We want to find weather race or age of property were responsible behind rejection of insurance policies in 1970s, which is illegal.

Dataset – Insurance availability in Chicago

The purpose of this report is to find whether insurance company rejected policies for legal reasons such as high theft or fire rates, or whether inappropriate factors, such as race or age of house were also responsible. The data was collected by US Commission on Civil rights from December 1977 to February 1978. It includes number of cancellations, non-renewal, new policies, and renewals of home and fire policies for various neighbourhood of Chicago. Each neighbourhood can be identified from its zip-code. The dataset includes zipcode, minority race, fire-rate, theft-cases, age of houses, accepted voluntary market activity and accepted involuntary activity. Insurance companies can deny insurance on basis of high fire-rate or high theft-rate for a neighbourhood but denying on basis of race or age of property is illegal. We aim to find weather race or age of property were responsible behind rejection of insurance policies in 1970s, given that fire-rate and theft-rate were controlled. A total of 47 neighborhoods were observed in this study.

One of the first observation on dataset reveal that there are two columns representing acceptance and rejection of voluntary market insurance. As neighborhoods of different sizes might have varying numbers of accepted policies, the current representation is not valid. We are interested in probability of voluntary market insurance acceptance for a given neighbourhood. We can combine the two columns into single columns by finding probability of accepted insurance application given total number of applications are provided. The reason for this is that rejected policies are already subtracted from voluntary market activity. The new column will follow this formula

$$Accepted = \frac{Volun * 100}{Volun + Invol}$$

We can find relation between new variables with help of scatter matrix (*Figure-A*). There are various types of relationship between each variable, some of these relationships are not significant as they do not convey meaningful insights. For example, relationship between race and fire or race and theft doesn't make much sense, as they are not related to each other. All the scatter plots are *LOWESS-fitted,* allowing us to easily describe the relationship between two variables.

The relationship between race and income, as well as race and accepted, reveal interesting trends. The graph of race vs income graph shows a negative curvature relation, suggesting that neighborhoods with a higher minority population have lower median incomes compared to those with a lower minority population. However, we cannot definitively say that race is the only factor responsible for variations in income. Other factors influencing the median income of a particular neighborhood could include the types of jobs available and the cost of living. Posh neighborhoods may have higher median incomes. Regarding the relationship between race vs accepted, there is a clear decreasing trend. This suggests that as minority of neighbourhood decreases, chances of acceptance for insurance increases. Another important observation is that the data is right-skewed; many neighbourhoods with minority population close to 0 have acceptance rate of over 90%. In contrast, the data becomes more scattered across the LOWESS-fitted line as the minority population increases.

Regarding the relationship between fire and age, the weak trend suggest that older homes have a higher chance of catching fire compared to newer ones. However, this relationship is not strong, which could be attributed to different home designs, as historically, older homes were not built with the same considerations for natural disasters. The relationship between fire rates and acceptance shows a weak decreasing trend. As the fire rate in a neighborhood decreases, the acceptance of insurance increases. It is important to note that there are some outliers, which may indicate that other factors are also influencing the acceptance of insurance policies. This makes sense, as companies tend to reject insurance policies for neighborhoods with high fire rates, as they generally avoid risk-prone areas.

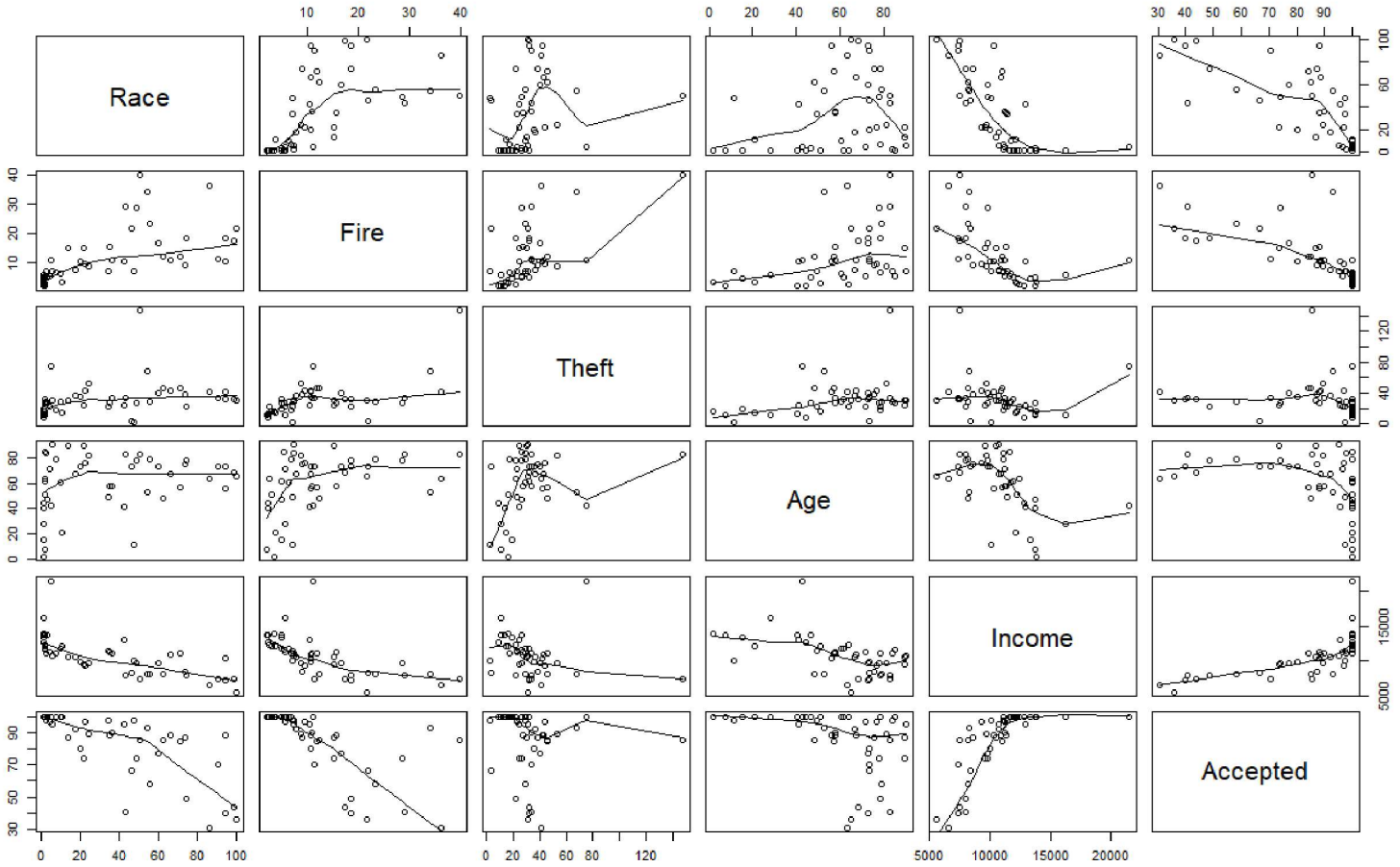

Figure – A: Scatter Matrix

The relationship between theft and accepted is one of the most perplexing. Even though high theft is a legal reason for companies to reject insurance, the graph suggests something different. The points appear randomly scattered, and there is no clear relationship. This could indicate two possibilities: 1) Theft rate is not responsible for acceptance of insurance by company. 2) Theft is responsible but only when other variables such as fire-rate, age and income are considered. We could further explore the correct possibility by fitting a multiple linear regression model with other variables against acceptance when theft rate is already included in model. However, based on the graph, one might conclude that the theft rate in a neighborhood does not significantly affect the chances of obtaining insurance.

In the relationship between age and accepted, there is weak curvature decreasing relationship, suggesting age might be one factor influencing the rejection of insurance; the evidence is not strong. In contrast, the relationship between income and accepted has a strong increasing trend, suggesting that neighbourhood with higher median income has higher insurance acceptance rate. This is valid, as income is one of the legal reasons for company to decide whether to accept the insurance request or not.

After analysis of all the plots, we realize that there are strong relationships, weak relationships, and no relationships at all between the various insurance data variables Importantly, in most cases, multiple factors are responsible for the rejection or acceptance of insurance. We aim to test which variables affect the acceptance rate and how they perform when other variables are introduced. Symbol plots and co-plots can be particularly helpful in this analysis.

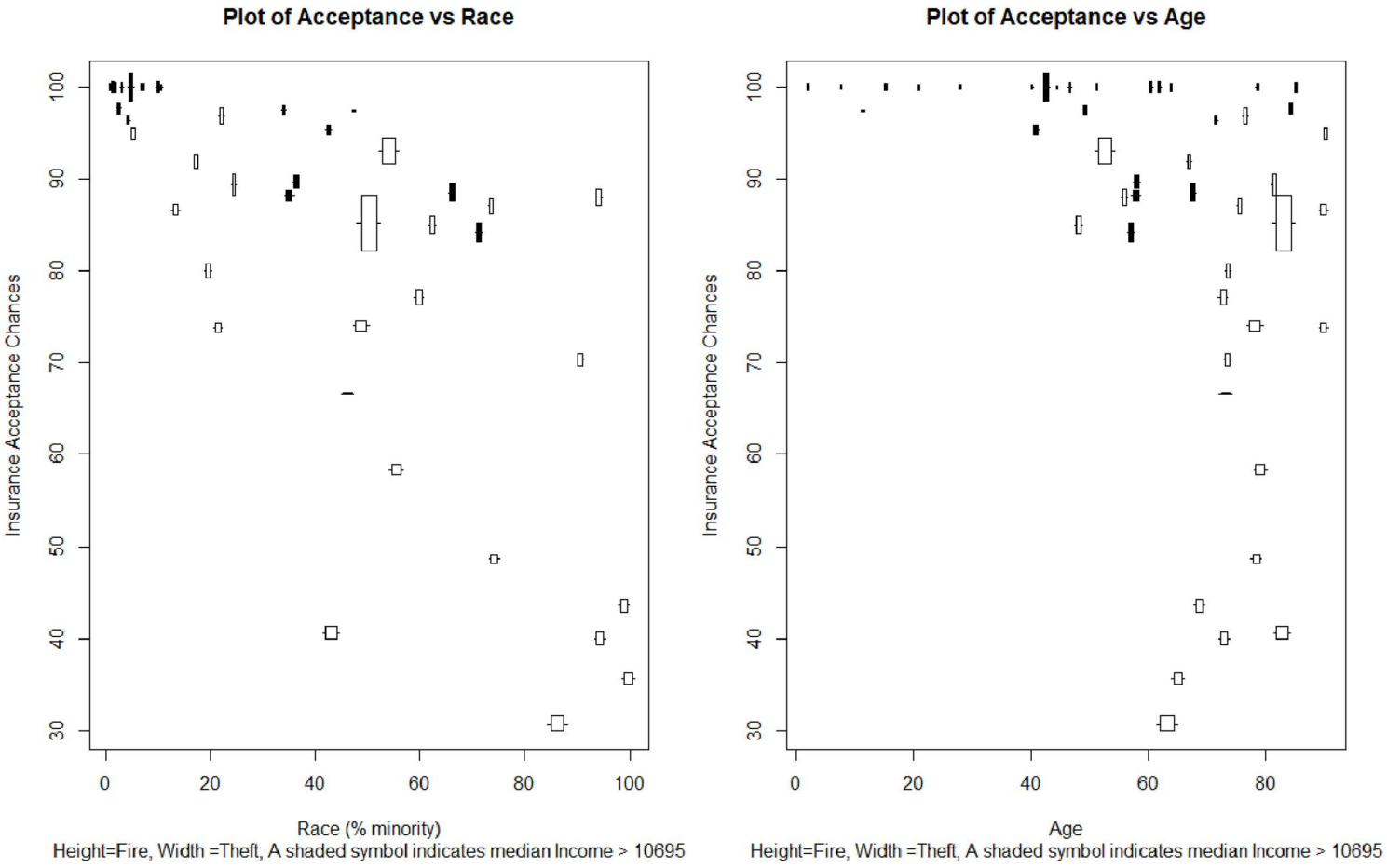

Figure B: Symbol Plot

Graph-1 in Figure-B represent relationship between race and insurance acceptance, incorporating fire, theft and income as additional variables to understand the overall effect. Each neighbourhood is represented by box, where the height of box indicates fire-rate and the width represents theft-rate for that neighbourhood. If median income of a neighbourhood exceeds $10,695 then box is shaded; this figure represents the median of median neighborhood incomes, so 50% of the boxes are shaded.

As seen in graph-1, there is a clear decreasing trend between insurance acceptance and the percentage of minority residents. Notably, most of the boxes with over 95% acceptance probability are shaded, indicating that higher-income neighborhoods have better chances of insurance acceptance. The widths of the boxes vary throughout the graph, showing no clear trend, which suggests that theft is not significantly important for

insurance acceptance. However, the height of the boxes tends to be smaller when acceptance is higher and larger when acceptance is lower, suggesting that fire rates also play a role in insurance acceptance.

Coming to Graph-2, which depicts the relationship between the age of houses and insurance acceptance, the trend is less clear, and the boxes are more scattered. An important observation is that as the age of the house increases, the number of rejections also increases. However, this relationship cannot be definitively concluded, as the age data is skewed—there are significantly more houses over 40 years old than those under 40. Nevertheless, it is noteworthy that almost all insurance applications from houses under 40 were accepted, likely due to their higher median income.

To further explore the relationships between race, insurance, fire, and income, we can also utilize co-plots.

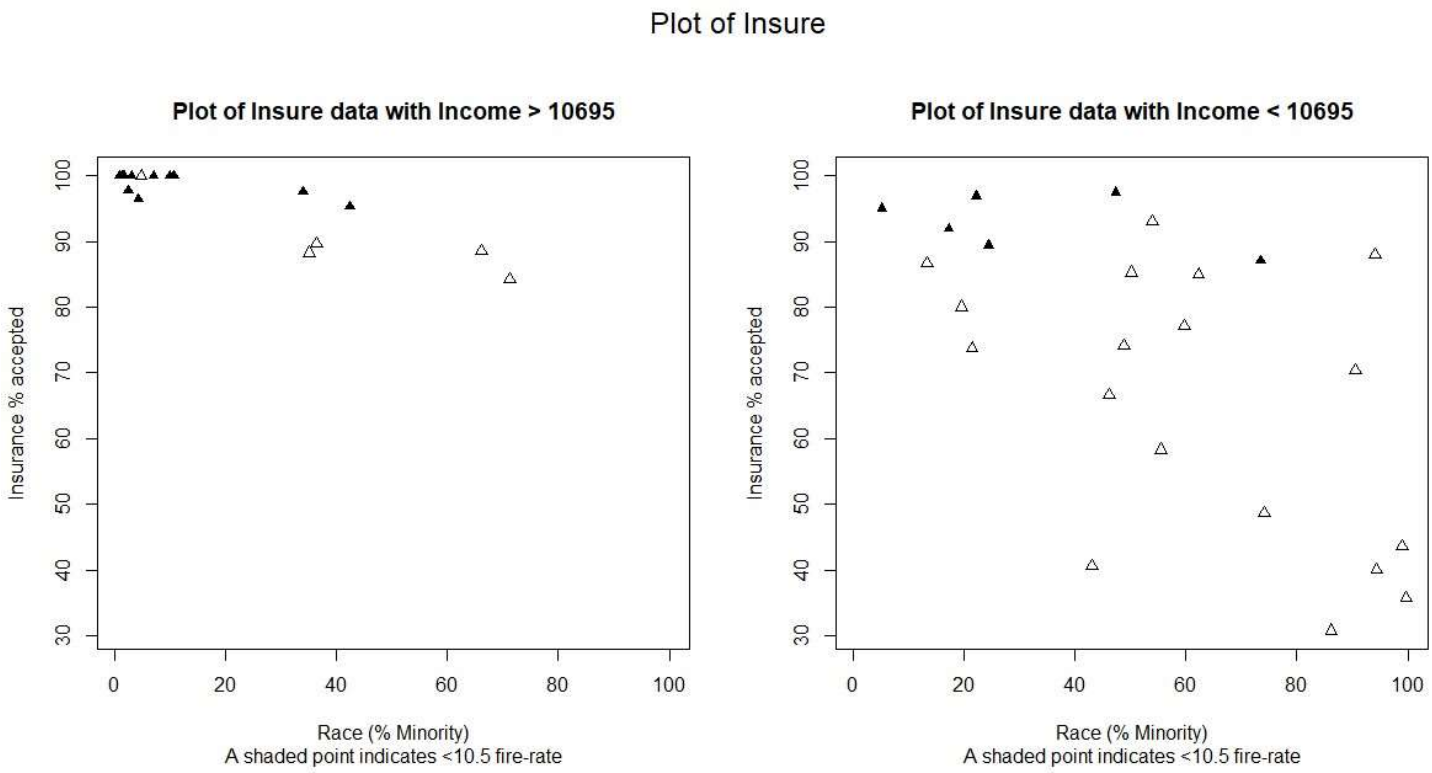

Figure C: Race co-plot

In Graph-1 of Figure-C, the trend is easy to understand. Neighbourhood with a median income greater than $10,695 and minority population of less than 20% have nearly perfect acceptance rates close to 100 In contrast, neighborhoods with lower acceptance rates either have higher minority populations or fire rates exceeding 10.5.

Comparing this with the second graph in Figure-C, the acceptance rates are significantly lower for neighborhoods with median incomes below $10,695. These neighborhoods are more scattered and exhibit lower insurance acceptance rates than those in the first graph. Notably, neighborhoods with fire rates below 10.5 still have acceptance rates greater than 90%, which is a positive finding. Thus, we can confirm the trend

observed in the symbol plot with the co-plot: fire rates, race, and income are indeed factors influencing insurance acceptance.

What we observed in second graph of Figure-B can be confirmed with Figure-D. As shown in the co-plot, there is no linear relationship between age of house and Insurance acceptance. The normal relation can be described from graph-2 of Figure-D The typical relationship can be described from Graph-2 of Figure-D: when the median income of a neighborhood is lower than $10,695 and the fire rate is higher than 10.5, older houses tend to have lower insurance acceptance rates.

Comparing this with Graph-1 of Figure-D, we can conclude that if the median income is greater than $10,695 and the fire rate is lower than 10.5, insurance is accepted by the company in more than 95% of cases. This suggests that age may be a factor, but it is only strongly connected to insurance acceptance when fire rates and neighborhood income are also taken into account.

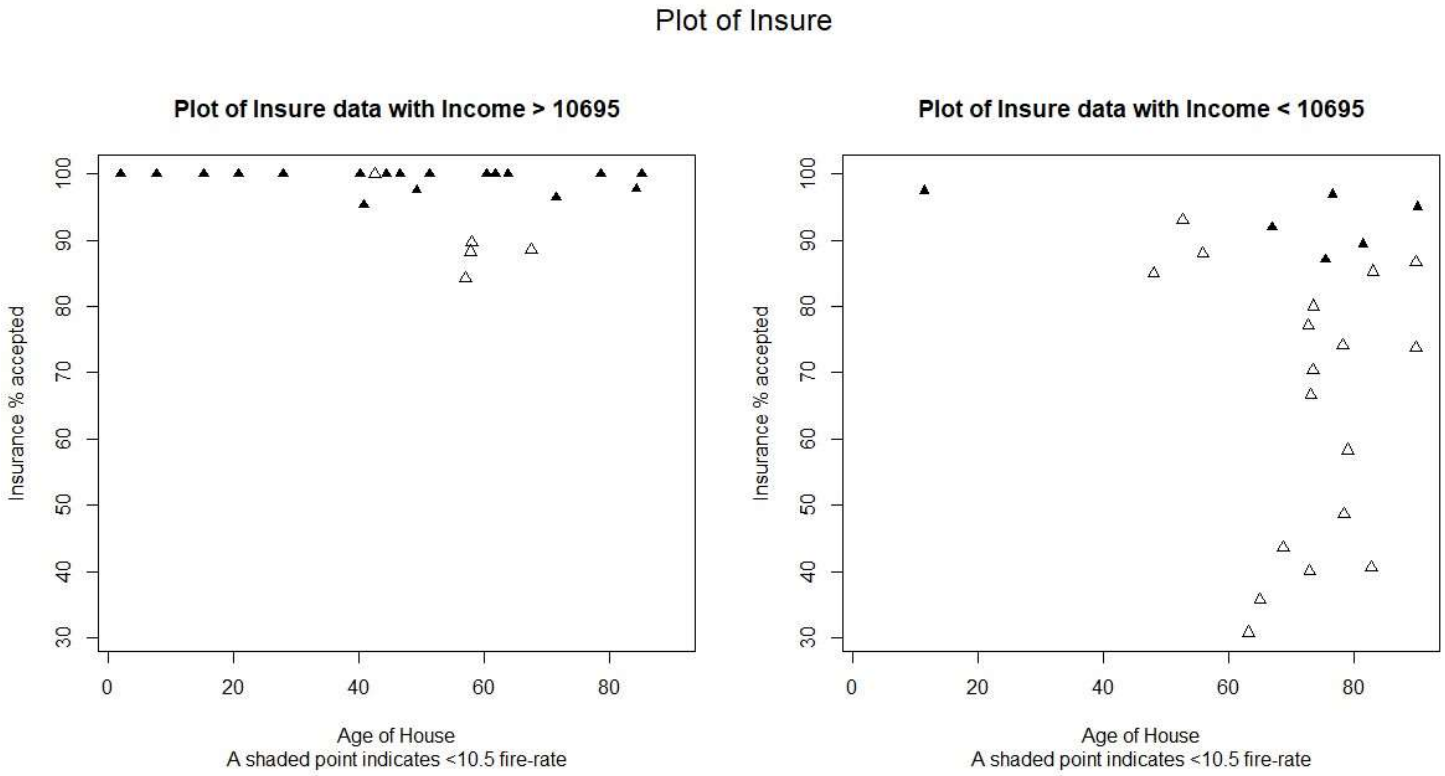

Figure D: Age Co-plot

As we discussed the relationships involving age and race, we would also like to explore similarities among Chicago neighborhoods to ensure we haven't overlooked anything. We can use a map of Chicago for this analysis.

The map is shown in Figure-E. Here, red indicates that the minority population of a neighborhood is less than 25% (the median), while blue represents neighborhoods with an insurance acceptance rate higher than 91% (the median). As visible on the map, most neighbourhoods marked in red colour also include blue areas. The only outliers are neighborhoods with zip codes 60651 and 60629. Conversely, exceptions include 60610,

60626, 60643, 60627, and 60333, which have over 25% minority populations but also exhibit acceptance rates higher than 91%.

One can observe that most of the neighborhoods with lower minority populations are in the northern part of the city The central and southern areas contain more neighborhoods with higher minority populations. This suggests a tendency for white residents to prefer living among other white residents. Most neighborhoods in the northern part also have higher acceptance rates, indicating that residents in this area are more likely to successfully obtain insurance. Thus, we can conclude that, according to the graphics, race was an important factor influencing insurance acceptance in the 1970s.

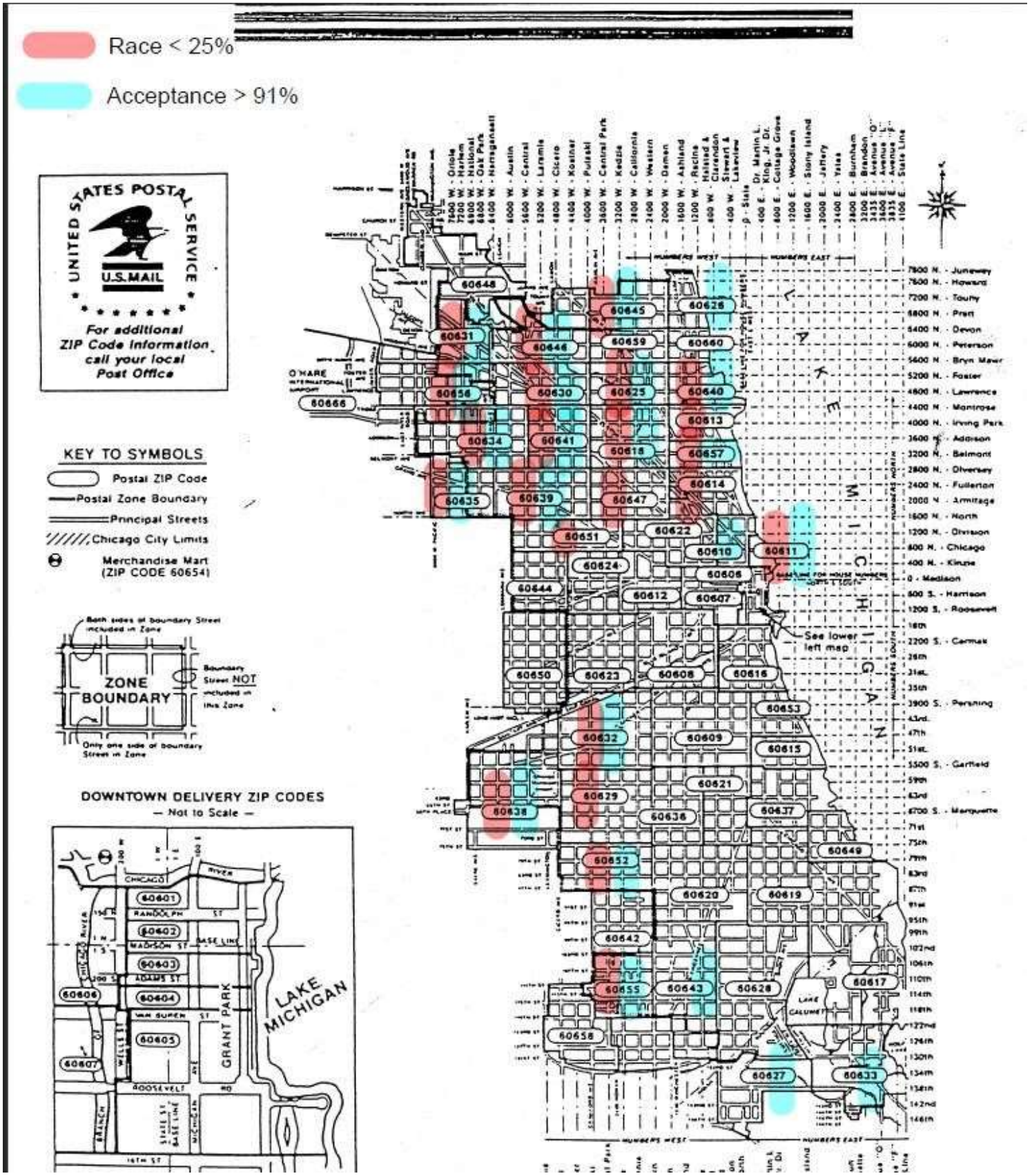

Figure E: Chicago Map

Now, we would like to fit a model for prediction following our discussion. From our previous analysis, we concluded that age is not a significant addition to the model. However, we can further validate this by including age in the model while controlling for fire, theft, income, and race. After that, we can perform a t-test to determine whether the inclusion of age is significant in the model. The current model looks like this:

```
Call:
lm(formula = Acceptance ~ Race + Age + Fire + Income + Theft)

Residuals:
     Min       1Q   Median       3Q      Max
-18.5210  -4.8854   0.0084   4.5174  25.8889

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.070e+02  1.474e+01   7.259 7.12e-09 ***
Race        -2.964e-01  6.892e-02  -4.301 0.000103 ***
Age         -1.571e-01  8.278e-02  -1.897 0.064845 .
Fire        -1.190e+00  2.510e-01  -4.742 2.56e-05 ***
Income       1.533e-05  9.433e-04   0.016 0.987117
Theft        3.659e-01  8.490e-02   4.310 9.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.973 on 41 degrees of freedom
Multiple R-squared:  0.7796,    Adjusted R-squared:  0.7527
F-statistic:    29 on 5 and 41 DF,  p-value: 1.845e-12
```

As seen in picture, the p-values for age and income are not significant because they are higher than α = 0.05.

Therefore, we need to modify the model by removing the age and income. We accept the null hypothesis, resulting in the coefficients for age and income becoming 0 in our model.

Our new and final model will like,

$E(Insurance\_Accpetance) = \beta_0 + \beta_1 * Race + \beta_2 * Theft + \beta_3 *$ Fire

The new summary table will look like:

```
Call:
lm(formula = Acceptance ~ Race + Fire + Theft)

Residuals:
     Min       1Q   Median       3Q      Max
-19.9000  -4.4473   0.0534   4.4967  30.6002

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.00009    2.91079  34.355  < 2e-16 ***
Race         -0.29974    0.05819  -5.151 6.16e-06 ***
Fire         -1.31624    0.23717  -5.550 1.65e-06 ***
Theft         0.34549    0.08244   4.191 0.000136 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can translate this to final model in numerical value.

E(Acceptance) = 100.00 – 0.3 * Race – 1.32 * Fire + 0.35 * Theft

Now, we need to check diagnostic of our selected model and confirm if it is appropriate or not.

As shown in Figure-F, the first graph indicates that the residual points are randomly scattered around the 0 line, providing evidence for the assumption of linearity. The points are also scattered along the curve, further supporting the equal variance among error terms. Notable unusual observations include points 6, 11, and 45, which are also visible in the Normal Q-Q plot of Figure-F. This suggests that these are high-influence points.

From graph-4, we can see that observation 6,11 and 34 has high cook's distance, indicating they are outliers. All the other observations follow the normal line with approximate close distances, but light tails can be seen in both ends suggesting model may need transformation; however, this is not strictly required since they are not heavily skewed. In Graph-3, observation 24 exhibits a high leverage value, suggesting it is another outlier.

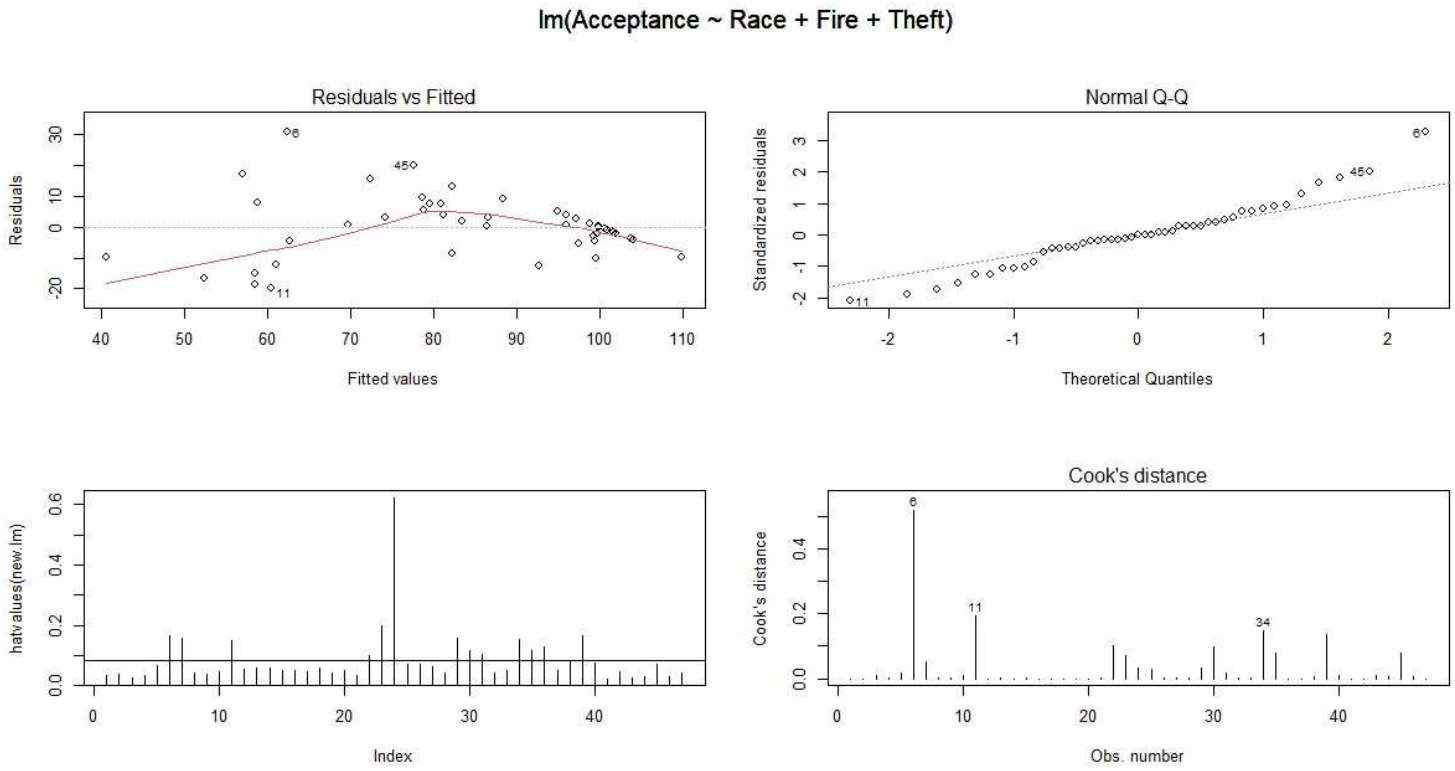

Figure F: Diagnostic Check

Overall, we can use this model as there are very few outliers, and it satisfies all the conditions of a multiple linear regression model. To conclude, due to the lack of evidence, we can say that age was not a significantly important factor in the rejection of insurance when fire, theft, and income were controlled. However, based on our observations, we can conclude that race was one of the major factors responsible for insurance acceptance, with odds favoring neighborhoods with lower minority populations, as they had a higher chance of successful insurance applications.

This is illegal, as companies are not allowed to reject applications based on race. The provided data suggests that insurance officers were more likely to assist white applicants compared to those in neighborhoods with higher minority populations.