Portfolio project

Bolormaa Mendbayar- x23176725

National College of Ireland

MSc in Data Analytics

Data Mining & Machine Learning 2023/2024

***Abstract*— This paper introduces five distinct data mining models applied to three diverse datasets, each addressing critical societal and public health concerns. The first dataset, air pollution, a significant environmental issue globally, poses health risks due to various pollutants generated by household combustion, vehicular emissions, industries, and natural causes like forest fires. This study focuses on analyzing air pollution datasets employing Multiple Linear Regression (MLR) and Random Forest Regression (RFR) methodologies to predict air quality index(AQI) and their impact on public health.**

The second dataset, encompassing San Francisco city employee's annual compensation from 2011 to 2014, employing K-Nearest Neighbors (KNN) regression, this research aimed to predict total pay based on various employee attributes and compensation details. Despite dataset constraints, the finalized regression model demonstrated substantial explanatory power, evident from the adjusted-R^2 squared measure.

The third dataset, heart disease and stroke rates, specifically focusing on a dataset spanning 2013 to 2015, features a three-year age-standardized average. Utilizing Decision Tree Classification and Logistic Regression models, this research aimed to predict gender based on demographic and health indicators. The Interactive Atlas of Heart Disease and Stroke, serving as an additional data repository, facilitated comprehensive analyses and methodological insights. Evaluation metrics like precision, recall, and F1-score underscored the model's effectiveness in categorical prediction, offering insights into factors driving distinct classifications.

I. Introduction

In conducting this analysis, we employed a systematic methodology known as Knowledge Discovery in Databases (KDD). KDD, as a comprehensive process for knowledge extraction from data, guided our approach through distinct stages encompassing data selection, preprocessing, transformation, mining, pattern evaluation, and knowledge presentation. This methodological framework enabled a structured and comprehensive exploration of the datasets, ensuring a methodical extraction of valuable insights and patterns essential for addressing the research objectives.

The increasing accessibility of a wide range of datasets offers a chance to use thorough data analysis to solve intricate societal problems. In order to gain a deeper understanding of three important areas—workforce dynamics, health demographics, and environmental sustainability—this project will investigate three different datasets.

This research is driven by the need to unravel complex patterns in air pollution, worker compensation, and health metrics. The importance of utilizing a variety of datasets and cutting-edge analytical techniques is highlighted by the urgent requirement to forecast and comprehend the subtleties of these domains. In this project we will research below the questions.

Research questions:

- How do different sources contribute to air pollution, and what is their impact on Air Quality Index (AQI) variations and subsequent public health implications?
- What factors influence employee compensation dynamics within the context of San Francisco city employees, and how accurately can total pay be predicted?
- Can demographic and health indicators effectively predict gender within the context of heart disease and stroke rates, and what insights can these predictions offer?

The objective of this research is to simplify complex data and reveal practical insights that are essential for making well-informed decisions across several domains of society.

II. Related work

1. Acito's KNIME-Based Predictive Analytics ([1]):
Acito's work serves as a valuable manual for utilizing KNIME, an analytics platform that empowers citizen data scientists. It is an essential tool for comprehending how KNIME may streamline intricate analytics procedures. With its focus on feature engineering, data preprocessing, and model selection, the book offers helpful advice that is necessary for managing a variety of datasets.

2. Hands-on Machine Learning with R by Boehmke and Greenwell ([2]):
This thorough manual provides an in-depth exploration of R-based machine learning methods. The breadth of methods and approaches it covers makes it an indispensable resource for comprehending both basic and complex machine learning ideas. The hands-on activities in the book help apply machine learning models for various applications in an efficient manner.

3. Statistical Regression Modeling with R by Chen and Chen ([3]):
This reference is essential to understanding multi-level and longitudinal modeling in complicated data structures. For projects involving time series or hierarchical data, its emphasis on comprehending temporal patterns and

interactions between variables across different levels is essential.

4. Regression Analysis by Ciaburro with R ([4]):
The investigation of statistical nodes in R by Ciaburro is important because it reveals unique linkages in large-scale datasets. It is especially useful for projects with complicated data interactions since it provides help on feature selection, model interpretation, and how to handle multicollinearity.

5. Introduction to Statistics with R by Dalgaard ([5]):
This introductory book is crucial for learning statistical topics in R. It is an excellent resource for those new to statistics and data analysis because of its concise explanations and useful examples. Its uses are widespread in fields where data exploration and hypothesis testing necessitate a rudimentary understanding of statistics.

6. Random Forests with R by Genuer and Poggi ([6]):
This resource, which focuses on ensemble learning methods (Random Forests in particular), is essential for comprehending intricate relationships found in datasets. For projects needing precise prediction models in complicated data environments, it offers insights into managing nonlinear correlations.

7. Linear Regression Models by Hoffmann: Utilizations in R ([7]):
The importance of linear regression models in R is emphasized by this resource. It clarifies assumptions, diagnostics, and model interpretation, which makes it an invaluable tool for projects involving predictive modeling and comprehending linear relationships between variables.

8. A Handbook of Statistical Analyses with R by Hothorn and Everitt ([8]):
This comprehensive guide covers a broad range of R statistical analysis. It is helpful for projects requiring a thorough understanding of statistical methods because of its chapters on hypothesis testing, ANOVA, and non-parametric procedures.

9. Regression Modeling in People Analytics, McNulty's Handbook ([9]):
Although geared on people analytics, this manual provides information on regression modeling with R and Python. It can be applied to problems like multicollinearity and those requiring regression-based forecasts.

10. Practical Machine Learning in R by Nwanganga and Chapple ([10]):
The R machine learning algorithm repertoire is expanded by this resource. It helps with projects that need for a more thorough investigation of machine learning methods than just the fundamentals.

11. Rhys's Machine Learning with R, the tidyverse, and mlr ([11]): This book blends the principles of the tidyverse and the adaptability of R programming with machine learning techniques. It is useful for machine learning applications in the real world because of its practical applications and use of the mlr package.

12. Applied Generalized Linear Models and Multilevel Models in R by Roback and Legler ([12]):
This resource explores multilevel and generalized linear models, making it appropriate for tasks needing sophisticated modeling methods to comprehend how affecting factors interact.

13. Sheather's [13] A Contemporary Method for Regression Analysis:
Sheather's book is helpful for projects that need to place a lot of attention on these areas of predictive analytics because it highlights modern regression algorithms in R that concentrate on model diagnostics and inference.

III. DATA MINING METHODOLOGY

In approaching the research questions, the adopted methodology followed the Knowledge Discovery in Databases (KDD) process, a comprehensive framework enabling systematic data analysis and knowledge extraction.

Key Stages of Methodology:

1. Data Collection and Understanding:

Datasets pertaining to Statista, Kaggle, Run my code, Central Statistics office, Data.Gov etc were sourced from reliable repositories. Initial exploratory data analysis (EDA) involved understanding the structure, variables, and general characteristics of each dataset.

2. Data Preprocessing and Transformation:

The datasets underwent preprocessing stages including cleaning, handling missing values, and addressing outliers and relationships between variables to gain insights into their independence and identify potential patterns or associations. Transformation steps involved normalization, encoding categorical variables, and feature engineering to make the data suitable for analysis.

3. Model Development and Evaluation:

Various data mining models' regression and classification were applied to different datasets based on their suitability for the research objectives. Model development encompassed until getting the higher accuracy and lower error.

Evaluation metrics such as best R-squared, and lowest standard error, Mean squared error, Root squared error, Mean absolute error, Root mean absolute error for regression model, higher accuracy and lowest metrics such as Precision, Recall, F1 score for classification model were employed to assess model performance, ensuring reliability and validity of the results.

Preliminary Aspects of Methodology:

Data Preparation:

Each dataset required specific preparatory steps due to variations in formats, missing values, and diverse data types. Imputation techniques were applied for missing values, and scaling methods were employed for numeric features.

Feature Selection and Engineering:

Feature selection was conducted based on correlation analysis, and importance scores. Additionally, new features were engineered to enhance model performance and capture latent patterns. This methodology framework aimed to ensure the reliability and robustness of the analysis while extracting meaningful insights from the datasets pertaining to the research questions.

IV. Implementation

The methodology employed in this analysis embodies a systematic and comprehensive approach aimed at exploring the determinants of AQI utilizing multiple linear regression and random forest regression, Total pay employing KNN regression, and predicting gender making use of logistic regression and decision tree classification. The methodology unfolds through distinct phases, meticulously structured to facilitate a robust and insightful analysis. As below shown table 1 has dataset details.

Dataset	Source	Information
1. Global air pollution	Kaggle	23463 rows 12 columns
2. SF Salaries	Kaggle	148654 rows 13 columns
3. Heart disease	Data.gov	59076 rows 19 columns

Table 1: Dataset details

A. Dataset 1

https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

4.A.1 Data Understanding

This dataset contains AQI(Air Quality Index) values of different pollutants for many cities all over the world.

The initial phase of this analysis involved a comprehensive exploration and understanding of the datasets to gain insights into its structure, variables, and characteristics. Upon loading the dataset into the R environment, an initial inspection was conducted to understand its structure and contents.

The variables in the first dataset are:

Dataset 1		
No	**Variable**	**Type**
1	Country	factor
2	City	factor
3	AQI.Value	integer
4	AQI.Category	factor
5	CO.AQI.Value	integer
6	CO.AQI.Category	factor
7	Ozone.AQI.Value	integer
8	Ozone.AQI.Category	factor
9	NO2.AQI.Value	integer
10	NO2.AQI.Category	factor
11	PM2.5.AQI.Value	integer
12	PM2.5.AQI.Category	factor

Table 2: Variables of Dataset 1

To understand the data the following descriptive statistics were calculated.

Variable	Mean	Std Dev	Median	Skew	Kurtosis
3	72.01	56.06	55	3.29	17.56
5	1.37	1.83	1	23.08	1288.96
7	35.19	28.1	31	3.1	12.12
9	3.06	5.25	1	3.79	22.43
11	68.52	54.8	54	2.82	13.62

Table 3: Descriptive statistics

Several visualizations were created to illustrate the relationships and distributions within the all dataset, as an example, from dataset 1 variables 3 and 7 appear normally distributed based on skew and kurtosis statistics, graphics are shown in Figure 1.

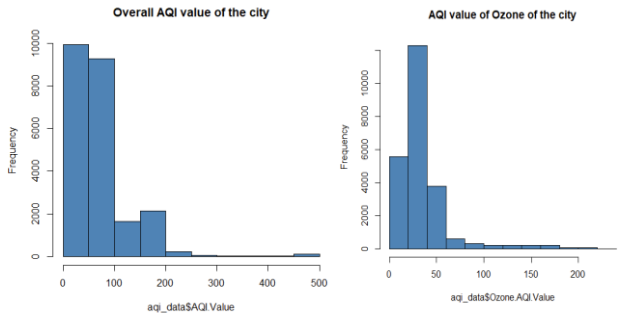


Figure 1. Distribution of the sample variables

Variable 5 is highly right skewed, see figure 2, and this may be an issue in the modelling phase.

We can see that from Figure 4, PM2.5.AQI.Value is highly correlated with target variable which is AQI.Value.

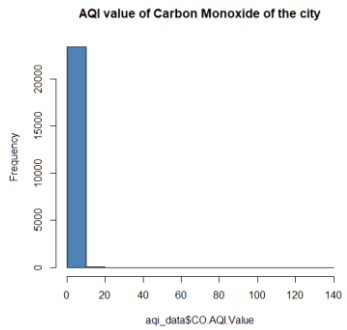


Figure 2. Distribution of variable 5.

4.A.2 Data Preprocessing

The first step in data cleaning was to identify missing values. We found that no missing values in the dataset. In this part didn't do hot encoding, since the categorical variables such as 4, 6, 8, 10, and 12 are category of the 3, 5, 7, 9 and 11 variables. The second step was analyzing the outliers, as a result #12472 rows remained. As an example, variable 7 before and after removing outliers are shown in figure 3.

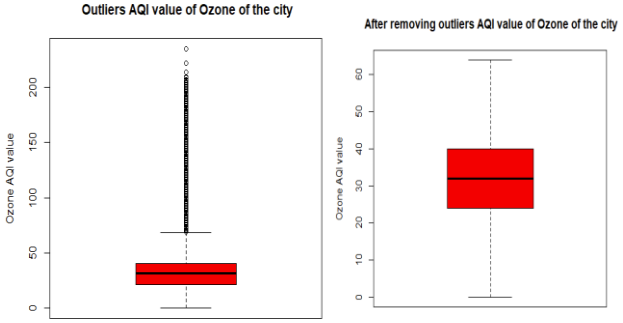


Figure 3. Comparison of the before and after outliers

To enhance the model's accuracy, a deliberate step was taken before constructing the correlation matrix shown in Figure 4, dataset 1 comprised #12472 records and 4 numerical variables such as 3, 7, 9 and 11 in shown Table 2. Note, variable 5 or 'CO.AQI.Value' was manually dropped from consideration due to its lack of range, thereby refining the dataset. Additionally, other irrelevant variables were meticulously identified and removed, ensuring a more focused and relevant set of predictors for the model's construction.

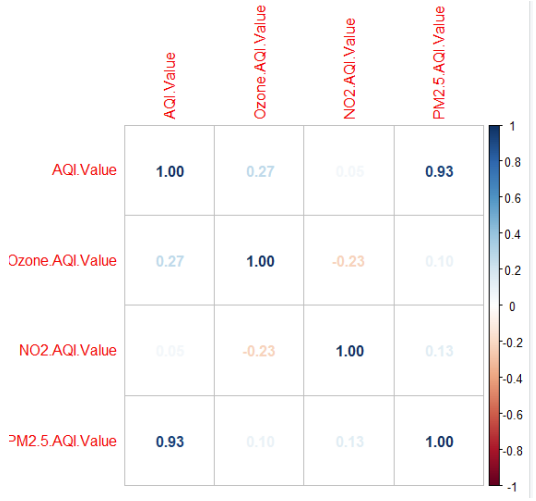


Figure 4: Correlation matrix – excluding no range variables

4.A.3 .Data Transformation

Inspection of the data showed that the following needs to be done:

- Transformation of independent variables is required to normalize.

Min-Max Normalization, also known as Min-Max Scaling, is a method used to rescale numerical data within a predetermined range, typically between 0 and 1. This technique is applied to standardize various features or attributes of a dataset, ensuring that all values lie within the specified range. Following Table 4 shows normalized features.

Dataset 1			
	Variable 7	**Variable 9**	**Variable 11**
Min.	0.0000	0.0000	0.0000
Median	0.5000	0.1429	0.4592
Mean	0.4968	0.2266	0.4508
Max.	1.0000	1.0000	1.0000

Table 4: Normalization of the features

4.A.4..Data Modellimg

4.A.4.1 Multiple linear regression model 1

The first model built to predict AQI.Value ,used the independent normalized variables, as shown in table 5. The output is as follows:

Linear regression model 1		
Test	**Result**	**Explanation**
Adjusted R-Squared	0.898	Passed: Higher than 0.75
Std Error	4.918	Not passed: Range 0 to 1
ncvTest	Chisquare= 1327.127, p= 2.22e-16	Not passed: Chisquare: /Range 0 to 1/ p: /Range 0 to 1/
Vif_model	Variable7: 1.072120 Variable9: 1.078618 Variable11: 1.035209	Passed: About 2 is best. Greater than 5 will be a problem.
Dubrin Watson Test	D-WStatistic: 2.002122 p-value: 0.94	Passed: A Durbin-Watson statistic close to 2 indicates no significant autocorrelat-

		ion. p-value is extremely small, indicating strong evidence against the null hypothesis.
Cook's distance	Min: 0.000 Max: 1.098e-02	Passed: Values should be <1 and close to zero. Above 1 is a potential problem.

Table 5: **Model 1 output**

Overall Model 1 failed the standard error and nonparametric covariance(ncv) test, and as shown *appendix A*/see appendix section/, dots in normality of residuals are not falling along the line it is deemed unsuccessful so a new model was prepared.

4.A.4.2 Multiple linear regression model 2

To address the homoscedasticity issue which is likely due to the skewness of the original variable and the logN of 'AQI.Value' was substituted as the dependent variable. All the independent variables as per model 1 were used. The output of model 2 is shown in *appendix B* and Table 6.

Model 2 has performed better than model 1. There was an approximately 0.03 decrease in the adjusted R^2 and the standard error of the estimate in model 2 has reduced from 4.918 to 0.119. In *appendix B* the normality of residuals and linearity were flat and horizontal than model 1. Homoscedasticity assumption appears to be met as the errors look to be randomly distributed. The curve in the Q-Q plot for the residuals is close to the cumulative plot line though there is some curving present.

Linear regression model 2		
Test	**Value**	**Explanation**
Adjusted R-Squared	0.8612	Passed:
Std Error	0.119	Passed:
ncvTest	Chisquare= 4852.689, p= 2.22e-16	Not passed: Increased by roughly 3525 chisqure value, p value remained same.
Vif_model	Values were same as model 1.	Passed:
Dubrin Watson Test	D-W Statistic: 2.012291 p-value: 0.49	Passed:
Cook's distance	Min: 0.000e+00 Max: 2.464e-02	Passed:

Table 6. **Model 2 output**

4.A.4.3 Multiple linear regression model 3

To tackle the problem of homoscedasticity likely arising from the skewed nature of the initial variable, we opted to replace the dependent variable 'AQI.Value' with its square root, aiming to mitigate the issue. All the features as per model 1 were used.

Linear regression model 3		
Test	**Value**	**Explanation**
Adjusted R-Squared	0.8915	Passed: Increased by about 0.03 than model 2.
Std Error	0.3585	Passed: Increased by roughly 0,24 than model 2.
ncvTest	Chisquare= 3290.072, p= 2.22e-16	Not passed: Reduced by approximately 1562 chisqure value, model 2.
Vif_model	Values were same as model 1.	Passed:
Dubrin Watson Test	D-W Statistic: 2.010097 p-value: 0.57	Passed:
Cook's distance	Min: 0.000e+00 Max: 9.871e-03	Passed:

Table 7. Model 3 output

Model 3 performed better than model 2 and has not passed the only ncvTest diagnostic test that got 3290.072 chisquare value. However, the adjusted R^2 and standard error in model 3 was higher than model 2.

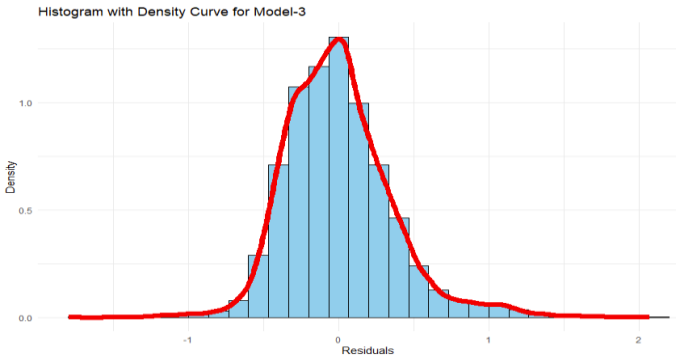


Figure 5. Distribution of model 3

We can see that as Figure 5, residuals follow a normal distribution.

4.A.4.4 Model 4

In the process of refining the model, we employed a stepwise backward elimination technique to enhance its performance. The selected model utilizes the square root of the 'AQI.Value' as the dependent variable, integrating in table 2, variables 7, 9 and 11 or 'Ozone.AQI.Value,' 'NO2.AQI.Value,' and 'PM2.5.AQI.Value' as predictor variables.

The summary of the final model demonstrates a robust fit. It showcases statistically significant coefficients for 'Ozone.AQI.Value' and 'PM2.5.AQI.Value,' indicating their substantial impact on the square root of the AQI. Notably, 'NO2.AQI.Value' also holds significance despite a smaller effect size. The model exhibits a high degree of explanatory power (Multiple R-squared: 0.8915), affirming its ability to explain the variability in the transformed AQI values.

4.A.5. Evaluation

The purpose of the evaluation is to assess the predictive capability of the multiple linear regression model using the designated test dataset. The dependent variable 'AQI.Value' was substituted with its square root ('sqrtN') in the test dataset, mirroring the transformation applied in the training dataset. This evaluation aimed to quantify the model's predictive accuracy by employing metrics such as mean absolute error(MAE), R-squared, and root mean squared error(RMSE).

Multiple linear regression	
Metrics	**Value**
RMSE	4.919524
R-squared	0.8979597
MAE	3.789789

Table8. Evaluation metrics

RMSE measures the average magnitude of the residuals, as Table 8, the model's predictions deviate from the actual values. R-squared(test) is higher than train(model 3), that means R-squared value of approximately 0.898 indicates that around 89.8% of the variance in the dependent is explained by the independent variables, which is better value. MAE represents the average absolute difference between predicted and actual values.

The best model produced that met the validation criteria has the following formula.

sqrt(AQI.Value) = 4.049 + 1.17 × Ozone.AQI.Value - 0.05 × NO2.AQI.Value + 5.16 × PM2.5.AQI.Value

4.A.6 Random Forest regression model

The second model utilized the initial dataset consisting of independent normalized variables, as presented in Table 4. In Table 8, the first default random forest (RF) regression model showcased a notably higher R-squared value compared to the multiple linear regression model. Moreover, metrics such as RMSE and MAE were marginally lower than those of the multiple regression model. Subsequently, an exploration was conducted to observe the impact of varying the 'k' value on the modeling results.

The second RF model employed a higher number of trees, specifically setting 'k' at 600. Consequently, this yielded a higher R-squared value than the initial RF model and resulted in decreased evaluation metrics. Further analysis revealed the 'k' value that produced the lowest MSE, which was determined to be 'k=146'.

In the final modeling phase, this identified 'k' value of 146 was utilized, leading to improved R-squared and overall metrics compared to the previous two models.

The output is as follows:

Random forest regression			
	Model 1	**Model 2**	**Model 3**
k tree numbers	500	600	146
R-squared	0.9704	0.9714	0.9729
mse	7.04	6.744	6.149
mae	1.612	1.581	1.556
rmse	2.653	2.597	2.480

Table9. Evaluation metrics

Created two types of plots commonly used in regression analysis to assess model's performance and distribution of residuals. Below illustrated in Figure 6, we can see that first plot shows the difference between the actual and predicted values, and the model is making predictions that are close to the actual values in most cases.

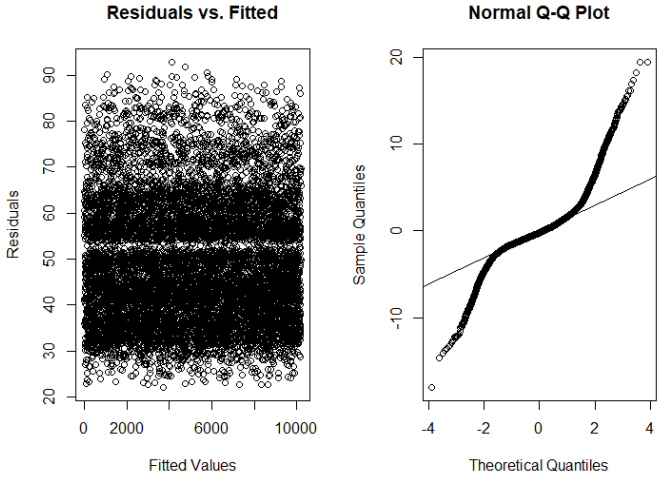


Figure 6. Residuals and QQ plot of *rf* model 3

In Figure 7, scatter plot compares the actual AQI values against the predicted AQI values obtained from the 3rd rf model. We can say that the red line closely aligns with the points.

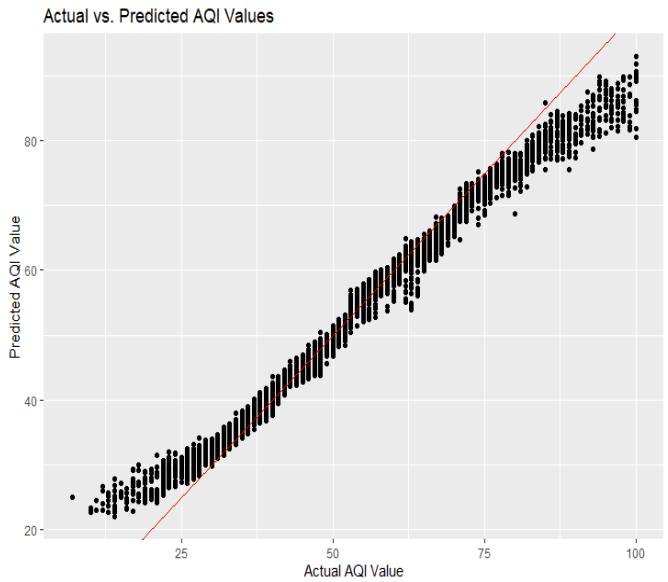


Figure 7 Scatter plot of *rf* model 3

Lastly, as we see Random Forest regression model has better performance than multiple linear regression model. Multiple linear regression has achieved 0.8915 R-squared, on the other hand Random Forest regression model's is 0.9729.

B. Dataset 2

https://www.kaggle.com/datasets/kaggle/sf-salaries

4.B.1 Data understanding and Preprocessing

This dataset examines the employment and compensation details of San Francisco city employees annually from 2011 to 2014. The variables included in the second dataset are:

The variables in the 2nd dataset are:

Dataset 2		
No	**Variable**	**Type**
1	Id	integer
2	EmployeeName	factor
3	JobTitle	factor
4	BasePay	factor
5	OvertimePay	factor
6	OtherPay	factor
7	Benefits	factor
8	TotalPay	numeric
9	TotalPayBenefits	numeric
10	Year	integer
11	Notes	logical
12	Agency	factor
13	Status	factor

Table 10: Variables of Dataset 2

Dataset 2 initially contained a large volume of records, totaling 148,654 entries. To facilitate modeling, I randomly sampled 10% of the data, resulting in a final dataset comprising 14,865 rows.

In the Data Understanding and Preprocessing stages, the same approach was adopted as in Dataset 1. During the Data Cleaning step, unnecessary variables such as 'notes,' 'agency,' and 'status' were dropped initially. Subsequently, checks were conducted for NA, NULL, and JUNK values. Outliers were identified and removed, identified outliers resulting in a dataset containing 14,752 rows and 10 columns.

A correlation matrix was then generated to assess the relationships between independent and dependent variables, and resulting Id and Year, Total pay and Total Pay benefits were correlated to each other. So, then dropped again some unnecessary variables such as Year, Id, Total pay benefits, Employee name and Job title to prepare modelling. Following this, the clean dataset underwent transformation using min-max normalization.

4.B.2 KNN regression model

In this modeling, we aimed to predict Total Pay based on variables such as Base Pay, Overtime Pay, Other Pay, and Benefits.

The initial step involved conducting an ANOVA test for these features, revealing that all of them are statistically significant.

The first KNN regression model was built using default k value 9. Among these, the optimal k value turned out to be 9. As illustrated in Table 11, the initial model achieved an R-squared value of about 0.61, which is moderately good. However, the other metrics exhibited a substantial increase.

In an attempt to reduce these metrics, I investigated the best k-value within the range of 1 to 20 for the modeling process. The best-performing k-value identified was 7. In the subsequent KNN modeling with this optimized k-value, there was a slight improvement in the R-squared value, while the metrics showed a slight decrease.

KNN regression model		
	Model 1	**Model 2**
k numbers	9	7
R-squared	0.615	0.648
RMSE	30306.86	29401.84
MAE	17387.15	16748.89
MSE	871236918	864468153

Table 11: Evaluation metrics

As below shown, conducted a thorough evaluation of our best-performing model using residual analysis. Figure 8 showcases the relationship between predicted and actual values. Prediction of red line align with most of the points diagonally.

.

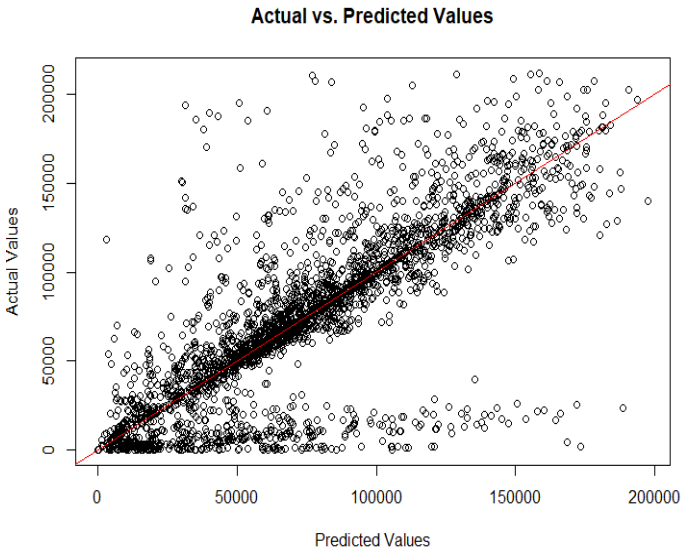


Figure 8. Actual VS Predicted Values on the best model

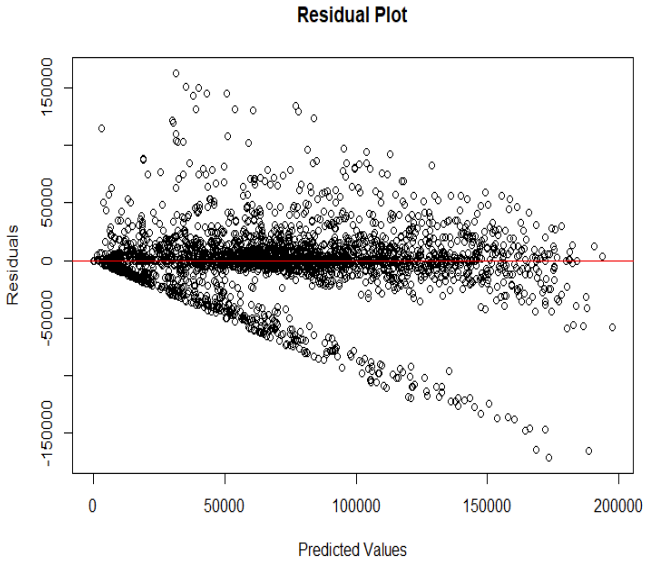


Figure 9. Residual plot on the best model

The Figure 9 displays the distribution of residuals against predicted values. A horizontal red line captures variability across the range of predictions.

C. Dataset 3

https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county?fbclid=IwAR1GMrbyJaxzeq8nIFd5m-nA3cuhdnF0mFhQu0sM-rzLbHBDjUHH0PZ-5W4

4.C.1 Data understanding and Preprocessing

This data is Heart Disease Mortality among US adults(+35) by state/territory and country from 2013 to 2015.

The variables in the 3rd dataset are:

Dataset 3		
No	**Variable**	**Type**
1	Year	integer
2	LocationAbbr	factor
3	LocationDesc	factor
4	GeographicLevel	factor
5	DataSource	factor
6	Class	factor
7	Topic	factor
8	Data_Value	numeric
9	Data_Value_Unit	factor
10	Data_Value_Type	factor
11	Data_Value_Footnote_Symbol	factor
12	Data_Value_Footnote	factor
13	StratificationCategory1	factor
14	Stratification1	factor
15	StratificationCategory2	factor
16	Stratification2	factor
17	TopicID	factor
18	LocationID	integer
19	Location.1	factor

Table 12: Variables of Dataset 3

In the Data Understanding and Preprocessing stages, the same approach was adopted as in Dataset 1. The first step of cleaning was excluded on of the level of Stratification1 which is Overall. Data is cleaned to classify gender of people who dead due to heart disease.

Subsequently, checks were conducted for NA, NULL, and JUNK values, and found 46619 NA values, after this dropped them, final data got 20633 records. During the Data Cleaning step, unnecessary variables such as "Year","DataSource","LocationAbbr","LocationDesc","StratificationCategory2","TopicID","Data_Value_Footnote_Symbol","Data_Value_Footnote","Data_Value_Unit","GeographicLevel","StratificationCategory1" identified and removed, resulting in a dataset containing 8 variables.

Identified factor variables which need to be numeric, used hot encoding to change data type to numeric, resulting in 13 variables. And then outliers were identified and removed, resulting in a dataset containing 12638 rows. After this checked the range for all variables, if there is no range dropped the variables, resulting a clean dataset containing 5 variables.

A correlation matrix was then generated to assess the relationships between independent and dependent variables, variables such as Stratification2Overall, Stratification2White are highly correlated to each other. Manually dropped Stratification2Overall from the cleaned dataset. Following this, the clean dataset underwent transformation using min-max normalization.

4.C.2 Decision tree classification

In this modeling, we aimed to predict Stratification1 which is gender based on variables such as Data_Value, LocationID, and Stratification2White. As shown in Table 13, modelling accuracy is 0.8282 which is quite good performance. The other metrics such as precision refer to the accuracy of positive predictions made by model, approximately 78.92% were actually correct. The Recall measures model's the model's ability to correctly identify positive instances from all actual positive instances, 89.44% of all positive instances.

Decision tree classification	
Accuracy	0.8282
Precision	0.7892
Recall	0.8944
F1 score	0.8385

Table 13: Evaluation metrics

In Figure 10, it generated a visual diagram of decision tree structure. This figure shows how the target variable is predicted by the model by dividing the data according to the predictor variables.

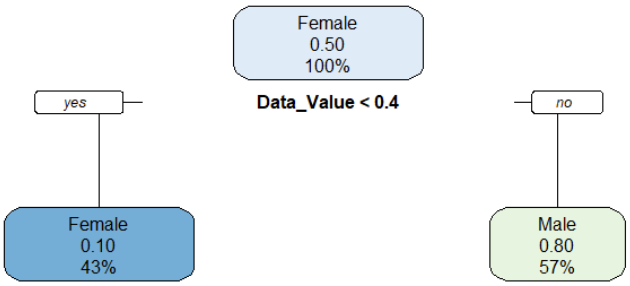


Figure 10. Decision tree model

Finally cross validation which to ensure the robustness generalizability of our predictive model, I employed a 10-fold cross-validation technique during the model training phase. Finding the ideal cp number to optimize model performance while avoiding overfitting or underfitting is the aim. It appears that in this instance, a better balance between accuracy and model complexity can be found in the middle range of cp values.

cp	Accuracy	Kappa
0.005277778 0.8673737 0.7347617	0.005277778 0.8673737 0.7347617	0.005277778 0.8673737 0.7347617

Table 14: Performance of decision tree model

4.C.3 Logistic regression

The Logistic regression model utilized the Dataset 3 consisting of independent normalized variables, as mentioned in *4.C.2.*

Logistic regression	
Accuracy	0.8385
Precision	0.8486
Recall	0.8230
F1 score	0.8356

Table 15: Evaluation metrics

In Table 15, roughly, 83.85% of the model's predictions across all classes were correct. The other metrics are more than 0.8 which means with higher values indicating better performance.

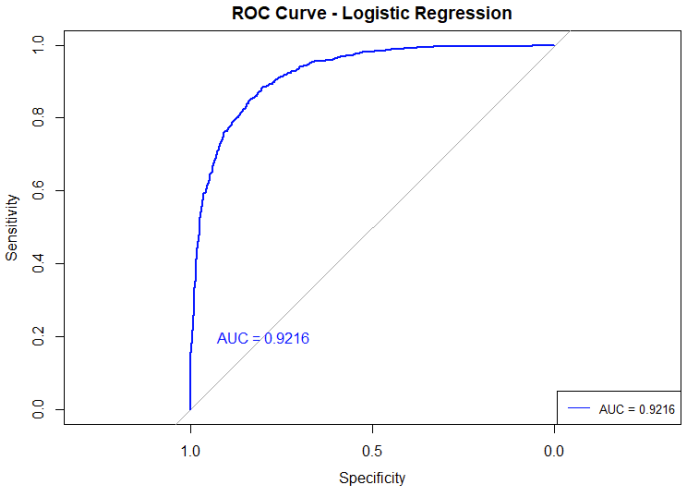


Figure 11. ROC curve

Figure 11 displays the corresponding Area Under the Curve (AUC), which is an essential metric for evaluating the performance of a binary classification model, such as the logistic regression model. As in the plot AUC value is about 0.92, in this situation indicates that the model has good discriminatory power to separate the two classes.

Lastly cross validation and resulting 0.8381 accuracy and 0.676 kappa values which has almost same performance as Table 12.

Comparison between Decision tree classification and Logistic regression model is Logistic regression has better performance than Decision tree. Both achieved higher than 80% accuracy, but Logistic regression is consistent with the metrics and analysis, emphasizing the advantages of the Logistic Regression model over the Decision Tree model in terms of accurate target variable prediction.

V. Conclusion

In conclusion, all 3 datasets were preprocessed and used for different 5 models.

The first multiple Linear regression model and second Random Forest regression models were used in the same first dataset and predicted overall AQI.Value. Random forest regression model performance was better than multiple linear regression models, because in order to decrease overfitting

and increase generalization, Random Forest constructs numerous decision trees and aggregates their predictions. The model coefficients exhibited significance for PM.2.5 predictor, which is Particulate Matter, also known as atmospheric aerosol particles among others. This coefficient indicates significant influence on air quality index.

The third KNN regression model used the second dataset and predicted TotalPay based on some payments and benefits. In order to achieve high R squared, built different models on unlike k numbers. Best k number was 7.

The fourth Decision tree classification and fifth Logistic regression models were used same third dataset and predicted gender of people who died due to heart disease. As a result, Logistic regression had greater performance than Decision tree classification. However, it's vital to note that the performance comparison between these models depends on the nature of the data, the complexity of relationships within the data, and specific problem being addressed.

Future work:

Feature Engineering and Selection: Investigate different feature transformations or create a new feature that could be a capture subtle correlation between different datasets.

Additional Data Sources: Investigate additional data sources that might offer more thorough insights or boost the model's capacity for prediction.

Other evaluation techniques might be needed and that all graphs serve only diagnostic tools, suggest further analysis or validation techniques to confirm the model's reliability.

Different models: To estimate the most suitable model for each dataset and compare results.

VI. REFERENCES

[1] [1] F. Acito, Predictive Analytics with KNIME: Analytics for Citizen Data Scientists. Springer Nature Switzerland, 2024.

[2] [2] B. Boehmke and B. M. Greenwell, Hands-On Machine Learning with R (Chapman & Hall/CRC The R Series). CRC Press, 2019.

[3] [3] D. G. Chen and J. K. Chen, Statistical Regression Modeling with R: Longitudinal and Multi-level Modeling (Emerging Topics in Statistics and Biostatistics). Springer International Publishing, 2021.

[4] [4] G. Ciaburro, Regression Analysis with R: Design and develop statistical nodes to identify unique relationships within data at scale. Packt Publishing, 2018.

[5] [5] P. Dalgaard, Introductory Statistics with R (Statistics and Computing). Springer New York, 2008.

[6] [6] R. Genuer and J. M. Poggi, Random Forests with R (Use R!). Springer International Publishing, 2020.

[7] [7] J. P. Hoffmann, Linear Regression Models: Applications in R (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences). CRC Press, 2021.

[8] [8] T. Hothorn and B. S. Everitt, A Handbook of Statistical Analyses using R, Third Edition (A Chapman et Hall book). Taylor & Francis, 2014.

[9] [9] K. McNulty, Handbook of Regression Modeling in People Analytics: With Examples in R and Python. CRC Press, 2021.

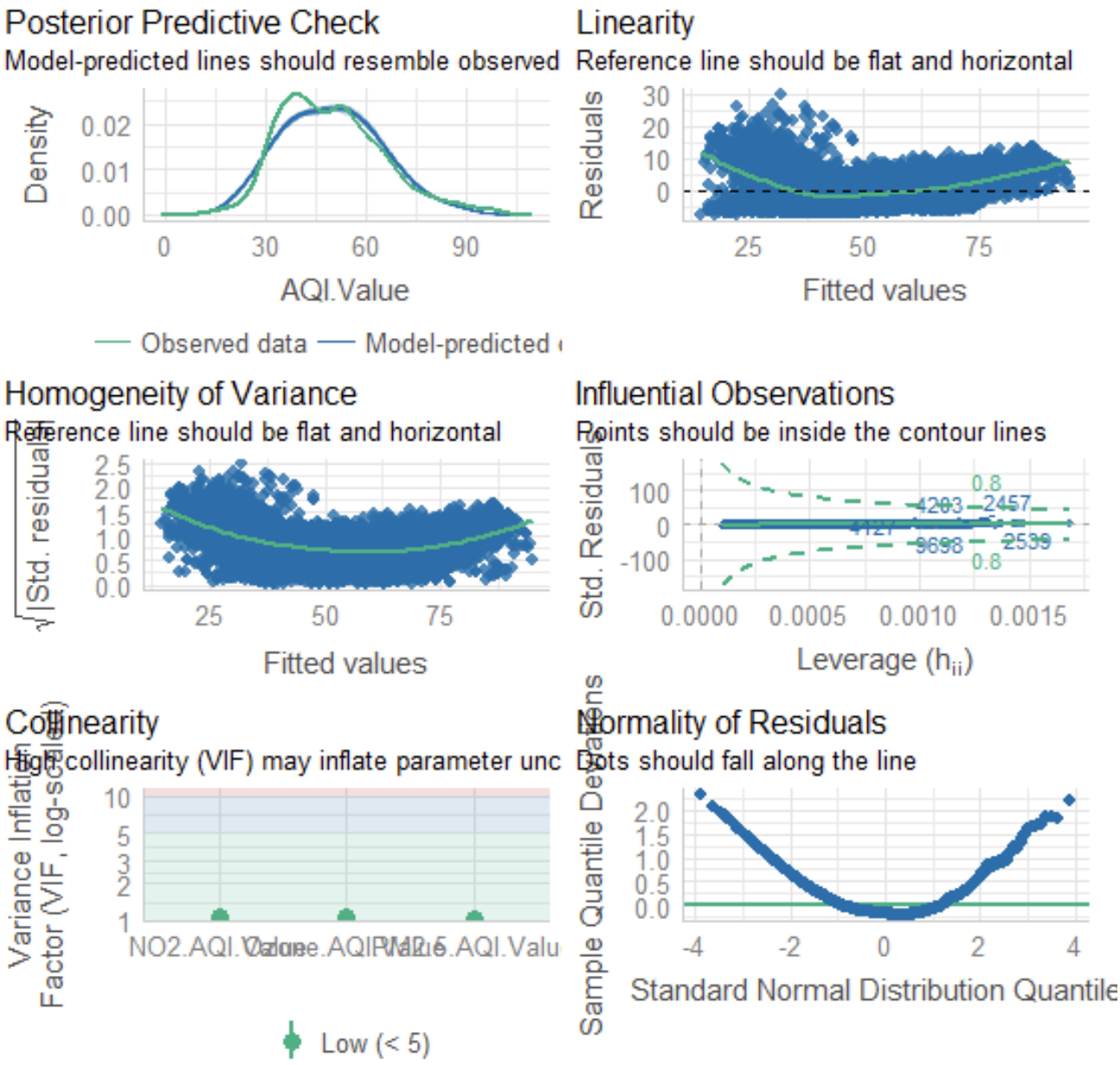
[10] [10] F. Nwanganga and M. Chapple, Practical Machine Learning in R. Wiley, 2020.

[11] [11] H. Rhys, Machine Learning with R, the tidyverse, and mlr. Manning, 2020.

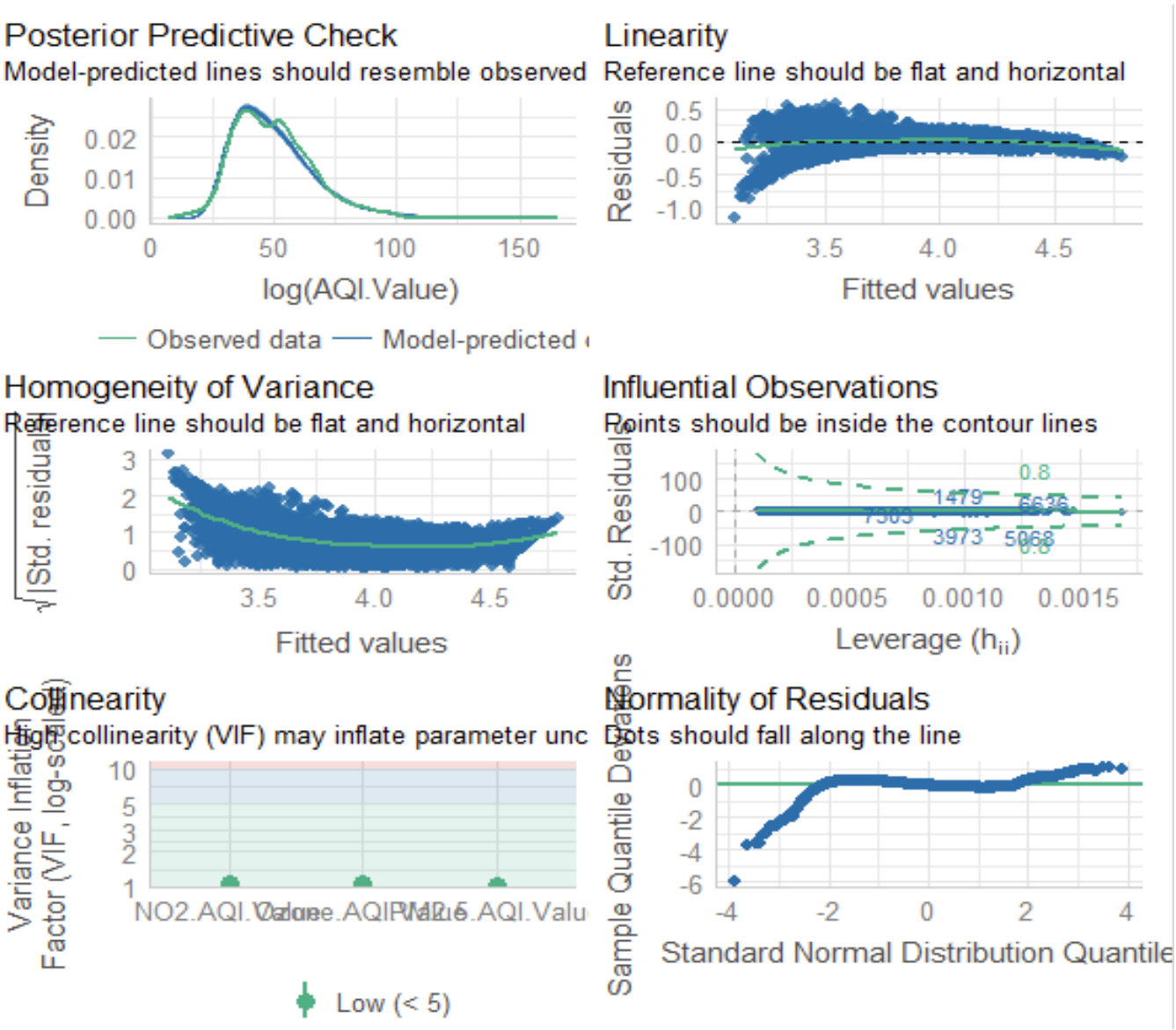
[12] [12] P. Roback and J. Legler, Beyond Multiple Linear Regression: Applied Generalized Linear Models And Multilevel Models in R (Chapman & Hall/CRC Texts in Statistical Science). CRC Press, 2021.

[13] S. Sheather, A Modern Approach to Regression with R (Springer Texts in Statistics). Springer New York, 2009.

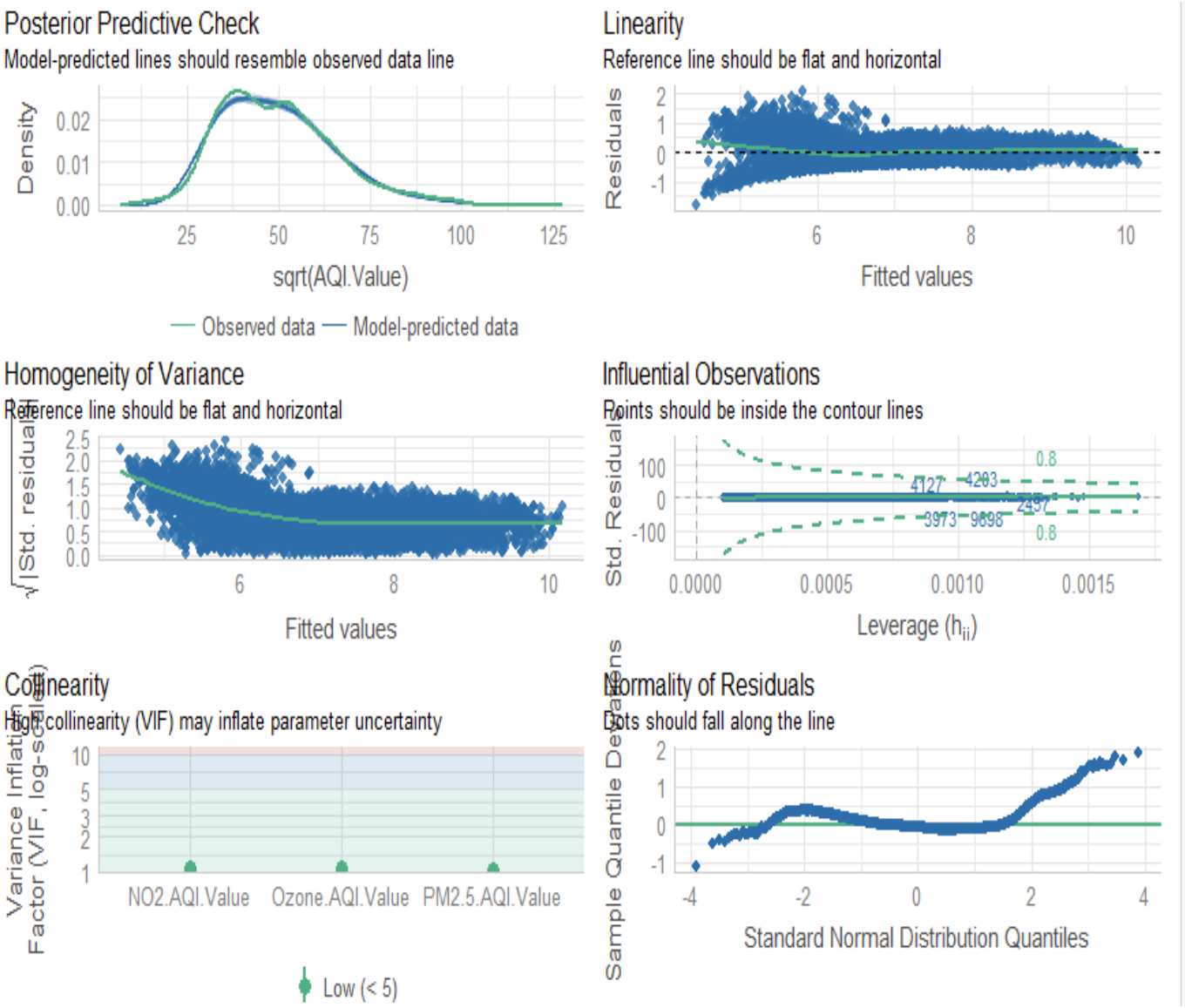
Appendix



A. ***Graph of model 1***



B. : Graph of model 2



C. Graph of model 3