Analysis of 311 hotline resolution

Factors that influence the number and type of cases and their resolution status

Bozhi Hao, Hongzhou Tao, Kaiyang Li, Yixun Tian

Data Science and Analytics Program, Georgetown University

Agenda:

1. Introduction

1.1 Background

In the United States, the concept of a "311" service refers to a non-emergency number that people can call to find service information, complain, or report problems such as graffiti, broken street lights, potholes, and garbage collection. The service is designed to help improve city management and maintenance while providing residents with an easy and convenient way to voice their concerns or requests for the local environment. 311 was first introduced by Baltimore, USA, aimed at relieving the burden on 911 and providing a centralized point of public service contact.

To better serve the public and ensure their satisfaction, municipal governments are committed to allocating resources and resolving problems more effectively. Citizens are also trying to describe issues more clearly and informatively in hopes of achieving better resolutions.

Therefore, this study is working on these data to find out the factors that influence the number of cases, the resolution rate and the resolution time. It also looks into the diversity of numbers and types of cases in different neighborhoods, and whether adding photo evidence will help with the speed and accuracy of resolving the problem. The result of this study may serve as a guidance for both the government and the residents.

1.2 About Data

neighborhood	year	mobile.dumm	case	month	request	notes	photo.dumm	income	population	num_cases	white	black	asian	resolution_time	agency
Bayview	2012	1	Street and	Septembe	General C	Case Com	0	56718	23467	2092	0.200239	0.294286	0.351941	1.456388889	DPW Ops Queue
Bayview	2012	0	Street and	December	Bulky Iten	Case Com	0	56718	23467	2092	0.200239	0.294286	0.351941	115.2247222	DPW Ops Queue
Bayview	2012	1	Street and	August	General C	Case Com	0	56718	23467	2092	0.200239	0.294286	0.351941	121.9111111	DPW Ops Queue
Bayview	2012	0	Street and	October	General C	Case Com	0	56718	23467	2092	0.200239	0.294286	0.351941	1.320833333	DPW Ops Queue
Bayview	2012	0	Abandone	March	Abandone	DPT Aba	0	56718	23467	2092	0.200239	0.294286	0.351941	40.70472222	DPT Abandoned Vehicles Work Queue
Bayview	2012	0	Streetligh	December	Streetligh	Case Resc	0	56718	23467	2092	0.200239	0.294286	0.351941	115.0166667	PG and E - Streetlights Queue
Bayview	2012	0	Sewer Issu	Novembe	Sewage_b	Case Com	0	56718	23467	2092	0.200239	0.294286	0.351941	193.4013889	PUC Sewer Ops
Bayview	2012	0	Graffiti	March	Graffiti o	See Notes	0	56718	23467	2092	0.200239	0.294286	0.351941	109.65	DPW Ops Queue
Bayview	2012	0	Sewer Issu	March	Sewage_b	See Notes	0	56718	23467	2092	0.200239	0.294286	0.351941	22.96444444	PUC Sewer Ops
Bayview	2012	1	Street and	July	General C	Case Com	0	56718	23467	2092	0.200239	0.294286	0.351941	37.10611111	DPW Ops Queue
Bayview	2012	1	Street and	April	Street and	See Notes	0	56718	23467	2092	0.200239	0.294286	0.351941	73.41277778	DPW Ops Queue

Table-1.2.1 Raw data

Neighborhood: Where does the request come from?
Year: In which year does the request happen?
Mobile.dumm: Whether the case was mobile-enabled (1) or not (0).
Case: Describes the type of case.
Month: Specifies the month when the case was reported (e.g., September, December).
Request: The type of help requested.
Notes: Status of the case.
Photo.dumm: whether a photo was provided.
Income: Average income of the neighborhood, constant across entries.

Population: The population of the neighborhood is also constant.
Num_cases: Number of cases, constant across entries.
White, Black, Asian: Demographic breakdown by race in percentages.
Resolution_time: Time taken to resolve the case in hours.
Agency: The agency handling the case.
Resolution_rate: Whether the case is resolved, 1 indicating resolved and 0 indicating unresolved.
CaseID: Unique identifier for each case.
Latitude, Longitude: Geographical coordinates of the cases.
Date: Exact date and time when the case was recorded.

1.3 Data science question and 10 supporting questions

Data science questions: Factors that influence the numbers and types of the cases and their resolution status(resolution rate, resolution time).

10 supporting questions:

Problem 1: What are the factors that influence the number of cases?
Problem 2: What are the factors that influence the resolution rate?
Problem 3: Is different race a factor influencing case resolution time?
Problem 4: Does the number of the same case change along with the area?
Problem 5: Does the resolution time change along with the month in the same area and year?
Problem 6: What is the case resolution rate, and how does it vary by case type or over time?
Problem 7: Analyze if cases with photo evidence (indicated by photo. dumm) are resolved faster or have a higher resolution rate than those without.
Problem 8: What's the trend of 311 cases through the years
Problem 9: Is there a specific relationship between the type of cases with race? how did the racial composition in different neighborhoods impact the type of most often events?
Problem 10: Is there an obvious relationship between income with types in a certain area? Does the average income of a neighborhood influence the type of service requests filed? For instance, do higher-income areas report different concerns than lower-income areas?

2. Methods

2.1 EDA

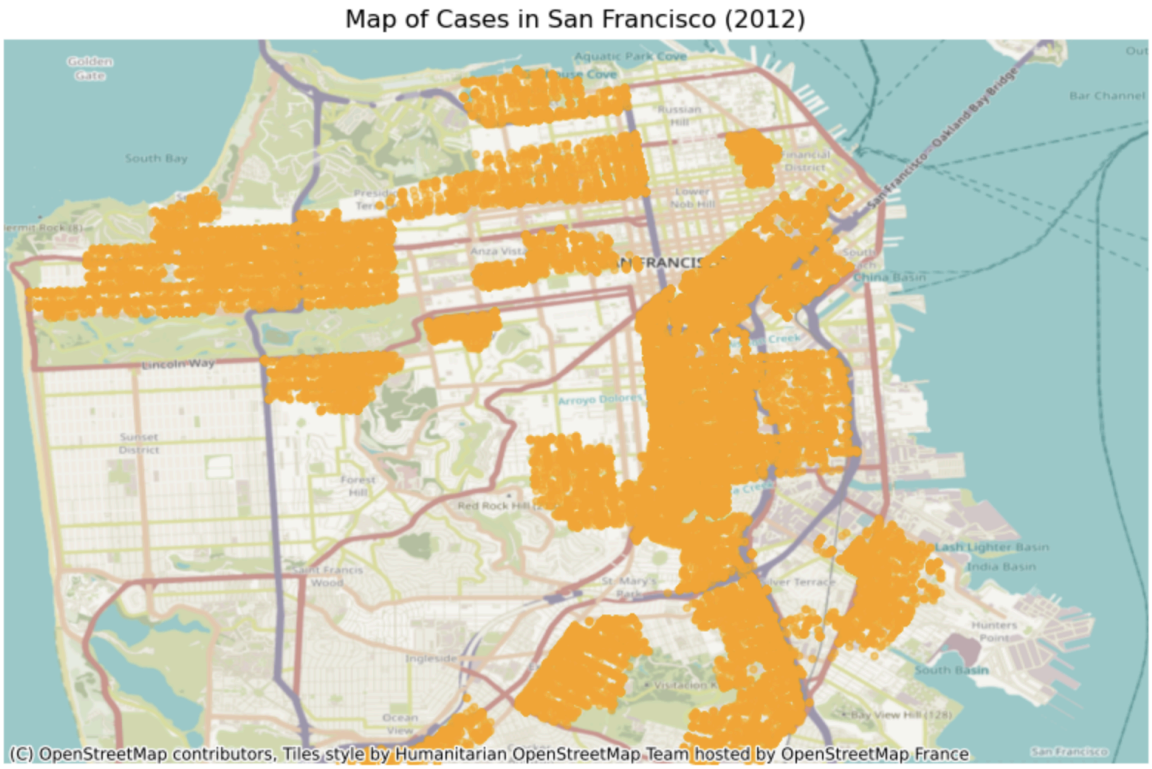


fig-2.1.1

In fig-2.1.1, we can have a preliminary look at the distribution of each case location on the map of San Francisco. The dataset does not include all communities and neighborhoods in San Francisco, but it contains various communities with different household incomes. Most of the data are concentrated in the community or neighborhood having a higher household income. With only limited data in the community or neighborhood having a rather lower household income, like the central area of Tenderloin and Chinatown.

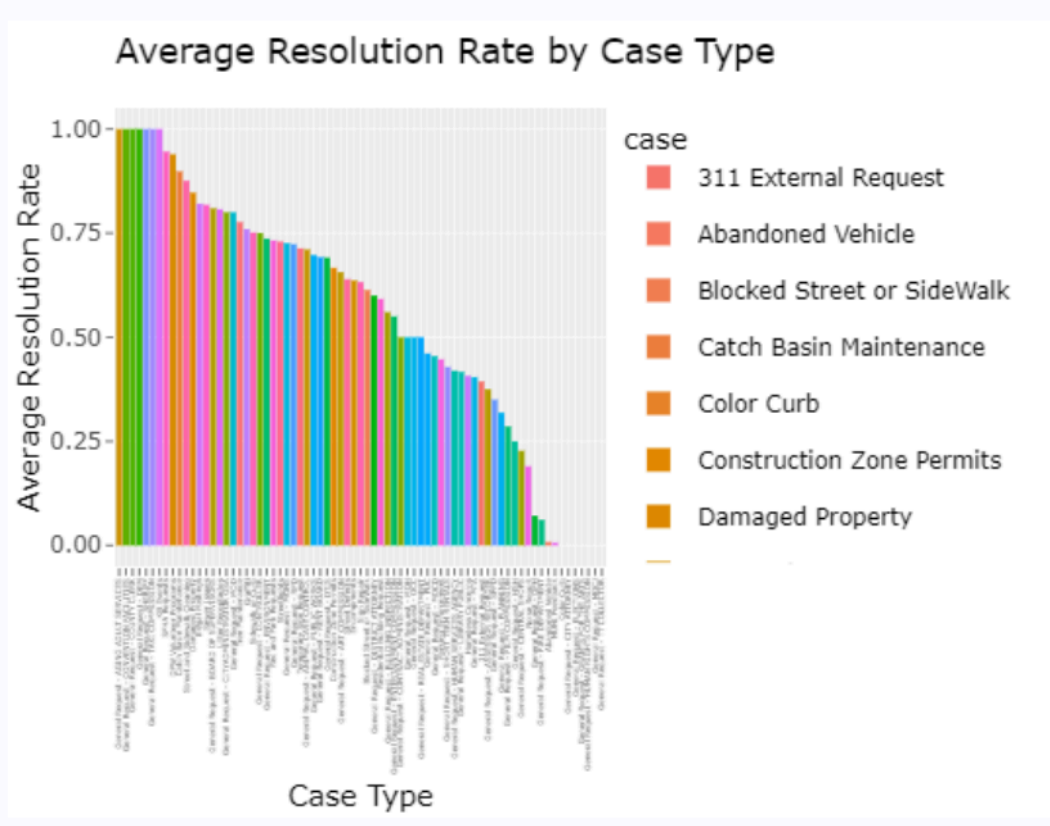


fig-2.1.2

The results of the analysis of 311 service request data allowed us to construct a bar chart shown in fig-2.1.2 – average resolution rate by case type. As one can in the graph, different case types significantly vary in their resolution efficiency. The graph is color-coded according to case type, which allows a quick and easy representation of the information for optimal city management service. Finally, the picture is in descending order of average resolution rate for case types, making more and less efficiently handled case types immediately apparent. For example, 311 External Request, Abandoned Vehicle, and Blocked Street or Sidewalk have a high resolution rate, while Damaged Property is poorly resolved.

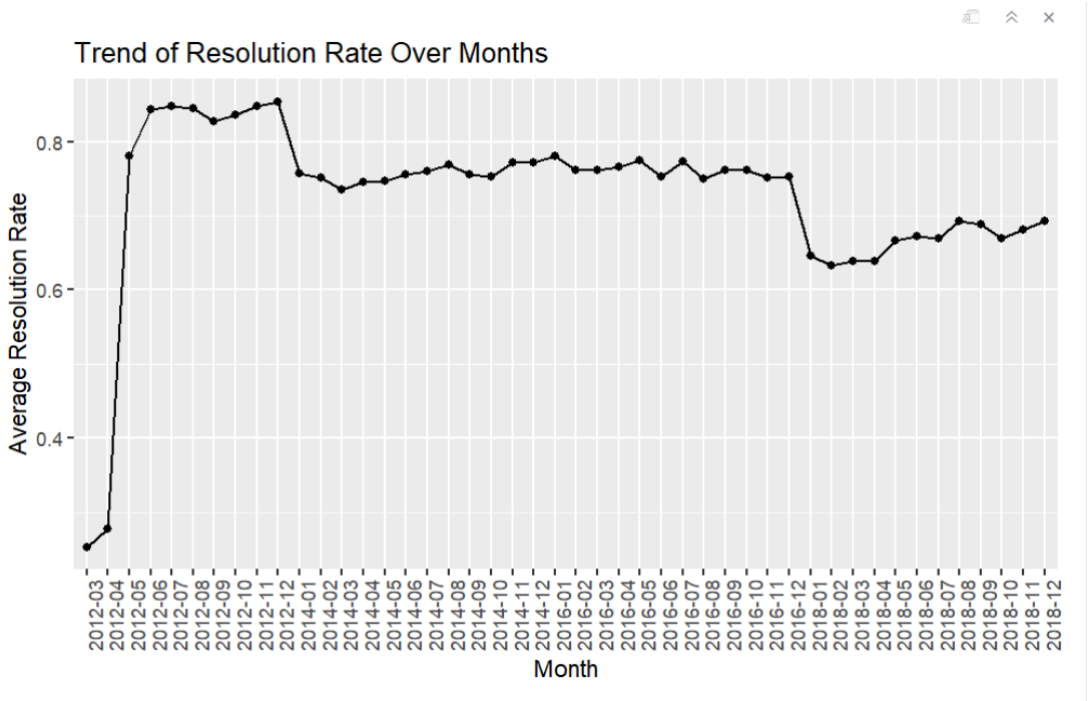


fig-2.1.3

Fig-2.1.3 depicts the trend of the average resolution rate of 311 service requests monthly from the start of 2012 to the end of 2018. By organizing the 311 service dataset and plotting a time-series line graph using ggplot2, we can reflect the change in resolution rate over time. As shown in the graph, the average resolution rate decreases at the start of 2012 and then stabilizes to low and major fluctuation in 2014 and 2016. In 2017, the average rate declined drastically, and it may relate to changed policies or an event at that time. Then it keeps rising until 2018.

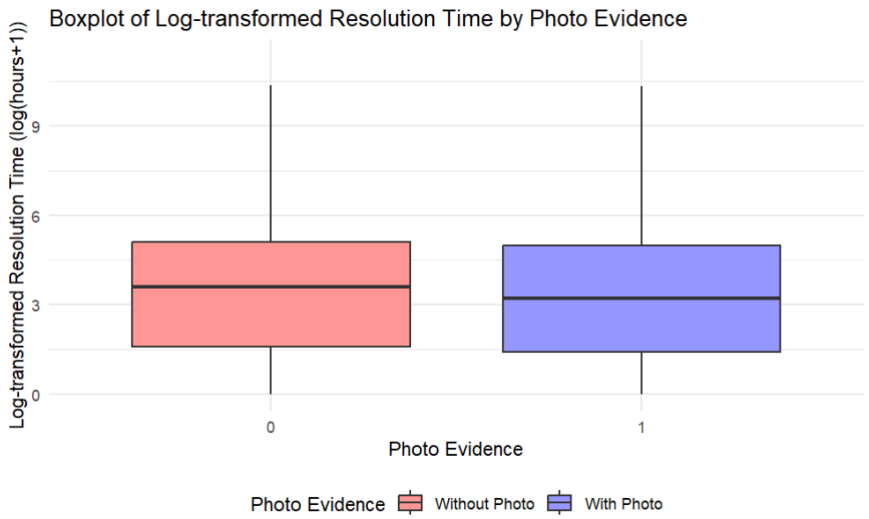


fig-2.1.4

In Fig-2.14, we show a box plot comparison of case resolution times after logarithmic transformation based on the presence or absence of photographic evidence. This transformation of the data enables us to more meaningfully address the extremes of resolution times and make fairer, more consistent comparisons between groups. As can be observed from Fig-2.14, there is a difference in resolution times between cases with photographic evidence and cases without photographic evidence. More specifically, while the two groups have similar medians, cases with photographic evidence have a more centralized rhythm of resolution times, whereas cases without photographic evidence exhibit more drastic fluctuations across the distribution.

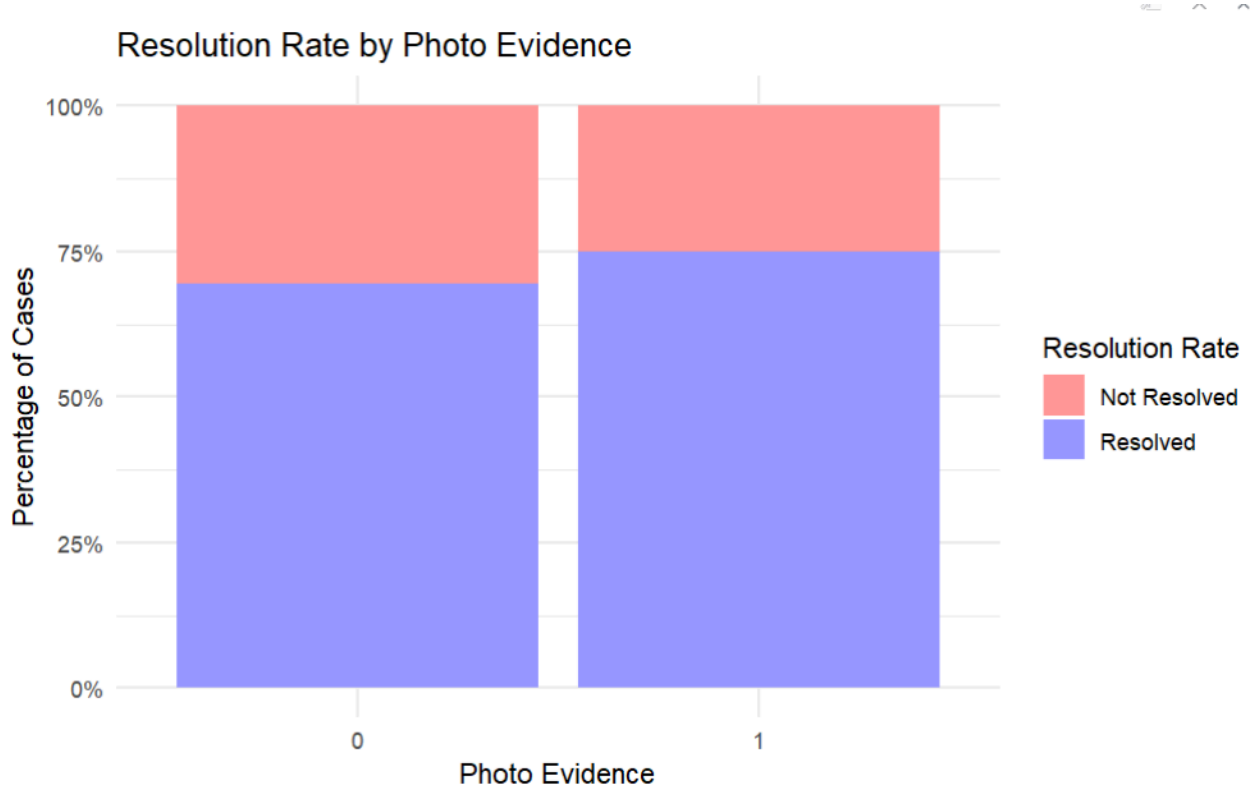


fig-2.1.5

It is clear from the above chart fig-2.15 that stacked bars have been used, with red representing cases that have not been resolved and blue representing resolved cases. Stacked bar charts Test the impact of the presence or absence of photographic evidence on the resolution rate score of the cases. The percentage shown on the y-axis enables us to estimate the rate of resolution in two cases. The graphical representation indicates that case resolution scores for the case with photographic evidence 1 draft are higher compared to that without photographic evidence. This implies that photographic evidence likely shortens the time taken to report a case since it provides the actual evidence to determine the diagnosis and removal of the problem.

2.2 Hypothesis test

Hypothesis test is often used to figure out whether there is a relationship between variable A and variable B, or whether group A and group B have differences. First the researchers generate a null hypothesis, which is often “A has no impact on B” or “A has no difference with B”. Then we calculate the t-statistic and the p-value. The t-statistic is a measure of how many standard errors a sample mean is away from the population mean, and p-value is a measure of evidence against the null hypothesis. If the p-value is small (smaller than the alpha-value) and the t-statistic is large, it is suggested that there is strong enough evidence to reject the null-hypothesis.

2.3 Negative Binomial Model

Analyzing the hypothesis of whether the case number of a specific type of case varied across different neighborhoods would require a statistical method that can account for variability and overdispersion. Under these requirements, the Negative Binomial is the best statistical approach to use since it is robust in dealing with the problem of variance exceeding the mean, which is quite common in real-world datasets. Another typical statistical method used when trying to solve the count problem is the Poisson distribution, however, the Poisson distribution has the attribute of assuming an evenly distributed rate of event occurrence independently and expects the mean and variance to be equal, therefore, making it hard to predict the true result of the dataset having non-even distributed rate, mean, and variance such like the 311 hotline dataset. On the contrary, the Negative Binomial distribution has a dispersion parameter that can make the variance a function of the mean, thus, providing a more accurate data analysis.

In the 311 hotline dataset, we used the Negative Binomial regression to analyze the question of how the number of specific types of cases varied across different neighborhoods during the year 2012. Whether a linear relationship between the number of cases and the neighborhood could be a crucial consideration when the municipality makes development policies and plans. Also, this model could potentially help the 311 hotlines adjust their responding and solving time according to the case number of neighborhoods. We found that there exists the problem of overdispersion during the data cleaning and munging procedure then the Negative Binomial regression would be the most suitable statistical method to use on this hypothesis. Each neighborhood’s impact on the count of cases was estimated while holding other factors constant.

2.4 Linear regression on race factor

We used Linear Regression when dealing with the question of whether race is a factor influencing the average time taken to resolve 311 service requests. The timely resolution time of service requests is critical for maintaining community satisfaction. Linear regression is a statistical method used to model the relationship between one dependent variable and one or more independent variables by fitting a linear regression on the dataset. We transformed the race variable in the dataset into three race variables, White, Black, and Asian, and then used the model to analyze the relationship between the proportion of these three races and the average resolution time of service requests. Therefore, whether race composition is a factor that could influence the average resolution time can be estimated by fitting the linear regression model.

2.5 Subset selection

The subset selection part and the next session (regression) are to figure out the factors that influence the number of cases and the resolution rate. First, it needs to be explained how these two dependent variables are calculated since they are not originally in the data. Originally, each case forms a row, and the resolution rate is either 0 (indicating the case has not been resolved) or 1 (indicating the case has been resolved). To study the factors that influence the number of cases in a certain neighborhood in a certain year, the data munging process is to group the data by neighborhood and year (and all other variables that remain consistent within the same area and year), count the number of rows in each group to be num_cases, and average the resolution rate of all rows in each group to be resolution_rate. All variables besides “year” and “neighborhood” should be transformed into numeric variables.

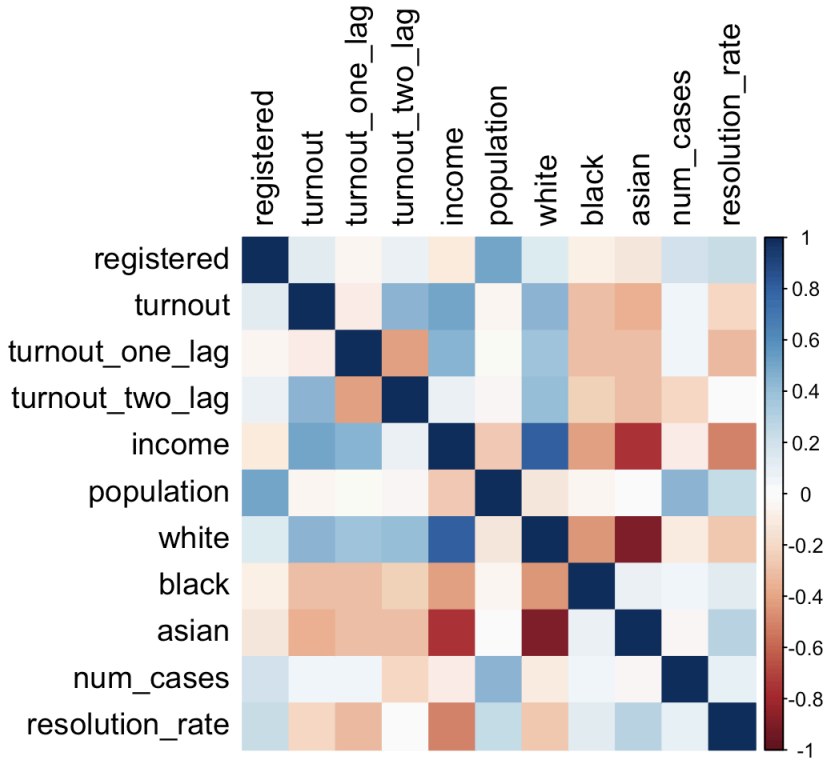


Fig- 2.5.1 Correlation heatmap of variables in the 311 dataset

From Fig-2.5.1, we can have a general view of how these variables correlate with each other. The number of cases seems to be related to population, turnout_two_lag, registered, and income, and the resolution rate seems to be related to turnout_one_lag, income, population, and the proportion of different races.

2.5.1 Best subset selection

The first step is to use the best subset selection to get 10 possible models. Best subset selection is an exhaustive method that attempts all possible combinations of features to select the one that performs the best. To select the best-performing model, 4 different criteria to evaluate:

- R-squared: A variable ranging from 0 to 1, explains the extent to which the independent variables explain the change in the dependent variable. The bigger the R-squared, the closer the variation in the dependent variable is to the variation predicted by the model. It intuitively and accurately shows the performance of the model, but the drawback is, as the variables are more or less related to the dependent variable from the data (although not logically), the R-square always gets bigger when more variables are taken into account, so it is not suitable for feature selection.

- Adjusted R-squared: Compared to R-squared, the adjusted R-squared introduced a penalty term to get rid of the unimportant variables, and take the number of variables into account, but it can still have the problem of overfitting.
- Mallows' Cp: Cp tries to find a balance between the model complexity and the performance in fitting by considering the residual sum of squares and the number of free parameters. It can prevent overfitting to some extent. However, it may not be suitable if the sample is too large or the dimension is too high.
- Bayesian Information Criterion(BIC): BIC is similar to Cp, but the penalty term is stronger. It can prevent overfitting more effectively, but the strict penalty on overfitting may result in too small models.

Generally, the number of features selected by the 4 criteria may be smaller and smaller, and it is up to the researcher to find a balance point or use other methods to validate.

2.5.2 Stepwise methods

Stepwise methods are also ways to select the appropriate model. There are two different methods:

- Forward stepwise selection: A method that starts with an empty model and adds variables incrementally.
- Backward stepwise selection: A method that starts with a model with all variables and removes variables incrementally.

2.5.3 Cross Validation

In cross-validation, the data set is split into the training set and the validation set, trained on the training set and uses the validation set to evaluate its performance. In each iteration, 1 of the 10 possible feature subsets is used to train the model, and the performance is recorded. The criterion used to evaluate the performance is the MSE, the mean squared error between the predicted dependent variable and the actual value of the dependent variable of the validation test. The point where the MSE is the smallest is the point where the feature subset is most suitable to be used for regression or classification.

2.5.4 K-fold Validation

K-fold validation works basically the same way with the cross-validation, but instead of one training set and one validation set, the data is split into k subsets, and each time one of them is taken out to be the validation set, with the rest used as the training set. This process will be repeated k times. K-fold validation is more robust, but when the scale of the data and the value of K is large, it can have great computational cost.

In feature selection, we put more emphasis on the result of K-fold validation, but when the conclusion given by the K-fold validation contains too few features, and other criteria reach consensus in some way, we may also discard the result of K-fold validation, and accept the result of other methods.

2.6 Regression

2.6.1 Linear Regression

Linear regression is a simple regression method that assumes the data follows a linear pattern and uses the method of least squares to figure out the best coefficients.

2.6.2 General Linear Regression

The general linear model is an extension of the linear regression, the pattern of the data can be more flexible, like Poisson distribution and binomial distribution. Besides single and simple variables, the input feature can also be the product of 2 variables, thus we can look into how the variables interact with each other.

2.6.3 Ridge Regression

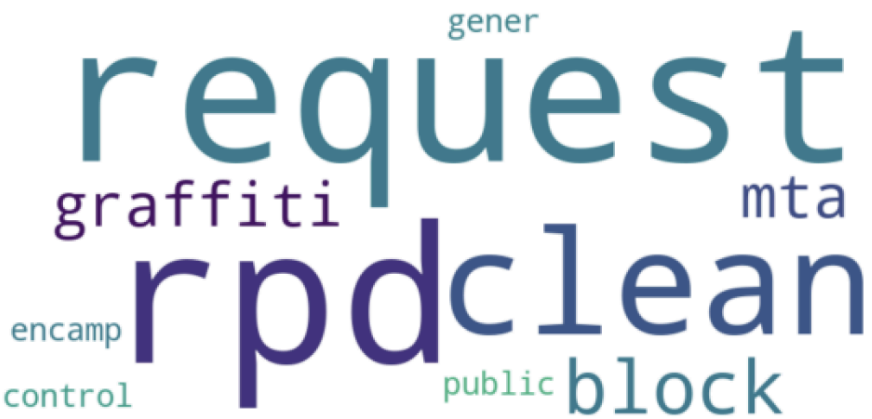
Ridge regression introduces a regularization parameter to reduce the model's variance. It is suitable when features are highly correlated with each other. It uses an L2 regularizer to prevent overfitting.

2.6.4 Lasso Regression

Lasso Regression is similar to Ridge regression, but it can reduce some coefficients to 0, thus automatically making feature selection. It uses an L1 regularizer to prevent overfitting.

2.7 Topic Modeling

We applied LDA to the case and request column, to get each optimized topic number of certain areas we chose the log-likelihood method, however, you can use elbow-method as well. After generating the topic and saved into a dictionary we created a word cloud for each topic, let's take the following as an example:



Words like "graffiti," "control," "public," "block," "general," and "camp" suggest that the data may be related to issues within city management or public Spaces, such as requests to clean up graffiti, control public areas, or manage block-level issues. The term "camp" may refer to camps and refers to problems associated with homeless people or temporary accommodation in public places.

3. Results and discussion

3.1 Hypothesis test

```
        Wilcoxon rank sum test with continuity correction

data:  resolution_time by photo.dumm
W = 3.5024e+10, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

fig-3.1.1

To test the likelihood that there is a difference in resolution times between cases that contain and do not contain photos, we used the Wilcoxon rank sum test. This approach compares the distribution of two data groups without the need to assume the normality of the data. We log-transformed resolution times to make the extreme values more normally distributed. W statistic returns the value of 3.5024e+10 with a highly significant p-value of less than 0.05, $p < 2.2e-16$, therefore confidently rejecting the null hypothesis that the medians are similar between two groups of cases and confirming the alternative hypothesis that days to resolution differ. Our results disclose that there is a solid association between the presence of photos and the faster resolution of cases. Additionally, histograms of log-transformed resolution times for the two categories clearly indicate the dissimilarity in the distributions between cases vs. no-cases, underlining the importance of photos in the efficiency of case resolution.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  resolution_rate_table
X-squared = 1925.1, df = 1, p-value < 2.2e-16
```

fig-3.1.2

Therefore, the statistical analysis based on Pearson's Chi-squared test with Yates' continuity correction revealed a highly statistical association between the case resolution rate and the existence of photo evidence. The result very convincingly speaks about the absence of similar outcomes solely by a random set of cases. Therefore, the resolution of cases differed with or without photo evidence in this analysis. As a result, it can be concluded that there is a correlation in which cases with photo evidence tend to achieve resolution compared to cases without photo evidence. Particularly, improved visual contextual attributes would enhance case resolution.

```
Series: resolution_ts
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1      ar2    mean
      1.0717  -0.4595  0.7221
s.e.  0.1380   0.1979  0.0300

sigma^2 = 0.005813:  log likelihood = 54.05
AIC=-100.1   AICc=-99.13   BIC=-92.79

Training set error measures:
                     ME       RMSE        MAE       MPE     MAPE      MASE
ACF1
Training set 0.002655479 0.07371691 0.04011616 -1.891985 7.329857 0.4699802
-0.1870509
```

fig-3.1.3

We used the ARIMA model, starting in R with the time series object for the monthly trends in 311 service-request resolution (not shown here). The auto.arima function automatically selected an optimal ARIMA model based on sample size and AICc criteria, in the forecast package in R.

ARIMA(2,0,0) the summary of the chosen model contains two autoregressive terms, their coefficients for AR1 and AR2 are 1.0717 and -0.4595 respectively, it shows that the term 0.4911 indicates a significant short-term correlation in the plot, mean value of that data is 0.7221 which indicate that probably that average resolution rate is high for the period.

The statistical indicators determining the adequacy of model fitting are, calculated the AICc as −99.13 and the BIC as −92.79. Other diagnostic checks necessary to determine the accuracy of the forecast are defined as Root Mean Squared Error (RMSE) of 0.0401161 and Mean Absolute Percentage Error (MAPE) of 7.329857.

3.2 Negative Binomial Model

In fig-3.2, The Negative Binomial model output included the coefficient estimate value for each neighborhood with their respective standard error and Z-values, we can use the result to analyze the significance of each neighborhood's effect on the number of cases.

	Estimate	Std.error	Z value
neighborhood: Mission	-1.25747	0.17085	-7.360
case: MOH	-5.10384	0.37518	-13.604

fig-3.2

We took an outstanding neighborhood Mission as an example, which has a coefficient estimate value of -1.25747 with a Z-value of -7.360 and the case of MOH has an estimated value of -5.10384 with a Z-value of -13.604. Given the large absolute value of both Zvalue, we can reject the null hypothesis for both the Mission neighborhood and MOH case, indicating that both neighborhood and case type have a statistically significant association with the number of service requests.

3.3 Linear regression on race factor

In fig-3.3 we can see that all three races have large coefficients(the reason for these large coefficients is because the time is recorded in seconds) which means that all three races are significant predictors of resolution time. They also have substantial t-values and low p-values. This means that we can reject the null hypothesis that race is not a factor influencing the resolution time.

	Coefficients	Standard Errors	t values	p values
white	3302.559826	231.615099	14.258828	1.337552e-11
black	5596.298518	1492.118149	3.750573	1.354206e-03
asian	2878.458237	369.738635	7.785116	2.505272e-07

fig-3.3

Therefore, race is a factor that could influence the average resolution time for service requests. This could help the municipality adjust and better distribute its resources according to the race composition in a community or neighborhood.

3.4 Subset selection

3.4.1 Subset selection for number of cases

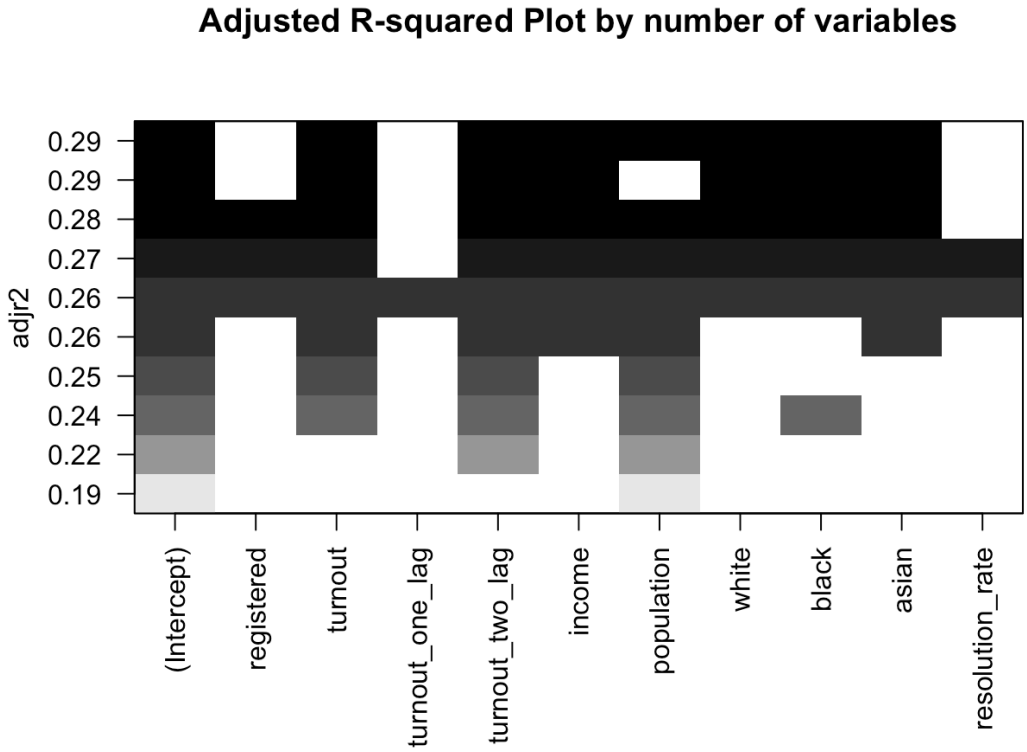


Fig-3.4.1 Adjusted R-squared of different models predicting the number of cases

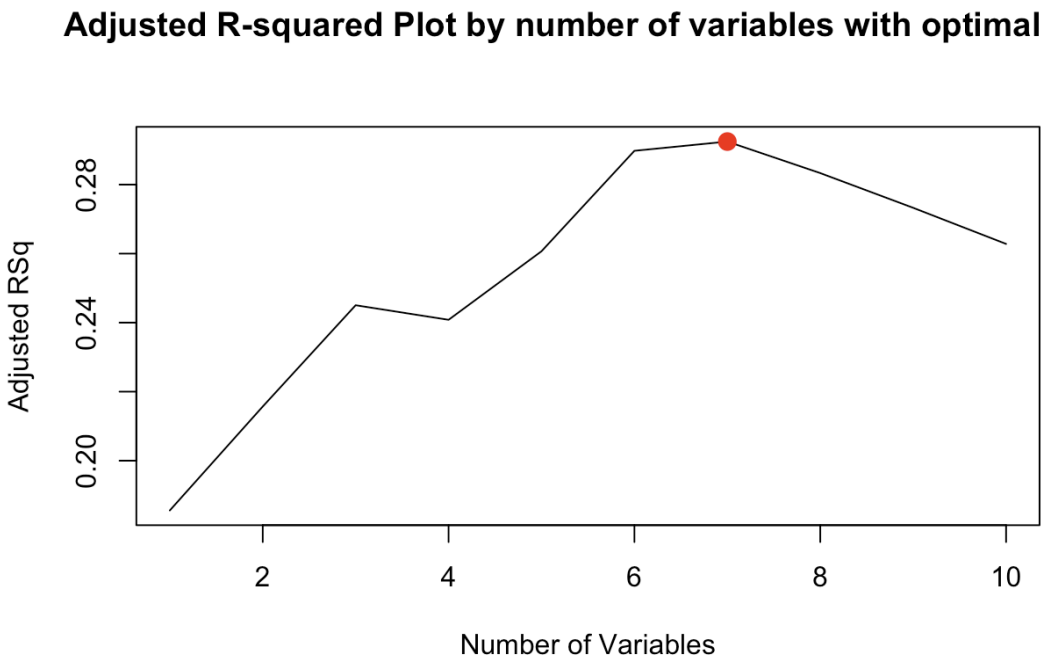


Fig-3.4.2 Another figure of Adjusted R-squared of different models predicting the number of cases

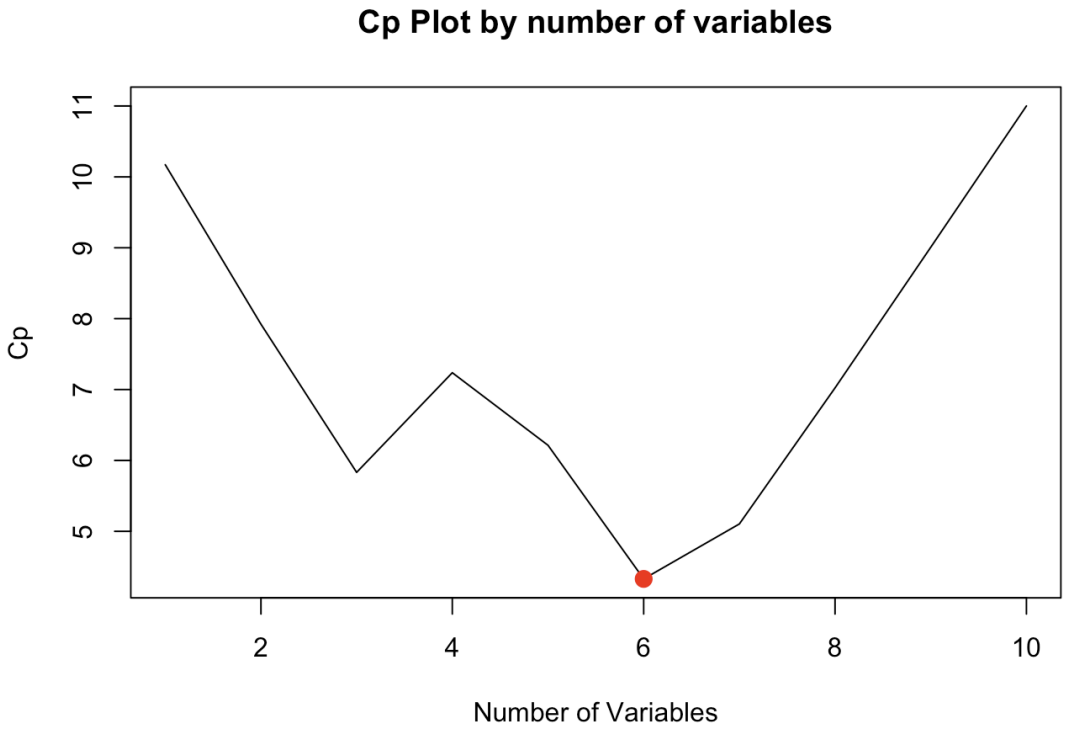


Fig-3.4.3 Figure of Cp of different models predicting the number of cases

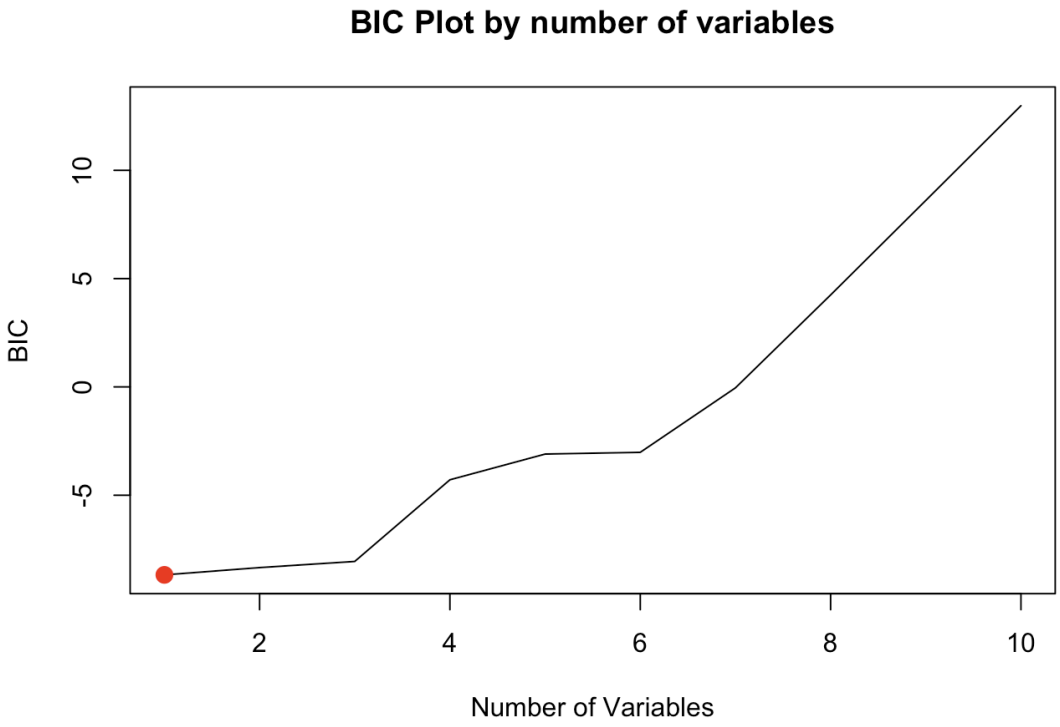


Fig-3.4.4 Figure of BIC of different models predicting the number of cases

```
print(coef(regfit_311_full,6))
```

```
  (Intercept)        turnout turnout_two_lag         income          white
 8.965564e+04   3.259007e+02  -3.812014e+02  -2.104522e-01  -5.346838e+04
        black          asian
-9.640393e+04  -8.693920e+04
```

```
print(coef(regfit.fwd,6))
```

```
  (Intercept)        turnout turnout_two_lag         income     population
 2.500865e+04   2.773188e+02  -3.644127e+02  -1.267524e-01   2.625069e-01
        black          asian
-1.350285e+04  -2.090668e+04
```

```
print(coef(regfit.bwd,6))
```

```
  (Intercept)        turnout turnout_two_lag         income          white
 8.965564e+04   3.259007e+02  -3.812014e+02  -2.104522e-01  -5.346838e+04
        black          asian
-9.640393e+04  -8.693920e+04
```

Table-3.4.1 Best subset selection, Forward stepwise method and Backward stepwise method disagree at the point of 6 features on selected features and coefficients

```
print(coef(regfit_311_full,7))
```

```
  (Intercept)        turnout turnout_two_lag         income     population
 6.868570e+04   2.939276e+02  -3.446700e+02  -1.606715e-01   1.299670e-01
        white          black          asian
-4.220675e+04  -7.204985e+04  -6.845102e+04
```

```
print(coef(regfit.fwd,7))
```

```
  (Intercept)        turnout turnout_two_lag         income     population
 6.868570e+04   2.939276e+02  -3.446700e+02  -1.606715e-01   1.299670e-01
        white          black          asian
-4.220675e+04  -7.204985e+04  -6.845102e+04
```

```
print(coef(regfit.bwd,7))
```

```
  (Intercept)        turnout turnout_two_lag         income     population
 6.868570e+04   2.939276e+02  -3.446700e+02  -1.606715e-01   1.299670e-01
        white          black          asian
-4.220675e+04  -7.204985e+04  -6.845102e+04
```

Table-3.4.2 Best subset selection, Forward stepwise method and Backward stepwise method reach consensus at the point of 7 features on selected features and coefficients

Cross-validation selected 3 features: turnout, turnout_two_lag, and population.
K-fold cross-validation selected 4 features: turnout, turnout_two_lag, population, and black.

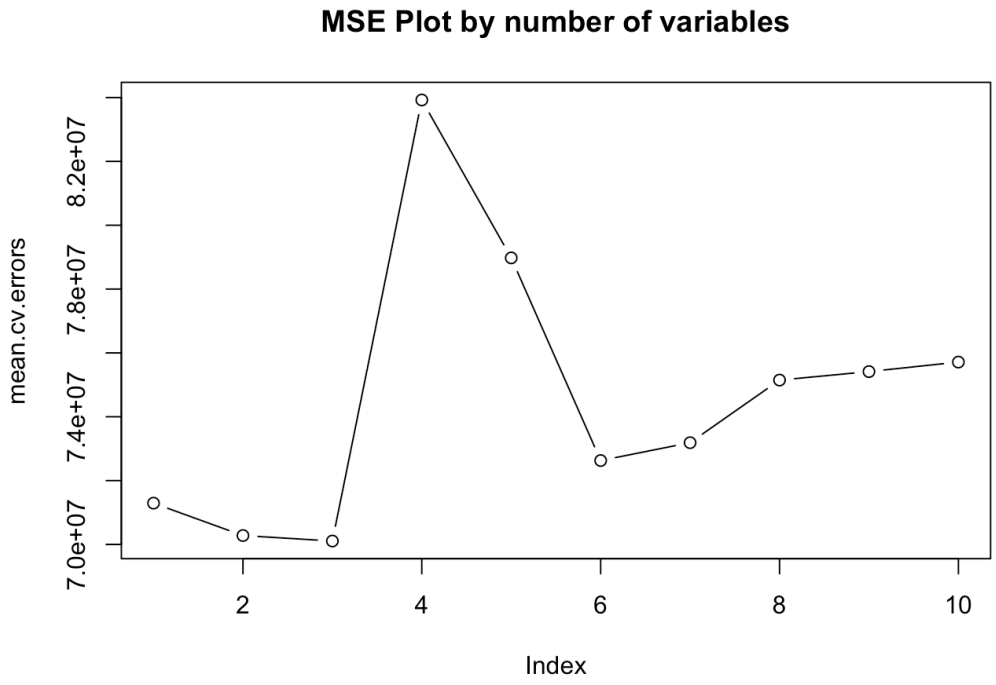


Fig-3.4.5 Figure of MSE of K-fold CV of different models predicting the number of cases

Adjusted R^2	Cp	BIC	R^2	CV	K-fold CV
7	6	1	10	3	4

Table-3.4.3 Different Criterias and their conclusions on the number of features the model should contain

So, the final conclusion is that the model with 4 features should be selected.

3.4.2 Subset selection for resolution rate

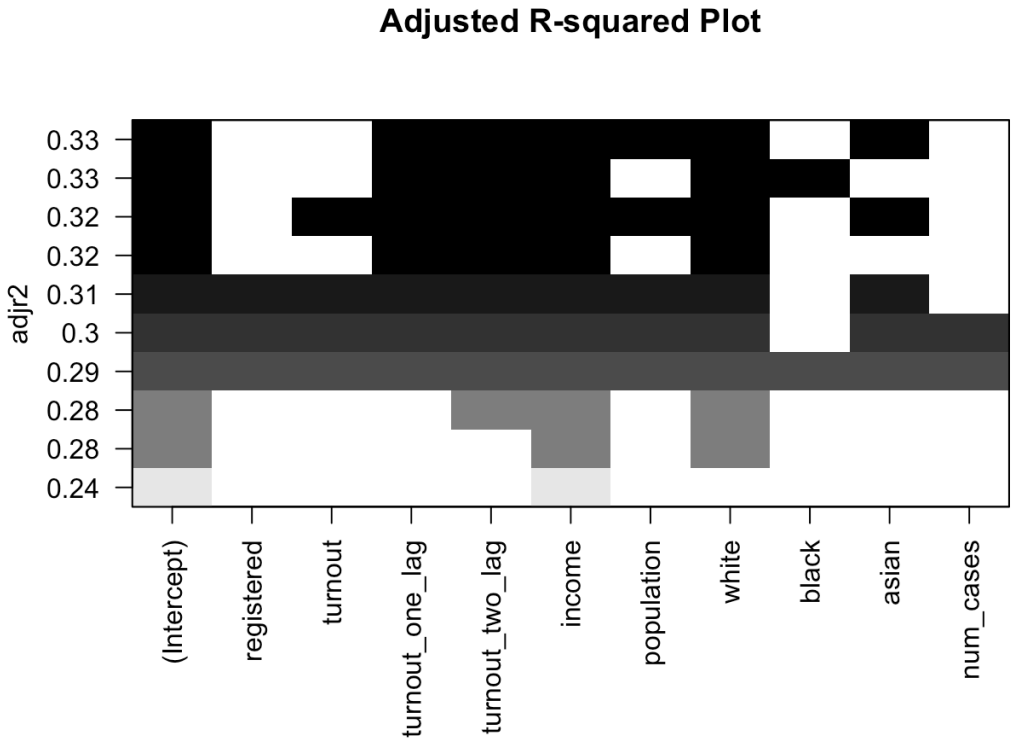


Fig-3.4.6 Adjusted R-squared of different models predicting the resolution rate

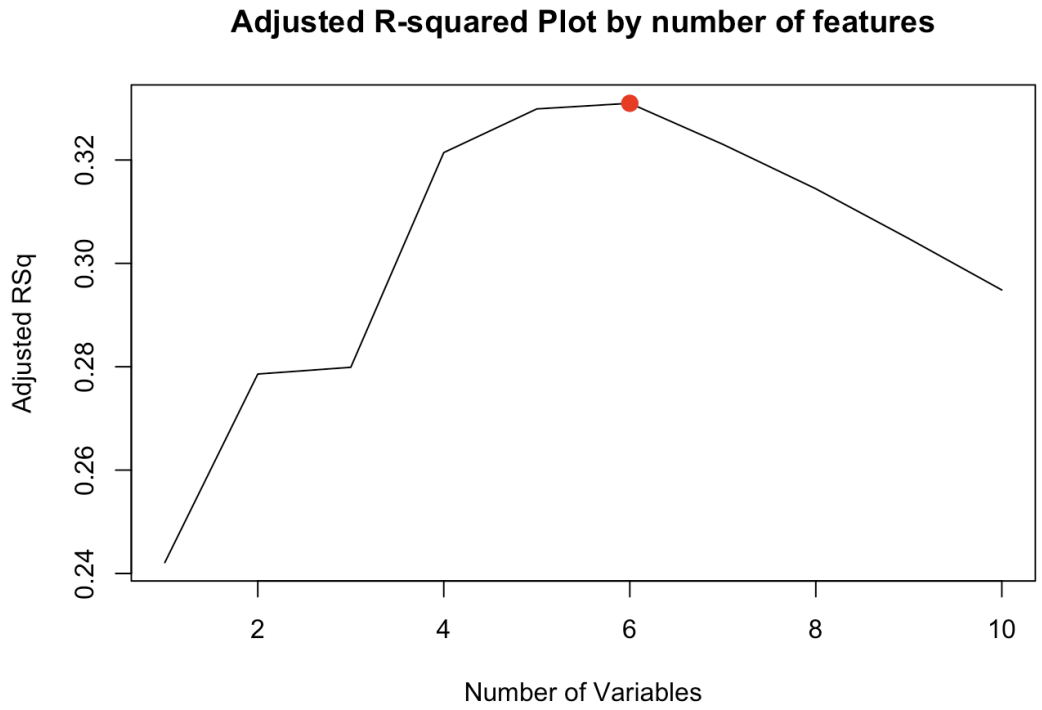


Fig-3.4.7 Figure of Adjusted R-squared of different models predicting the resolution rate with optimal point

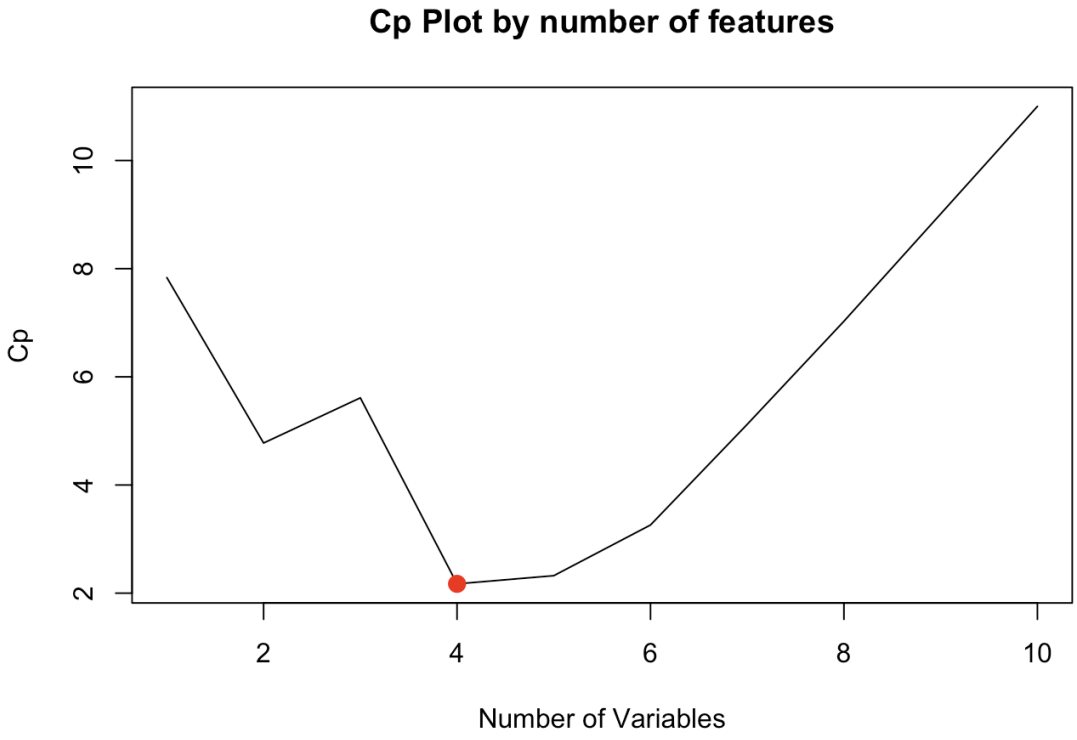


Fig-3.4.8 Figure of Cp of different models predicting the resolution rate

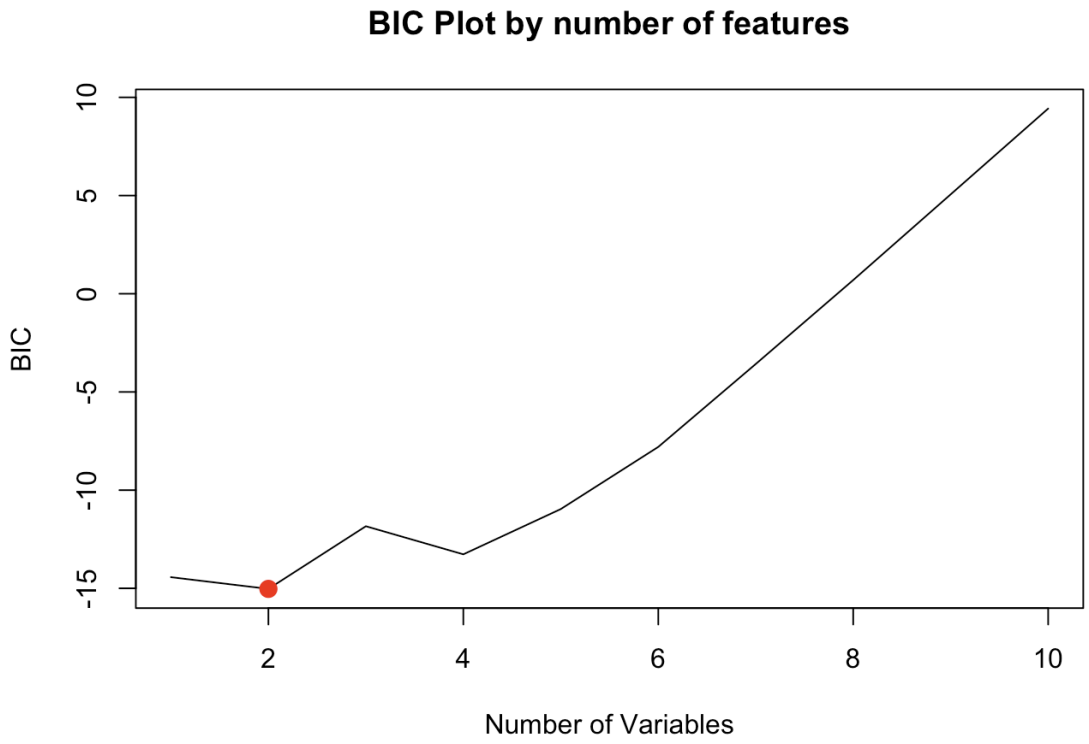


Fig-3.4.9 Figure of BIC of different models predicting the resolution rate

```
print(coef(regfit_311_full,6))
```

```
   (Intercept) turnout_one_lag turnout_two_lag          income      population
  1.007026e+00   -2.917751e-03   -2.981254e-03   -1.948924e-06    9.627568e-07
         white           asian
  4.002117e-01    1.663316e-01
```

```
print(coef(regfit.fwd,6))
```

```
   (Intercept)         turnout turnout_one_lag turnout_two_lag          income
  1.157198e+00    6.122932e-04   -2.718241e-03   -3.294136e-03   -2.426789e-06
         white           black
  2.949362e-01   -1.945515e-01
```

```
print(coef(regfit.bwd,6))
```

```
   (Intercept) turnout_one_lag turnout_two_lag          income      population
  1.007026e+00   -2.917751e-03   -2.981254e-03   -1.948924e-06    9.627568e-07
         white           asian
  4.002117e-01    1.663316e-01
```

Table-3.4.4 Best subset selection, Forward stepwise method and Backward stepwise method disagree at the point of 6 features on selected features and coefficients

```
print(coef(regfit_311_full,4))
```

```
  (Intercept) turnout_one_lag turnout_two_lag          income           white
1.105830e+00   -2.401964e-03   -2.610953e-03   -2.179858e-06    2.809511e-01
```

```
print(coef(regfit.fwd,4))
```

```
  (Intercept) turnout_one_lag turnout_two_lag          income           white
1.105830e+00   -2.401964e-03   -2.610953e-03   -2.179858e-06    2.809511e-01
```

```
print(coef(regfit.bwd,4))
```

```
  (Intercept) turnout_one_lag turnout_two_lag          income           white
1.105830e+00   -2.401964e-03   -2.610953e-03   -2.179858e-06    2.809511e-01
```

Table-3.4.5 Best subset selection, Forward stepwise method and Backward stepwise method reach consensus at the point of 4 features on selected features and coefficients

Cross Validation and K-fold Cross Validation Both select 4 features: turnout_one_lag, turnout_two_lag, income, and white.

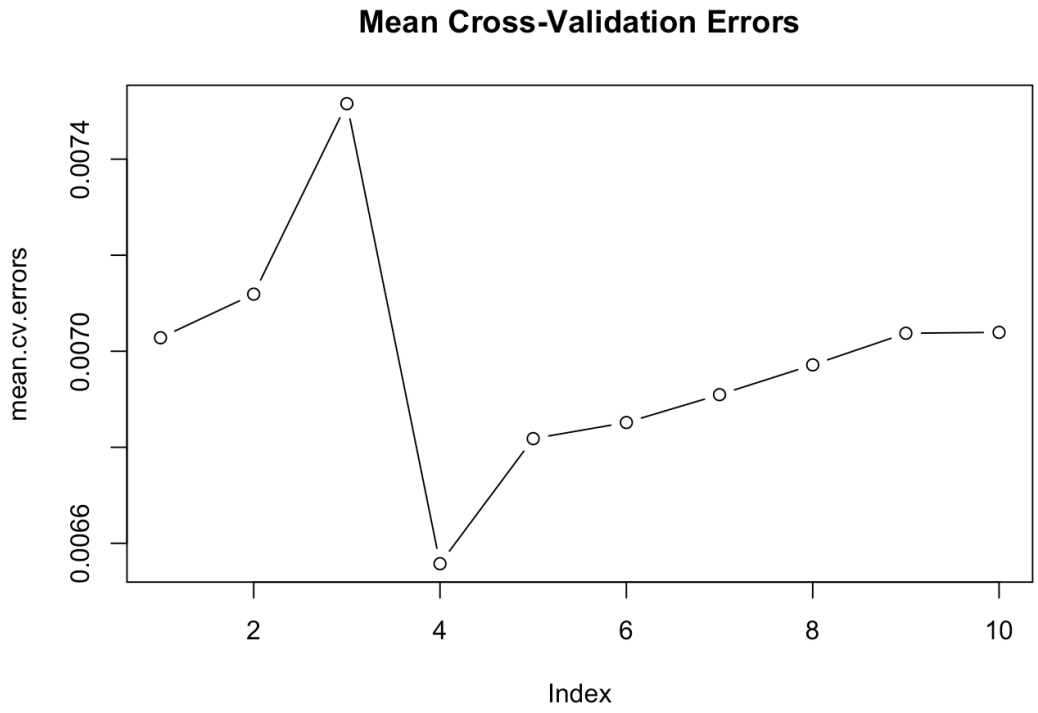


Fig-3.4.10 Figure of MSE of K-fold CV of different models predicting the resolution rate

Adjusted R^2	Cp	BIC	R^2	CV	K-fold CV
6	4	2	10	4	4

Table-3.4.6 Different criteria and their conclusions on the number of features the model should contain

So, the final conclusion is that the model with 4 features should be selected.

3.5 Regression

3.5.1 Regression for number of cases

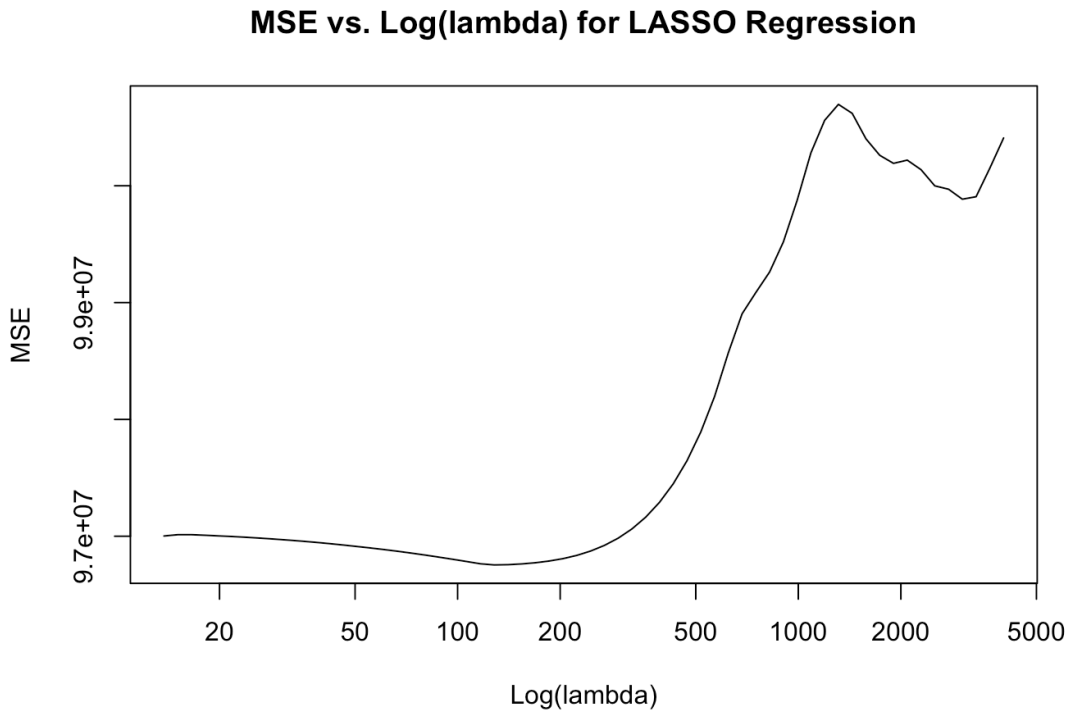


Fig-3.5.1 Parameter selection for lasso regression of the model for number of cases

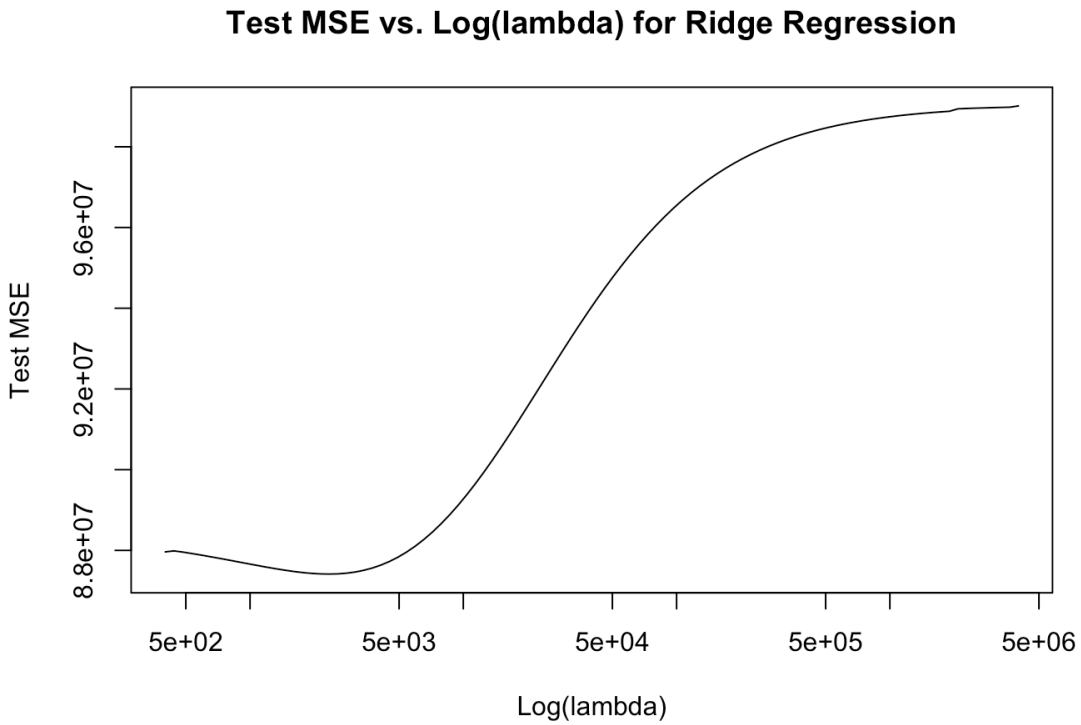


Fig-3.5.2 Parameter selection for ridge regression of the model for number of cases

Linear Model	General Linear Model	Ridge	Lasso	selected
91004262	91004262	89834339	89712556	Ridge Regression

Table-3.5.1 The MSE of different models and model selection

Conclusion: The features used for regression should be turnout, turnout_two_lag, population, and black, and the model used should be ridge regression.

3.5.2 Regression for resolution rate

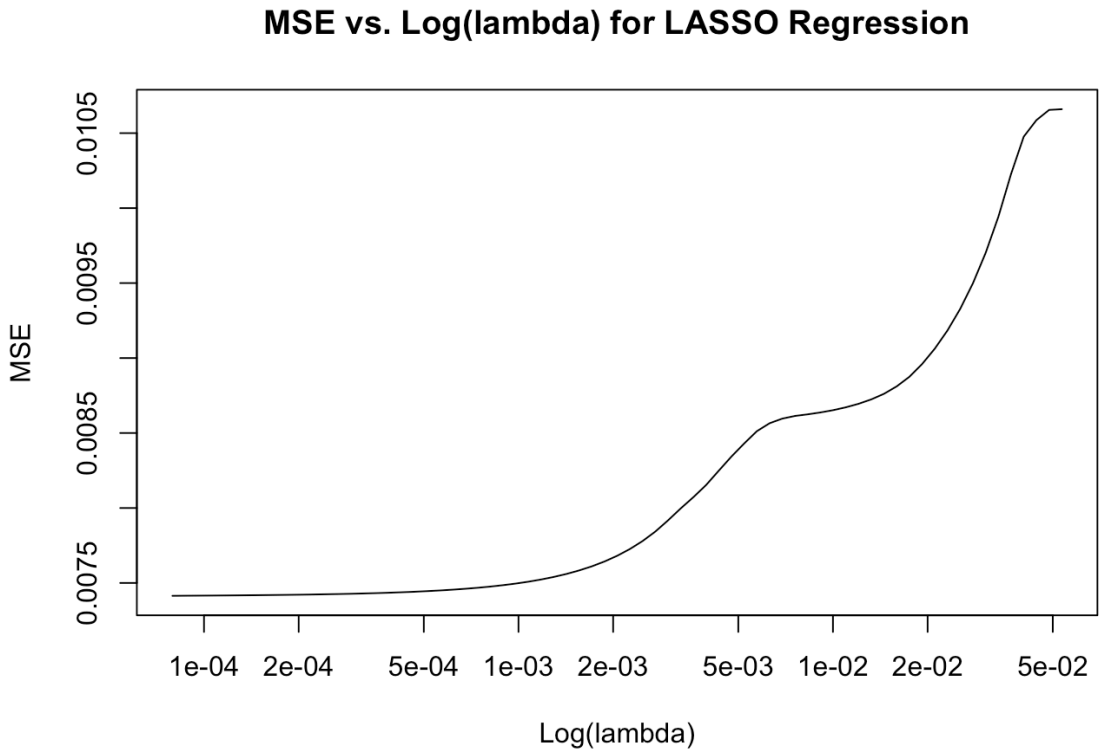


Fig-3.5.3 Parameter selection for lasso regression of the model for the resolution rate

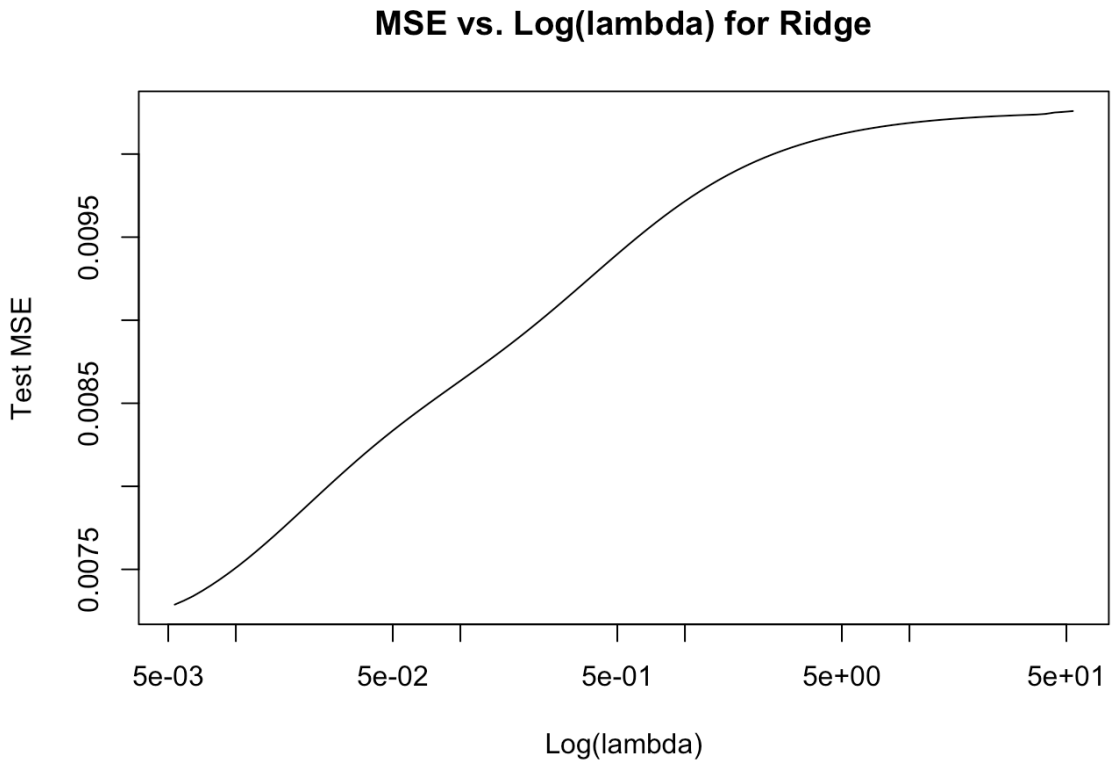


Fig-3.5.4 Parameter selection for ridge regression of the model for the resolution rate

Linear Model	General Linear Model	Ridge	Lasso	selected
0.004052783	0.004052783	0.003238037	0.004021761	Ridge Regression

Table-3.5.2 The MSE of different models and model selection

Conclusion: The features used for regression should be tturnout_one_lag, turnout_two_lag, income, and white, and the model used should be ridge regression.

3.6 Discussion

3.6.1 About feature selection

For the features used to predict the number of cases, we originally decided to use the 7 feature one because we thought the 3 feature model selected by cross-validation is too small, but we found that the K-fold cross-validation selects a 4 feature model, and K-fold cross-validation is more robust and reliable, so we turned to the 4 features one for the report.

3.6.2 About the GLM

Originally we thought the phenomenon that the GLM produces the same MSE with the linear model indicates the data follows a linear pattern. However, after discussion with the professors, we realized that it's more likely due to the fact that we do not give GLM product variables as input.

3.6.3 About model selection

For both the "number of cases" problem and the "resolution rate" problem, the ridge regression produces the least MSE. Ridge regression is suitable especially when multicollinearity exists, that is, independent variables are not actually independent of each other, but rather correlated. This is just the case with this dataset. The proportion of white, black and asian should sum up to roughly the same amount (not exactly one because there are other races), so they are likely to be negatively correlated with each other. Turnout rate, turnout_one_lag and turnout_two_lag both indicate the resource richness during that year and neighborhood, so they are likely to be positively correlated with each other. Registered and population are also likely to positively correlate with each other. Next time, if we calculate some product and quotient of the variables and use them to fit a general linear model, maybe we can get better results.

4. Conclusions

Therefore, we make a reasonable conclusion that income, population, race proportion, and turnout time are the factors that influence the number of cases and the resolution rate. Out of these factors, income, turnout time, and race proportion of whites are the main factors that greatly influence the resolution rate. And the population, the turnout time, and the race proportion of blacks are the main factors that greatly influence the resolution rate. Different neighborhoods are likely to have different most frequent types of cases, and as the test results from the regression model indicated, the racial composition does influence case resolution time. The resolution time also varies with months and years. For residents, it is a good idea to attach a photo when requesting help because having a photo in description helps with resolving the problem faster and more successfully.

Reference:

Code source: https://jfh.georgetown.domains/dsan5300/