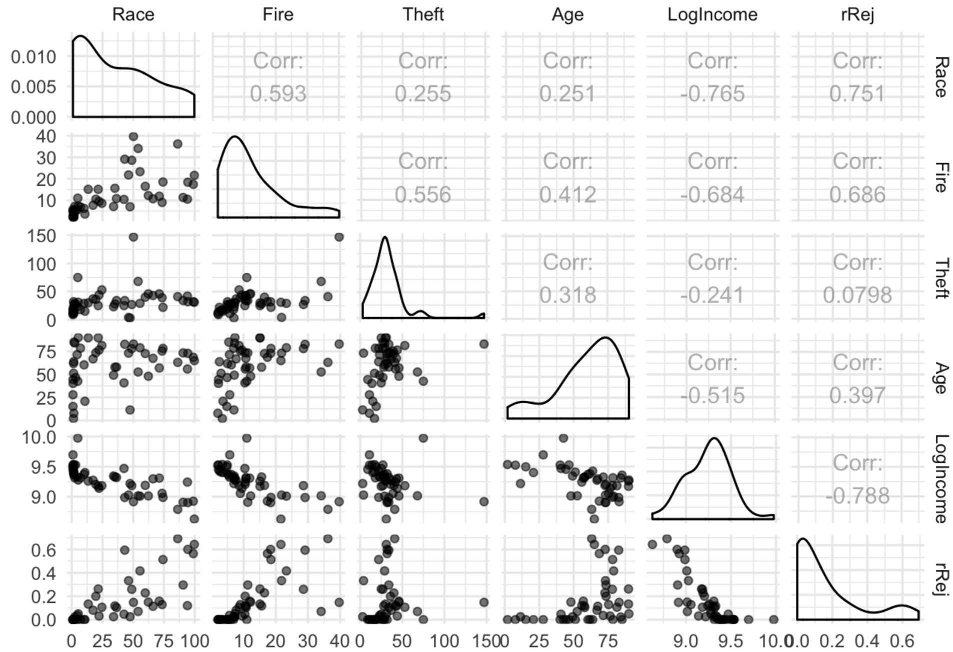
Lastly, based on a map of community areas in Chicago [3], we divide 1970s zip code map into 9 major communities and save this information in a new variable `Suburb`. On a larger scale, we cluster the communities into 3 regions: *North side*, *West side* and *South side*. The classification of regions is saved as a new variable `NSW`. Additionally, we add an alternative classification which combines North side and West side together as the new *North side* and *South side* as before.



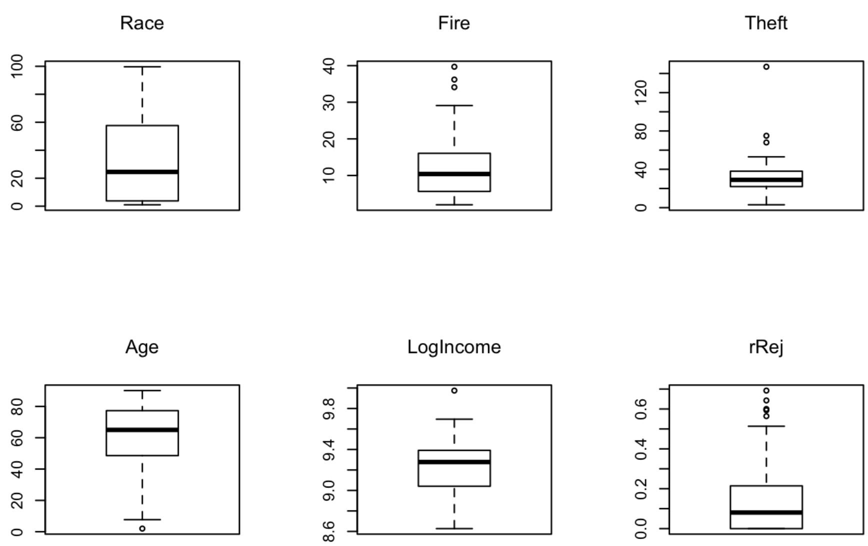
Exploratory Analysis

We use scatter matrix to get a general idea about how variables are correlated.

[3] Community areas in Chicago, Wikipedia, https://en.wikipedia.org/wiki/Community_areas_in_Chicago.



We notice that the response `rRej` is postively correlated with `Race` and `Fire`, which is not surprising. We doubt that insurance companies are indeed redlining, and traditionally, fire risk is a concern when evaluating an insurance case. As for `Race`, if it is the main cause of `rRej` then the redlining is confirmed. `LogIncome` is negatively correlated with `rRej`. Although may sound cruel, refusing selling or renewing insurance to poor people is legal. Another secret here is that `LogIncome` is highly correlated with `Race`. This entwinement can really produce some problems as high rejection rate can be explained by low income which is legal, or it can be explaiend by high percentage of minority races.



Then we plot the boxplots of response and predictors. Note that the mid 50-percentile of `Race` has a rather large range approximately from 0 to 60. To conclude, the level of racial diversity of Chicago is very high. Districts vary from districts. Since we are studying the effects of `Race`, we really need to take locations into consideration.

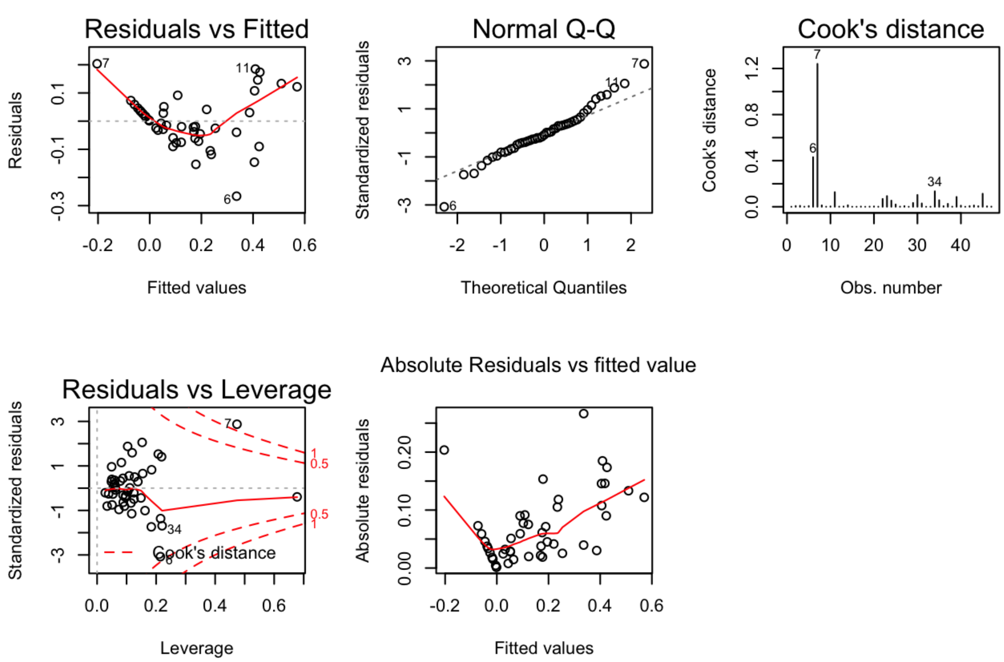
Model Building

Our goal in this part is to fit an appropriate linear model for rejection rate. And the target predictor is `Race`. We can simply fit in a reduced model with one predictor `Race` only. The result is quite satisfying, `Race` shows its significance when other variables are ignored. (Detailed summary table is left out.)

The next step is trying to fit a full model with all predictors in an order of `Fire`, `Theft`, `LogIncome`, `Race` and `Age`. We particularly pick this order because we tend to treat `Race` and `Age` as extra reasons explaining variabilities when fitting the model. That is to say, we believe those insurance companies made their decisions first based on those legit reasons, then (probably) based on the those redlining-related factors.

```
##
## Call:
## lm(formula = rRej ~ Fire + Theft + LogIncome + Race + Age, data = insure)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.266383 -0.052177 -0.007906  0.043626  0.203596
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4162387  1.1262226   1.258 0.215684
## Fire         0.0104839  0.0025595   4.096 0.000193 ***
## Theft       -0.0033497  0.0008227  -4.071 0.000208 ***
## LogIncome   -0.1545067  0.1168250  -1.323 0.193316
## Race         0.0023457  0.0007269   3.227 0.002461 **
## Age          0.0010742  0.0008012   1.341 0.187358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09767 on 41 degrees of freedom
## Multiple R-squared:  0.7886, Adjusted R-squared:  0.7628
## F-statistic: 30.59 on 5 and 41 DF,  p-value: 7.963e-13
```

For this model, `Race` is significant along with `Fire` and `Theft`, while `LogIncome` and `Age` seem to be not that significant. Then we check the diagnostic plots of this model.



In residuals vs fitted plot, the alignment of some points in the left part is a perfect linear line. We will investigate further later.

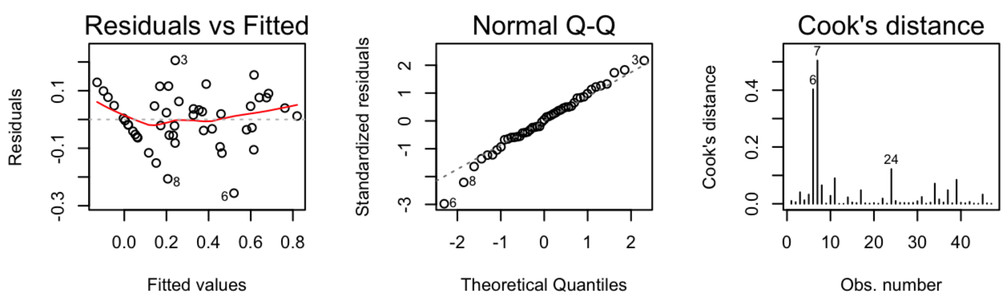
It also suggests a huge problem of heteroscadasticity, then we might need to do some transformation on the response to stablize the variances. In this case, we transform the response `rRej` by taking a square root. So our full model is:

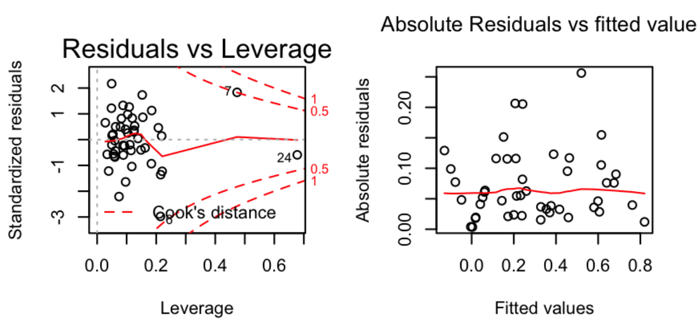
$$\sqrt{\text{rRej}} = \underset{(1.120)}{1.230} + \underset{(0.003)}{0.012}\,\text{Fire} - \underset{(0.001)}{0.003}\,\text{Theft} - \underset{(0.116)}{0.139}\,\text{LogIncome} + \underset{(0.001)}{0.004}\,\text{Race} + \underset{(0.001)}{0.003}\,\text{Age}$$

```
##
```

```
## Call:
## lm(formula = sqrt(rRej) ~ Fire + Theft + LogIncome + Race + Age,
##     data = insure)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.256243 -0.053366  0.004027  0.055390  0.205400
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2303861  1.1195508   1.099 0.278180
## Fire         0.0117327  0.0025443   4.611 3.88e-05 ***
## Theft       -0.0032041  0.0008179  -3.918 0.000332 ***
## LogIncome   -0.1390405  0.1161329  -1.197 0.238084
## Race         0.0039643  0.0007226   5.486 2.32e-06 ***
## Age          0.0027995  0.0007964   3.515 0.001088 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09709 on 41 degrees of freedom
## Multiple R-squared:  0.8809, Adjusted R-squared:  0.8664
## F-statistic: 60.66 on 5 and 41 DF,  p-value: < 2.2e-16
```

The summary statistics of this model is even better. Only `LogIncome` and the intercept stay insignificant.





The linear alignment of data points in residuals vs fitted plot still exists. After some trials we identify that these points belong to data with zero `rRej`. Another common feature of these data is high `LogIncome`.

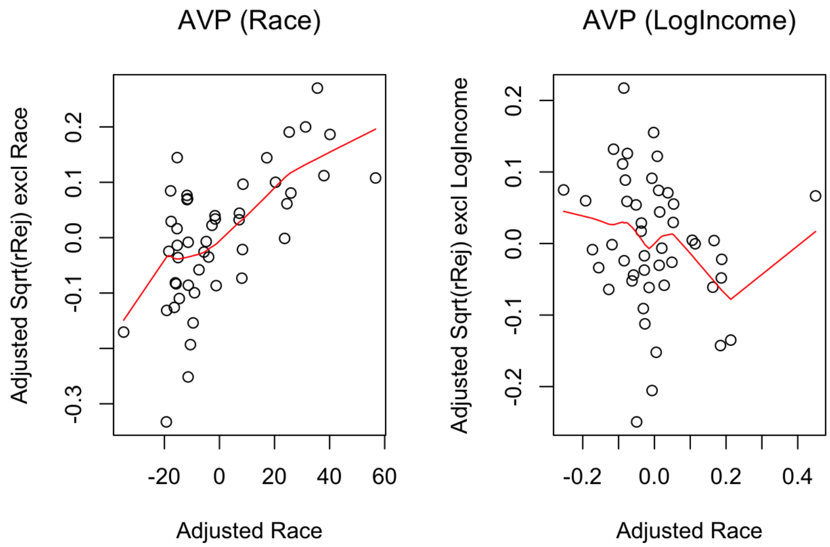
Addtionally, we can find out that those three influential points standing out are zip codes 60610, 60611, 60607 respectively. Coincidently, they are all zip codes near or in *Central* area. And what make them speical are

- 60610 has quite high level of `Race` (near 3rd quantile), very high level of `Fire`, very high level of `Theft`, low level of `LogIncome`, but a level of `rRej` below 50 percentile.
- 60611 has low level of `Race`, high level of `Theft` and the highest level of `LogIncome`, and zero `rRej`. This also belongs to that weird straight line in residual plot.
- 60607 has high `Race`, the highest values of both `Theft` and `Fire` and low `LogIncome` and a rather low level of `rRej`.

As discussed before, even with a significant p-value of `Race` and an insignificant p-value of `LogIncome`, one cannot simply separate the two from each other without doing some extra analysis. So far, the existence of redlining still remains as a myth.

Showdown of Two Nemesis

A great tool to resume our redlining detection odyssey is the added variable plot which is capable of showing the relationship between `rRej` and `Race`/`LogIncome` adjusted for the other explanatory variables in the model.



The added variable plots indicate that, the information of what `Race` knows about `rRej` is more than that of what `LogIncome` knows about `rRej`. The linearity in the left hand side plot is more obvious. Besides, the non-constant variance in the right hand side plot introduces some outliers, and may be responsible for the false impression that `LogIncome` is truly a competitor with `Race`. To conclude, **`Race` is the main cause of variabilities in our model instead of `LogIncome`, we can confirm that those insurance companies indeed redlined some of their cases.**

Considering Locations

As mentioned, location seems to matter in our model. The auto-detected outliers in our model also outlines a certain part of Chicago City, which is the *Central* area. We pick the North-South-West division and fit the previously refined model into different subsets, which only contains data from the same region. Then we check summary statistics to see if our prior knowledge still works.

Surprisingly, `Race` is significant in none of these three models. Here we only show the model of *South side*, it is the only model with some significant terms as well.

```
##
## Call:
## lm(formula = sqrt(rRej) ~ Fire + Theft + LogIncome + Race + Age,
##     data = insure[insure$NSW == "S", ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.088862 -0.048209 -0.001366  0.037047  0.111292
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.3954888  1.3764992   2.467  0.02831 *
## Fire         0.0139333  0.0034516   4.037  0.00141 **
## Theft        0.0002151  0.0022793   0.094  0.92626
## LogIncome   -0.3674316  0.1455652  -2.524  0.02540 *
```

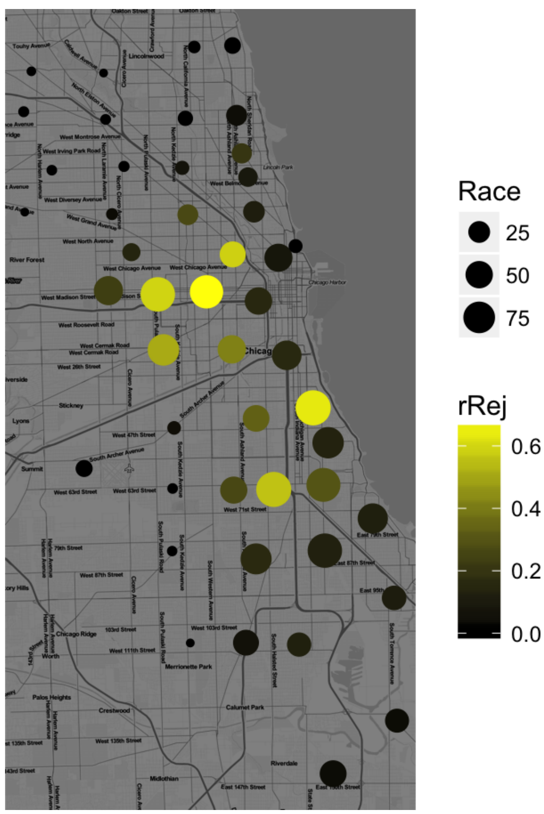
```
## Race         0.0021882  0.0010260   2.133  0.05259 .
## Age          0.0008455  0.0011904   0.710  0.49010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06573 on 13 degrees of freedom
## Multiple R-squared:  0.9474, Adjusted R-squared:  0.9272
## F-statistic: 46.86 on 5 and 13 DF,  p-value: 7.315e-08
```

Basically, the model we selected performs rather badly in small isolated regions of Chicago. Hence, another problem is presented in front of us: a unified model can work in a macroscopic view, but it might fail when constrained to partial data. Realistically, different regions of a city differentiate in the process of urbanization. Due to administrative reasons, they form different neighbourhoods. While due to more cultrual or economical reasons, they form communities. And various classes of people inhabit in those areas based on their social status, thus there might be tremendously diverging social problems in those areas. In our case, redlining prevails in Chicago, but its severity in a smaller section of city can be either more or less than the whole city.

A further study can be initiated with alternative separation methods, for example, splitting the area into smaller parts.

Mapping Data

A final act before wrapping up is to see our results plotted on a map. High rejection rate and high proportion of racial minority have overlaps in the *West side* and and northern part of the *South side*. These are the suspicious "Redlining zone". But not all large points are painted with bright color. *Far Southeast side* of Chicago have high `Race` values but does not seem to have a high rejection rate. And this agrees with our discovery when fitting the modified full model with subsetted data.



Conlusion and Beyond

Human sees this world as binary. No matter how various factors we take into consideration, when it comes to make a judgement, the ultimate question always remains, *to be or not to be?* Even though we can list all the pros and cons of an single object, our mind pushes us to determine if it is good or bad. We would like to respond to the question "if there is redlining in Chicago" with an universal and deterministic answer. What makes us disappointed is that the reality has never been binary. On different scales, distinct judgements have to be made.

As we stated, if we treat Chicago as an inseparable entity, **`Race` is responsible for insurability, i.e. redlining exists.** We almost happily concluded the case with an agreement that `Race` played a significant role in affecting the insurability rejection rate. Nevertheless, for some parts of Chicago, redlining is not obsious. Luckily, this can be explained by common sense. The redlining phenomenon targets residents from minority communities, so it is not very likely that redlining happens in a district with few racial minorities.

A proprer terminology to describe such dilemma is *ecological fallacy* [4]. In details, we have a stereotype that all individuals behave identically to the entity. After all, it is a family, a person who gets redlined. We believe this is the hidden gem in insurance data and we should handle those similar cases more carefully.

Last but not least, aggregating data by zip codes is debatable. A zip code is usually assigned to one or several streets, sometimes it is shared by multiple communities. Therefore, the inter-community correlations can be possibly diluted. Thus our analysis is less interpretable under such circumstance.

References

- Robert K. Nelson, LaDale Winling, Richard Marciano, Nathan Connolly, et al. Mapping Inequality. Retrieved October 25, 2017, from https://dsl.richmond.edu/panorama/redlining/#loc=0/-58/-148&opacity=0.8.
- Encyclopedia of Chicago. Redlining, Retrieved October 25, 2017 from http://www.encyclopedia.chicagohistory.org/pages/1050.html.
- Faraway, J. J. (2015). Linear models with R. Boca Raton: CRC Press, Taylor & Francis Group.
- Stamen Maps. (n.d.). Retrieved October 25, 2017, from http://maps.stamen.com/#toner/12/41.8790/-87.6606
- Multiple graphs on one page (ggplot2). (n.d.). Retrieved October 25, 2017, from http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

[4] Ecological fallacy, Wikipedia, https://en.wikipedia.org/wiki/Ecological_fallacy.