UNT UNIVERSITY OF NORTH TEXAS

“Analyzing the Impact of Urban Air Quality on Public Health: A Data Mining and Machine Learning Approach”

GROUP 09

TEAM MEMBERS

1. ACHYUTHA RAVULA

2.LIKHITHA GANDI

3.POOJA SRI KANAKAMEDALA

4. RAVI TEJA PATTESAM

5. VEENA MADHURI PILLI

Executive Summary

- Analyzed the impact of environmental factors on Health Risk Score using machine learning techniques to support public health initiatives.

Models Implemented

- ➢ **Linear Regression:** Provided a baseline understanding of the relationship between environmental factors and health risk.
- ➢ **Decision Tree:** Captured non-linear interactions, with Model 2 preferred for its complex pattern identification.
- ➢ **Random Forest:** Enhanced prediction accuracy by averaging multiple trees, reducing overfitting.
- ➢ **Auto Neural Network:** Explored deep non-linear relationships but had higher error in comparison to other models.

- Heat Index and Humidity emerged as primary drivers of Health Risk Score, highlighting the health impact of high temperatures and air moisture.
- **Best Model:** The Linear Regression model with stepwise selection achieved the lowest average squared error (0.013), offering the most accurate and interpretable results.

Findings underline the importance of monitoring heat and humidity for health risk management in urban planning and healthcare policies.

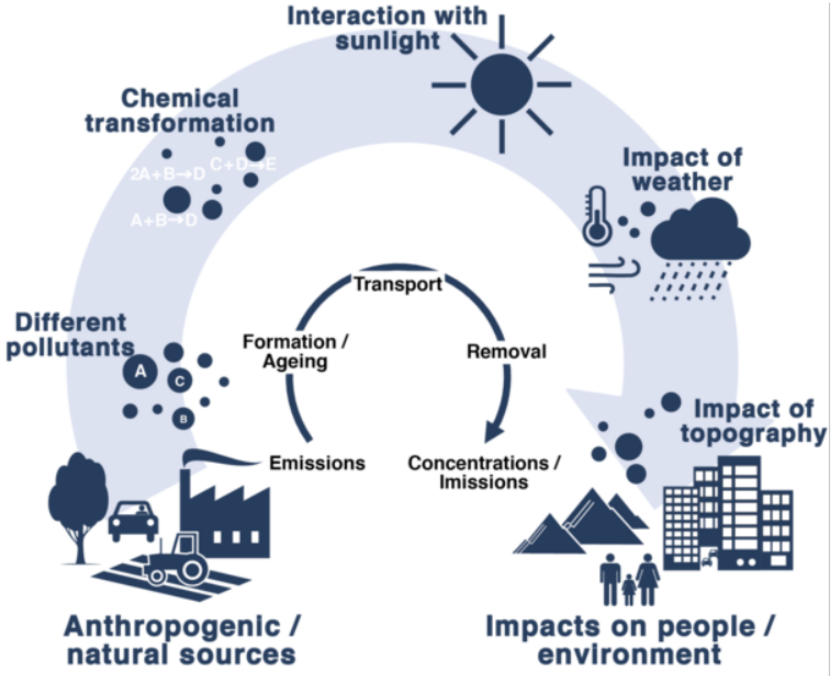
Project Motivation/Background

Why We Chose This Project

- ➢ Increasing awareness and concern about how urban air quality impacts health, especially in dense, industrialized cities.
- ➢ This project allows us to leverage data mining and machine learning to uncover patterns and trends in air quality data and correlate them with health impacts.
- ➢ Results can inform policy-making, health advisories, and urban planning for healthier cities.

Importance of the Project

- ➢ With growing urban populations, the importance of monitoring and managing air pollution is crucial for sustainable living.
- ➢ Poor air quality is linked to several health issues, including respiratory diseases, cardiovascular problems, and even premature mortality.
- ➢ By building predictive models, we aim to proactively address health risks associated with urban air pollution.



Dataset Overview

- The chosen dataset, **Urban Air Quality and Health Impact Analysis**, contains variables that capture various indicators of air quality—such as pollutant levels (e.g., PM2.5, NO2, O3), weather conditions, and demographic or socioeconomic factors—and links these to health outcomes in urban populations.
- This dataset allows us to explore the direct and indirect effects of air pollution on public health, making it ideal for predictive modeling in public health and environmental sciences.

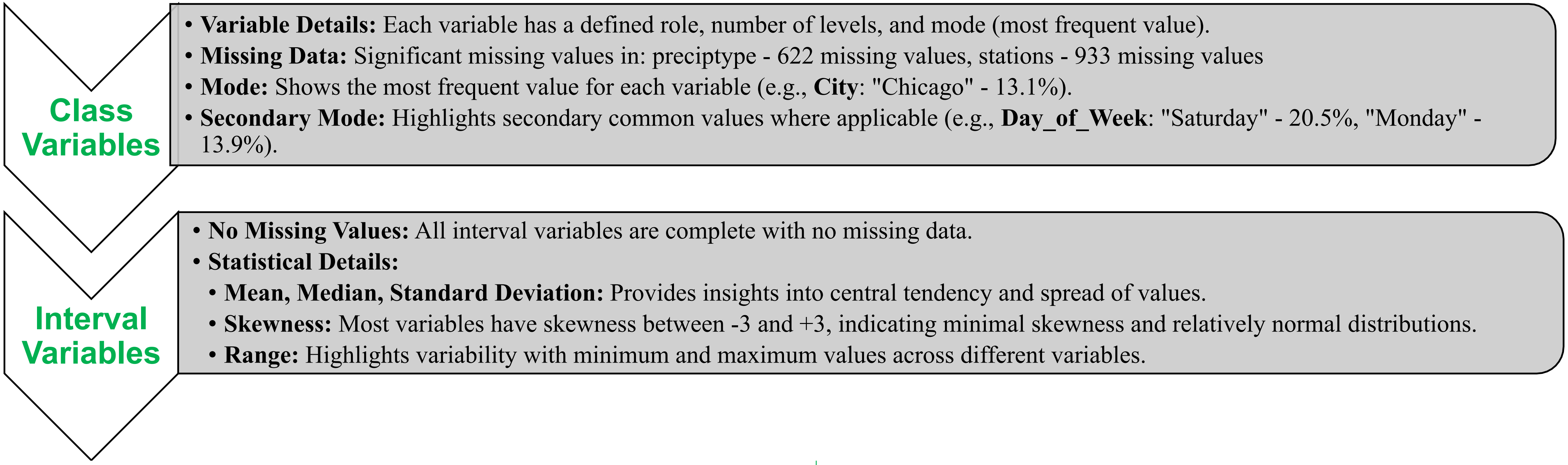
The dataset contains 46 columns/variables and a significant number of rows representing various weather and health-related variables.

Predictive Variables:

- **Health_Risk_Score:** Index representing the effects of urban air quality on public health.
- The goal is to predict how air pollutant concentrations affect health based on weather and air quality data.

Category	**Variables**	**Description**
Target Variable	Health Risk Score	Quantifies overall health risks based on environmental factors.
Independent Variables		
Weather Attributes	Temperature (Max, Min, Avg, Feels Like), Dew Point, Humidity, Precipitation, Wind Speed, Pressure, Cloud Cover, Visibility and more	Key weather indicators impacting pollutant dispersion and health outcomes.
Additional Fields	Solar Radiation, UV Index, Moon Phase and more	Extra environmental factors for a comprehensive analysis.

Summary Statistics

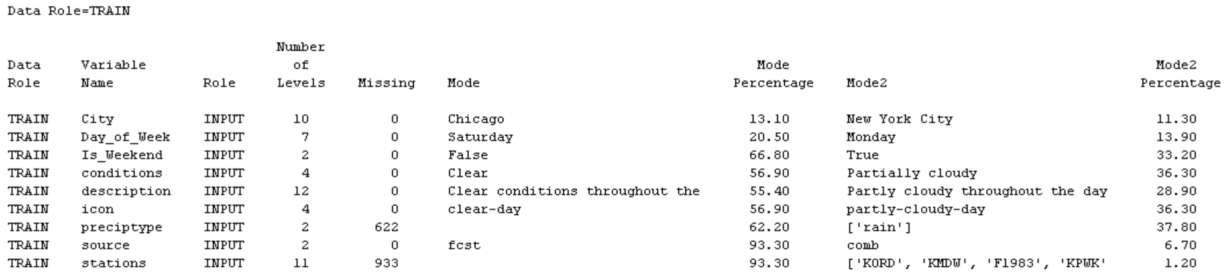


Class Variables

- **Variable Details:** Each variable has a defined role, number of levels, and mode (most frequent value).
- **Missing Data:** Significant missing values in: preciptype - 622 missing values, stations - 933 missing values
- **Mode:** Shows the most frequent value for each variable (e.g., **City**: "Chicago" - 13.1%).
- **Secondary Mode:** Highlights secondary common values where applicable (e.g., **Day_of_Week**: "Saturday" - 20.5%, "Monday" - 13.9%).

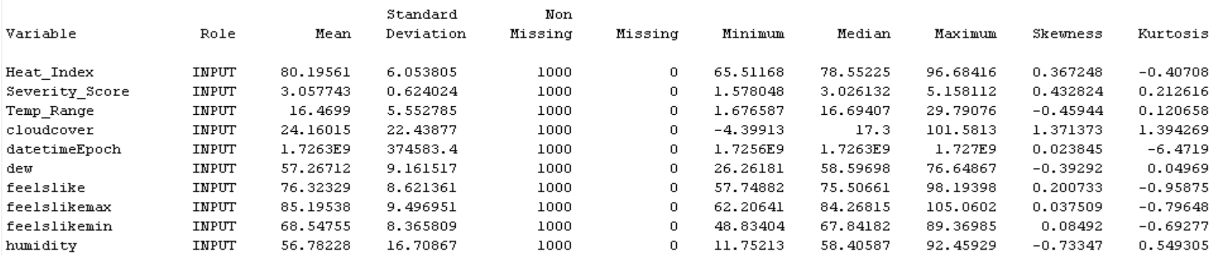
Interval Variables

- **No Missing Values:** All interval variables are complete with no missing data.
- **Statistical Details:**
  - **Mean, Median, Standard Deviation:** Provides insights into central tendency and spread of values.
  - **Skewness:** Most variables have skewness between -3 and +3, indicating minimal skewness and relatively normal distributions.
  - **Range:** Highlights variability with minimum and maximum values across different variables.



```
Data Role=TRAIN

                                Number
Data     Variable                 of                                                   Mode                                          Mode2
Role     Name           Role    Levels    Missing    Mode                            Percentage    Mode2                            Percentage

TRAIN    City           INPUT     10          0      Chicago                            13.10      New York City                       11.30
TRAIN    Day_of_Week    INPUT      7          0      Saturday                           20.50      Monday                              13.90
TRAIN    Is_Weekend     INPUT      2          0      False                              66.80      True                                33.20
TRAIN    conditions     INPUT      4          0      Clear                              56.90      Partially cloudy                    36.30
TRAIN    description    INPUT     12          0      Clear conditions throughout the    55.40      Partly cloudy throughout the day    28.90
TRAIN    icon           INPUT      4          0      clear-day                          56.90      partly-cloudy-day                   36.30
TRAIN    preciptype     INPUT      2        622                                         62.20      ['rain']                            37.80
TRAIN    source         INPUT      2          0      fcst                               93.30      comb                                 6.70
TRAIN    stations       INPUT     11        933                                         93.30      ['KORD', 'KMDW', 'F1983', 'KPWK'     1.20
```

Class Variables Summary



```
                                              Standard        Non
Variable             Role          Mean      Deviation    Missing     Missing     Minimum      Median     Maximum    Skewness    Kurtosis

Heat_Index          INPUT      80.19561       6.053805       1000           0    65.51168    78.55225    96.68416    0.367248    -0.40708
Severity_Score      INPUT      3.057743       0.624024       1000           0    1.578048    3.026132    5.158112    0.432824    0.212616
Temp_Range          INPUT       16.4699       5.552785       1000           0    1.676587    16.69407    29.79076    -0.45944    0.120658
cloudcover          INPUT      24.16015       22.43877       1000           0    -4.39913        17.3    101.5813    1.371373    1.394269
datetimeEpoch       INPUT      1.7263E9       374583.4       1000           0    1.7256E9    1.7263E9     1.727E9    0.023845     -6.4719
dew                 INPUT      57.26712       9.161517       1000           0    26.26181    58.59698    76.64867    -0.39292     0.04969
feelslike           INPUT      76.32329       8.621361       1000           0    57.74882    75.50661    98.19398    0.200733    -0.95875
feelslikemax        INPUT      85.19538       9.496951       1000           0    62.20641    84.26815    105.0602    0.037509    -0.79648
feelslikemin        INPUT      68.54755       8.365809       1000           0    48.83404    67.84182    89.36985     0.08492    -0.69277
humidity            INPUT      56.78228       16.70867       1000           0    11.75213    58.40587    92.45929    -0.73347    0.549305
```

Interval Variables Summary

Data Preparation

STEP 1: Initial Data Cleaning

Variable Rejection: Three variables, condition_code, snow_depth, and snow, were identified as containing only 0 values, which provide no analytical value. These variables were rejected during file import in SAS Enterprise Miner.

STEP 2: Handling Missing Values

For variables with missing values, we applied imputation techniques to ensure a complete dataset for analysis.

→ **Tree Surrogate Method:** For class variables with missing values, we used the Tree Surrogate Method to intelligently fill in missing data based on patterns observed in other variables. This approach preserved relationships within the dataset for enhanced model accuracy.

STEP 3: Outlier Treatment

Outlier Detection: Used box plots to identify outliers in significant variables, including **Humidity**, **Dew**, **Temp**, and **UV Index….**

→ **Handling Strategy:** We treated outliers using methods like Imputation and Transformation.

STEP 4: Addressing Non-Normal Distributions

- Applied **log transformations** to a few slightly skewed variables to improve interpretability and ensure model accuracy, particularly for **Linear Regression** and **Neural Networks**.
- **No Transformation is applied to Decision Tree** as it is robust it can handle as splits are created based on data ranges rather than specific values, means outliers do not significantly influence model structure.

We Ensured that the data met assumptions required for effective modeling.

Data Partitioning

- Splitting the data is essential for training **supervised machine learning algorithms**.
- In supervised learning, the model learns from **historical data** (training set), identifying patterns and rules to apply to new, unseen data.

Why is Data Splitting Important?

- It helps **reduce overfitting**, where a model may perform well on training data but poorly on new data.
- Ensures the model's **predictions are reliable** when applied to real-world scenarios.

Data Split for This Project:

By dividing the dataset into different subsets, we can ensure the model **generalizes well** and performs reliably:

- **60% for Training**: The portion of data used to build the model.
- **20% for Validation**: Helps in model tuning and selection.
- **20% for Testing**: Final check on model performance

Data Modelling

Target Variable: Health Risk Score
Type: Continuous
Model Suitability: We selected models that are well-suited for predicting continuous outcomes to ensure accurate health risk predictions.

Chosen Models

1. Linear Regression

- Provides a baseline by modeling linear relationships between predictors and the health risk score.
- Ideal for continuous target variables, offering insights into the direct effects of environmental factors.

2. Decision Tree

- Captures non-linear relationships and interactions among predictors to predict continuous outcomes.
- Well-suited for continuous targets, helping identify key factors impacting health risks.

3. Random Forest

- Ensemble of decision trees that improves accuracy and reduces overfitting for continuous predictions.
- Effective for continuous targets, as it handles feature interactions and variability in complex datasets.

4. Auto Neural Network

- Models complex, non-linear relationships to predict continuous target values.
- Appropriate for continuous outcomes, especially useful for capturing intricate patterns in data.

Model 1 – Linear Regression

We performed regression using different methods to identify the best fit model for predicting health risk.

Model Comparisons:

1. **Regression Model 1 - All Variables**

   R-Squared: 0.9749, **Adjusted R-Squared:** 0.9728

   Not Valid, Contains insignificant variables, which can inflate R-squared values.

2. **Regression Model 2 - Only Significant Variables**

   R-Squared: 0.9702, **Adjusted R-Squared:** 0.9696

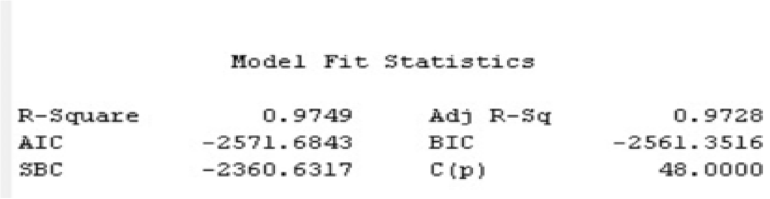
   Simplifies the model by focusing on important factors, reducing potential errors.

3. **Regression Model 3 after transformation - Stepwise Selection**

   R-Squared: 0.9728, **Adjusted R-Squared:** 0.9718

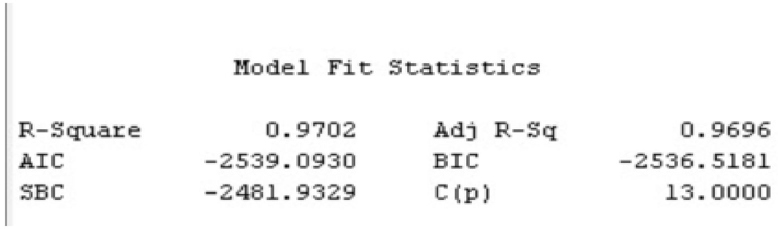
   Chosen Model: Stepwise regression due to efficient variable selection.

MODEL 1



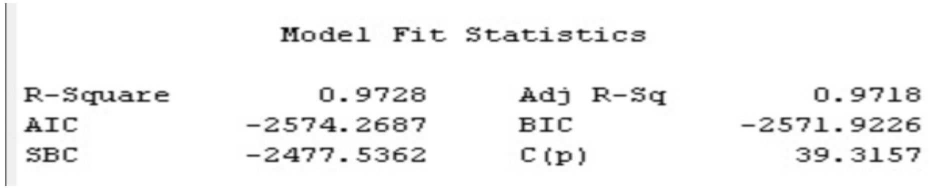
```
                    Model Fit Statistics

R-Square            0.9749      Adj R-Sq              0.9728
AIC             -2571.6843      BIC               -2561.3516
SBC             -2360.6317      C(p)                 48.0000
```

MODEL 2



```
                    Model Fit Statistics

R-Square            0.9702      Adj R-Sq              0.9696
AIC             -2539.0930      BIC               -2536.5181
SBC             -2481.9329      C(p)                 13.0000
```

MODEL 3



```
                    Model Fit Statistics

R-Square            0.9728      Adj R-Sq              0.9718
AIC             -2574.2687      BIC               -2571.9226
SBC             -2477.5362      C(p)                 39.3157
```

Why Stepwise Regression?

Model 1 is not valid because of insignificant variables, Between Model 2 and Model 3 - Stepwise Regression has Slightly more R-square Value.

- Automatically selects significant variables, improving interpretability.
- Focuses on key predictors, enhancing model robustness. Eliminates irrelevant variables, generalizing better to new data.
- Minimizes human bias in model selection.
  From all these models we used Regression model 3(stepwise) for model comparison with other models.

Findings from linear regression model

Model Performance

- **R-Squared:** 0.9728, **Adjusted R-Squared:** 0.9718
- Indicates a strong fit, explaining over 97% of the variance in the Health Risk Score.

Important Predictors

- **Heat Index:** Most significant predictor (F Value = 2377.95), suggesting a strong influence on health risk.
- **Humidity** High impact on health risk scores, with F values of 316.39 respectively.
- **Other Influential Variables:** Conditions, Pressure, UV Index, and Wind Speed also show significant effects on health risk.

Influence on Health Risk Score

- **Positive Impact:** Variables like Heat Index and Humidity are associated with increased health risk scores.
- **Negative or Neutral Impact:** Some environmental variables like Visibility and Wind Direction have less influence but still contribute to overall risk assessment.

Regression Equation for Health Risk Score:

Health_Risk_Score = β_0+ 0.7143 * city + 31.4239 * Heat_Index + 0. 6313 * Severity_Score ….+ 0.2811 * windgust

- A one-unit increase in **Heat Index** leads to a **31.42** increase in the Health Risk Score, holding all other variables constant.
- Other factors like **Severity Score**, **Conditions**, and **Dew** have similar effects on Health Risk Score, but their impact is smaller compared to Heat Index

Model 2 – Decision Tree

Decision Tree Model 1: Autonomous Tree with Subtree Assessment

- Automatically created tree using average square error for subtree assessment.
- **Optimal Leaves:** 35
- **Method:** Tree grown by minimizing average square error and pruned to achieve an optimal structure with 35 leaves.

Decision Tree Model 2: Maximum Branches with Three-Way Splits

- Tree model allowing maximum branches with three-way splits.
- **Optimal Leaves:** 50
- **Method:** Provides greater flexibility in splitting, resulting in a deeper structure with more leaves, capturing more complex patterns in the data.

Preferred Model: Decision Tree Model 2

- **Reason:** Captures more intricate patterns with three-way splits and a larger number of leaves, making it better suited for identifying complex relationships in the dataset.

Findings from Decision Tree

The model provides a hierarchical structure that shows the most influential variables in predicting **Health_Risk_Score**.

Top Splitting Variable:

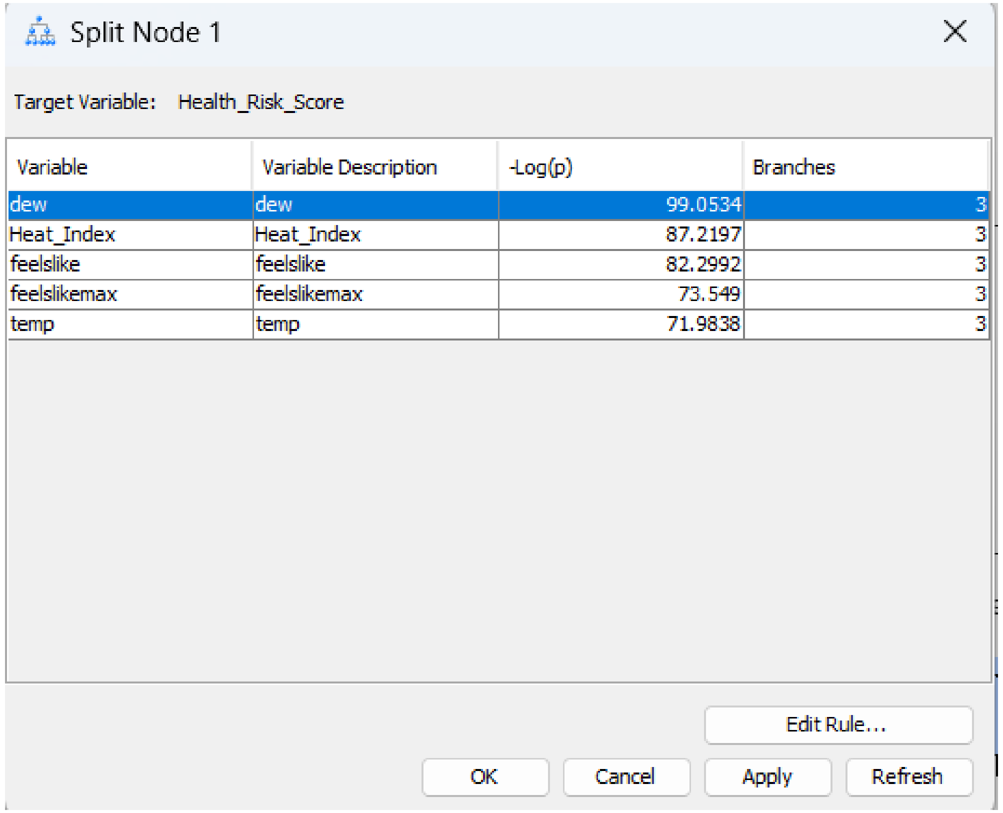
- **Dew**: Primary split variable with highest predictive value, indicating a strong relationship between humidity levels and health risks.

Other Important Predictors:

- **City**: Different cities show distinct health risk profiles, likely due to local pollution and climate conditions.
- **Heat Index and FeelsLike**: Both suggest that perceived temperature and heat stress are key factors impacting health risk.

Influence on Dependent Variable:

- **Dew** has the most significant impact, followed by **City** and **Heat Index**.
- Variables like **FeelsLike** indicate that both actual and perceived temperatures are relevant.



Split Node 1

Target Variable: Health_Risk_Score

Variable	Variable Description	-Log(p)	Branches
dew	dew	99.0534	3
Heat_Index	Heat_Index	87.2197	3
feelslike	feelslike	82.2992	3
feelslikemax	feelslikemax	73.549	3
temp	temp	71.9838	3

Model 3 – Random Forest

A Random Forest is an ensemble method, combining multiple decision trees to improve predictive performance and reduce overfitting by averaging predictions across many trees.

Created a **Random Forest** model using **average square error** as the evaluation measure on the **validation data**.

Key Findings

- **windspeed**, **dew**, **Severity**, and **solarradiation** appear to be among the most influential variables, as they have the highest number of splitting rules and comparatively low error rates.
- **windspeed** has the highest count, indicating it plays a significant role in prediction accuracy.
- Variables like windspeed and dew, which have high importance scores, suggest that these factors have a strong influence on the **Health_Risk_Score** (the dependent variable).

Model Performance Metrics

- **Average Squared Error (ASE)**: Training: **0.0158 ,**Validation: **0.0248**
- **Root Average Square Error (RASE)**: Training: **0.1253,** Validation: **0.1573**

Random Forest model has low error rates, suggesting good fit for this dataset.

Model 4 – Auto Neural Network

Why AutoNeural?

- AutoNeural simplifies the creation and optimization of neural networks, making it easier to configure and train models.
- It automatically selects the best configuration for the neural network, reducing manual tuning efforts.

Neural Network Architecture:

- **Input Layer:** Contains all independent variables (predictors) used in the model.
- **Hidden Layer:** Contains 3 hidden units, providing flexibility to capture complex relationships between input variables and the output. This choice strikes a balance between model complexity and overfitting.
- **Output Layer:** The dependent variable (health score risk) is the predicted output.

Key Benefits:

- Simplifies network design while maintaining flexibility.
- Efficiently captures non-linear patterns for accurate predictions.

Findings for Auto Neural Network

Performance Metrics:

- **Training ASE**: 0.4256 | **Validation ASE**: 0.4571 | **Test ASE**: 0.3901
- **Insight**: Low error rates indicate effective prediction of Health Risk Score.

Influential Variables

- Due to the nature of neural networks, specific variable importance isn't directly displayed; however, the model likely captured complex relationships among variables like **Heat Index**, **Humidity**, and **UV Index** that drive the Health Risk Score.

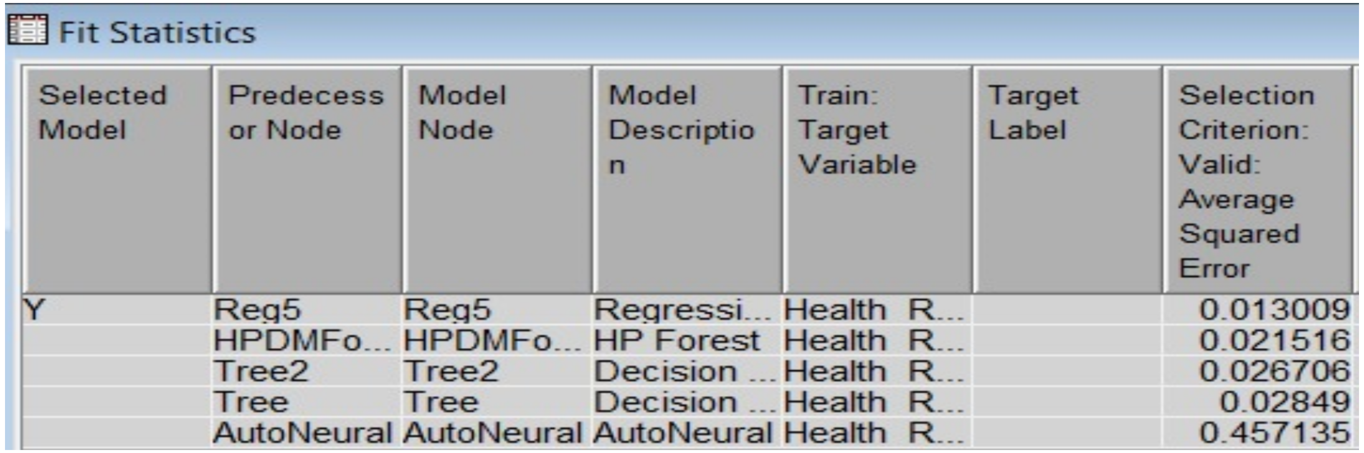
Complex Relationships Captured: Model captures multi-variable influences on health risk, though individual variable impacts aren't directly visible.

- **Hidden Variable Interactions**: Unlike decision trees, neural networks analyze combinations of variables, making them effective for capturing complex relationships affecting health risk scores.
- Specific variable impacts aren't directly visible, but the model's structure ensures all variables contribute to accurate predictions.

Model Comparision

Based on the Average Square Error (ASE) in the Fit Statistics table:

- Reg 5 has an ASE of 0.013009 for validation.
- HP Forest has an ASE of 0.021516 for validation.
- The Tree 2 Decision Tree has an ASE of 0.026706
- The Tree 1 Decision Tree has an ASE of 0.02849
- Auto Neural has an ASE of 0.457135



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Y	Reg5	Reg5	Regressi...	Health R...		0.013009
	HPDMFo...	HPDMFo...	HP Forest	Health R...		0.021516
	Tree2	Tree2	Decision ...	Health R...		0.026706
	Tree	Tree	Decision ...	Health R...		0.02849
	AutoNeural	AutoNeural	AutoNeural	Health R...		0.457135

The Reg5 Regression Model has the lowest ASE (0.013009) on the validation data, making it the best model among the all based on average square error.

Business Recommendations

- **Implement Real-Time Public Health Alerts**

Use model predictions to issue alerts for high-risk air quality, helping vulnerable populations reduce exposure.

- **Guide Resource Allocation for Health Initiatives**

Prioritize deployment of resources (e.g., air purifiers, cooling stations) based on environmental factors influencing health risks.

- **Develop Targeted Preventative Health Campaigns**

Educate residents in high-risk areas on protective measures like mask-wearing and reduced outdoor activity during peak pollution.

- **Collaborate with Urban Planners for Long-Term Solutions**

Share insights with city planners to support zoning and green space initiatives that reduce pollution exposure.

- **Optimize Health Resource Planning**

Use risk predictions to help hospitals prepare for increased demand during pollution spikes.

Conclusion

- Analyzed the relationship between urban air quality and health impacts using machine learning models.
- **Heat Index, Humidity, and UV Index** are major contributors to Health Risk Score.
- **Stepwise Linear Regression** achieved the highest accuracy with minimal error, providing a clear and interpretable solution.
- Highlighted the need for monitoring key environmental factors to mitigate health risks, supporting informed decision-making in urban planning and healthcare.
- This model can guide policymakers and urban planners to develop targeted interventions for high-risk environmental conditions.
- This project demonstrates how machine learning and data analytics can provide actionable insights, underscoring the importance of environmental monitoring for public health.

References

Data Sources

- U.S. Environmental Protection Agency. **Air Quality Data**. epa.gov
- Centers for Disease Control and Prevention. **Environmental Health Tracking**. cdc.gov

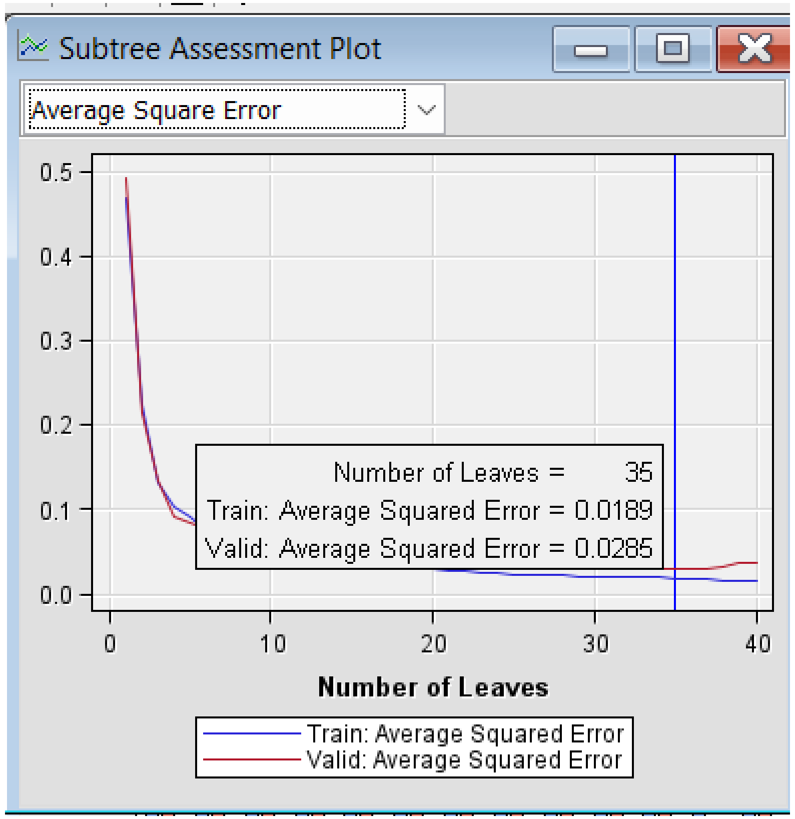
Machine Learning Techniques

- Hastie, T., Tibshirani, R., & Friedman, J. **The Elements of Statistical Learning**. Springer, 2009.
- Breiman, L. **Random Forests**. Machine Learning, 2001.

Environmental Health Studies

- Brook, R.D., et al. **Air Pollution and Cardiovascular Disease**. Circulation, 2010.
- Hoek, G., et al. **Long-term Air Pollution and Mortality**. Environmental Health, 2013.

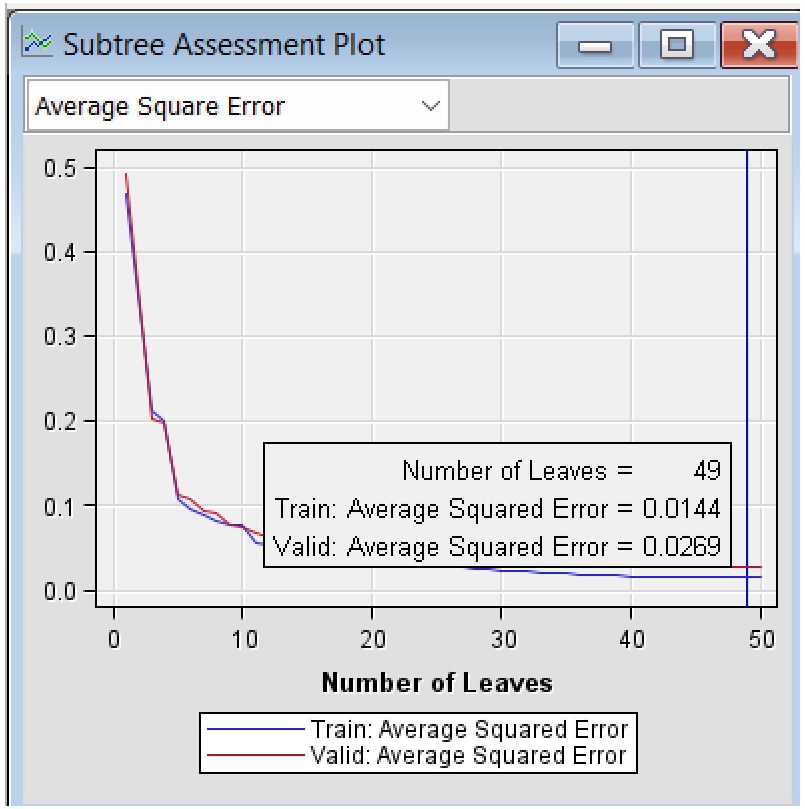
IMPORTANT OUTPUT SCREENSHOTS


Subtree Assessment Plot

Average Square Error

Number of Leaves = 35
Train: Average Squared Error = 0.0189
Valid: Average Squared Error = 0.0285

Decision tree 1

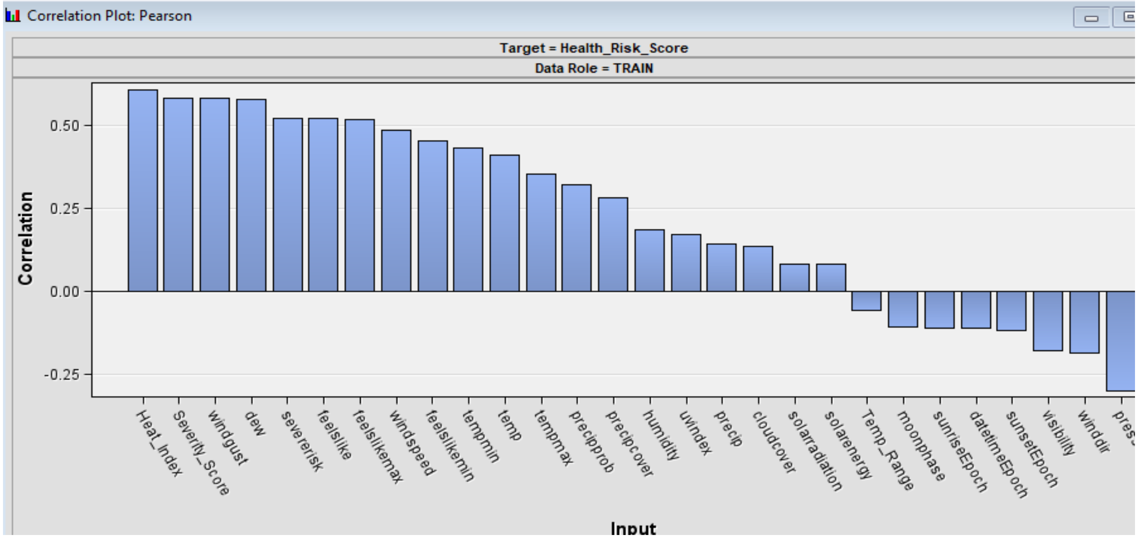

Subtree Assessment Plot

Average Square Error

Number of Leaves = 49
Train: Average Squared Error = 0.0144
Valid: Average Squared Error = 0.0269

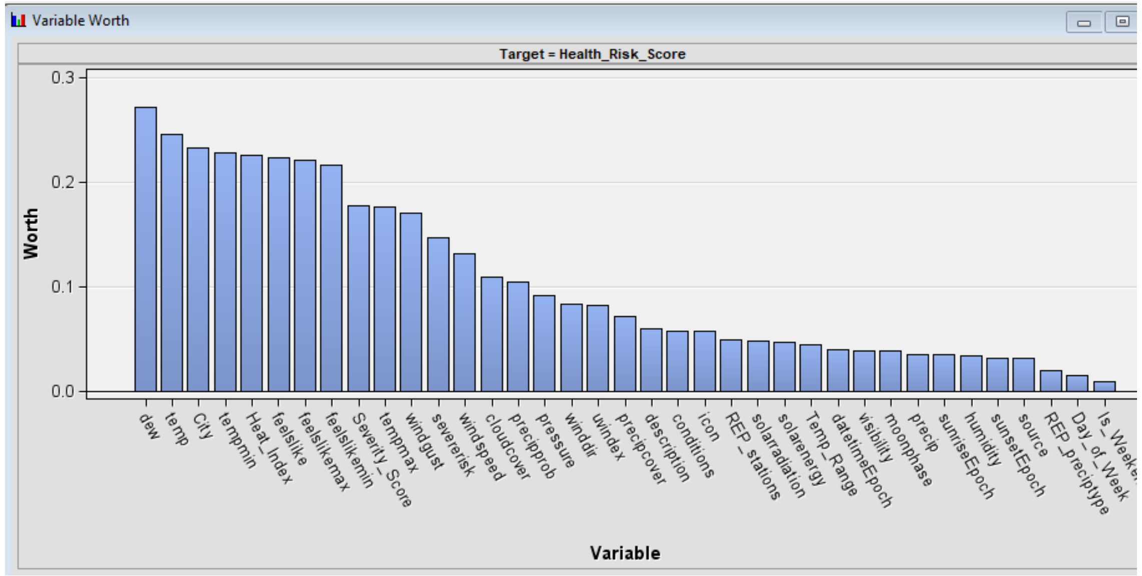
Decision tree 2

PLOTS

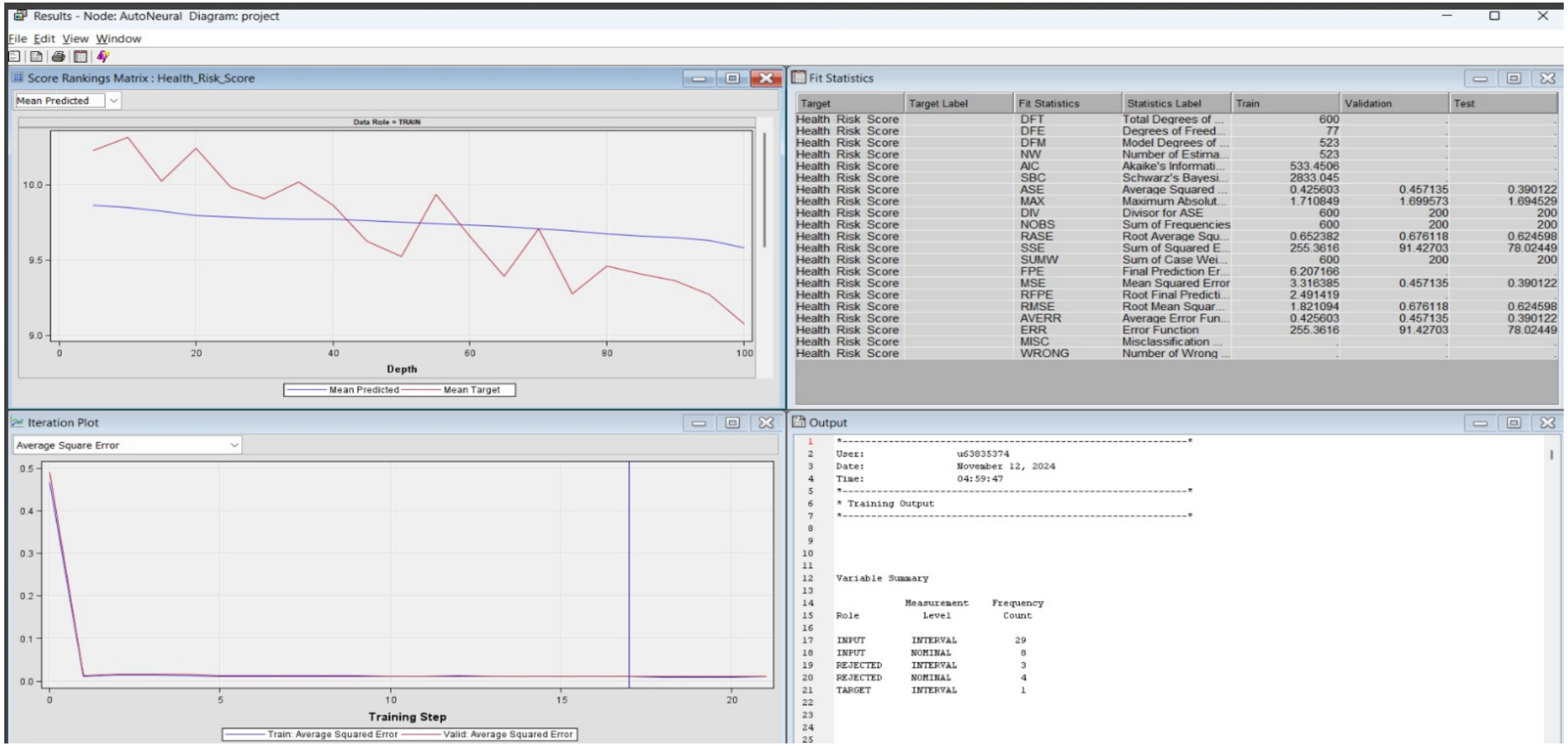


Correlation Plot – Correlation with Target Variable

Two variables are considered correlated if they change in the same direction. For example, the **heat index**, which represents the heat on a particular day, is strongly positively correlated with health risk, which makes sense given the impact of temperature on health. In contrast, **pressure** is negatively correlated with health risk, indicating that as pressure increases, health risk tends to decrease.



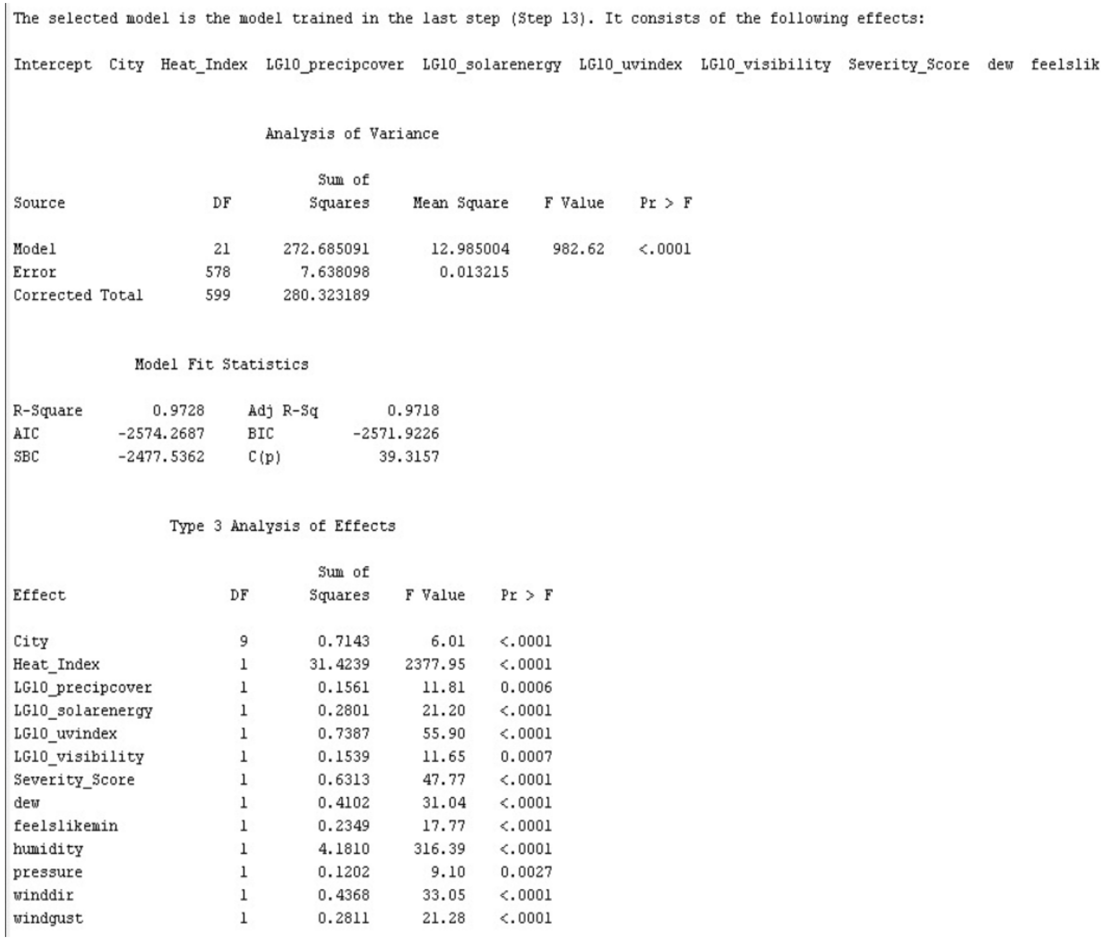
We also have a plot called **Variable Worth** in SAS Enterprise Miner, which evaluates the importance of each variable in the model. It shows how significant each variable is in predicting the target variable.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Health Risk Score		DFT	Total Degrees of ...	600	.	.
Health Risk Score		DFE	Degrees of Freed...	77	.	.
Health Risk Score		DFM	Model Degrees of ...	523	.	.
Health Risk Score		NW	Number of Estima...	523	.	.
Health Risk Score		AIC	Akaike's Informati...	533.4506	.	.
Health Risk Score		SBC	Schwarz's Bayesi...	2833.045	.	.
Health Risk Score		ASE	Average Squared ...	0.425603	0.457135	0.390122
Health Risk Score		MAX	Maximum Absolut...	1.710849	1.699573	1.694529
Health Risk Score		DIV	Divisor for ASE	600	200	200
Health Risk Score		NOBS	Sum of Frequencies	600	200	200
Health Risk Score		RASE	Root Average Squ...	0.652382	0.676118	0.624598
Health Risk Score		SSE	Sum of Squared E...	255.3616	91.42703	78.02449
Health Risk Score		SUMW	Sum of Case Wei...	600	200	200
Health Risk Score		FPE	Final Prediction Er...	6.207166	.	.
Health Risk Score		MSE	Mean Squared Error	3.316385	0.457135	0.390122
Health Risk Score		RFPE	Root Final Predicti...	2.491419	.	.
Health Risk Score		RMSE	Root Mean Squar...	1.821094	0.676118	0.624598
Health Risk Score		AVERR	Average Error Fun...	0.425603	0.457135	0.390122
Health Risk Score		ERR	Error Function	255.3616	91.42703	78.02449
Health Risk Score		MISC	Misclassification ...	.	.	.
Health Risk Score		WRONG	Number of Wrong ...	.	.	.

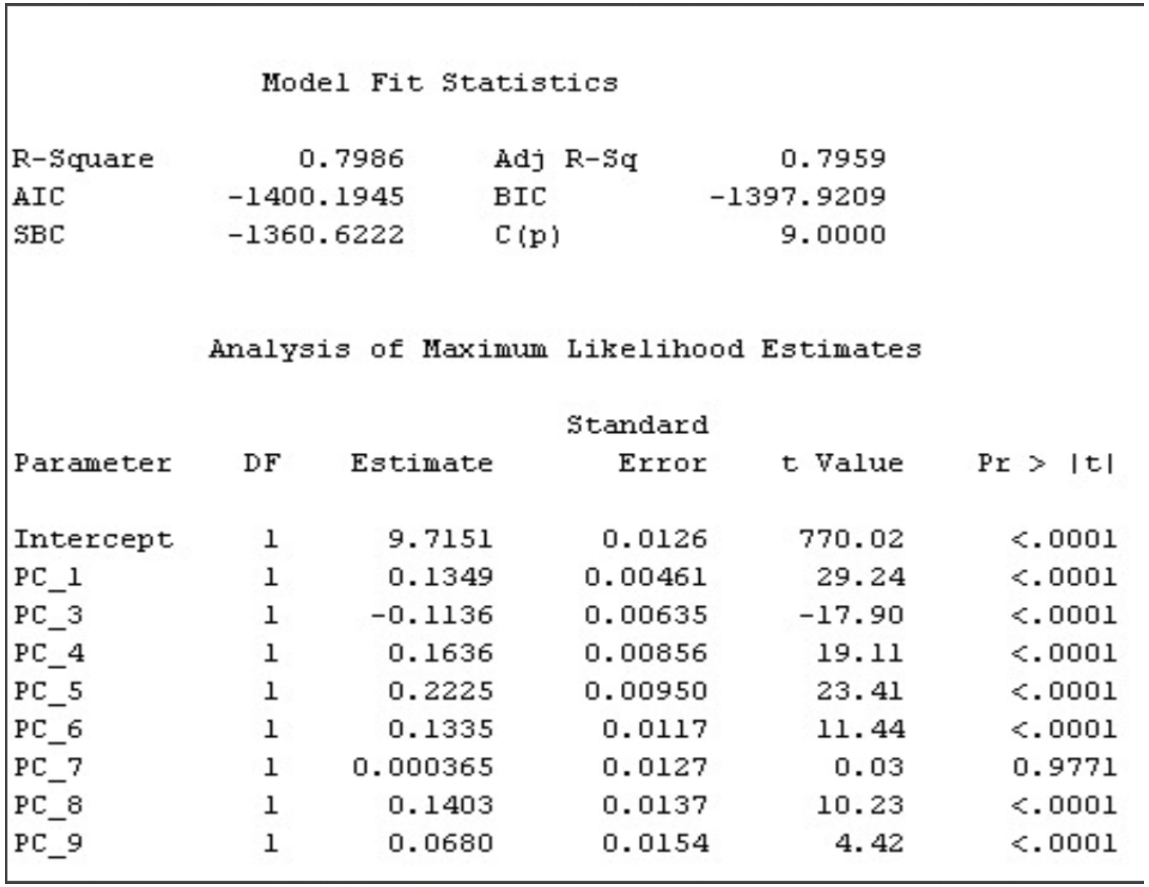
```
*------------------------------------------------------------*
User:                   u63835374
Date:                   November 12, 2024
Time:                   04:59:47
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT        INTERVAL         29
INPUT        NOMINAL           8
REJECTED     INTERVAL          3
REJECTED     NOMINAL           4
TARGET       INTERVAL          1



```

Auto Neural Network



```
The selected model is the model trained in the last step (Step 13). It consists of the following effects:

Intercept  City  Heat_Index  LG10_precipcover  LG10_solarenergy  LG10_uvindex  LG10_visibility  Severity_Score  dew  feelslikemin  humidity  pressure  winddir  windgust


                          Analysis of Variance

                                      Sum of
Source                   DF          Squares      Mean Square    F Value    Pr > F

Model                    21       272.685091        12.985004     982.62    <.0001
Error                   578         7.638098         0.013215
Corrected Total         599       280.323189


              Model Fit Statistics

R-Square         0.9728     Adj R-Sq         0.9718
AIC          -2574.2687     BIC          -2571.9226
SBC          -2477.5362     C(p)            39.3157


                  Type 3 Analysis of Effects

                                      Sum of
Effect                     DF        Squares    F Value    Pr > F

City                        9         0.7143       6.01    <.0001
Heat_Index                  1        31.4239    2377.95    <.0001
LG10_precipcover            1         0.1561      11.81    0.0006
LG10_solarenergy            1         0.2801      21.20    <.0001
LG10_uvindex                1         0.7387      55.90    <.0001
LG10_visibility             1         0.1539      11.65    0.0007
Severity_Score              1         0.6313      47.77    <.0001
dew                         1         0.4102      31.04    <.0001
feelslikemin                1         0.2349      17.77    <.0001
humidity                    1         4.1810     316.39    <.0001
pressure                    1         0.1202       9.10    0.0027
winddir                     1         0.4368      33.05    <.0001
windgust                    1         0.2811      21.28    <.0001
```

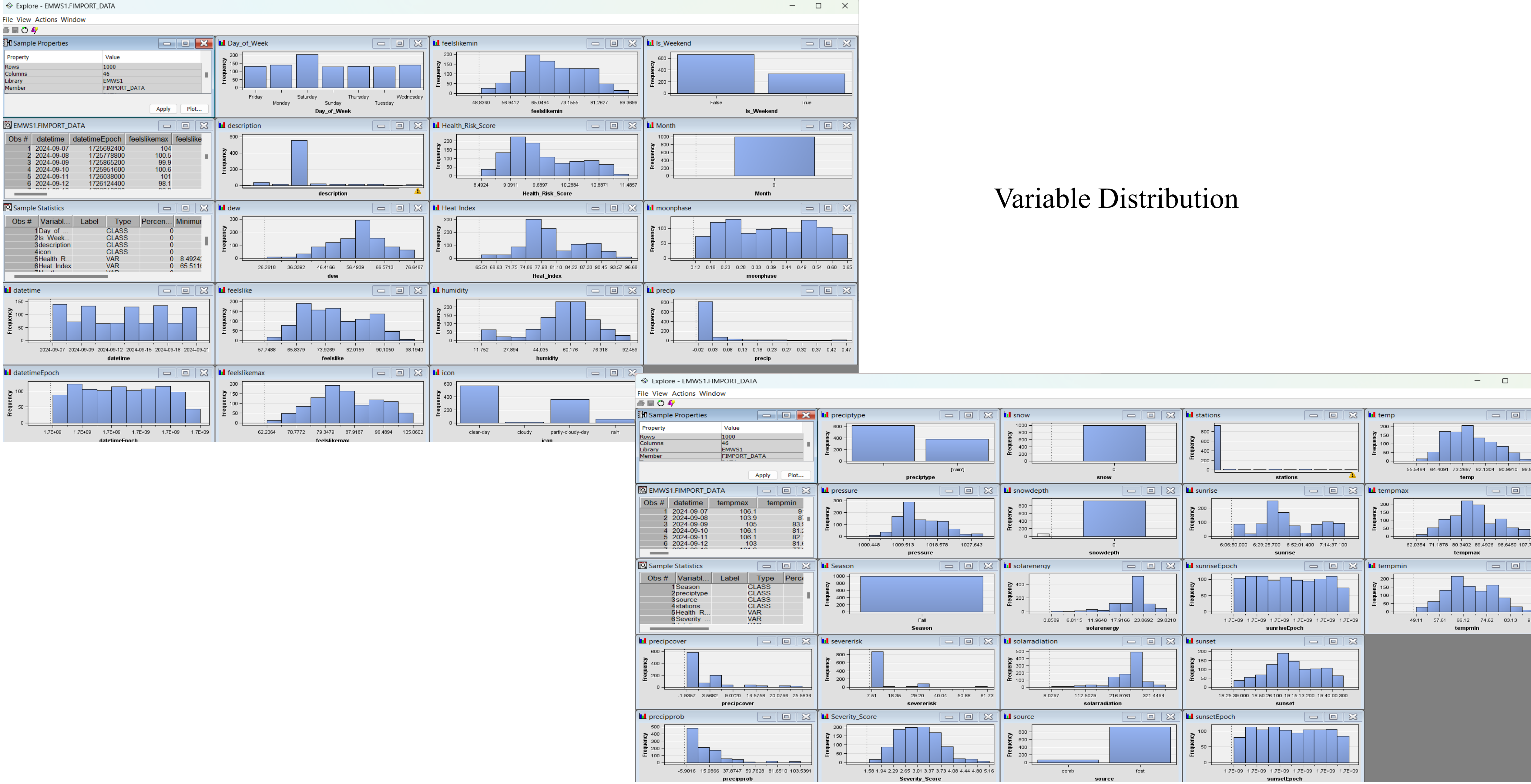
Linear Regression - Stepwise



```
            Model Fit Statistics

R-Square          0.7986     Adj R-Sq          0.7959
AIC           -1400.1945     BIC           -1397.9209
SBC           -1360.6222     C(p)              9.0000


        Analysis of Maximum Likelihood Estimates

                                     Standard
Parameter    DF      Estimate           Error     t Value     Pr > |t|

Intercept     1        9.7151          0.0126      770.02       <.0001
PC_1          1        0.1349         0.00461       29.24       <.0001
PC_3          1       -0.1136         0.00635      -17.90       <.0001
PC_4          1        0.1636         0.00856       19.11       <.0001
PC_5          1        0.2225         0.00950       23.41       <.0001
PC_6          1        0.1335          0.0117       11.44       <.0001
PC_7          1      0.000365          0.0127        0.03       0.9771
PC_8          1        0.1403          0.0137       10.23       <.0001
PC_9          1        0.0680          0.0154        4.42       <.0001
```

Principal Component Analysis Regression

Variable Distribution



Explore - EMWS1.FIMPORT_DATA

Sample Properties

Property	Value
Rows	1000
Columns	46
Library	EMWS1
Member	FIMPORT_DATA

EMWS1.FIMPORT_DATA

Obs #	datetime	datetimeEpoch	feelslikemax	feelslike
1	2024-09-07	1725692400	104	
2	2024-09-08	1725778800	100.5	
3	2024-09-09	1725865200	99.9	
4	2024-09-10	1725951600	100.6	
5	2024-09-11	1726038000	101	
6	2024-09-12	1726124400	98.1	

Sample Statistics

Obs #	Variabl...	Label	Type	Percen...	Minimur
1	Day of ...		CLASS	0	
2	Is Week...		CLASS	0	
3	description		CLASS	0	
4	icon		CLASS	0	
5	Health R...		VAR	0	8.49243
6	Heat Index		VAR	0	65.5116

Explore - EMWS1.FIMPORT_DATA

Sample Properties

Property	Value
Rows	1000
Columns	46
Library	EMWS1
Member	FIMPORT_DATA

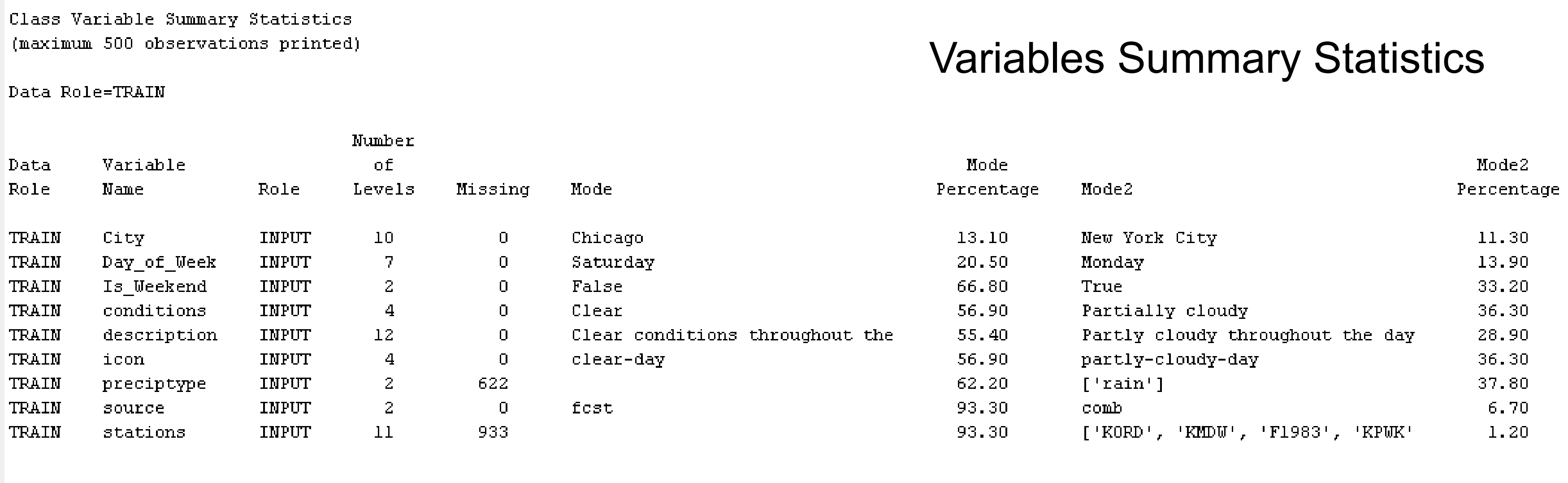
EMWS1.FIMPORT_DATA

Obs #	datetime	tempmax	tempmin
1	2024-09-07	106.1	91
2	2024-09-08	103.9	87
3	2024-09-09	105	83.9
4	2024-09-10	106.1	81.2
5	2024-09-11	106.1	82.1
6	2024-09-12	103	81.6

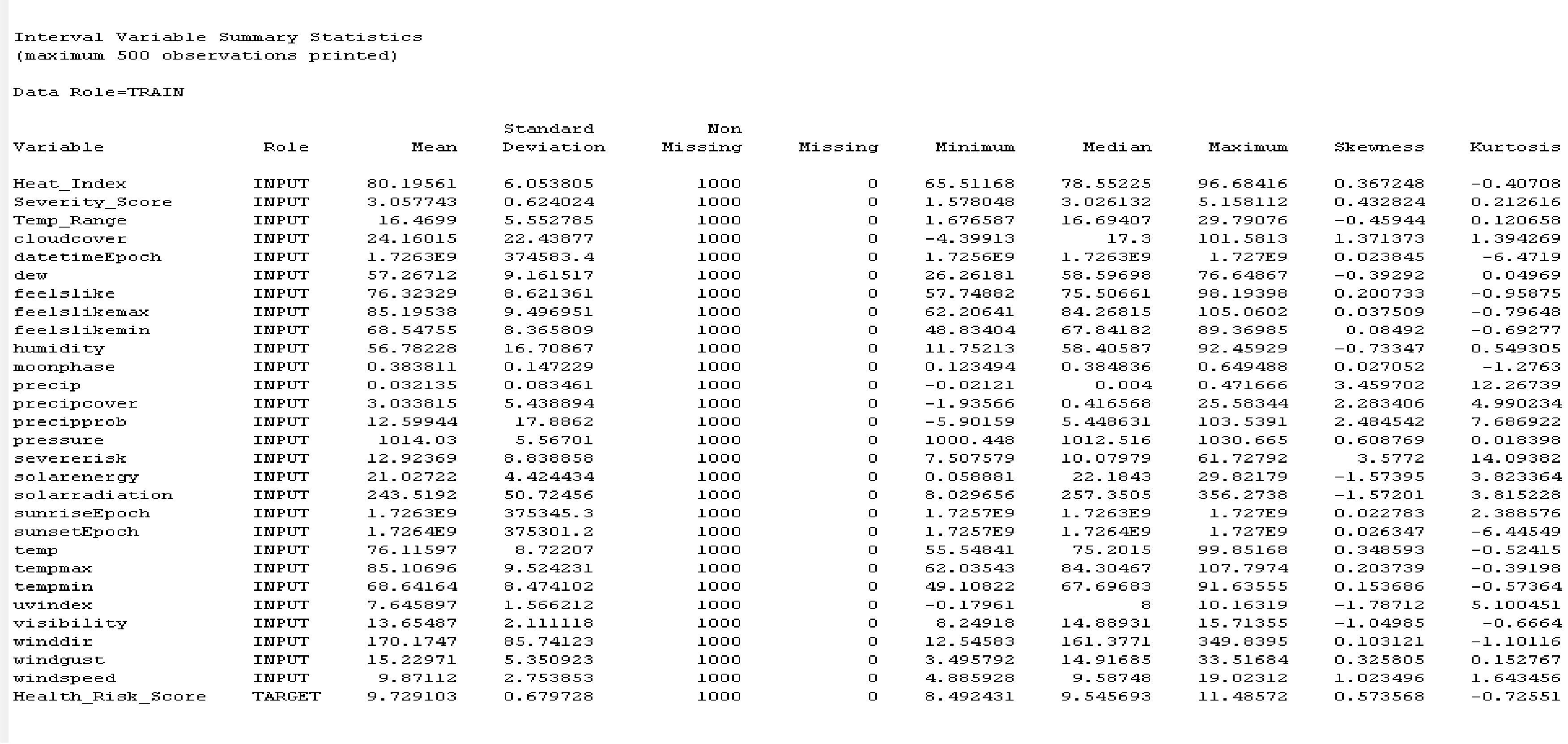
Sample Statistics

Obs #	Variabl...	Label	Type	Perce
1	Season		CLASS	
2	preciptype		CLASS	
3	source		CLASS	
4	stations		CLASS	
5	Health R...		VAR	
6	Severity ...		VAR	

Variables Summary Statistics

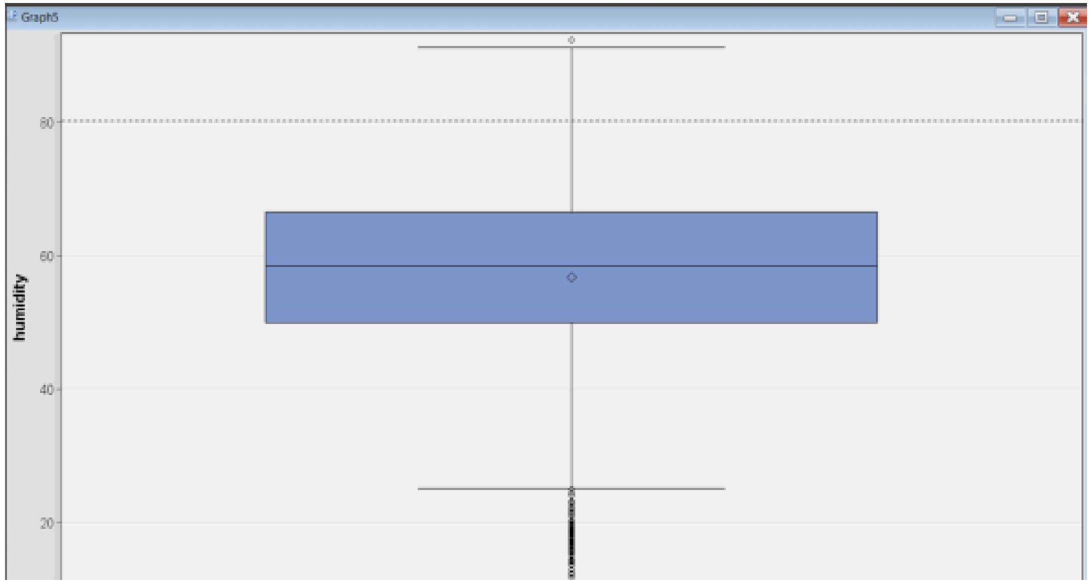


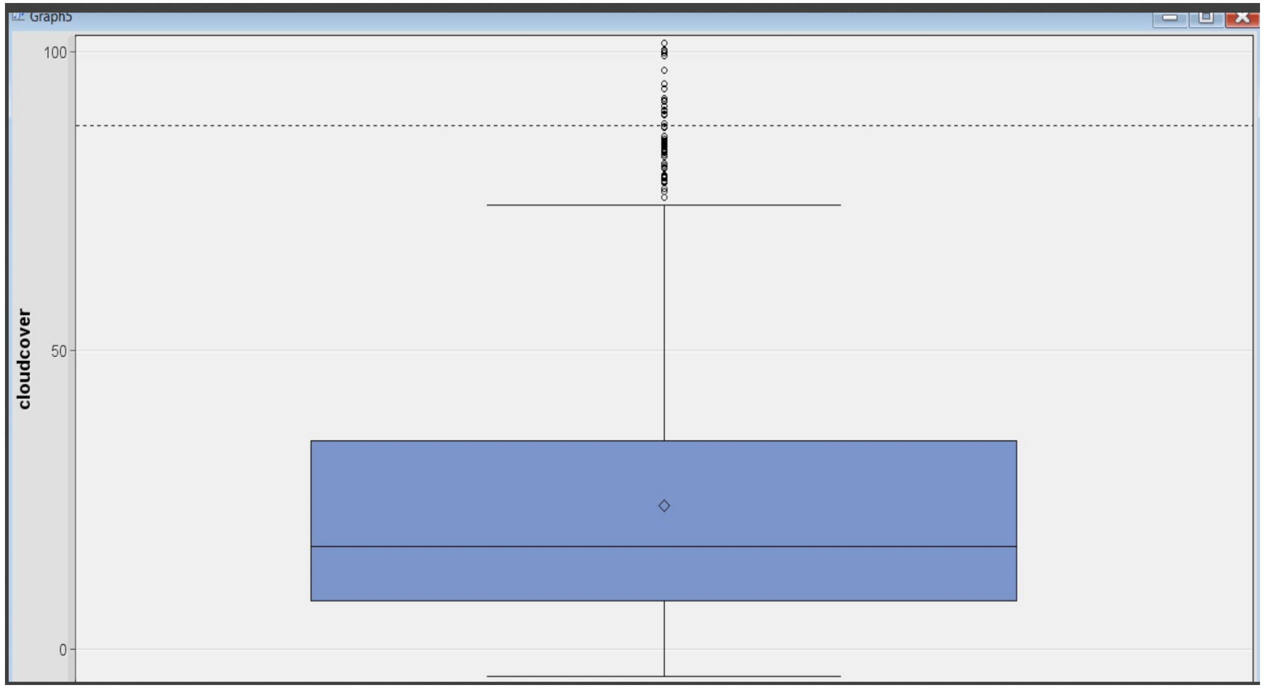
```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                Number
Data     Variable                 of                                                     Mode                                              Mode2
Role     Name           Role    Levels   Missing   Mode                               Percentage   Mode2                                  Percentage

TRAIN    City           INPUT     10          0    Chicago                               13.10     New York City                             11.30
TRAIN    Day_of_Week    INPUT      7          0    Saturday                              20.50     Monday                                    13.90
TRAIN    Is_Weekend     INPUT      2          0    False                                 66.80     True                                      33.20
TRAIN    conditions     INPUT      4          0    Clear                                 56.90     Partially cloudy                          36.30
TRAIN    description    INPUT     12          0    Clear conditions throughout the       55.40     Partly cloudy throughout the day          28.90
TRAIN    icon           INPUT      4          0    clear-day                             56.90     partly-cloudy-day                         36.30
TRAIN    preciptype     INPUT      2        622                                          62.20     ['rain']                                  37.80
TRAIN    source         INPUT      2          0    fcst                                  93.30     comb                                       6.70
TRAIN    stations       INPUT     11        933                                          93.30     ['KORD', 'KMDW', 'F1983', 'KPWK'          1.20
```



```
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                           Standard        Non
Variable              Role         Mean    Deviation    Missing    Missing     Minimum      Median     Maximum    Skewness    Kurtosis

Heat_Index            INPUT    80.19561     6.053805       1000          0    65.51168    78.55225    96.68416    0.367248    -0.40708
Severity_Score        INPUT    3.057743     0.624024       1000          0    1.578048    3.026132    5.158112    0.432824    0.212616
Temp_Range            INPUT     16.4699     5.552785       1000          0    1.676587    16.69407    29.79076    -0.45944    0.120658
cloudcover            INPUT    24.16015     22.43877       1000          0    -4.39913        17.3    101.5813    1.371373    1.394269
datetimeEpoch         INPUT    1.7263E9     374583.4       1000          0    1.7256E9    1.7263E9     1.727E9    0.023845     -6.4719
dew                   INPUT    57.26712     9.161517       1000          0    26.26181    58.59698    76.64867    -0.39292     0.04969
feelslike             INPUT    76.32329     8.621361       1000          0    57.74882    75.50661    98.19398    0.200733    -0.95875
feelslikemax          INPUT    85.19538     9.496951       1000          0    62.20641    84.26815    105.0602    0.037509    -0.79648
feelslikemin          INPUT    68.54755     8.365809       1000          0    48.83404    67.84182    89.36985     0.08492    -0.69277
humidity              INPUT    56.78228     16.70867       1000          0    11.75213    58.40587    92.45929    -0.73347    0.549305
moonphase             INPUT    0.383811     0.147229       1000          0    0.123494    0.384836    0.649488    0.027052     -1.2763
precip                INPUT    0.032135     0.083461       1000          0    -0.02121       0.004    0.471666    3.459702    12.26739
precipcover           INPUT    3.033815     5.438894       1000          0    -1.93566    0.416568    25.58344    2.283406    4.990234
precipprob            INPUT    12.59944      17.8862       1000          0    -5.90159    5.448631    103.5391    2.484542    7.686922
pressure              INPUT     1014.03      5.56701       1000          0    1000.448    1012.516    1030.665    0.608769    0.018398
severerisk            INPUT    12.92369     8.838858       1000          0    7.507579    10.07979    61.72792      3.5772    14.09382
solarenergy           INPUT    21.02722     4.424434       1000          0    0.058881     22.1843    29.82179    -1.57395    3.823364
solarradiation        INPUT    243.5192     50.72456       1000          0    8.029656    257.3505    356.2738    -1.57201    3.815228
sunriseEpoch          INPUT    1.7263E9     375345.3       1000          0    1.7257E9    1.7263E9     1.727E9    0.022783    2.388576
sunsetEpoch           INPUT    1.7264E9     375301.2       1000          0    1.7257E9    1.7264E9     1.727E9    0.026347    -6.44549
temp                  INPUT    76.11597      8.72207       1000          0    55.54841     75.2015    99.85168    0.348593    -0.52415
tempmax               INPUT    85.10696     9.524231       1000          0    62.03543    84.30467    107.7974    0.203739    -0.39198
tempmin               INPUT    68.64164     8.474102       1000          0    49.10822    67.69683    91.63555    0.153686    -0.57364
uvindex               INPUT    7.645897     1.566212       1000          0    -0.17961           8    10.16319    -1.78712    5.100451
visibility            INPUT    13.65487     2.111118       1000          0     8.24918    14.88931    15.71355    -1.04985     -0.6664
winddir               INPUT    170.1747     85.74123       1000          0    12.54583    161.3771    349.8395    0.103121    -1.10116
windgust              INPUT    15.22971     5.350923       1000          0    3.495792    14.91685    33.51684    0.325805    0.152767
windspeed             INPUT     9.87112     2.753853       1000          0    4.885928     9.58748    19.02312    1.023496    1.643456
Health_Risk_Score     TARGET   9.729103     0.679728       1000          0    8.492431    9.545693    11.48572    0.573568    -0.72551
```

Box Plot for Outlier Detection





Question?

Thank you