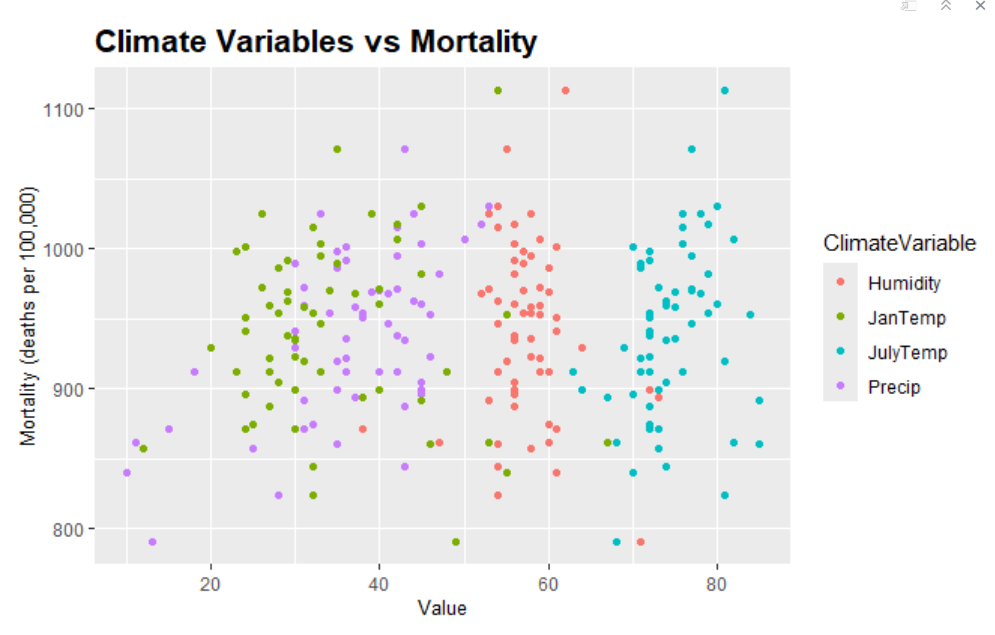
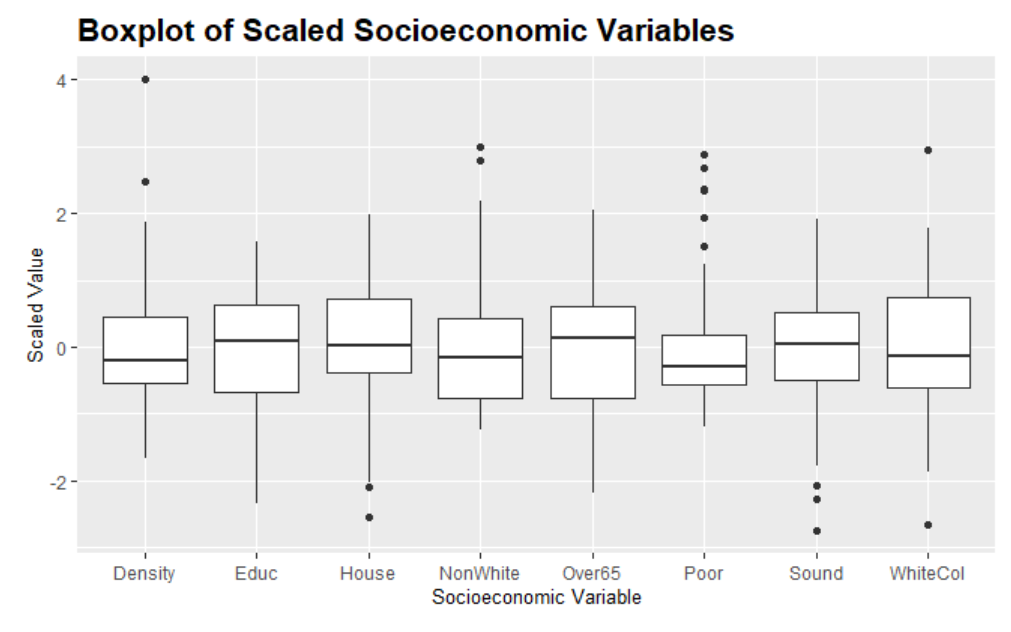
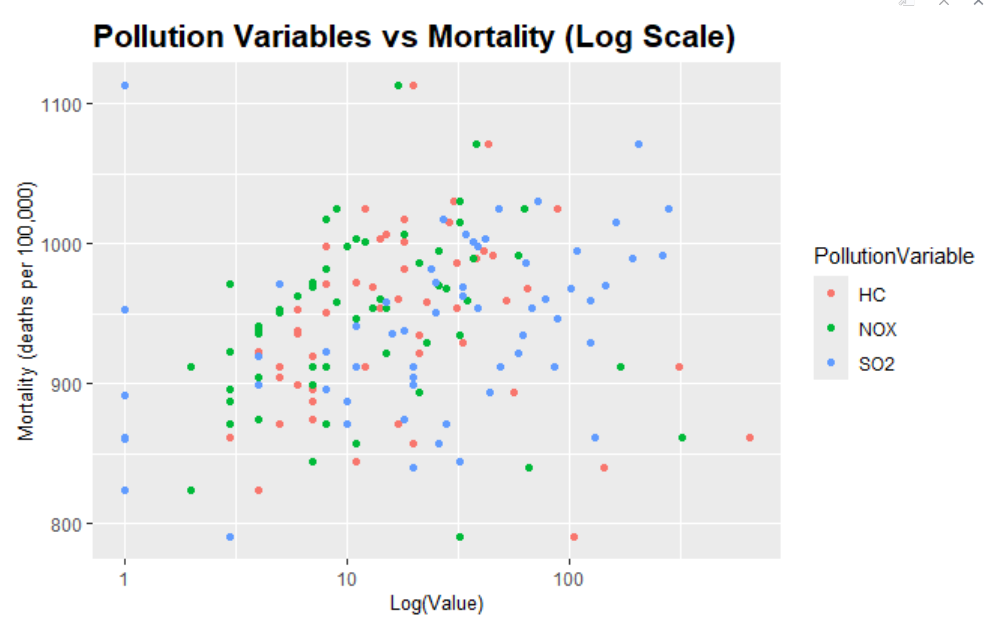
# 1a i



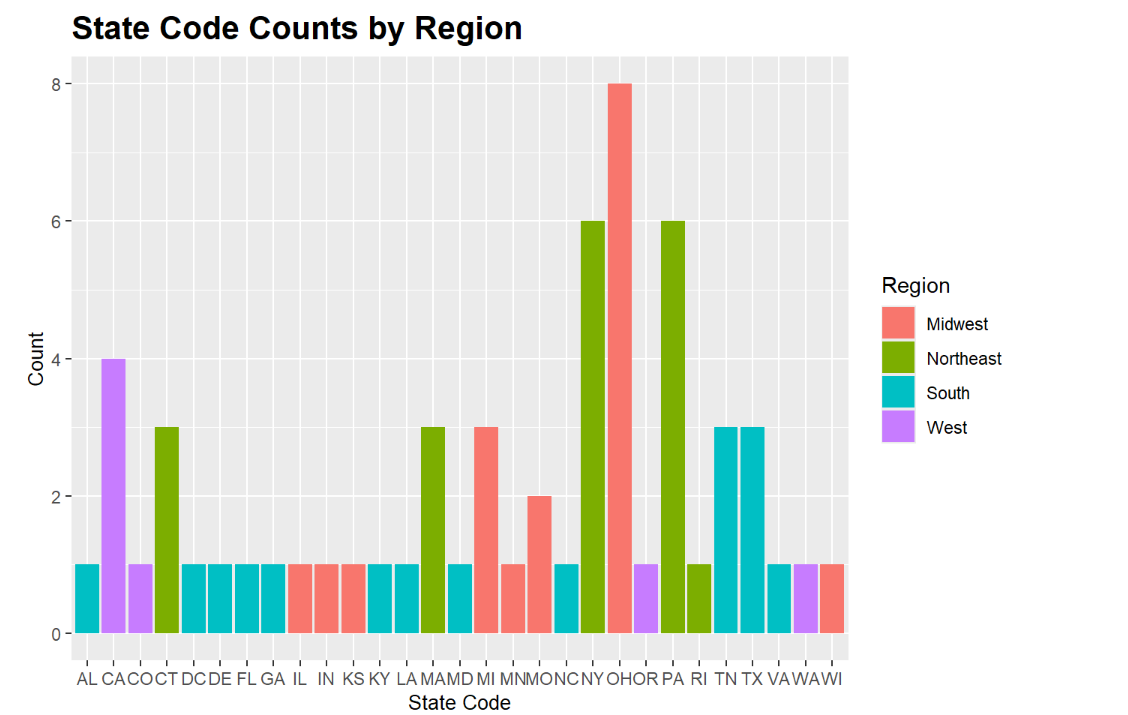
# 1 a ii



# 1 a iii

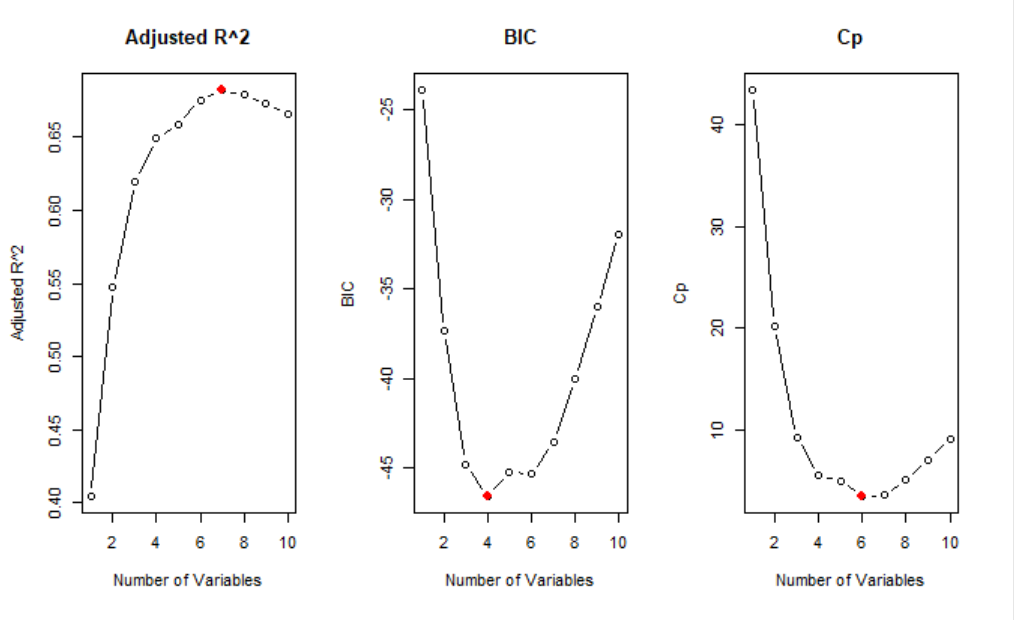


# 1 a iv

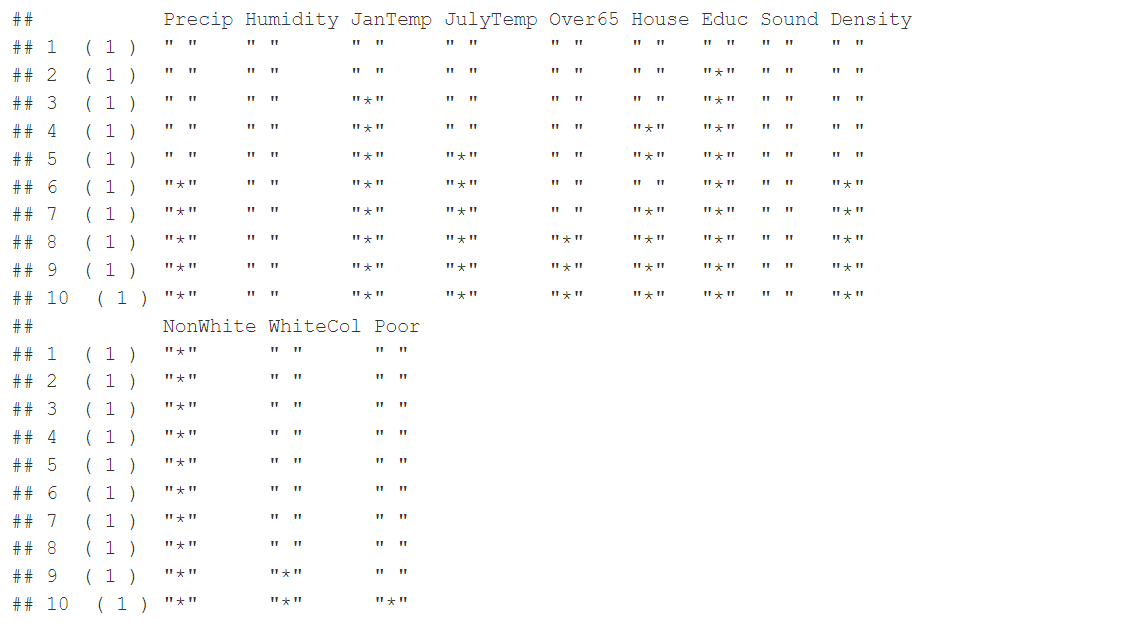


# 1 b

**Plot for model selection -**



**Outmat-**



The outmat shows the inclusion (denoted by \*) or exclusion of variables in the models, of different sizes.

1. **Model with 1 Variable**:
   * Includes only **NonWhite**.
2. **Model with 2 Variables**:
   * Includes **JanTemp** and **NonWhite**.
3. **Model with 3 Variables**:
   * Includes **JanTemp**, **JulyTemp**, and **NonWhite**.
4. **Model with 4 Variables**:
   * Includes **JanTemp**, **JulyTemp**, **House**, and **NonWhite**.
5. **Model with 5 Variables**:
   * Includes **JanTemp**, **JulyTemp**, **House**, **Educ**, and **NonWhite**.
6. **Model with 6 Variables**:
   * Includes **Precip**, **JanTemp**, **JulyTemp**, **House**, **Educ**, and **NonWhite**.
7. **Model with 7 Variables**:
   * Includes **Precip**, **JanTemp**, **JulyTemp**, **Educ**, **Density**, and **NonWhite**.
8. **Model with 8 Variables**:
   * Includes **Precip**, **JanTemp**, **JulyTemp**, **House**, **Educ**, **Density**, and **NonWhite**.
9. **Model with 9 Variables**:
   * Includes **Precip**, **Humidity**, **JanTemp**, **JulyTemp**, **House**, **Educ**, **Density**, **NonWhite**.
10. **Model with 10 Variables**:
    * Includes all variables except **Over65**, **Sound**, and **WhiteCol**.

**Optimal Models based on**:

1. **Adjusted R-squared** - optimal subset size is 7 Variable Selected is House, Education, Density, NonWhite, Precip, JanTemp, JulyTemp
2. **Cp** - subset size is 6 Variables Selected are Education, Density, NonWhite, Precip, JanTemp, JulyTemp
3. **BIC** - optimal subset size is 4 Variables Selected are House, Education, NonWhite, janTemp

# 1c

**Model Equation for optimal model with lowest Cp** :

***Mortality=1242+1.401×Precip−1.684×JanTemp−2.840×JulyTemp−16.16×Educ+0.00757×Density+5.275×NonWhite***

Intercept (1242): Baseline mortality rate.

Precip (1.401): Mortality increases with more precipitation.

JanTemp (-1.684): Mortality decreases with higher January temperatures.

JulyTemp (-2.840): Mortality decreases with higher July temperatures.

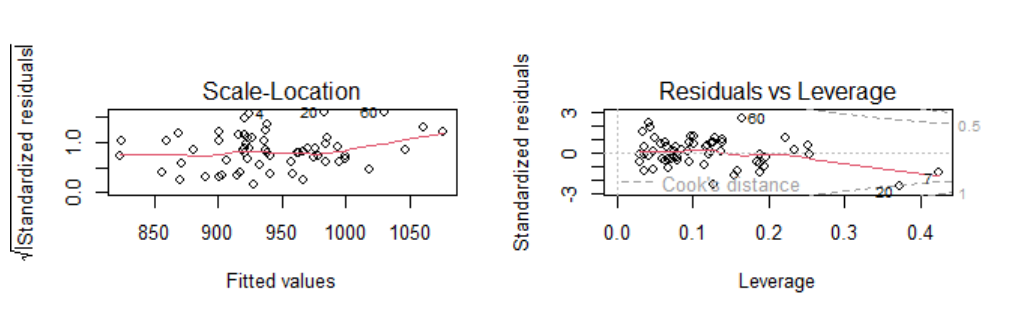
Educ (-16.16): Mortality decreases with higher education levels.

Density (0.00757): Slight increase in mortality with higher population density.

NonWhite (5.275): Mortality increases with a higher percentage of the non-white population.

Number in brackets shows the degree of impact of that variable in predicting Mortality

# 1 d



Based on both diagnostic plots, we can see that points 4, 7, 20, 60 are influential points, however we would need to dive deeper to make sure, and also find if there are more influential points.

High leverage points:

7 20

**Influential cities by Leverage**:

1) Miami, FL 2) York, PA

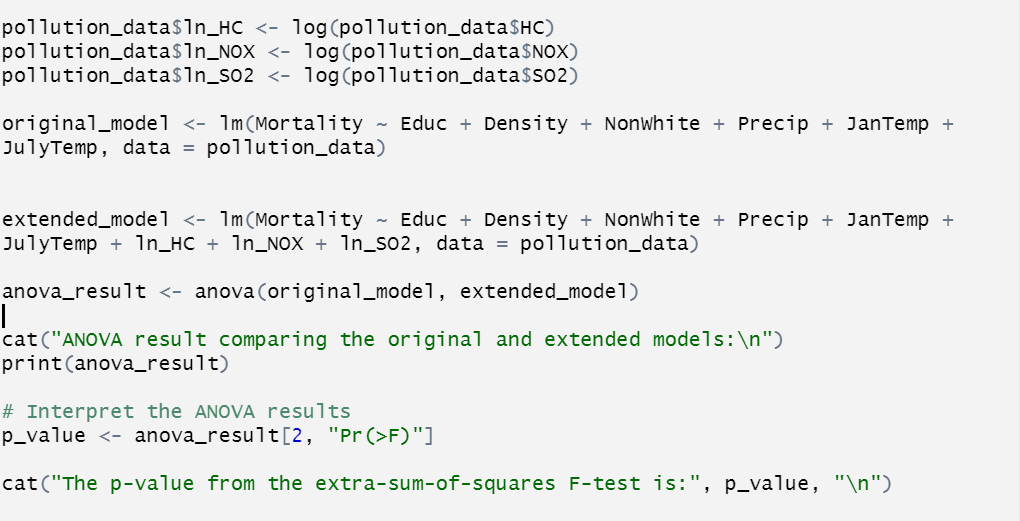
Influential points by Cook's distance:

4 7 20 53 58 60

**Influential cities by Cook's Distance**:

1) Lancaster, PA 2) Miami, FL 3) York, PA 4) Memphis, TN 5) Birmingham, AL 6) New Orleans, LA

# 1e



Adding the log-transformed pollution variables, log(HC), log(NOX), log(SO2), significantly improves the model as shown by the p-value obtained (0.0083), using Anova table. This indicates the logarithmic version of pollution variables has meaningful impact on mortality rates. The extended model provides a better explanation of the variation in mortality between cities.