Kayode Okunola
(002845768)
Computational Methods Final Project

Introduction

In recent years the number and complexity of statistical tools used to predict has grown significantly however, having analyzed research results obtained with these tools, one comes to a conclusion that when varied analytical methods are not used, the global impact of results might be limited (Hegel et al., 2010). Therefore I decided to compare the performance of two models, Generalized Addictive Model (GAM) and Random Forest (RF) to predict the non linear relationship between neighborhood factors and home prices. The principal predictors in our model are Crime Rate, Industrial Zone, Distance, Rooms and Pupil Teacher Ratio, these five variables are classified as neighborhood factors. We analyze the model accuracy, and use the best model make a conclusion.

Research Goal

To investigate and predict the non-linear relationship between neighborhood factors and Home Prices

Data Exploration

Dataset: Data were aggregated from housing data to predict median home values in Boston suburbs. Source: Available on Kaggle and sklearn.datasets.

- **Missing Value:** Utilized a missing map to examine the dataset for missing values; no missing values were identified, as shown in Figure 1 of the appendix.
- **Summary**: The mean of each variable of interest are as follows: Crime Rate (M=3.6, sd=8.6), Industrial Zone (M=11, sd=6.9), Distance (M=3.8, sd=2.1), Rooms (M=6.3, sd=0.7), Pupil-Teacher Ratio (M=18, sd=9.2), and Home Price (M=23, sd=9.2), as detailed in Table 1 in the appendix.
- **Correlation Analysis:** The correlation shows a likely non-linear relationship between crime rate and Distance against home price. However, a strong relationship tends to exist between rooms and home price while a moderate negative relationship tends to exist between industrial zone, pupil teacher ratio against home price, the correlation matrix is in table 2 of the appendix.
- **Correlation Plot:** correlation plot, the correlation plot in figure 2 of the appendix give further visualization on the relationship.

- **Scatter Plot:** We use a scatter plot to check for non-linearity trend between the predators variable and response variable. As shown in the scatter plots in appendix, the plots shows that the points are highly deviated from the regression line which suggest a non-linear relationship among the variables. Rooms shows a clear positive trend with home prices, Crime Rate and Distance show a dispersed pattern with potential outliers. Hence, this affirmed the use of non-linear modelling techniques to capture the non-linear relationships. Scatter plots displayed in figure 3.1 to figure 3.5 of the appendix.

Methodology

To adequately address our research goal, we will employ non-linear approaches, specifically the Generalized Additive Model (GAM) and Random Forest, utilizing the mgcv package for GAM and rpart for implementation. The response variable, referred to as "House Price," is the key component of our predictive model. The predictive model is constructed utilizing a framework of Generalized Additive Models (GAMs) and Random Forest, with careful consideration given to fitness evaluation and model accuracy, which serves as a measure of the model's explanatory power.

Findings

Using R, I wrote the algorithms for both GAMs and random forest

Generalized Additive Models (GAMs)

The F-statistics of our model measures the strength of the relationship between the predictor and response after accounting for the smoothness penalty. However, our model reported a high significant terms ($p<0.05$). The model summary is shown in figure 4 of the appendix.

- s(CrimeRate): edf = 3.834, indicating a non-linear relationship between crime rate and home prices. The p-value (<0.05) shows this effect is statistically significant.
- s(IndustrialZone): edf = 3.491, showing a weaker non-linear effect with marginal significance (p = 0.0603).
- s(Distance): edf = 8.653, suggesting a highly non-linear relationship, and the F-statistic/p-value confirms its importance.
- s(Rooms): edf = 8.016, suggesting a highly non-linear relationship, and the F-statistic/p-value confirms its importance.

- s(PupilTeacherRatio): edf = 3.433, indicating a non-linear relationship between crime rate and home prices. The p-value (<0.05) shows this effect is statistically significant.

Model fit

R-sq.(adj): The adjusted R-squared of 0.74, indicating that 74% proportion of variance in the response variable explained by the model. 0.74 Values closer to 1 which indicate a good fit for our model.

Random Forest Model

An RMSE of 4.89 suggests that the Random Forest model predicts home prices within approximately ±4.89 units of the actual price on average.

Model Performance Comparison

GAM: RMSE = 4.535, Adjusted R^2 = 0.742, GAM effectively captures complex non-linear relationships for predictors like Distance and Rooms, which exhibit smooth but intricate trends. While **Random Forest** RMSE = 4.89, random forest models some non-linearity, its performance is limited by high RMSE value. GAMs has the lowest RMSE value. Hence GAMs model perform better.

Discussion and Conclusion

The GAMs results shows that there is a non-linear relationship with home prices, suggesting home prices decrease sharply at low crime rates but stabilize at higher rates. A strong, highly non-linear relationship where home prices increase with the number of rooms, but the relationship is not uniform. In conclusion, There are significant non-linear relationships between home prices and key predictors such as CrimeRate, Rooms, Distance, and PupilTeacherRatio. The GAM model (Generalized Additive Model) outperforms Random Forest in both predictive accuracy and interpretability. Therefore, the GAM model effectively answers the research question, revealing important non-linear relationships between neighborhood factors and home prices. It serves as a robust tool for understanding and predicting home price dynamics in response to neighborhood attributes.

Appendix

Table 1: Summary Statistic

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
CrimeRate	506	3.6	8.6	0.0063	0.082	3.7	89
IndustrialZone	506	11	6.9	0.46	5.2	18	28
Distance	506	3.8	2.1	1.1	2.1	5.2	12
Rooms	506	6.3	0.7	3.6	5.9	6.6	8.8
PupilTeacherRatio	506	18	2.2	13	17	20	22
HomePrice	506	23	9.2	5	17	25	50

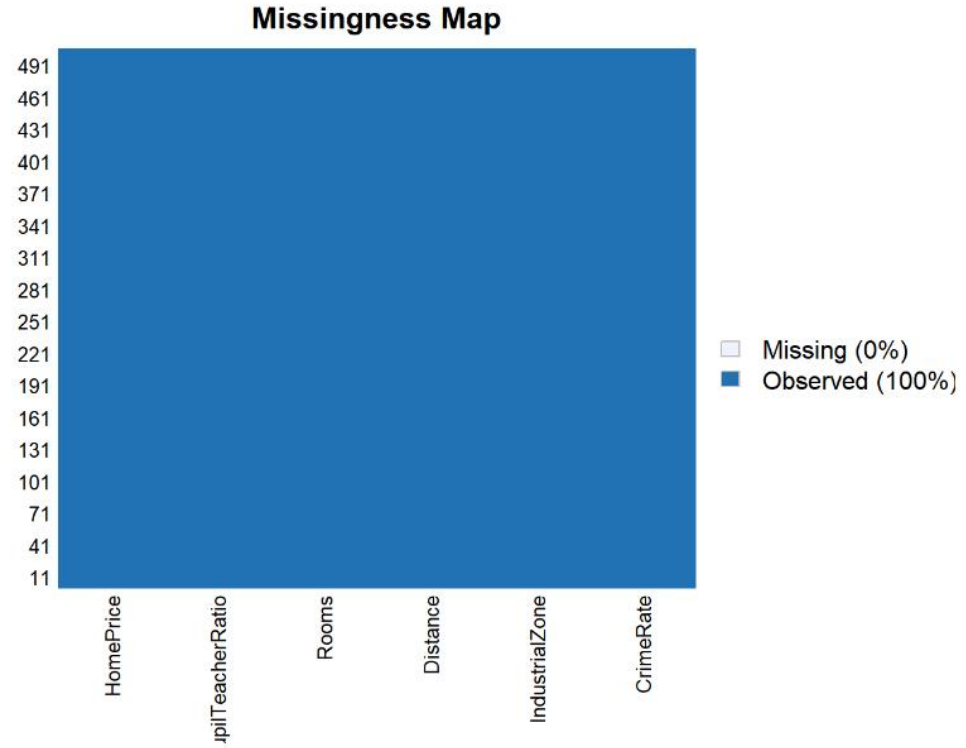


Figure 1: missing plot

```
##                    CrimeRate IndustrialZone   Distance      Rooms
## CrimeRate          1.0000000      0.4065834 -0.3796701 -0.2192467
## IndustrialZone     0.4065834      1.0000000 -0.7080270 -0.3916759
## Distance          -0.3796701     -0.7080270  1.0000000  0.2052462
## Rooms             -0.2192467     -0.3916759  0.2052462  1.0000000
## PupilTeacherRatio  0.2899456      0.3832476 -0.2324705 -0.3555015
## HomePrice         -0.3883046     -0.4837252  0.2499287  0.6953599
##                   PupilTeacherRatio  HomePrice
## CrimeRate                 0.2899456 -0.3883046
## IndustrialZone            0.3832476 -0.4837252
## Distance                 -0.2324705  0.2499287
## Rooms                    -0.3555015  0.6953599
## PupilTeacherRatio         1.0000000 -0.5077867
## HomePrice                -0.5077867  1.0000000
```

Table 2: Correlation matrix

Correlation Matrix

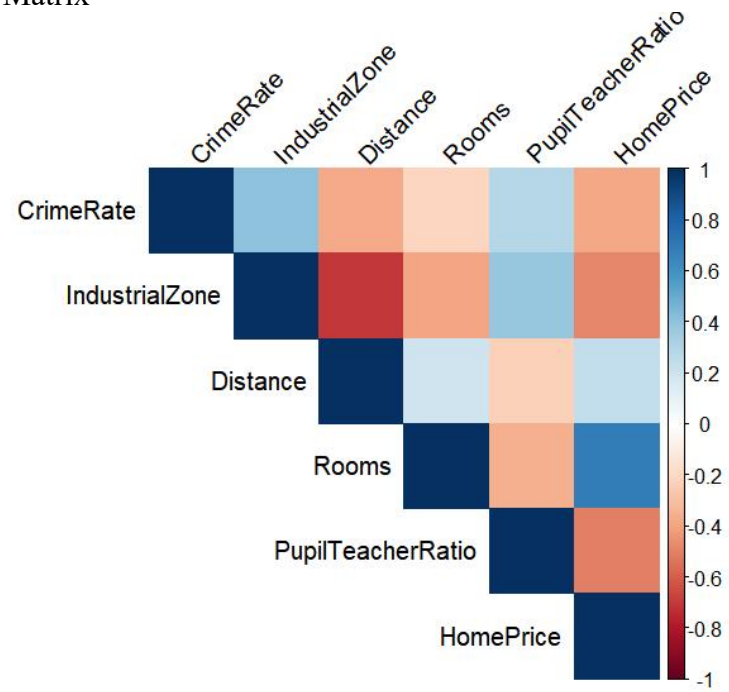


Figure 2: Correlation plot

Scatter Plot

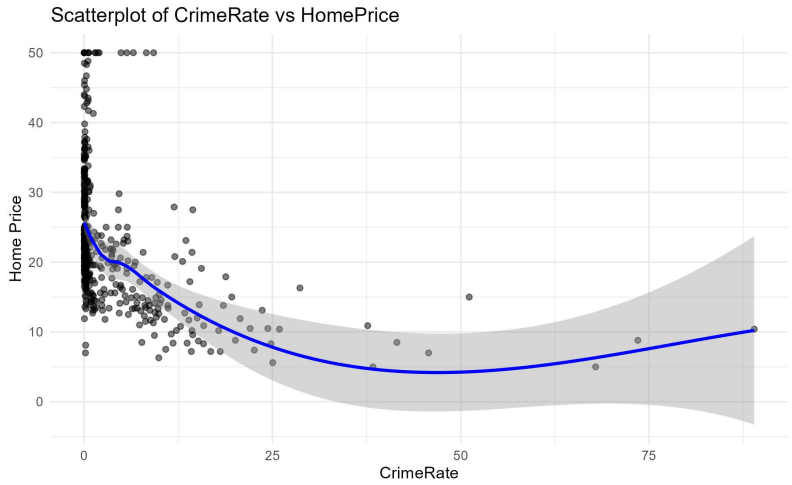


Figure 3.1: Scatter Plot of Crime rate against Home Price

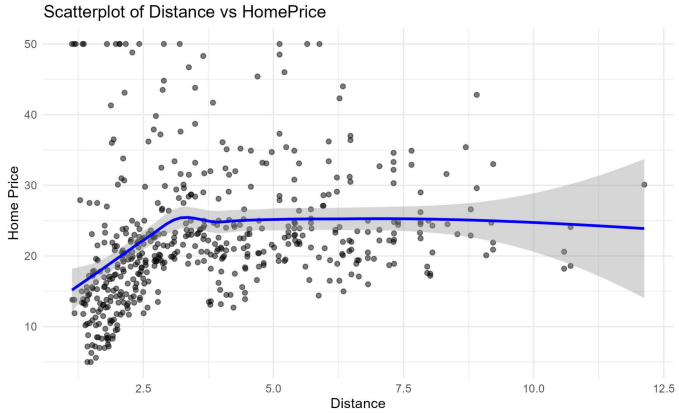


Figure 3.2: Scatter Plot of Distance against Home Price

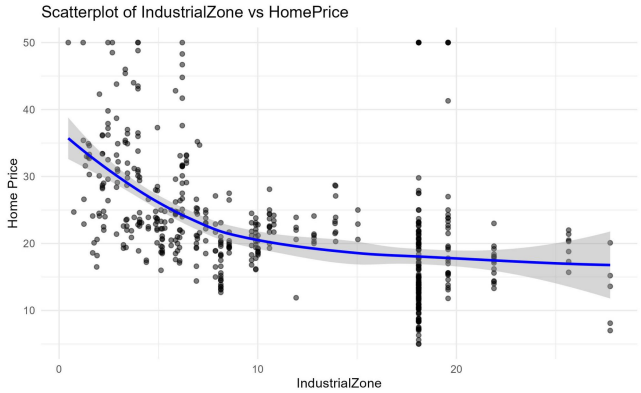


Figure 3.3: Scatter Plot of Industrial Zone against Home Price

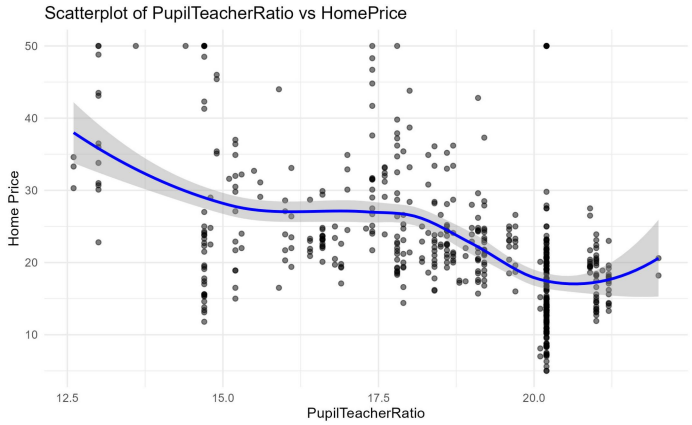


Figure 3.4: Scatter Plot of Pupil Teacher Ratio against Home Price

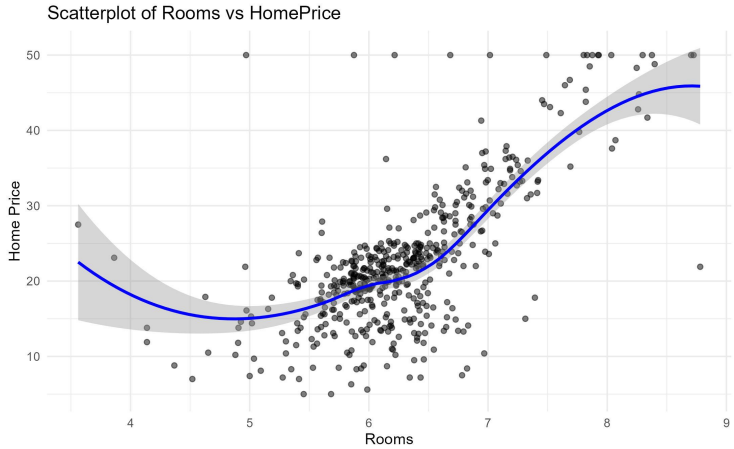
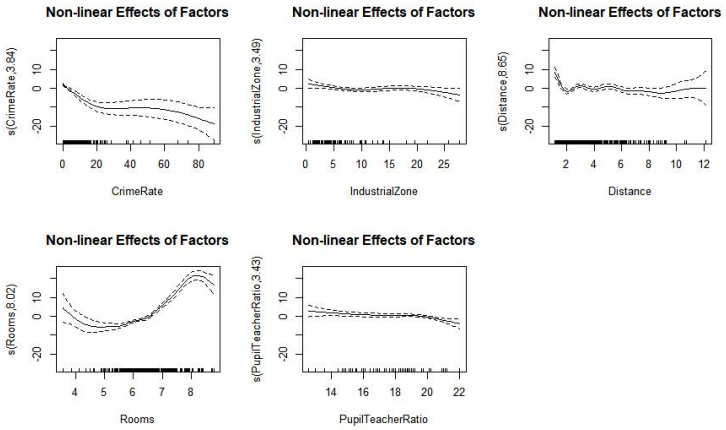


Figure 3.5: Scatter Plot of Rooms against Home Price

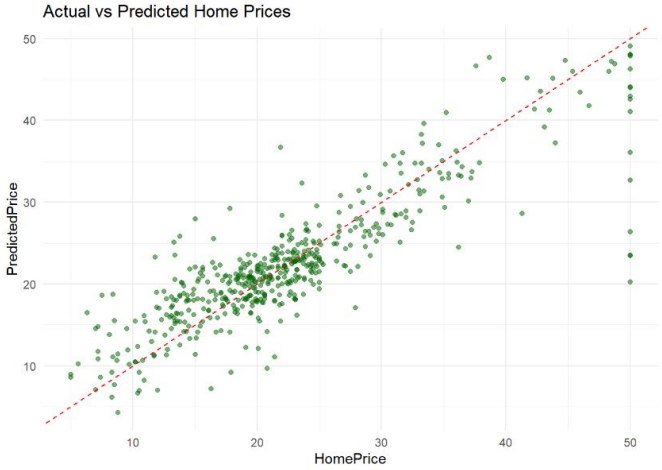
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## HomePrice ~ s(CrimeRate) + s(IndustrialZone) + s(Distance) +
##     s(Rooms) + s(PupilTeacherRatio)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.5328     0.2075   108.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df      F  p-value
## s(CrimeRate)         3.838  4.716 17.964  < 2e-16 ***
## s(IndustrialZone)    3.491  4.283  2.230 0.060337 .
## s(Distance)          8.653  8.961  5.783 2.27e-07 ***
## s(Rooms)             8.016  8.739 61.355  < 2e-16 ***
## s(PupilTeacherRatio) 3.433  4.198  4.859 0.000654 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.742   Deviance explained = 75.6%
## GCV = 23.092  Scale est. = 21.795    n = 506
```

Figure 4: GAM Model output

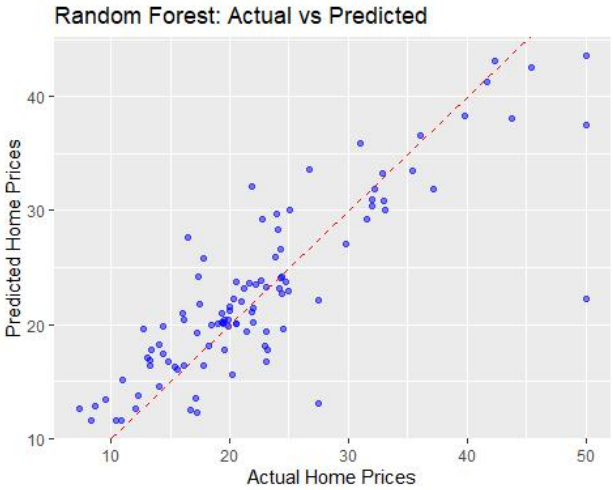
Visualized Non-linear effect of GAMs



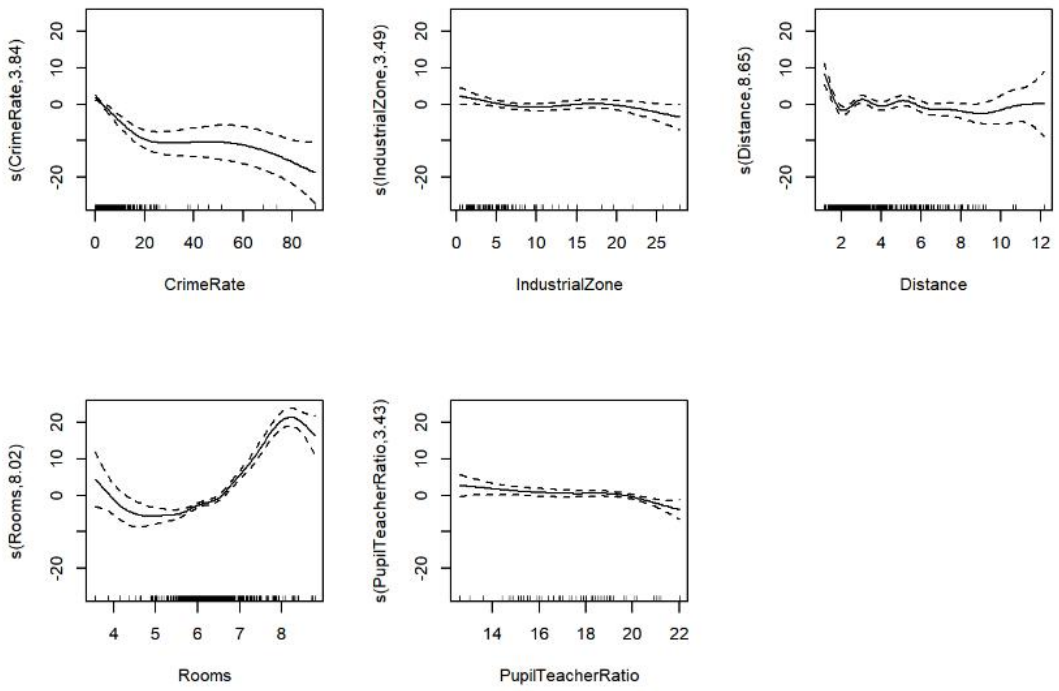
Model Evaluation



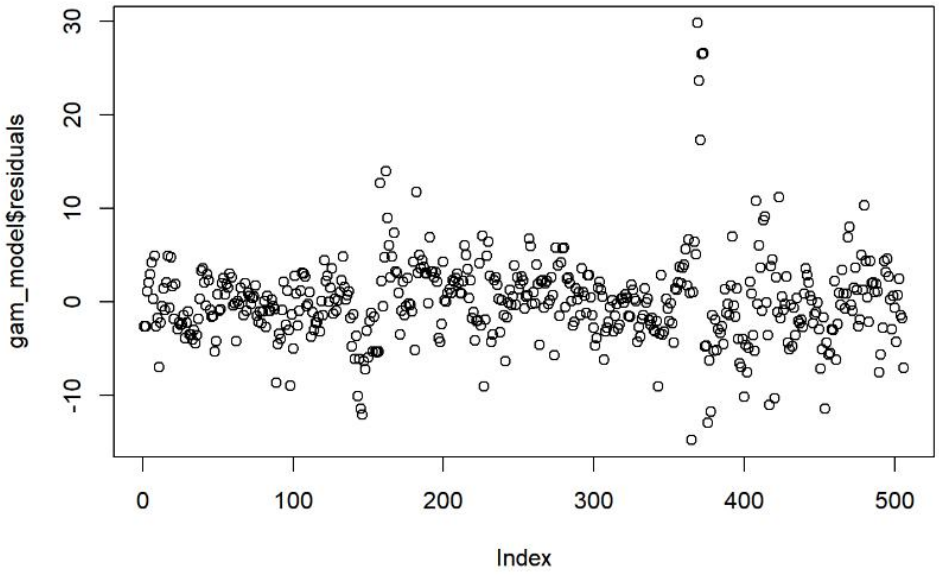
Random Forest



Plot individual smooth terms for significant factors from GAM to visualize their effects on home prices.



Check residual plots for both models to identify areas of under-/over-prediction.

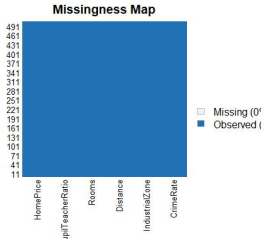


Reference

Hegel TM, Cushman SA, Evans J, Huettmann F (2010) Current state of the art for statistical modelling of species distributions. In: Spatial complexity, informatics, and wildlife conservation. Springer, Japan, pp. 273–311

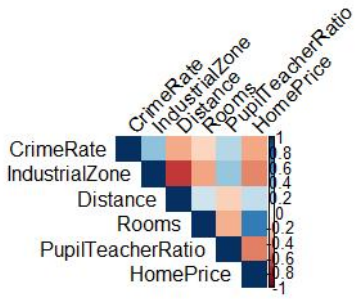
Appendix 2: Rcode and Output

```
data=read.csv("C:/Users/hp/Desktop/GSU/COMPUTATIONAL/HW6/data.csv")
data= data[, c("crim", "indus", "dis", "rm", "ptratio","medv")]
#shapiro.test(data$zn)
# Rename columns for easier interpretation
colnames(data) <- c("CrimeRate", "IndustrialZone","Distance","Rooms","PupilTeach
erRatio","HomePrice")
# Check for missing values
colSums(is.na(data))
##         CrimeRate    IndustrialZone          Distance             Rooms
##                 0                 0                 0                 0
## PupilTeacherRatio         HomePrice
##                 0                 0
#visualize the missing data using the missmap
missmap(data)
```




```
# Initial Data Exploration
summary(data)
##    CrimeRate       IndustrialZone     Distance         Rooms
##  Min.   : 0.00632   Min.   : 0.46   Min.   : 1.130   Min.   :3.561
##  1st Qu.: 0.08205   1st Qu.: 5.19   1st Qu.: 2.100   1st Qu.:5.886
##  Median : 0.25651   Median : 9.69   Median : 3.207   Median :6.208
##  Mean   : 3.61352   Mean   :11.14   Mean   : 3.795   Mean   :6.285
##  3rd Qu.: 3.67708   3rd Qu.:18.10   3rd Qu.: 5.188   3rd Qu.:6.623
##  Max.   :88.97620   Max.   :27.74   Max.   :12.127   Max.   :8.780
##  PupilTeacherRatio   HomePrice
##  Min.   :12.60     Min.   : 5.00
##  1st Qu.:17.40     1st Qu.:17.02
##  Median :19.05     Median :21.20
##  Mean   :18.46     Mean   :22.53
##  3rd Qu.:20.20     3rd Qu.:25.00
##  Max.   :22.00     Max.   :50.00
st(data)
# Identify key variables
# Assuming 'HomePrice' represents home prices and other columns represent neighbo
rhood factors
home_price <- "HomePrice" # Replace with the actual column name for home prices

neighborhood_factors <- setdiff(names(data), home_price)

#home_price <- data["HomePrice"]
```

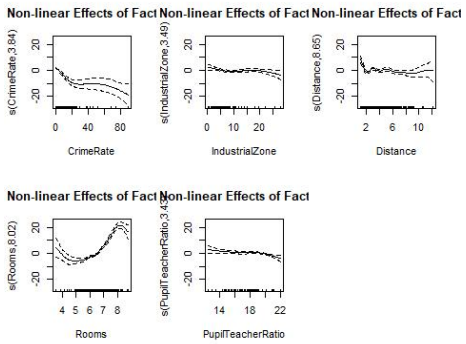
```
# Initial visualization: Scatterplots for each neighborhood factor vs. HomePrice
for (factor in neighborhood_factors) {
  ggplot(data, aes_string(x = factor, y = home_price)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "loess", col = "blue") +
    labs(title = paste("Scatterplot of", factor, "vs HomePrice"),
         x = factor,
         y = "Home Price") +
    theme_minimal() #+
    ggsave(paste0("scatter_", factor, ".jpeg"))
}
# Correlation matrix
cor_matrix <- cor(data[, sapply(data, is.numeric)], use = "complete.obs")
corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45)
```




```
# Fitting Generalized Additive Models (GAMs) for Non-linear
# Home Price is the dependent variable and 'neighborhood_factors' are predictors
gam_formula <- as.formula(
  paste("HomePrice ~", paste(paste0("s(", neighborhood_factors, ")"), collapse = " +
"))
)

# Fit the GAM model
gam_model <- gam(gam_formula, data = data)

# Summary of the model
summary(gam_model)
##
## Family: gaussian
## Link function: identity
##
## Formula:
## HomePrice ~ s(CrimeRate) + s(IndustrialZone) + s(Distance) +
##     s(Rooms) + s(PupilTeacherRatio)
##
## Parametric coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.5328     0.2075   108.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df      F  p-value
## s(CrimeRate)        3.838  4.716 17.964  < 2e-16 ***
## s(IndustrialZone)   3.491  4.283  2.230 0.060337 .
## s(Distance)         8.653  8.961  5.783 2.27e-07 ***
## s(Rooms)            8.016  8.739 61.355  < 2e-16 ***
## s(PupilTeacherRatio) 3.433  4.198  4.859 0.000654 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.742   Deviance explained = 75.6%
## GCV = 23.092  Scale est. = 21.795    n = 506
```

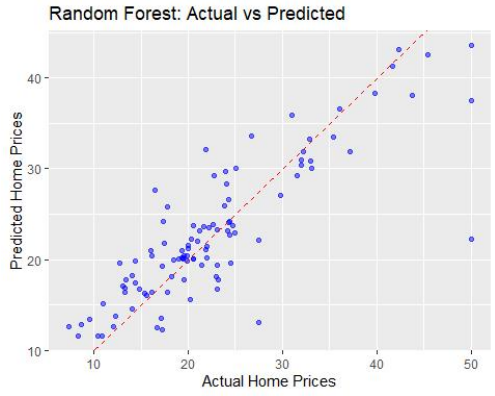
```
# Visualize Non-linear Effects
plot(gam_model, pages = 1, rug = TRUE, se = TRUE, main = "Non-linear Effects of Factors")
```




```
# Predictions for Model Evaluation
data$PredictedPrice <- predict(gam_model)

# Plot Actual vs Predicted Prices
ggplot(data, aes(x = HomePrice, y = PredictedPrice)) +
  geom_point(alpha = 0.5, color = "darkgreen") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  ggtitle("Actual vs Predicted Home Prices") +
  theme_minimal()
```


```
# Load necessary library
library(randomForest)
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:dplyr':
##
##     combine
## The following object is masked from 'package:ggplot2':
##
##     margin
# Set a seed for reproducibility
set.seed(123)

# Split the data into training and testing sets (80/20 split)
train_index <- sample(1:nrow(data), 0.8 * nrow(data))
train_data <- data[train_index, ]
test_data <- data[-train_index, ]

# Fit Random Forest model
rf_model <- randomForest(HomePrice ~ CrimeRate + IndustrialZone + Distance + Ro
oms + PupilTeacherRatio,
                         data = train_data, importance = TRUE, ntree = 500)

# Print model summary
print(rf_model)
##
## Call:
##  randomForest(formula = HomePrice ~ CrimeRate + IndustrialZone +      Distance
+ Rooms + PupilTeacherRatio, data = train_data,      importance = TRUE, ntree = 500)

##               Type of random forest: regression
##                     Number of trees: 500
## No. of variables tried at each split: 1
##
##          Mean of squared residuals: 20.63257
##                    % Var explained: 75.56
```

```
# Predictions on test data
rf_predictions <- predict(rf_model, newdata = test_data)

# Calculate RMSE
rf_rmse <- sqrt(mean((rf_predictions - test_data$HomePrice)^2))
cat("Random Forest RMSE:", rf_rmse, "\n")
## Random Forest RMSE: 4.890353
# Scatterplot of Actual vs Predicted values

test_data$RF_Predicted <- rf_predictions

ggplot(test_data, aes(x = HomePrice, y = RF_Predicted)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Random Forest: Actual vs Predicted",
       x = "Actual Home Prices",
       y = "Predicted Home Prices")
```

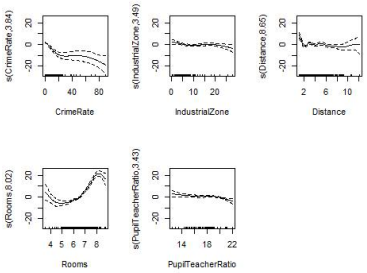



```
# Calculate RMSE
rmse_gam <- sqrt(mean((data$HomePrice - data$PredictedPrice)^2, na.rm = TRUE))
rf_rmse <- sqrt(mean((rf_predictions - test_data$HomePrice)^2))

print(paste("GAM RMSE:", rmse_gam))
## [1] "GAM RMSE: 4.53545683943156"
cat("Random Forest RMSE:", rf_rmse, "\n")
## Random Forest RMSE: 4.890353
```

Visualization:

Plot individual smooth terms for significant factors from GAM to visualize their effects on home prices.

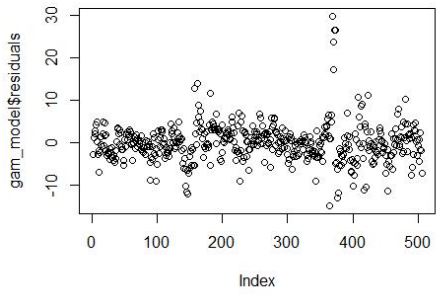
```
plot(gam_model, pages = 1)
```



Residual Diagnostics:

Check residual plots for both models to identify areas of under-/over-prediction.

```
plot(gam_model$residuals)
```



Visualizing GAM Smooth Terms and Visualizing GAMs Results

```
# Plot smooth terms for GAM model
plot(gam_model, pages = 1, shade = TRUE, seWithMean = TRUE)
```