A Machine Learning Approach to Predict the Total Age-Adjusted Mortality Rate Using Air Pollution Data and Socioeconomic Status in the United States

Oon Yu Yang[1]

School of Physical and Mathematical Sciences, NTU (Singapore).

Abstract

In this study, we apply a machine learning (ML) approach to predict the total age-adjusted mortality rate (AAMR) using the air pollution data and the population's socioeconomic status across sixty metropolitan areas of the United States. We use three different ML regression techniques, namely linear regression (LR), k-Nearest-Neighbours (kNN), and the random forests (RF) models to achieve our aim. The result of this study suggests that the RF model has the best predicting performance. We report that the most significant factor which affects the AAMR is the percentage of the non-white population in the studied areas. This knowledge and approach are particularly useful for air quality management authority, city development board, and for other demographic, sociologic, and humanitarian study purposes.

Keywords: Machine learning, mortality, air pollution, particulate matter, black community.

1. Introduction

Air pollution is a major environmental and ecological challenge, resulting in potential loss of human lives, disruption of biological diversity, and creating huge economical losses. Since the beginning of the industrial revolution, air pollution has been continually threatening human populations. In recent years, it is well-known that the poorest communities embrace the largest impact as they do not have access to enough medical resources and financial aid[1, 2]. In this study, we determine which factors affect the age-adjusted mortality rate (AAMR) the most? Either they are environmental (e.g., particulate matter combined with temperature and rain precipitation), socioeconomical (e.g., employment rate in white-collar occupations, median school years completed, salary), or both. By comparing and selecting the best model out of three different ML techniques, namely k-Nearest Neighbours (kNN), linear regression (LR), and random forests (RF), we want to predict an urbanised area's AAMR using regression method since mortality rate is a continuous, numerical variable.

2. Data Source

The dataset used is from McDonald and Schwing (1973), which is now available online on Statlib (http://lib.stat.cmu.edu/datasets/)[3]. This dataset consists of 16 independent variables (see table below) is used to train our ML models and the total AAMR on 60 U.S. metropolitan areas in 1959-1961 is our target variable.

`age`	Total Age-Adjusted Mortality Rate (AAMR).
`prep`	Mean annual precipitation in inches.
`jan.temp`	Mean January temperature in degrees Fahrenheit.
`jul.temp`	Mean July temperature in degrees Fahrenheit.
`older.65`	Percent of 1960 SMSA population that is 65 years of age or over.
`ppl.household`	Population per household, 1960 SMSA.
`school.year`	Median school years completed for those over 25 in 1960 SMSA.
`housing.unit`	Percent of housing units that are found with facilities.
`ppl.sqmile`	Population per square mile in urbanized area in 1960.
`ppl.nonwhite`	Percent of 1960 urbanized area population that is non-white
`white.collar`	Percent employment in white-collar occupations in 1960.
`income`	Percent of families with income under 3,000 in 1960 urbanized area.
`hc`	Relative pollution potential of hydrocarbons (HC).
`nox`	Relative pollution potential of oxides of nitrogen (NO_x).
`so2`	Relative pollution potential of sulfur dioxide (SO_2).
`rel.humidity`	Percent relative humidity, annual average at 1 p.m.
`risk`	Considered high if the age-adjusted mortality rate is higher than the 3rd quartile (Q3) of the recorded rates; considered low if it is lower than the 1st quartile (Q1); otherwise, it is considered medium.

Table 1: Variables and their attributes in the dataset.

[1] Email address: P190001@e.ntu.edu.sg. Matriculation number: U1940100K.

The age-adjusted rate (AAR) is a method to make fair comparisons between groups with different age distributions[4]. For instance, a state with more elderly people tends to have higher rate of death or hospitalisation than a state with younger population, merely because the elderly is more likely to die or to be hospitalised. Age adjustment can make the different groups more comparable. A "standard" population is used to adjust the mortality rates in this context, which are the rates that would have existed if the population under study had the same age distribution as the "standard" population[5]. The National Center for Health Statistics of the United States recommends that the U.S. 1940, 1970 or 2000 standard population should be used when calculating age-adjusted rates. Thus, the U.S. 1940 standard population was used in this dataset.

The mortality rate of each U.S. metropolitan area was computed for each age group by dividing the number of mortalities in that age group by the estimated population of the same age group in that area and then multiplying by a constant of 100,000. This results in an age-specific mortality rate (ASMR) per 100,000 population for each age group. Each ASMR is then multiplied by the proportion of the standard population that same age group. The age-specific results are summed to get the AAMR for each state. The formula is:

$$\text{AAMR} = \sum (\text{ASMR} \times \text{standard population})$$

Using the dataset, we hypothesize that areas with more non-white, less highly educated populations, and worse environmental pollutions tend to have higher AAMR because the community did not have access to better medical facilities or higher living standards[6, 7].

3. Preliminary Analysis

First, we analyse the distribution of the total age-adjusted mortality rates across all 60 areas (Figure 1B). It is found that the histogram peaks at 940-960. Second, we analyse the correlations between all variables in the dataset, excluding the categorial variable `risk`. In the correlation plot, it is found that variables `school.year`, `housing.unit`, and `white.collar` have the most negative correlations with the mortality rate, which are all socioeconomical factors. This implies that population which had better education, higher-paid jobs, and better living conditions had lower mortality rate, which agrees with our hypothesis. On the other hand, variables `ppl.nonwhite`, `prep`, and `so2` had the most positive correlations with the mortality rate. It implies that urban areas with higher non-white population had higher mortality rate in general. While for the environmental factors `prep` and `so2`, it implies that urban areas with more acidic rain of higher frequency tend to have population with higher mortality rate, which again, agrees with our hypothesis.

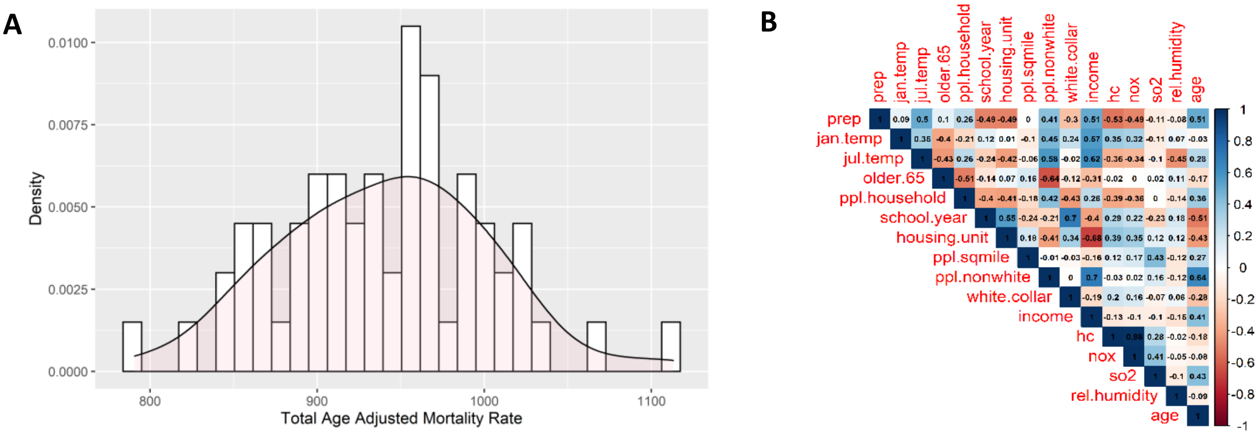


Figure 1. (A) The histogram for the total-age adjusted mortality rate, showing a slightly left-skewed normal distribution. (B) The correlation plot showing the correlations between all continuous variables in the dataset.

4. Predictions using ML Models and their Interpretations

In the dataset, there are no NA values. We transform the `risk` variable from low, medium, and high to the factors 1, 2, and 3, respectively. Then, we split the dataset (60 entries) into the training set (48 entries) and the testing set (12 entries) according to the 80:20 rule. To predict the AAMR, which is the variable `age`, we also include other environmental factors (e.g., rain precipitation and temperature, etc.) because study reveals the interactions between particulate matter in the air and the temperature can greatly influence the human mortality[8].

4.1 k-Nearest-Neighbour (kNN) Model

We perform the prediction by training the kNN model twice. First, we use all numerical variables in linear order other than the variable `age`. Second, all numerical variables other than `age`, but now we raise the three most negatively correlated variables (`school.year`, `housing.unit`, and `white.collar`) and the three most positively correlated variables (`ppl.nonwhite`, `prep`, and `so2`) to the second order to enhance the performance of the model. To find the best `k` parameter, we design a for-loop to tune this parameter. The best parameter found is `k=5` (Figure 5, Appendix) since it reports the lowest root-mean-squared error (RMSE). Moreover, we set the following training parameters to improve the kNN performance:

`trControl`	`"cv"`, with `number = 3`
`preProcess`	`"center"` and `"scale"`
`tuneLength`	`10`

4.2 Linear Regression (LR) Model

Next, we train a LR model using the method identical to the training of the kNN model. From the summary details of the trained LR model (Figure 5B, Appendix), it shows that the non-white population variable has the most significant impact on the AAMR. The R-squared value of the LR model is approximately 0.75, which means that the proposed LR model can explain 75% of variations in the mortality rate. We also present the residual plot of the better LR model (Figure 6A, Appendix) to check the linearity. The pattern shows a roughly symmetrical distributed shape, which means that the LR model is acceptable, yet it still can be improved.

4.3 Random Forests (RF) Model

Lastly, we train a RF model using the method identical to the training of the kNN model. The summary details of the RF model shows that the best tuned RF model has 500 trees and 12 variables tried at each split (Figure 6B, Appendix).

4.4 Prediction Performance and Results

After having trained all the model, we evaluate the performance of all three models using the testing set. The RMSE of each model, using different training methods, are recorded in Table 2 and their best prediction performance are shown in Figure 2. According to the results tabulated, the best result (lowest RMSE) is 13.5517 from the RF model. In addition, the performance of all models increases except for the kNN model. Thus, we concluded that the three most negatively correlated variables (`school.year`, `housing.unit`, and `white.collar`) and the three most positively correlated variables (`ppl.nonwhite`, `prep`, and `so2`) are really the important factors to determine the AAMR while the most significant variable among them is the `ppl.nonwhite`, which is the percentage of non-white population in the urbanised area.

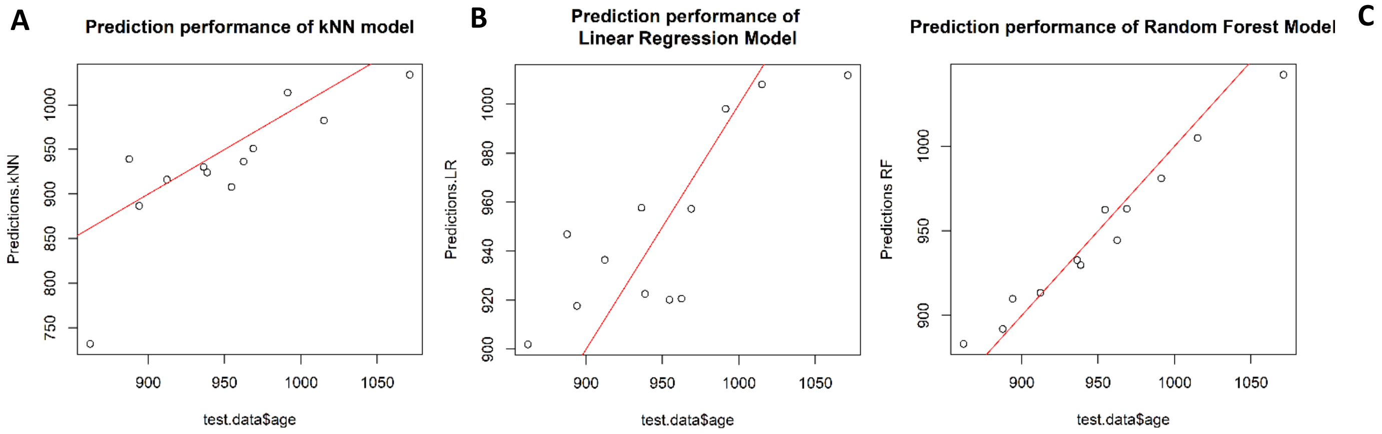


Figure 2. The best prediction performance of the (A) kNN model, (B) linear regression (LR) model, and (C) random forest (RF) model.

ML (Regression) Models	RMSE (All variables in first order)	RMSE (Significant variables in second order; otherwise, in first order)
k-Nearest-Neighbour (kNN)	32.3143	33.7568
Linear Regression (LR)	47.5076	46.4968
Random Forest (RF)	13.8154	13.5517

Table 2. Comparison of RMSE of different regression models using different choices of predictor variables.

5. k-Means Clustering

In this section, we perform unsupervised learning using the k-Means clustering method predicting on the categorical variable `risk`, which is a factor based on the AAMR, using all the predictor variables. Initially, we set the number of clusters to 3, since there are three levels of factors in this variable: low, medium, and high, correspond to 1, 2, and 3, respectively and we use all the predictor variables. The result of the clustering is shown in Table 3 and Figure 7A (Appendix):

Table 3	Clusters		
`Risk`	1	2	3
Low	1	11	8
Medium	4	8	3
High	1	23	4

Clearly, the results are not perfect as they are not distributed along the main diagonal of the table. Hence, we want to find an optimal number of clusters `k` using the elbow method. The motivation is to minimise the within-cluster sum of square (WSS) as small as possible. We plot the curve of total WSS against the number of clusters `k` (Figure 3), and the location of a bend (knee) in the plot considered as an indicator of the appropriate number of clusters. The result suggests that 6 is the optimal number of clusters as it appears to be the bend in the knee. Hence, the optimal number of levels for assigning the `risk` factor is 6, instead of 3 as we considered initially. Then, we perform the final analysis using the most significant factor we found in the previous section: `ppl.nonwhite` and obtain the results by setting the number of clusters to 6 (Figure 4, Figure 7B in Appendix).

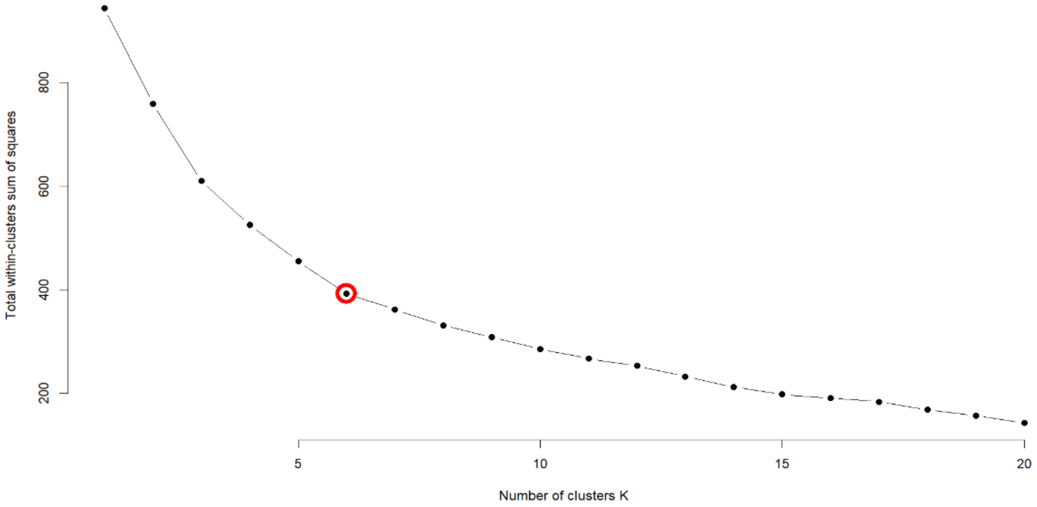


Figure 3. The plot of total WSS against the number of clusters `k`. It suggests that the optimal number of clusters is 6 as it appears to be the bend in the knee (circled in red).

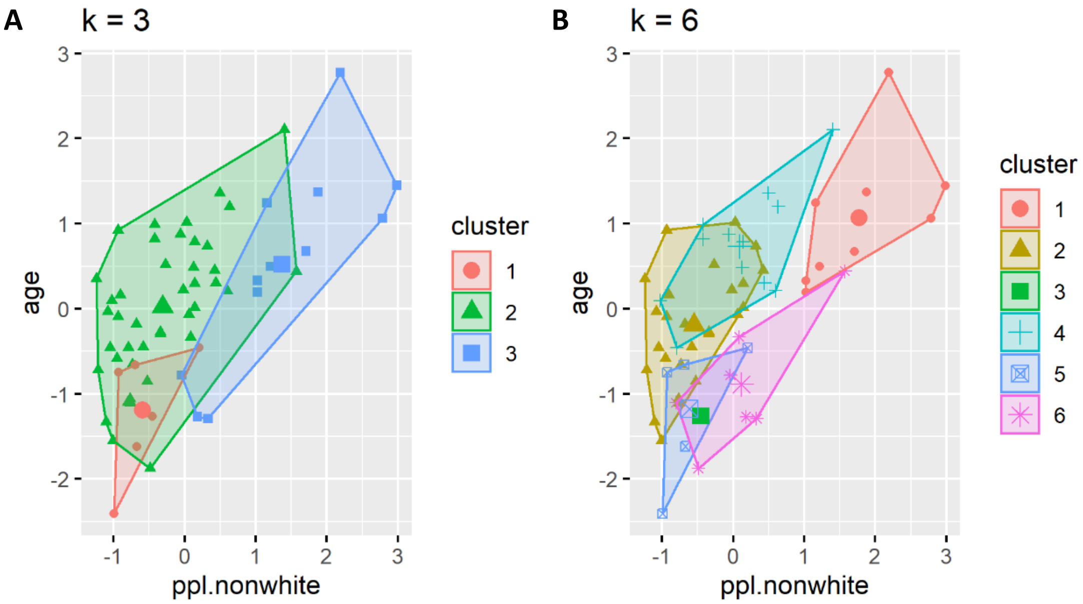


Figure 4. Unsupervised learning using k-Means clustering based on (A) 3 clusters, since in the `risk` variable, there are three levels of factors considered: low, medium, and high. However, the optimal value of clusters `k` is found to be 6 from the WSS plot. (B) Therefore, we repeated the clustering but now using 6 clusters instead of 3.

6. Conclusion

In this study, we conclude that the best ML model for predicting the age-adjusted mortality rate is the random forest (RF) model, compared with the kNN and the linear regression model while the significant variables are `school.year`, `housing.unit`, `white.collar`, `ppl.nonwhite`, `prep`, and `so2`. Among these significant variables, `ppl.nonwhite`, the percentage of the non-white population has the most profound effect on the age-adjusted mortality rate. Consequently, training the RF model using the significant variables raised to the second power produce the best prediction. Finally, from the unsupervised learning using the k-Means clustering method, the result suggests that the optimal number of levels assigned for the mortality `risk` categorical factor is 6, instead of 3 as we considered (low, medium, and high) at the beginning.

However, there are still some limitations of this study. First, the dataset used in this study is outdated. We use the 1959-1961 total age-adjusted mortality rate based on the U.S. 1940 standard population. Second, the data size is limited in scope. In this study, only 60 metropolitan areas from the U.S. are considered. Third, there are rooms for improvement on the performance of the regression model. Not all parameters for training the model are considered and tuned systematically. Therefore, future research studies shall focus on the most recent data, not to mention that more metropolitan areas across the U.S. shall also be included as well. Moreover, future study shall focus on using the `tuneGrid` method from the Caret package to search the algorithm parameters by specifying a tune grid manually[9]. In the grid, each algorithm parameter can be specified as a vector of possible values. These vectors combine to define all the possible combinations to try.

Data Availability Statement and Supplementary Material

All datasets, R code file, R library packages and their respective versions presented and applied in this study are available online at https://github.com/Isomorpfishm/NTU-PS0002-AY2021-MLProject.

Acknowledgement

The author would like to express his sincere gratitude and appreciation to the course instructor Assoc Prof Xiang Liming for her support and instructions while conducting this study and Mr Tiang Bi Hong from the School of Physical and Mathematical Sciences for his review, comments, and suggestions on this paper.

Appendix

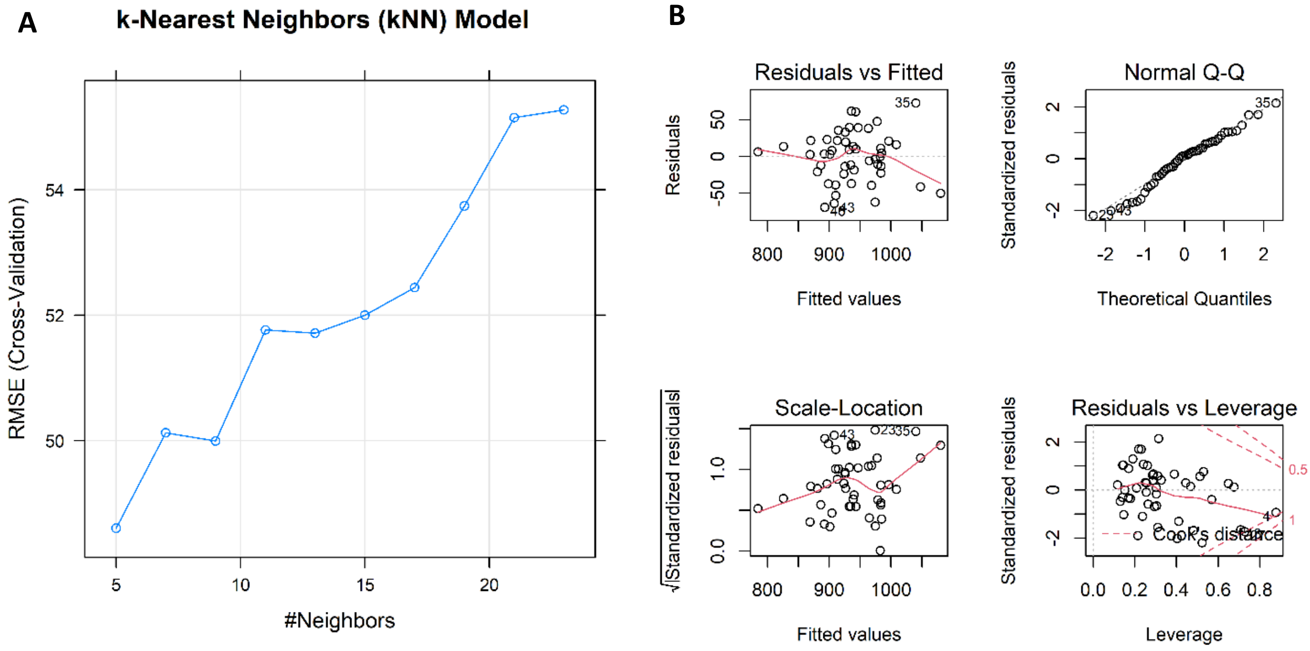
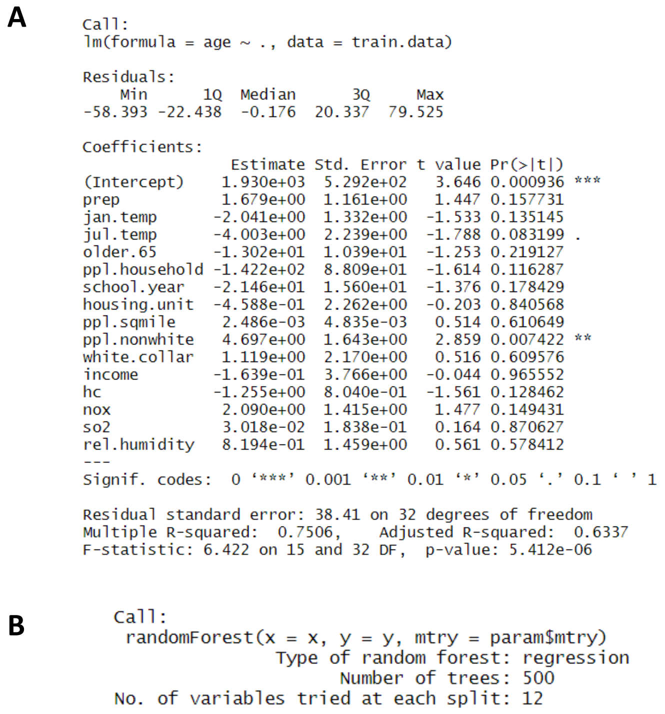


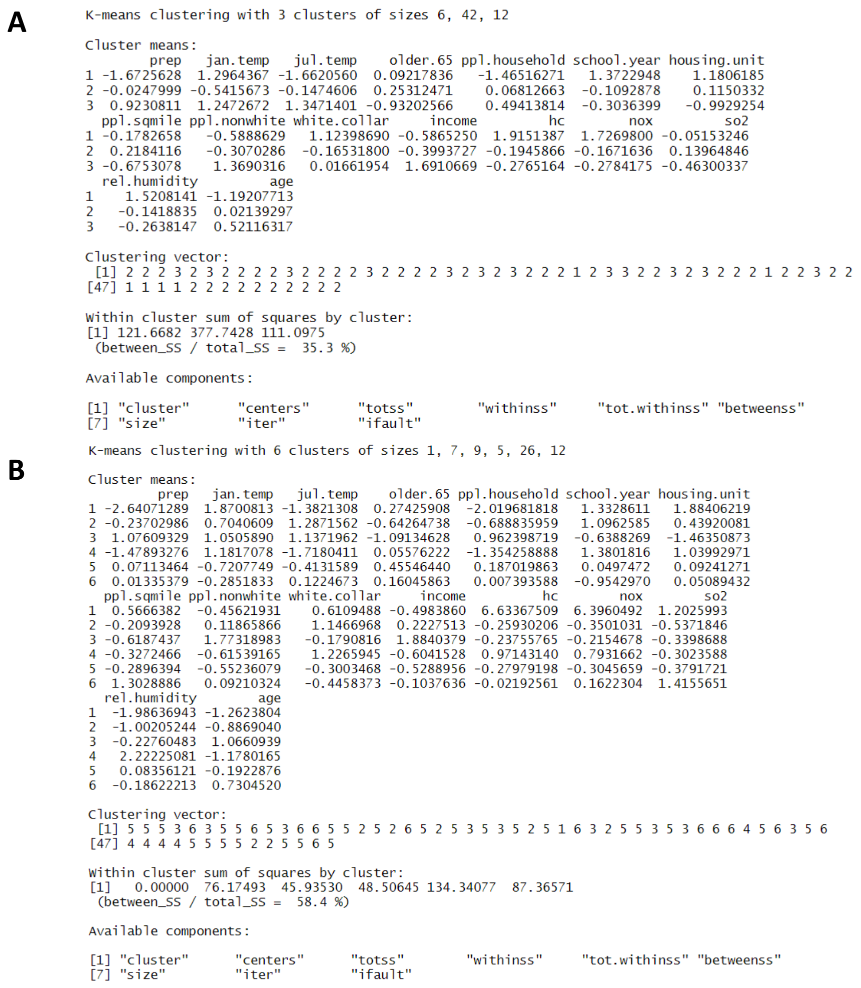
Figure 5. (A)The optimal value for the parameter k in the kNN model, and (B) the residual plot for the linear regression (LR) model.



```
Call:
lm(formula = age ~ ., data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-58.393 -22.438  -0.176  20.337  79.525 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.930e+03  5.292e+02   3.646 0.000936 ***
prep           1.679e+00  1.161e+00   1.447 0.157731    
jan.temp      -2.041e+00  1.332e+00  -1.533 0.135145    
jul.temp      -4.003e+00  2.239e+00  -1.788 0.083199 .  
older.65      -1.302e+01  1.039e+01  -1.253 0.219127    
ppl.household -1.422e+02  8.809e+01  -1.614 0.116287    
school.year   -2.146e+01  1.560e+01  -1.376 0.178429    
housing.unit  -4.588e-01  2.262e+00  -0.203 0.840568    
ppl.sqmile     2.486e-03  4.835e-03   0.514 0.610649    
ppl.nonwhite   4.697e+00  1.643e+00   2.859 0.007422 ** 
white.collar   1.119e+00  2.170e+00   0.516 0.609576    
income        -1.639e-01  3.766e+00  -0.044 0.965552    
hc            -1.255e+00  8.040e-01  -1.561 0.128462    
nox            2.090e+00  1.415e+00   1.477 0.149431    
so2            3.018e-02  1.838e-01   0.164 0.870627    
rel.humidity   8.194e-01  1.459e+00   0.561 0.578412    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.41 on 32 degrees of freedom
Multiple R-squared:  0.7506,	Adjusted R-squared:  0.6337 
F-statistic: 6.422 on 15 and 32 DF,  p-value: 5.412e-06
```

```
Call:
 randomForest(x = x, y = y, mtry = param$mtry) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 12
```

Figure 6. The summary details of the results for (A) the linear regression model. The result suggests that among all the significant factors that we consider, the percentage of the non-white population has the most profound influence on the prediction outcome. (B) The random forest (RF) model for the regression task. The optimal number of trees is 500 while in each split, 12 variables are tried.



```
K-means clustering with 3 clusters of sizes 6, 42, 12

Cluster means:
        prep   jan.temp   jul.temp    older.65 ppl.household school.year housing.unit
1 -1.6725628  1.2964367 -1.6620560  0.09217836   -1.46516271   1.3722948    1.1806185
2 -0.0247999 -0.5415673 -0.1474606  0.25312471    0.06812663  -0.1092878    0.1150332
3  0.9230811  1.2472672  1.3471401 -0.93202566    0.49413814  -0.3036399   -0.9929254
  ppl.sqmile ppl.nonwhite white.collar     income         hc        nox         so2
1 -0.1782658   -0.5888629   1.12398690 -0.5865250  1.9151387  1.7269800 -0.05153246
2  0.2184116   -0.3070286  -0.16531800 -0.3993727 -0.1945866 -0.1671636  0.13964846
3 -0.6753078    1.3690316   0.01661954  1.6910669 -0.2765164 -0.2784175 -0.46300337
  rel.humidity         age
1    1.5208141 -1.19207713
2   -0.1418835  0.02139297
3   -0.2638147  0.52116317

Clustering vector:
 [1] 2 2 2 3 2 3 2 2 2 2 3 2 2 2 2 3 2 2 2 2 3 2 3 2 3 2 2 2 1 2 3 3 2 2 3 2 3 2 2 2 1 2 2 3 2 2
[47] 1 1 1 1 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 121.6682 377.7428 111.0975
 (between_SS / total_SS =  35.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"   
[7] "size"         "iter"         "ifault"      
```

```
K-means clustering with 6 clusters of sizes 1, 7, 9, 5, 26, 12

Cluster means:
         prep  jan.temp   jul.temp    older.65 ppl.household school.year housing.unit
1 -2.64071289 1.8700813 -1.3821308  0.27425908  -2.019681818   1.3328611   1.88406219
2 -0.23702986 0.7040609  1.2871562 -0.64264738  -0.688835959   1.0962585   0.43920081
3  1.07609329 1.0505890  1.1371962 -1.09134628   0.962398719  -0.6388269  -1.46350873
4 -1.47893276 1.1817078 -1.7180411  0.05576222  -1.354258888   1.3801816   1.03992971
5  0.07113464 -0.7207749 -0.4131589  0.45546440   0.187019863   0.0497472   0.09241271
6  0.01335379 -0.2851833  0.1224673  0.16045863   0.007393588  -0.9542970   0.05089432
  ppl.sqmile ppl.nonwhite white.collar     income          hc        nox        so2
1  0.5666382  -0.45621931    0.6109488 -0.4983860  6.63367509  6.3960492  1.2025993
2 -0.2093928   0.11865866    1.1466968  0.2227513 -0.25930206 -0.3501031 -0.5371846
3 -0.6187437   1.77318983   -0.1790816  1.8840379 -0.23755765 -0.2154678 -0.3398688
4 -0.3272466  -0.61539165    1.2265945 -0.6041528  0.97143140  0.7931662 -0.3023588
5 -0.2896394  -0.55236079   -0.3003468 -0.5288956 -0.27979198 -0.3045659 -0.3791721
6  1.3028886   0.09210324   -0.4458373 -0.1037636 -0.02192561  0.1622304  1.4155651
  rel.humidity        age
1  -1.98636943 -1.2623804
2  -1.00205244 -0.8869040
3  -0.22760483  1.0660939
4   2.22225081 -1.1780165
5   0.08356121 -0.1922876
6  -0.18622213  0.7304520

Clustering vector:
 [1] 5 5 5 3 6 3 5 5 6 5 3 6 6 5 5 2 5 2 6 5 2 5 3 5 3 5 2 5 1 6 3 2 5 5 3 5 3 6 6 6 4 5 6 3 5 6
[47] 4 4 4 4 5 5 5 5 2 2 5 5 6 5

Within cluster sum of squares by cluster:
[1]   0.00000  76.17493  45.93530  48.50645 134.34077  87.36571
 (between_SS / total_SS =  58.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"   
[7] "size"         "iter"         "ifault"      
```

Figure 7. The summary details of the results for the unsupervised learning using the k-Means clustering fixed with (A) three clusters, and (B) six clusters.

References

[1] A. P. Polednak, "Black-white differences in infant mortality in 38 standard metropolitan statistical areas," (in en), *Am J Public Health,* vol. 81, no. 11, pp. 1480-1482, 1991/11// 1991, doi: 10.2105/AJPH.81.11.1480.

[2] M. A. Winkleby and C. Cubbin, "Influence of individual and neighbourhood socioeconomic status on mortality among black, Mexican-American, and white women and men in the United States," (in en), *Journal of Epidemiology & Community Health,* vol. 57, no. 6, pp. 444-452, 2003/06/01/ 2003, doi: 10.1136/jech.57.6.444.

[3] G. C. McDonald and R. C. Schwing, "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics,* vol. 15, no. 3, pp. 463-481, 1973/08/01 1973, doi: 10.1080/00401706.1973.10489073.

[4] D. F. Austin and S. B. Werner, *Epidemiology for the health sciences: a primer on epidemiologic concepts and their uses*. Springfield: Thomas (in eng), 1974, p. 75.

[5] N. N. Naing, "Easy way to learn standardization : direct and indirect methods," (in eng), *Malays J Med Sci,* vol. 7, no. 1, pp. 10-15, 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22844209

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406211/.

[6] R. T. Anderson, P. Sorlie, E. Backlund, N. Johnson, and G. A. Kaplan, "Mortality Effects of Community Socioeconomic Status," *Epidemiology,* vol. 8, no. 1, pp. 42-47, 1997 1997. [Online]. Available: https://www.jstor.org/stable/3702421.

[7] R. A. Hummer, "Black-White Differences in Health and Mortality: A Review And Conceptual Model," (in en), *The Sociological Quarterly,* vol. 37, no. 1, pp. 105-125, 1996/01// 1996, doi: 10.1111/j.1533-8525.1996.tb02333.x.

[8] S. Roberts, "Interactions between particulate air pollution and temperature in air pollution mortality time series studies," (in en), *Environmental Research,* vol. 96, no. 3, pp. 328-337, 2004/11// 2004, doi: 10.1016/j.envres.2004.01.015.

[9] *caret: Classification and Regression Training*. (2020). Accessed: 2021/04/15/05:59:38. [Online]. Available: https://CRAN.R-project.org/package=caret