VIETNAM NATIONAL UNIVERSITY, HCM

UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

SUBJECT: CSC14004 DATA MINING

Report: Lab01: Preprocessing data

lecturer : Lê Hoài Bắc

Nguyễn Bảo Long

Nguyễn Ngọc Đức

Student : Hồ Thế Phúc – 21127670

Class : 21KHMT1

Contents

I.General:

1. file organization:

- 1 test file: “house-prices.csv”
- 1 source file: “21127670_preprocess.py”
- 1 example command lines commands file:” example_cmd.txt”
- Outputs file will follow format : “input_file_name_operationDone_columnAffected_timeStamp.csv”

2. Group information

Name	ID
Hồ Thế Phúc	21127670

3. Check list

Work	Percent of work completed
- complete requirement 1, requirement 2	100%
- Exploring Breast Cancer data set	100%
- code command line arguments	100%
- Exploring Weather data set	100%
- Exploring Credit in Germany data set	100%
- code functions: 1, 2, 3, 4, 5, 6, 7, 8	100%
- complete report	100%

II. Detail description

1.Install WEKA

1.1. Requirement 1

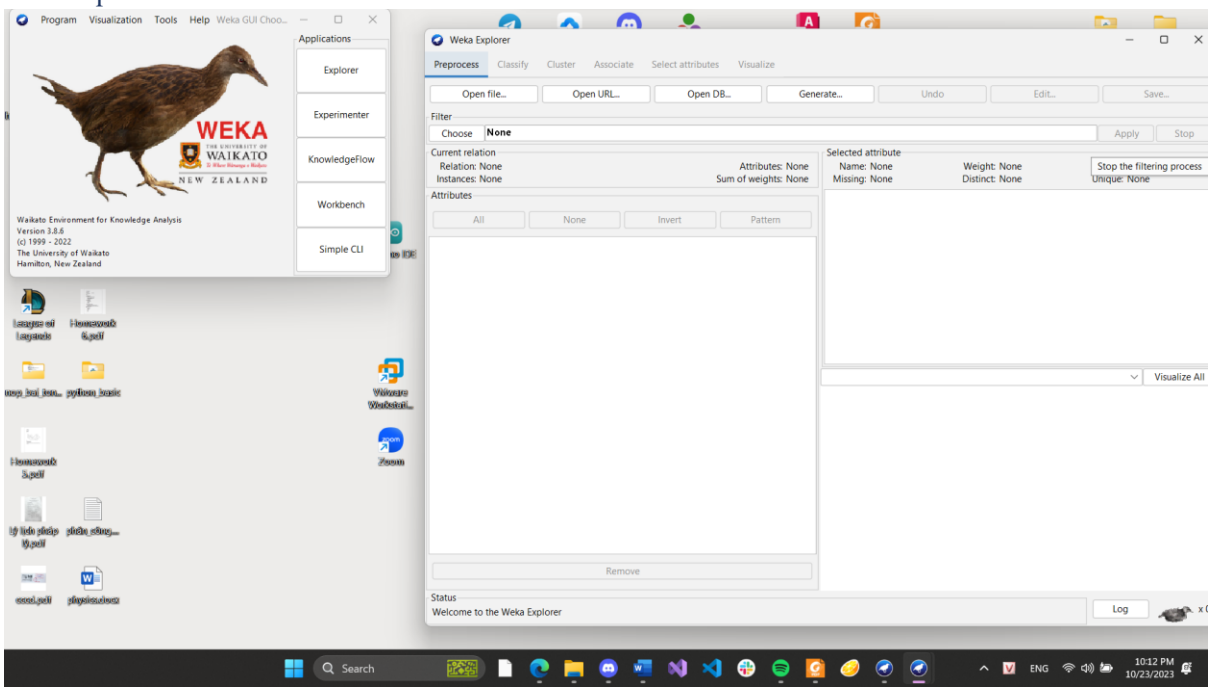


Image 1.After installation of weka and opening the 'explorer' function

1.2. Requirement 2

WEKA provides flexible options for loading dataset files into the platform. The data can originate from the local filesystem, a remote URL, a database connection, or be generated from within WEKA itself. Once loaded, the dataset instances can be edited directly - instances can be added or removed as needed. Any changes made to the active dataset can also be saved back to a file for reuse in subsequent analyses. This combination of versatile data input and manipulation capabilities allows users to readily prepare, customize, and experiment with datasets using the functionality provided by WEKA.

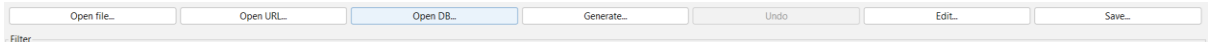


Image 1.2.a options for opening data set in WEKA explorer

The cpu.arff dataset has been loaded into WEKA for analysis. This indicates that WEKA will allow interactive exploration and modeling to be performed on the instances contained in this dataset file.

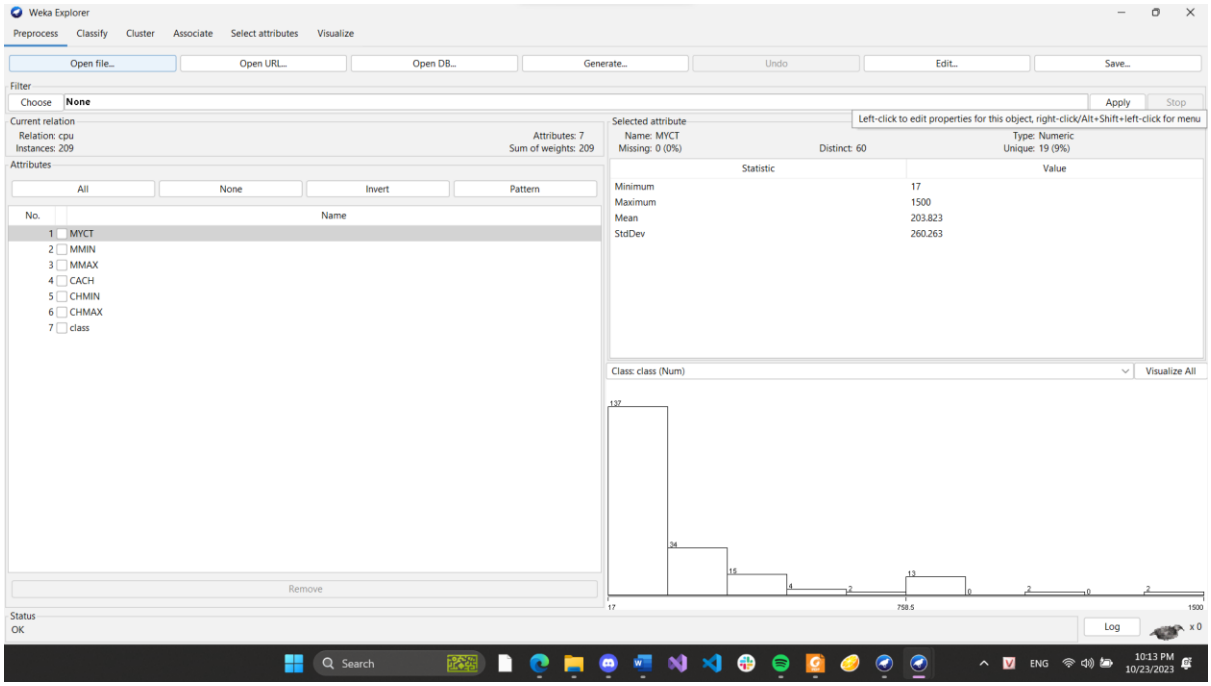


Image 1.2.b The GUI after opening cpu.arff file

- Current Relation section provides summary statistics about the active dataset. It displays the name assigned to the dataset, the total number of attributes, the count of instances included, and the aggregate weight of all instances. These details offer a quick overview of the loaded dataset's size and composition. Tracking this relation information allows confirming that the intended data has been properly imported into WEKA and facilitates understanding the breadth of available data for analysis.

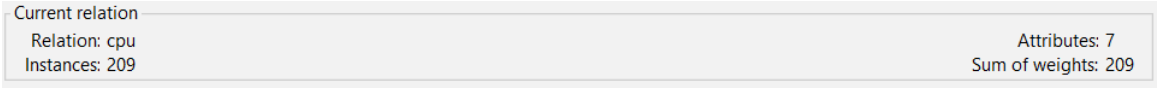


Image 1.2.c The current relation section

- Attributes section lists the names of attributes and gives you the option to select one of them. There are quick access keys for quick comparison of these attributes. The ALL option selects all present attributes, the NONE lets you deselect selected attributes, the Invert deselects current attributes and selects the remaining attributes, the Pattern button lets you enter a perl expression and explore the pattern of the data set.

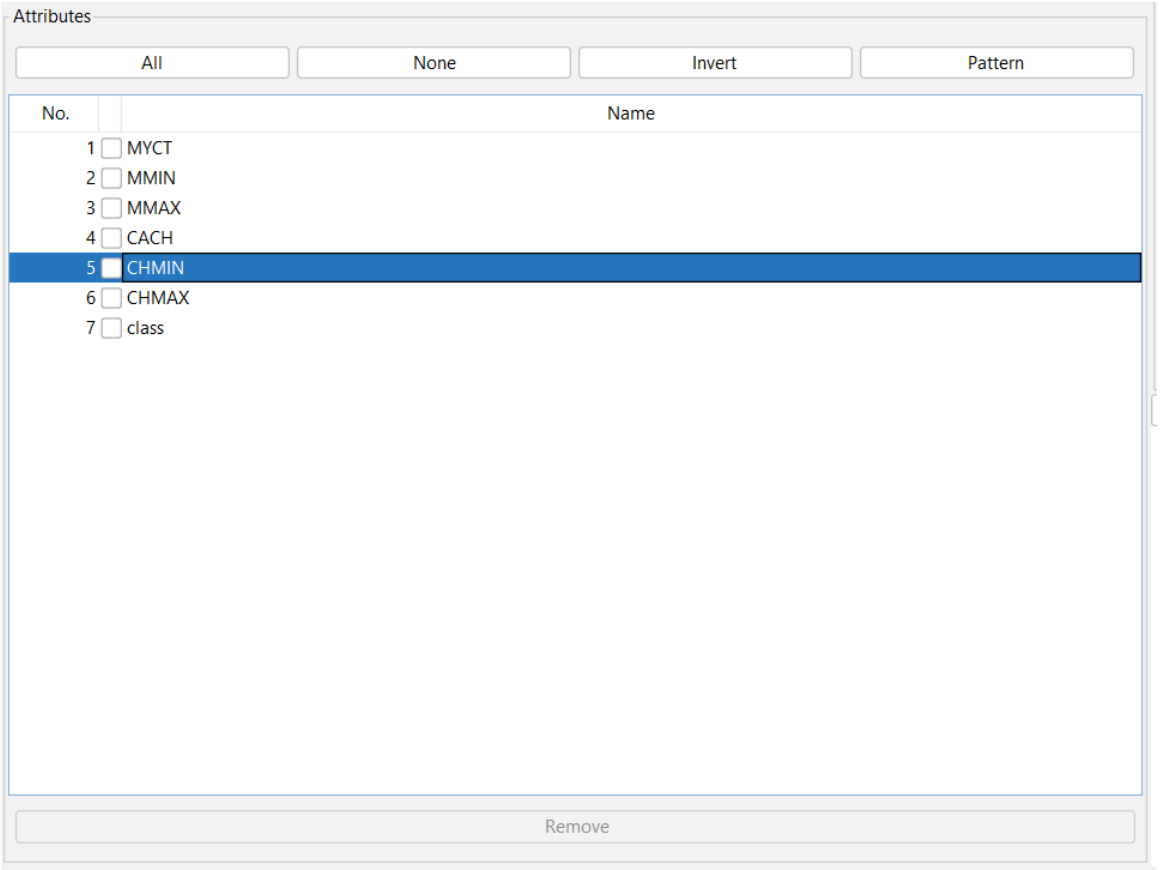


Image 1.2.d The attributes section

- Selected Attributes: pane displays in-depth information about the currently chosen attribute. It indicates the attribute's data type and provides summary statistics including the count of missing values, number of distinct values, and number of unique values. When an attribute is clicked, additional numeric statistics are shown such as minimum, maximum, mean, and standard deviation. A visualization is also generated in the form of a bar chart depicting the distribution of values for the selected attribute. This comprehensive attribute analysis and visualization enables deeper examination and understanding of individual attribute characteristics and patterns within the dataset.

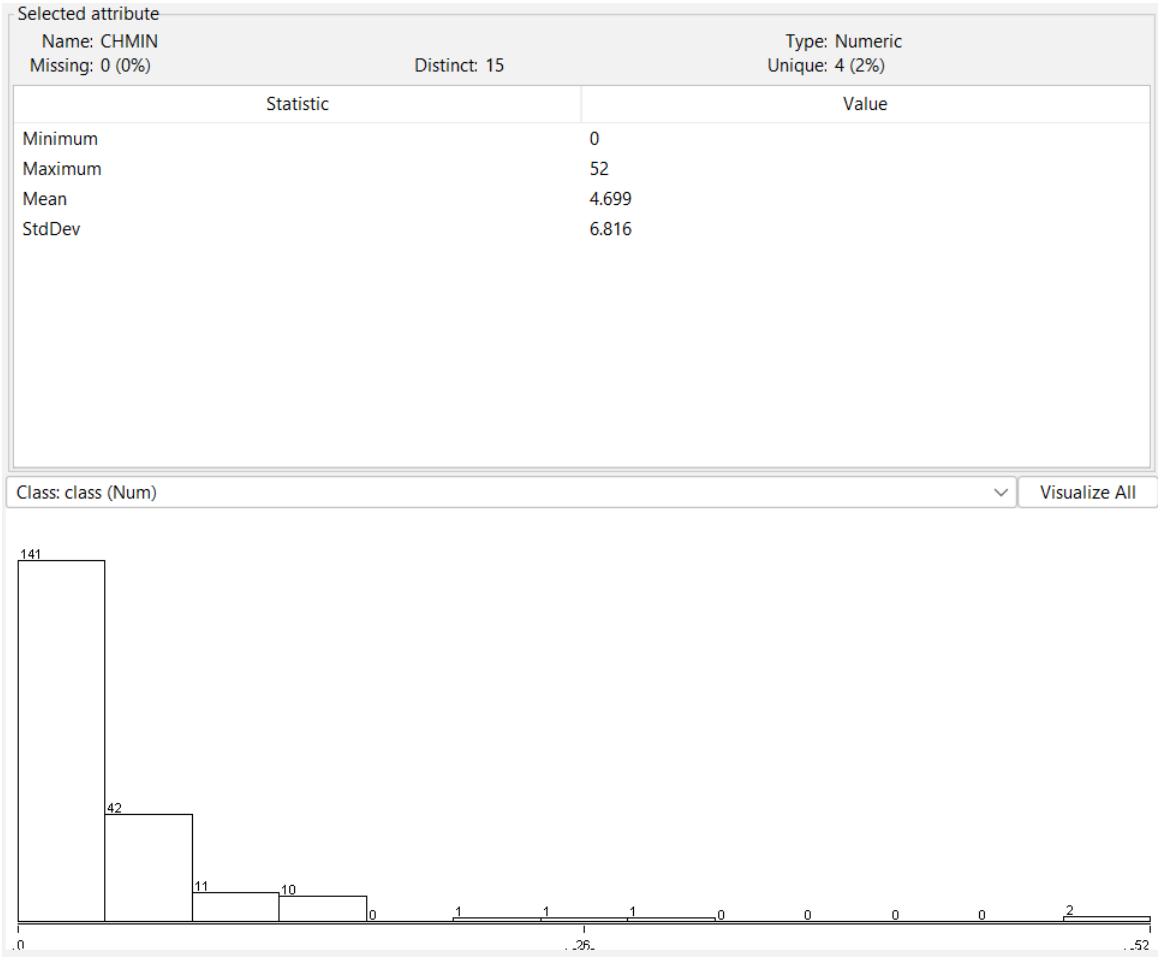


Image 1.2.e The selected attribute section

1.3. Other tabs:

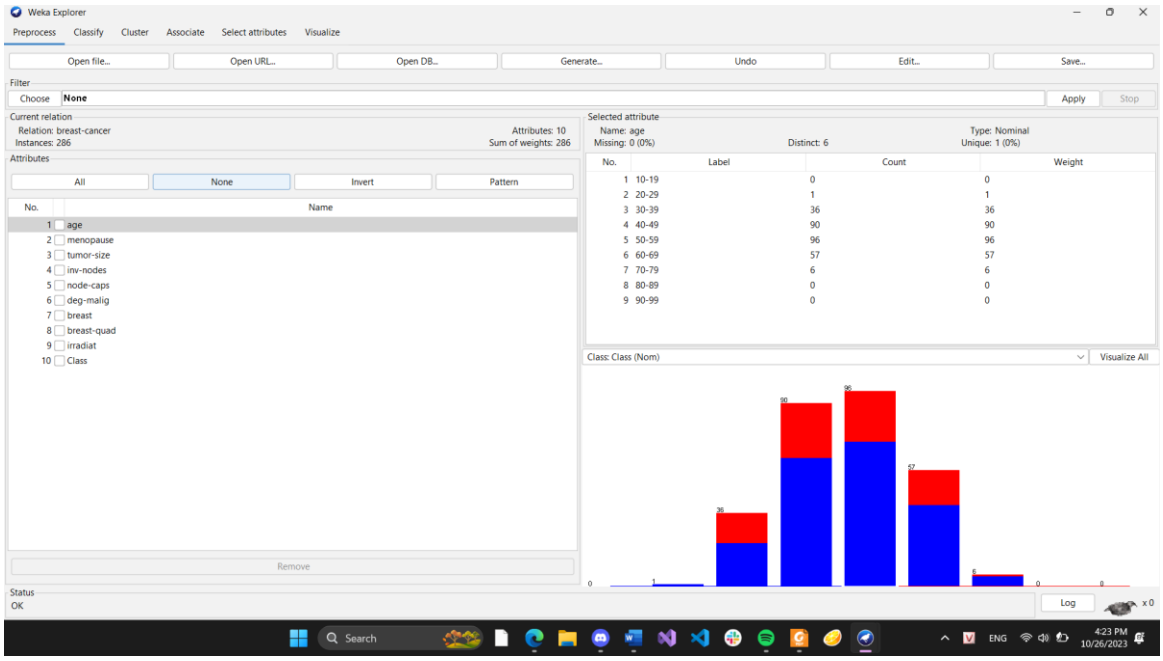
WEKA provides a range of built-in machine learning capabilities accessible through its interface tabs:

- The Classification tab contains algorithms like Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, Naive Bayes, and more. Both supervised and unsupervised techniques are included for predictive data modeling.
- The Clustering tab gives access to cluster analysis algorithms including K-Means, hierarchical clustering, and specialized clusterers. These group data points with similar characteristics.
- The Association tab enables association rule mining via algorithms like Apriori and FP-Growth to uncover relationships in data attributes.
- The Select Attributes tab provides filters and methods like Principal Component Analysis to select optimal attributes and reduce feature space dimensionality.

2. Getting Acquainted With WEKA:

2.1. Exploring Breast Cancer data set:

2.1.1. General findings:



2.1.1.a. The initial screen of Breast Cancer data

- The dataset contains 286 data instances.
- There are 10 attributes in this dataset.
- The 10th attribute, referred to as the class attribute, is used as the label for each instance and is responsible for categorizing instances into "no-recurrence-events" and "recurrence-events."

Meaning of all attributes are:

Age: Age is a categorical/nominal attribute where the categories are defined by age ranges. It is not a true numeric attribute, even though the values look like numbers.

Menopause :Indicates the patient's menopausal status with 3 possible values:

- lt40: Less than 40 years old, premenopausal
- ge40: 40 years old or older, perimenopausal/postmenopausal
- premeno: Had premature menopause before age 40

Tumor-size: It denotes the tumor's dimension and is sorted into different size intervals, including "0-4," "5-9," and similar categories.

Inv-nodes: Number of axillary lymph nodes invaded by cancer cells based on examination of the resected lymph nodes after surgery.

- Axillary lymph nodes are the lymph nodes located in the armpit/axilla region which drain the breast area
- When breast cancer spreads, these are often the first lymph nodes to which it can metastasize

- After breast cancer surgery, the removed lymph nodes are examined for signs of cancer cells
- Inv-nodes indicates the number of lymph nodes out of those examined that tested positive for cancer
- Higher numbers indicate larger spread of the breast cancer to the lymphatic system

Node-caps: Indicates whether cancer cells were found in the lymph node capsule. The possible values are:

- yes: Cancer cells were present in the capsule surrounding the involved lymph node(s)
- no: Cancer cells were not found in the capsule of the involved lymph node(s)

The lymph node capsule is a thick layer of connective tissue that surrounds lymph nodes. If cancer cells penetrate the capsule and spread to surrounding tissues, it is considered more advanced disease. The cancer is no longer contained in the node.

Therefore, the node-caps attribute provides additional information about the extent of cancer spread to the lymph nodes:

- yes means the cancer has started breaking out of the lymph node capsule, suggesting more aggressive spread
- no means the cancer was detected in the lymph node but has not yet penetrated the capsule

Deg-malig: Degree of malignancy of the breast tumor based on microscopic examination of the cells and tissue.

Reported on a scale of 1 to 3

Higher score indicates a more aggressive cancer and poorly differentiated cells

Scale:

- 1 - Well-differentiated cells (look mostly normal)
- 2 - Moderately differentiated cells (some abnormal features)
- 3 - Poorly differentiated cells (very abnormal, aggressive cancer)

The degree of malignancy gives information about how abnormal the cancer cells and tissue look under a microscope. Well-differentiated tumors still retain a lot of normal cell structure. Poorly differentiated indicates the cells and tissue have lost a lot of normal structure, suggesting a more aggressive cancer.

The deg-malig score is determined by a pathologist examining a tumor sample. It provides insight into the microscopic composition and behavior of the breast cance

Breast(left or right):

Breast - Indicates whether the breast tumor occurred in the left or right breast. The possible values are:

- left
- right

Knowing the laterality of the tumor is important because:

It allows correlating the tumor with other laterality-specific factors like breastfeeding history. Breastfeeding predominantly from one breast can change cancer risks.

It indicates if it is a new primary tumor or a metastasis from the opposite breast cancer.

Left/right breast anatomy can have slight differences that influence tumor properties.

Surgery and radiation therapy planning depends on tumor laterality.

So in summary, the breast attribute records the laterality of the primary tumor which can provide information about the origin, correlations to other risk factors, and determine treatment planning. Tracking if it is left vs right is meaningful both clinically and for data analysis.

Breast-quad: Indicates the specific quadrant of the breast where the tumor was located. The possible values are:

- left_up
- left_low
- right_up
- right_low
- central

The breast quadrants are defined as:

- Upper outer quadrant
- Lower outer quadrant
- Upper inner quadrant
- Lower inner quadrant
- Central portion behind the nipple

Knowing the breast quadrant helps determine tumor stage and surgery/radiation planning. Key points:

- Tumors in central/nipple area have higher risk of spreading to nipple and skin
- Outer quadrant tumors can spread to axillary lymph nodes
- Inner tumors can spread to internal mammary lymph nodes
- Quadrant impacts required surgical margins and radiation fields

In summary, the breast-quad attribute provides precise location information to correlate with tumor properties and determine proper treatment.

Irradiat: Indicates whether the patient received radiation therapy as part of their breast cancer treatment. The possible values are:

- yes: The patient did receive radiation therapy.
- no: The patient did not receive radiation therapy.

Radiation therapy is often given after surgery to lower the risk of cancer recurrence in the breast area. It may also be used to provide relief from symptoms or as the main treatment if surgery is not possible.

Key points about the irradiat attribute:

- Radiation therapy impacts prognosis and recurrence risk.
- The fields and dosage of radiation depend on tumor properties like location, size, and spread.
- Tracking radiation therapy is needed to correlate with treatment outcomes.
- Radiation side effects and interactions must be considered in ongoing care.

In summary, the binary irradiat attribute indicates if radiation therapy was part of the breast cancer treatment regimen, which provides useful clinical information about the patient's care and follow-up needs.

Class - The outcome classification of whether the breast cancer patient had a recurrence event during the follow-up period. The possible values are:

- no-recurrence-events: The patient did not have a recurrence of breast cancer during the follow-up window.
- recurrence-event: The patient did have a confirmed recurrence of breast cancer in that time frame.

Recurrence means the breast cancer has returned after previous treatment. This can occur months or years later, either in the original site or elsewhere in the body.

Some key points:

- The recurrence classification window is defined for each patient based on their follow-up duration.
- Recurrence indicates treatment failure and a worse prognosis overall.
- Predicting recurrence risk helps guide treatment intensity and follow-up schedules

2.1.2. Missing valua and solution:

Missing values:

- The node-caps attribute has missing values for 8 of the 286 instances.
- The breast-quad attribute has 1 missing value out of 286 instances.

Solution:

- Omitting rows with missing values (complete case analysis):
  - Drop any rows that have one or more missing values
  - Pros: Simple, unbiased analysis on complete cases
  - Cons: Loss of information, reduced sample size
- Imputing with mean/median/mode:
  - Replace missing numeric values with mean/median of that column
  - Replace missing categorical values with most frequent value (mode)
  - Pros: Retains more data, simple substitutions
  - Cons: Can introduce bias, inappropriate imputations
- Using a model to predict missing values:

- Build ML model to predict missing values from other attributes
- E.g. regression for numeric, classification for categorical
- Pros: Uses correlations and patterns in data
- Cons: Complex, still some bias, error in predictions

- Assigning a special value:
  - Set all missing values to a dummy value like -999 or 'NA'
  - Pros: Maintains completeness, accommodates analysis
  - Cons: Can still bias, special handling needed in analysis

The best approach depends on the dataset and analysis goals. Key is to consider tradeoffs between retaining data and introducing bias.

Fixing specific missing data in this dataset:

- For node-caps, we could try imputing based on the known correlation between node-caps and other attributes like node-inv and deg-malig. However, a quicker way that utilize the weka GUI is to open the filter option.
  - Below is the filter box before replacing missing value, and the node-caps before replacing the missing value.

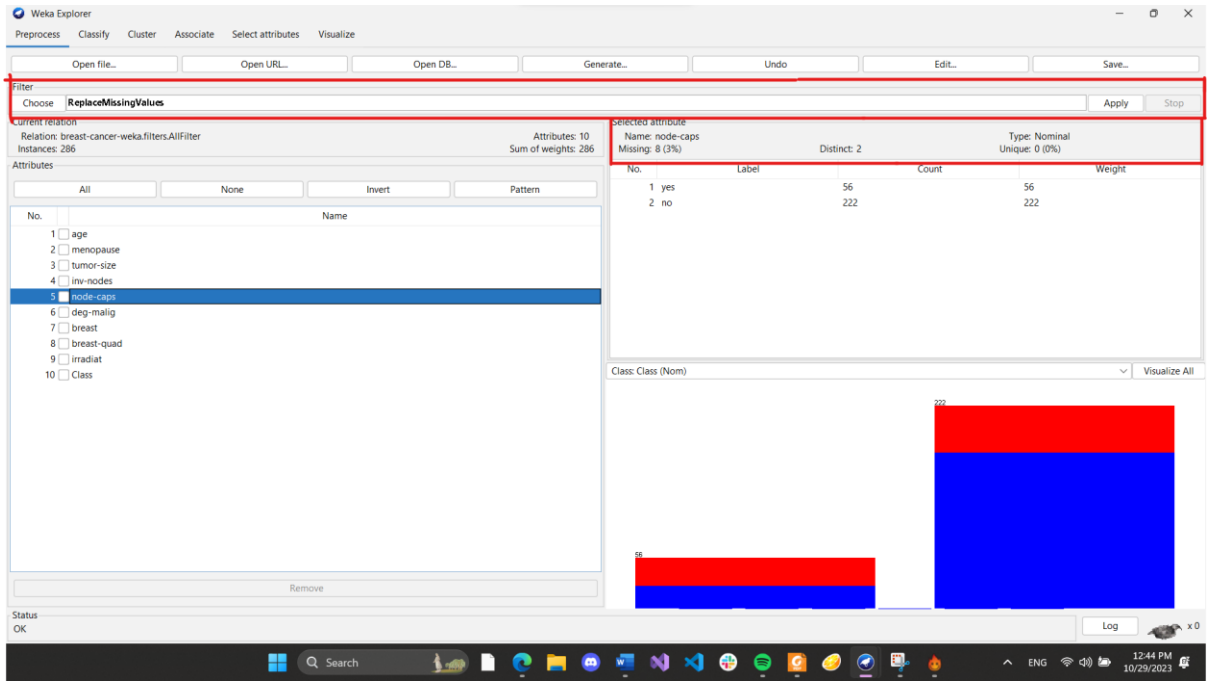


Image 2.1.2.a. Node-caps values before replacing missing value

- Open filter, open choose-> filters-> unsupervised->ReplaceMissingValues

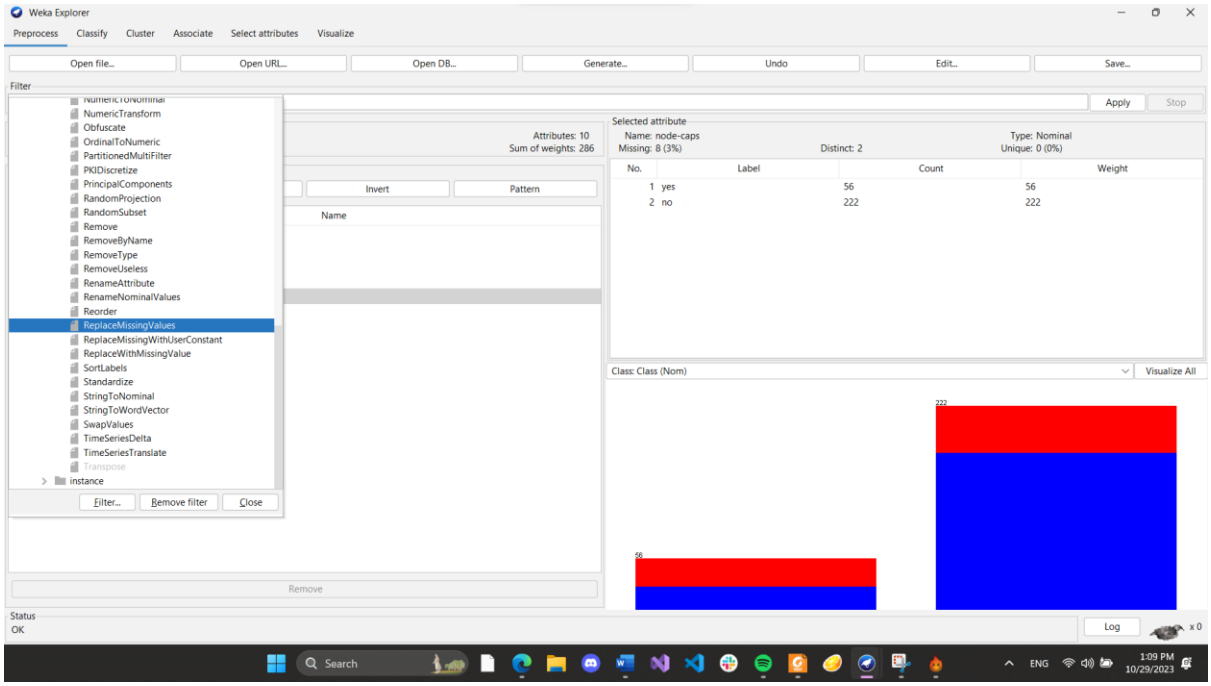


Image 2.1.2.b

- After applying filter:

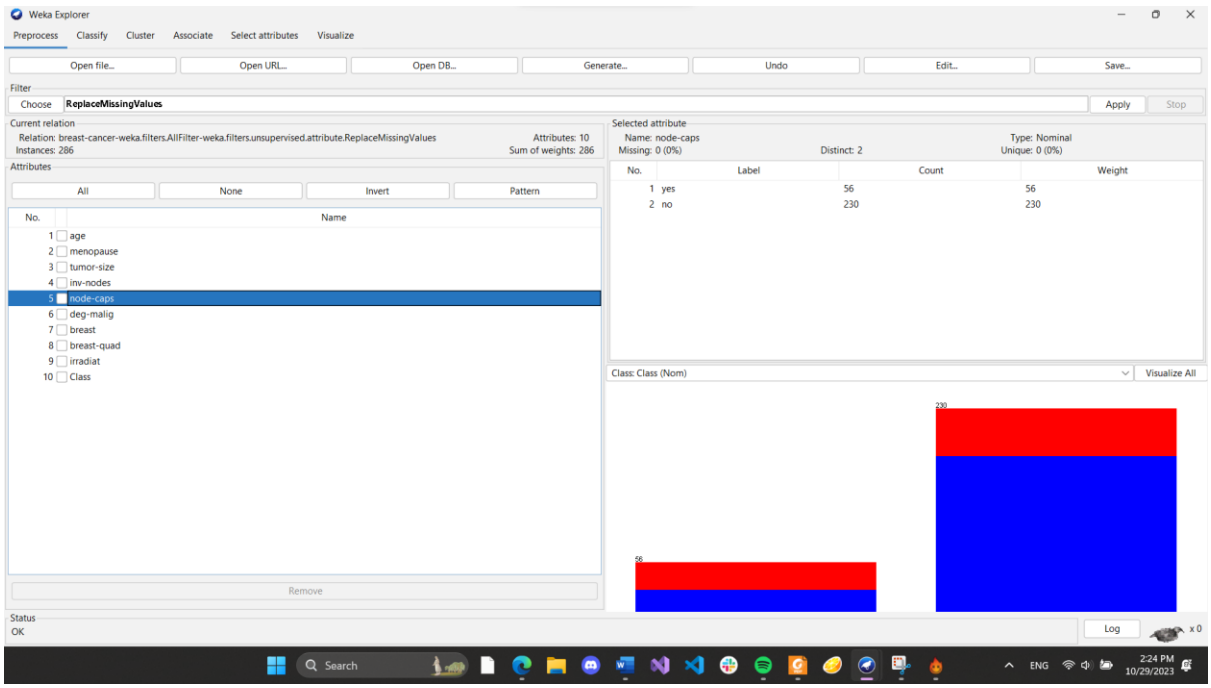


Image 2.1.2.c

- For breast-quad, with only 1 missing value, we could likely just omit that single row without much downside.
  - Click on ***edit*** button -> find the row has the missing value -> left click to choose that row -> right click -> ***delete selected instance***

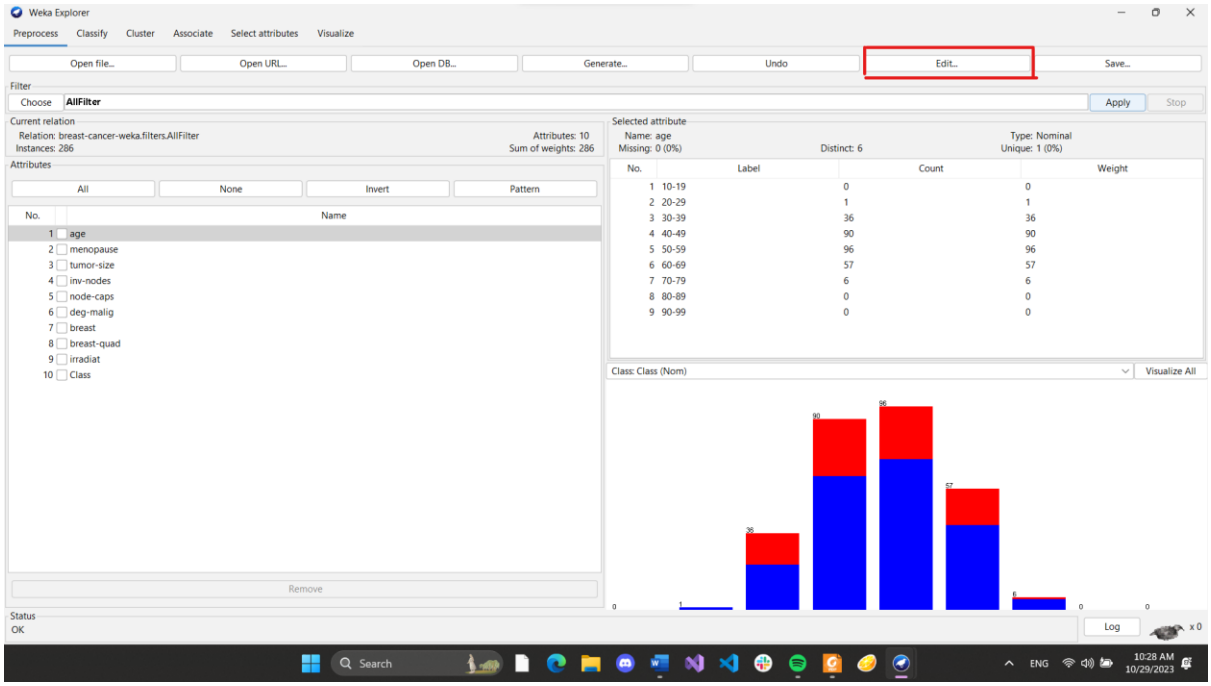


Image 2.1.2.d. Opening edit window

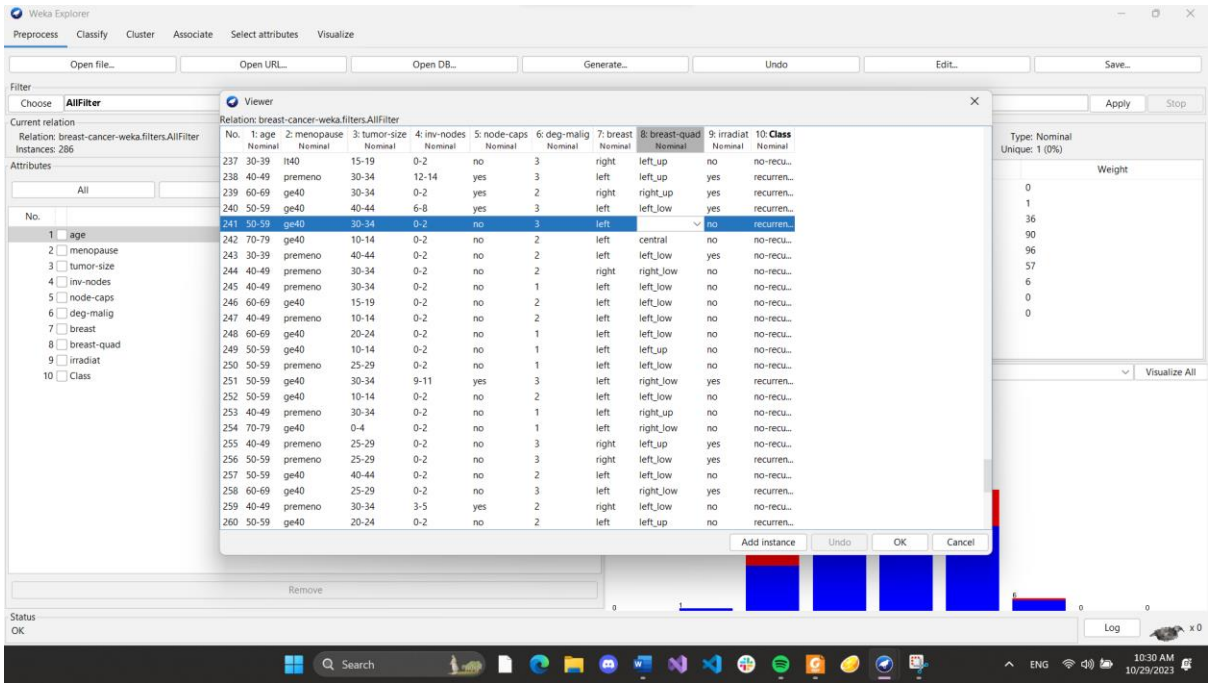


Image 2.1.2.e. Selecting instance with missing value.

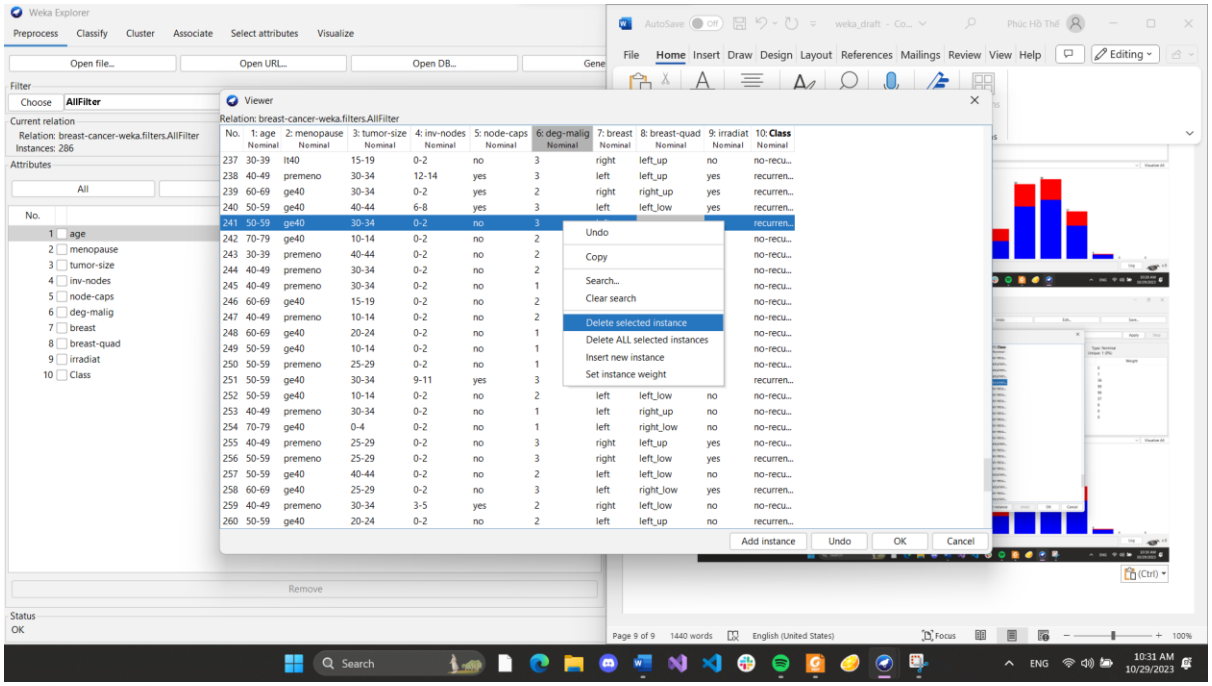


Image 2.1.2.e. right-click for more options to edit data and click on delete

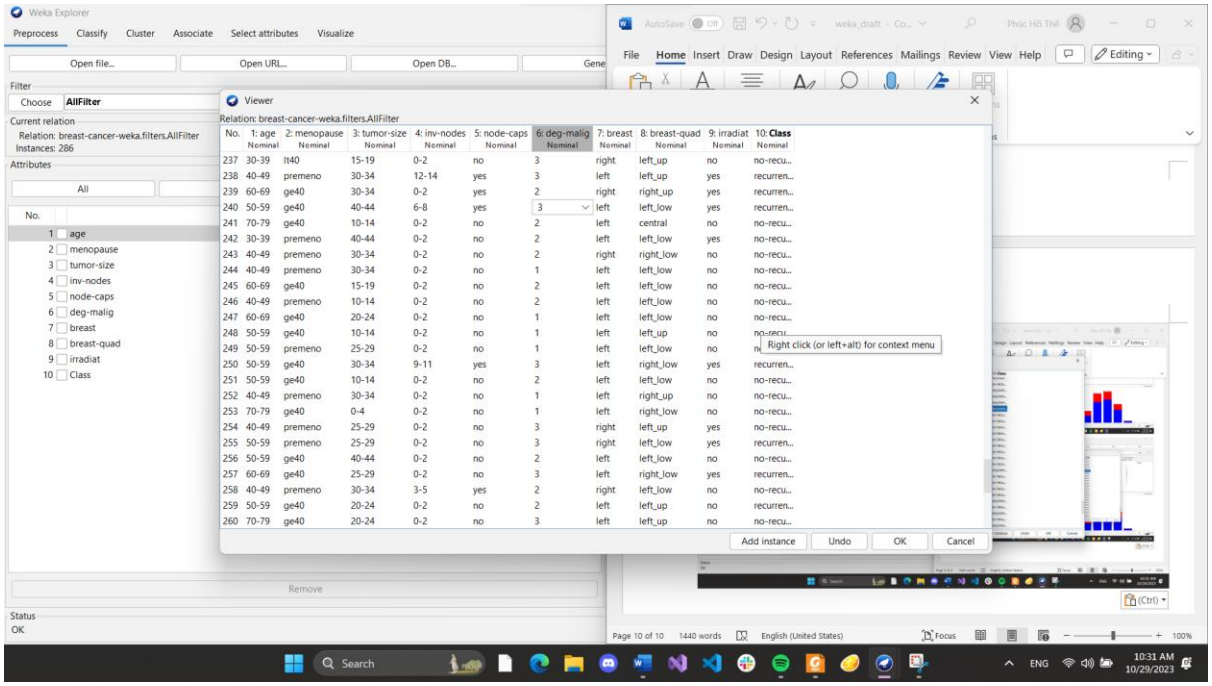


Image 2.1.2.f. Dataset after omission

Explains the meaning of the chart in the WEKA Explorer:

- Distribution plots are available for each attribute, offering a visual assessment of value distribution.
- Choosing an attribute in the "Attributes" panel generates a corresponding distribution plot in the "Selected attribute" panel.
- The x-axis illustrates the range of values within the selected attribute.
- The y-axis represents frequency, displaying the count of instances for each attribute value.

- For categorical attributes, the plot employs color-coding to distinguish positive instances in red and negative instances in blue, aiding in identifying class imbalances.
- These distribution plots aid in qualitatively evaluating attribute characteristics, supporting exploratory data analysis before quantitative modeling.
- In essence, WEKA's attribute distribution visualizations provide a graphical overview of value frequencies, spreads, and class balances, serving as a valuable preliminary tool for understanding dataset properties and guiding subsequent modeling steps.

2.2. Exploring Weather data set:

General information:

- The weather dataset contains 14 instances and 5 attributes.
- The **outlook** attribute is categorical, representing qualitative conditions like sunny or overcast.
- The **temperature** and **humidity** attributes are numerical, capturing numeric measurements.
- The **play** attribute serves as the binary target label, indicating if conditions are suitable for play (yes or no).
- 5- number summary of the dataset:

Five-number summary	**Temperature**	**Humidity**
The smallest observation	64.0	65.0
The lower quartile (Q1)	68.5	70.0
The median (Q2)	72.0	82.5
The upper quartile (Q3)	80.5	90.5
The largest observation	85.0	96.0

  - Weka does not directly output numeric summaries, but manual inspection suggests temperature ranges from 64 to 85 with lower and upper quartiles around 68.5 and 80.5. Meanwhile, humidity spans 65 to 96, with the bulk of values falling between 70 and 90.5.

- all charts of this data set are of type bar chart:

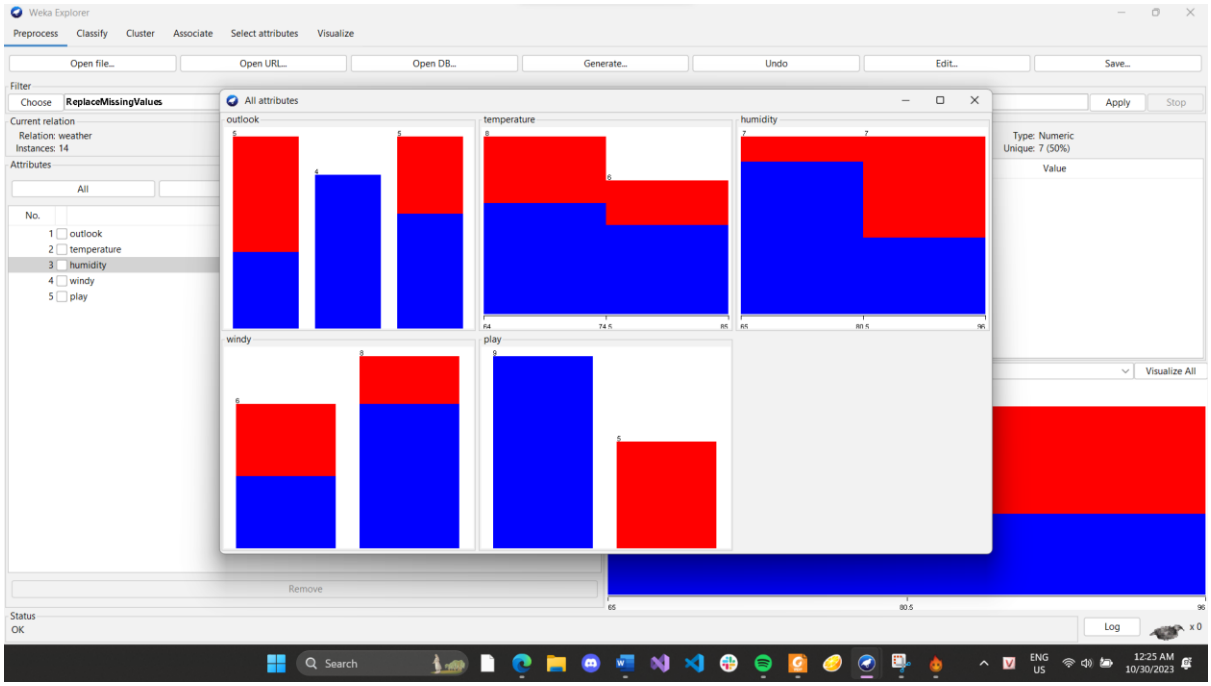


Image2.2.a. Visualization of all attributes

Explanation of all attributes:

- Outlook:
  - Categorical variable with 3 possible values: sunny, overcast, raining
  - Chart has 3 columns representing the distribution of instances across each value
  - Column color coding indicates play=no (red) and play=yes (blue) outcomes within each value
  - Sunny outlook has mostly play=yes (blue column)
  - Overcast also has predominantly play=yes (blue)
  - Raining outlook has primarily play=no (red column)
  - Therefore, sunny and overcast conditions strongly favor play=yes decisions
- Temperature:
  - Continuous numeric attribute
  - Chart splits full range into 2 bins: cool temps (64-74.5) and warm temps (74.5-85)
  - Compares play decisions between the 2 temperature ranges
  - Warm temps column has more play=yes (blue) than play=no (red)
  - Cool temps more balanced, with slightly more play=no (red)
  - Indicates warmer temperatures are associated with more play=yes choices
- Humidity:
  - Continuous numeric attribute

- Divided into 2 ranges: low humidity (65-80.5) and high humidity (80.5-96)
- Columns display play=no (red) and play=yes (blue) counts within each range
- High humidity column is predominantly play=no (red)
- Low humidity more balanced, but still slightly more play=no (red)
- Suggests high humidity strongly correlated with not playing

- Windy:
  - Categorical with 2 values: true, false
  - Columns represent distribution of instances across values
  - Windy=true column has majority play=no (red)
  - Windy=false column is mostly play=yes (blue)
  - Indicates windy conditions deter playing, while calm favors it

Visualize tag:

Chart name: scatter plot matrix (SPLOM) A scatter plot matrix is a grid of scatter plots where each combination of attributes is plotted against each other. This type of chart is useful for visualizing relationships and correlations between numerical attributes in your dataset. By examining the scatter plots, you can get a sense of how different attributes relate to each other and whether there are any patterns or correlations between them.

correlation :

after examining three attributes that are most likely correlated(**temperature, humidity and windy**), there are no correlation that can be reached.

2.3. Exploring Credit in Germany data set:

The content of the comment section in the ARFF file often contains metadata and additional information about the dataset:

- **Author's Information**: This section typically includes the name of the author, institute, university, department, and its address.
- **Dataset Information**: Information about the dataset itself, possibly describing the nature of the data, its source, purpose, or any unique characteristics it holds.
- **Algorithms Information and How to Use It**: Instructions or details about algorithms used with the dataset and guidance on how to utilize the data for analysis, machine learning, or other purposes.

Attribute used for the label: **Class**

Describing attributes:

1. Status of Existing Checking Account (Qualitative):
   - This attribute represents the status of an applicant's existing checking account.
   - It is a categorical variable with several categories:
     - Account balance less than 1,000 DM (Deutsche Mark).

        - Account balance between 1,000 and 2,000 DM.
        - Account balance greater than or equal to 2,000 DM.
        - No checking account (applicant does not have a checking account).
    - The status of an applicant's checking account can be an important indicator of their financial stability and their ability to manage credit.
2. Duration in Months (Numerical):
    - This attribute represents the duration of the credit in months.
    - It is a numerical variable that indicates how long the applicant intends to have the credit.
    - The duration is an essential factor in assessing the risk associated with the credit, as longer durations may indicate greater exposure to risk.
3. Credit History (Qualitative):
    - This attribute describes the applicant's credit history.
    - It is a categorical variable with several categories:
        - No credits taken or all credits paid back duly.
        - All credits at this bank have been paid back duly.
        - Existing credits have been paid back duly until now.
        - There has been a delay in paying off in the past.
        - Critical account with other credits existing (not at this bank).
    - Credit history is a critical factor in assessing the creditworthiness of an applicant, as it reflects their past behavior in managing credit obligations.
4. Purpose (Qualitative):
    - This attribute describes the purpose for which the credit is sought.
    - It is a categorical variable with various categories representing different purposes:
        - Car (new).
        - Car (used).
        - Furniture/equipment.
        - Radio/television.
        - Domestic appliances.
        - Repairs.
        - Education.
        - Vacation (possibly does not exist).
        - Retraining.
        - Business.

- Others.

- The purpose of the credit can provide insights into how the funds will be used and is relevant for risk assessment and decision-making.

5. Credit Amount (Numerical):
    - This attribute represents the amount of credit requested by the applicant.
    - It is a numerical variable and denotes the monetary value of the credit.
    - The credit amount is a fundamental factor in assessing the applicant's financial needs and the level of risk associated with the credit.

Determine form of Continuous attribute:

Attribute	**Mean**	**Median**	**Left / Right skewed**
duration	20.903	18.0	Left skewed
credit_amount	3271.258	2319.5	Right skewed
installment_commitment	2.973	3.0	Symmetric
residence_since	2.845	3.0	Symmetric
age	35.546	33.0	Left skewed
existing_credits	1.407	1.0	Left skewed
num_dependents	1.155	1.0	Left skewed

-Select Attribute tab explanation:

In the 'Select Attribute' tab, the goal is to identify the most effective attribute subset for predictive modeling. This process involves configuring two essential components: the Attribute Evaluator and the Search Method.

The Attribute Evaluator is responsible for assigning a value or worth to each potential attribute subset, which is crucial for ranking their contributions to the predictive task.

The Search Method dictates the strategy for exploring attribute subsets.

In the 'Attribute Selection Mode' box, you have two options:

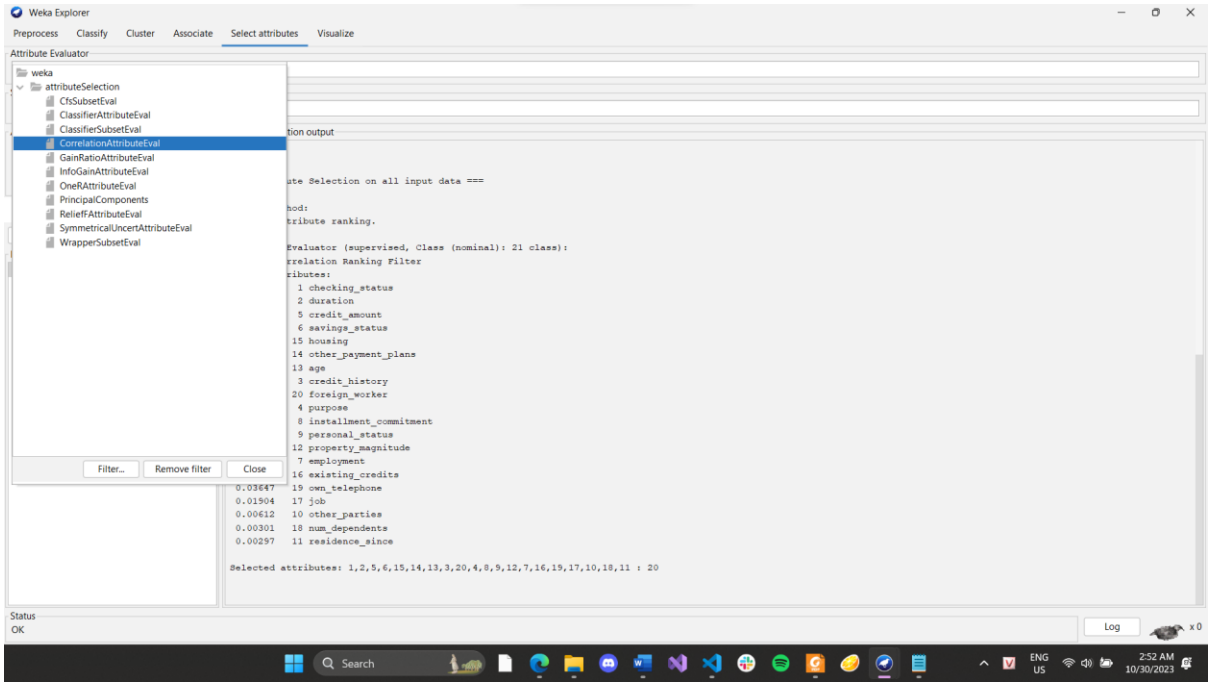
1. Use full training set: This method evaluates attribute subsets based on their performance with the entire training dataset.

2. Cross-validation: Cross-validation assesses attribute subsets by dividing the dataset into multiple subsets (folds) and testing their performance against various data partitions. The "Fold" and "Seed" parameters control the number of folds and data shuffling for cross-validation. This approach ensures robust generalization performance and mitigates overfitting risks.

-To select the 5 attributes with the highest correlation we should use CorrelationAttributeEval option in *Attribute Evaluator* box in *Select Attributes* tag

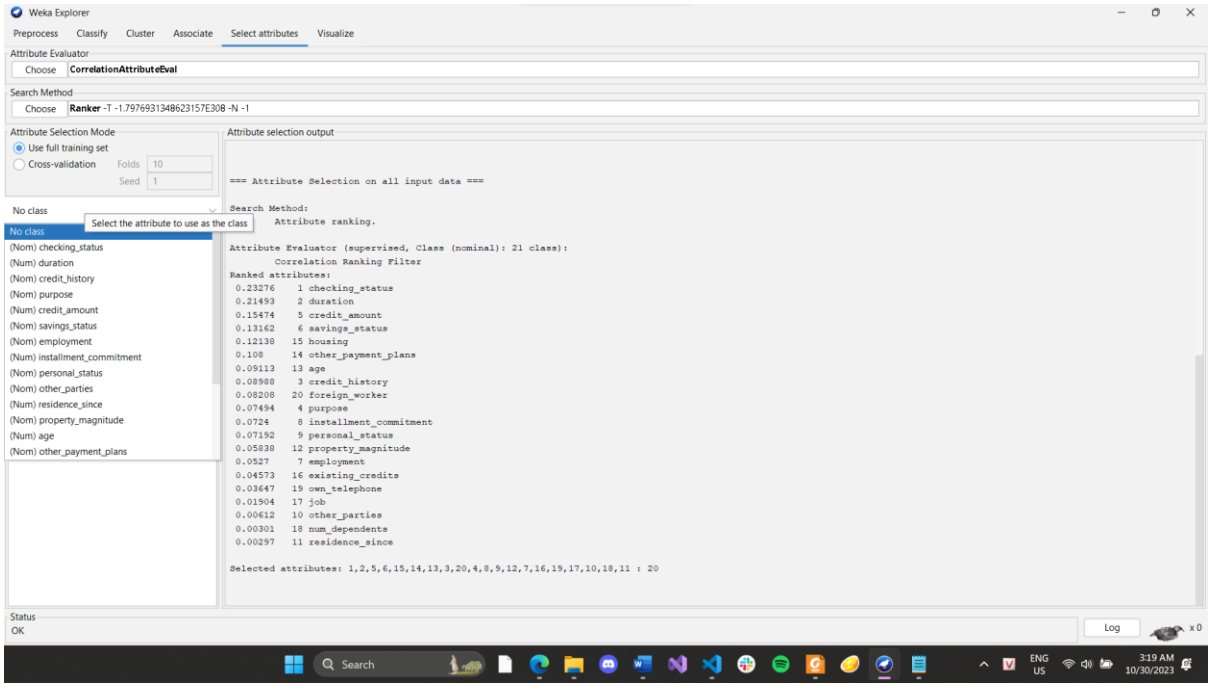
- -Step-by-step description:

Click on the ***Choose*** button in Attribute Evaluator box -> ***attributeSelection-> CorrelationAttributeEval***



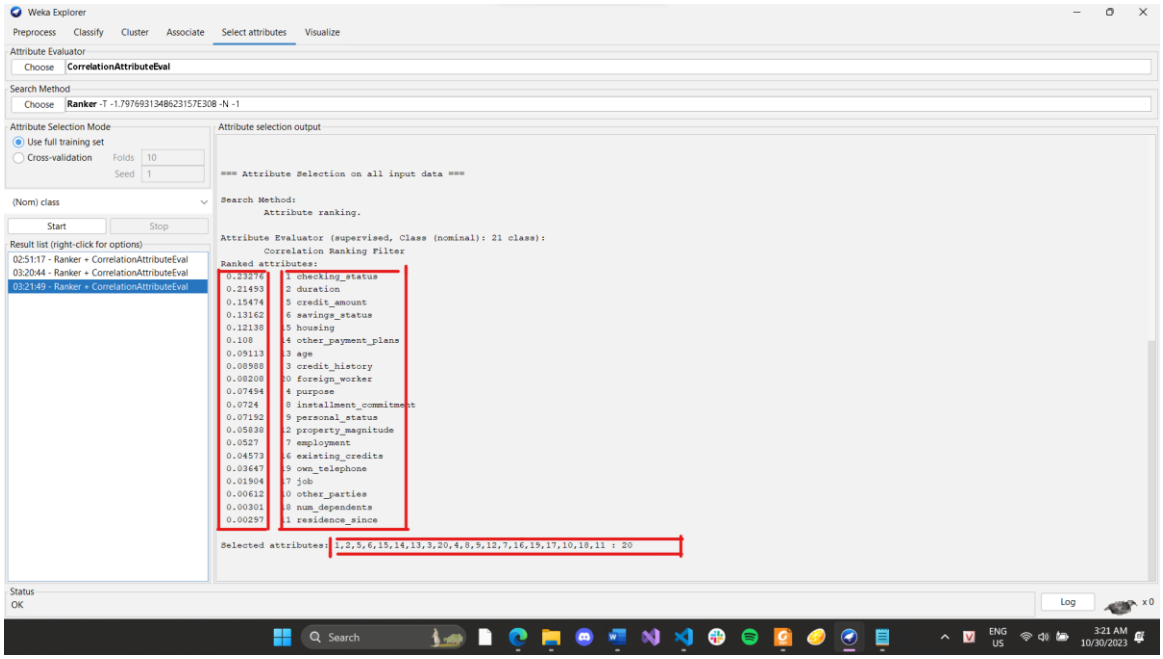
The choose options in ***Attribute Selection Mode*** has been explained before

Select the attribute that is used as the label.



After that, click on start to determine the rankings of attributes, based on correlation:

The five most correlated attributes are: checking_status, duration, credit_amount, savings_status, housing



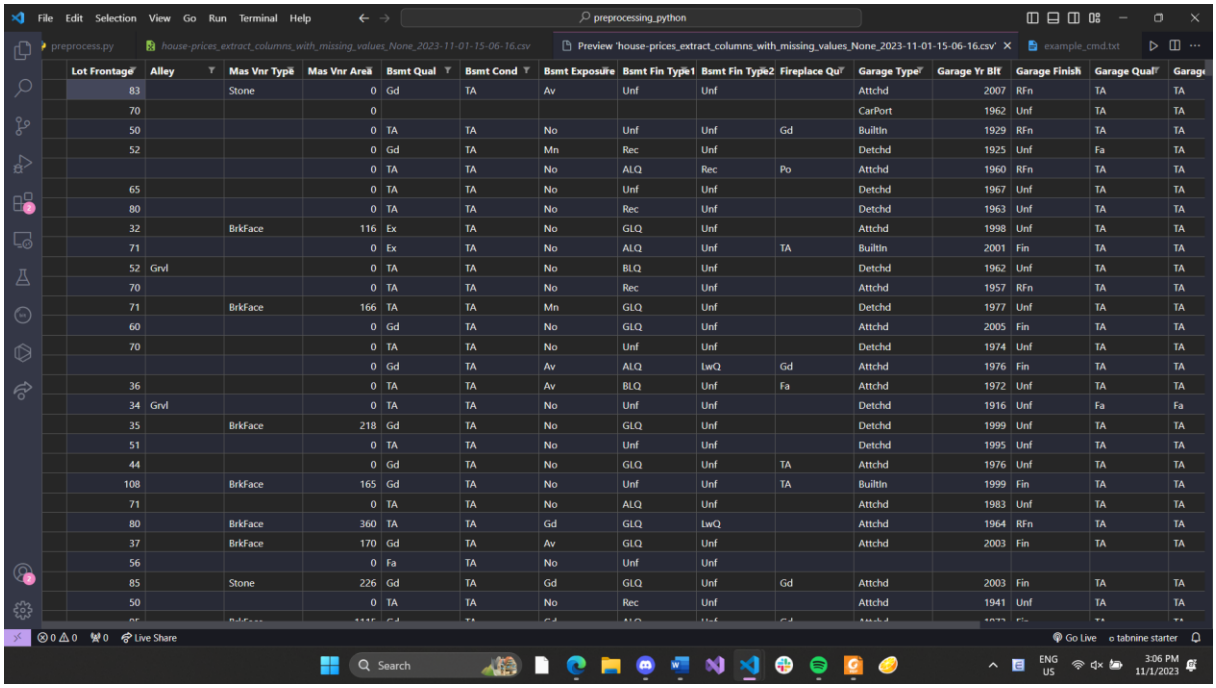
3. Preprocessing Data in Python

3.1. Extract columns with missing values

Prompt: *python preprocess.py house-prices.csv 1*

```
PS C:\Users\Admin\OneDrive\Desktop\preprocessing_python> python preprocess.py house-prices.csv 1
     LotFrontage Alley MasVnrType  MasVnrArea BsmtQual BsmtCond BsmtExposure  ... GarageYrBlt GarageFinish GarageQual GarageCond  PoolQC Fence MiscFeature
0           83.0   NaN      Stone         0.0       Gd       TA           Av  ...      2007.0          RFn         TA         TA     NaN   NaN         NaN
1           70.0   NaN        NaN         0.0      NaN      NaN          NaN  ...      1962.0          Unf         TA         TA     NaN   NaN         NaN
2           50.0   NaN        NaN         0.0       TA       TA           No  ...      1929.0          RFn         TA         TA     NaN   NaN         NaN
3           52.0   NaN        NaN         0.0       Gd       TA           Mn  ...      1925.0          Unf         Fa         TA     NaN   NaN         NaN
4            NaN   NaN        NaN         0.0       TA       TA           No  ...      1960.0          RFn         TA         TA     NaN  GdWo         NaN
..           ...   ...        ...         ...      ...      ...          ...  ...         ...          ...        ...        ...     ...   ...         ...
995         60.0   NaN        NaN         0.0       TA       TA           No  ...      1999.0          Fin         TA         TA     NaN   NaN         NaN
996          NaN   NaN    BrkFace       138.0       TA       TA           No  ...      1972.0          Fin         TA         TA     NaN   NaN         NaN
997         65.0   NaN        NaN         0.0       Gd       Gd           No  ...      2006.0          RFn         TA         TA     NaN   NaN         NaN
998         98.0   NaN    BrkFace       975.0       Gd       TA           No  ...      1998.0          Fin         TA         TA     NaN   NaN         NaN
999         75.0   NaN        NaN         0.0       TA       TA           Mn  ...      1965.0          Unf         TA         TA     NaN   NaN         NaN

[1000 rows x 18 columns]
```

3.1.a Commandline display of extracted values.



3.1.b Display of extracted values.

3.2. Count the number of lines with missing data
Prompt: *python preprocess.py house-prices.csv 2*

```
PS C:\Users\Admin\OneDrive\Desktop\preprocessing_python> python preprocess.py house-prices.csv 2
Number of instances with missing values: 1000
```

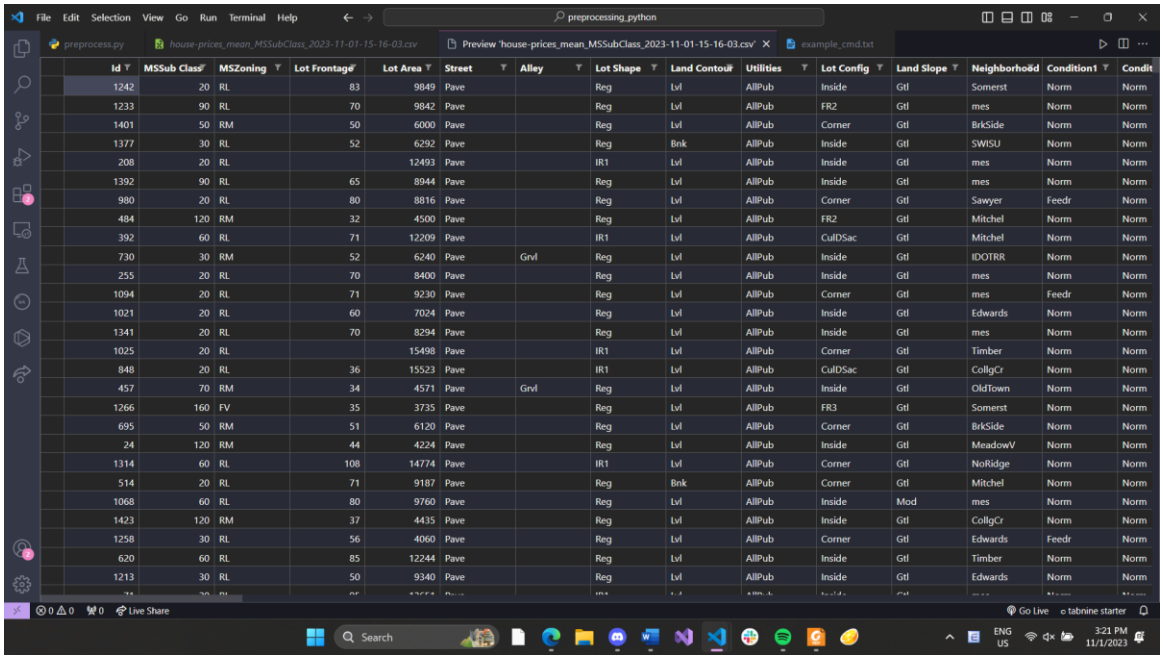
3.2 Commandline display of extracted values.

3.3. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute)
3.3.1. Example to fill missing data using the mean for the 'MSSubClass' column:

Prompt: *python preprocess.py house-prices.csv 3 --method mean --column MSSubClass*

The result file have this format: “house-prices_mean_MSSubClass_2023-11-01-15-16-03.csv”

- **input_file** is "house-prices.csv"
- **imputation_method** is "mean"
- **column_names** is "MSSubClass"
- **timestamp** is "2023-11-01-15-16-03"



3.3.1 Display of result file

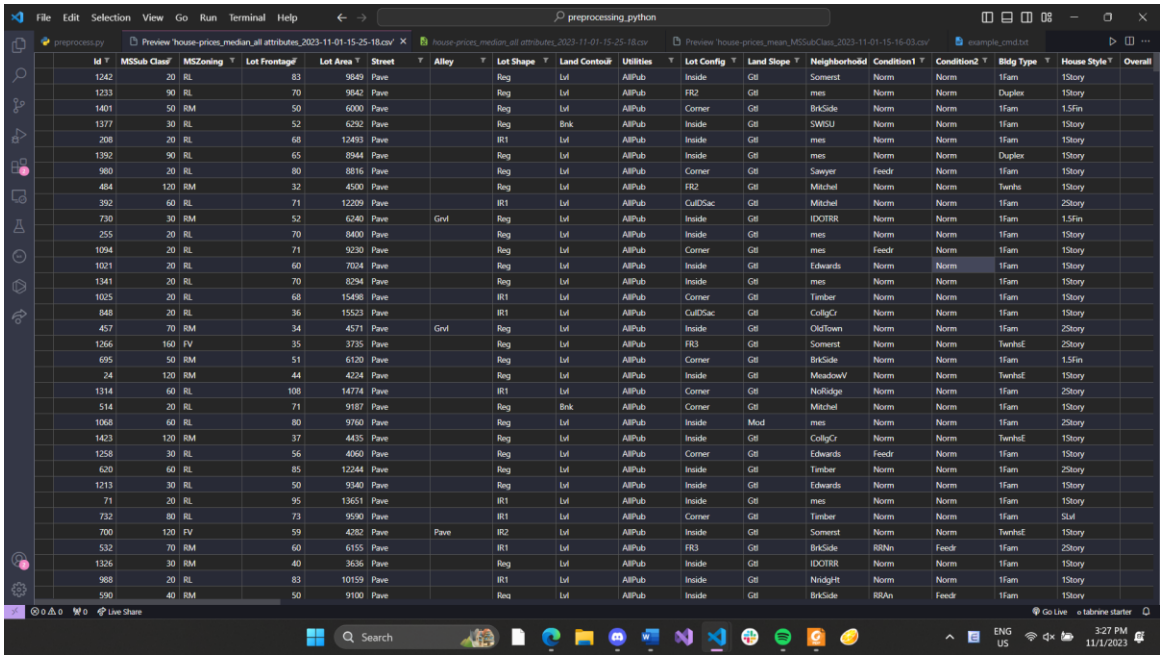
3.3.2. Example to fill missing data using the median for all column

Prompt: *python preprocess.py house-prices.csv 3 --method median*

The result file have this format: “house-prices_median_all_attributes_2023-11-01-15-25-18.csv”

Given the values:

- **input_file** is "house-prices.csv"
- **imputation_method** is "median"
- **column_names** is "all attributes"
- **timestamp** is "2023-11-01-15-25-18"



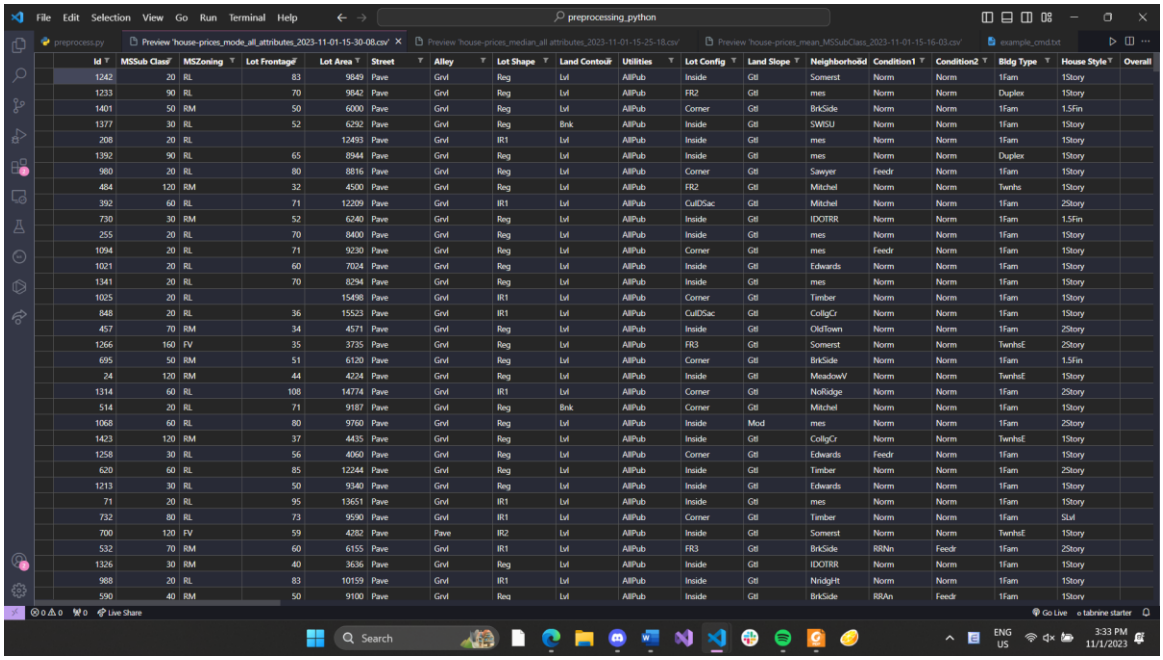
3.3.2 Display of result file

3.3.3. Example to fill missing data using the mode for all column

Prompt: python preprocess.py house-prices.csv 3 --method mode

house-prices_mode_all_attributes_2023-11-01-15-30-08

- **input_file** is "house-prices.csv"
- **imputation_method** is "mode"
- **column_names** is "all attributes"
- **timestamp** is "2023-11-01-15-30-08"



3.3.3. Display of result file

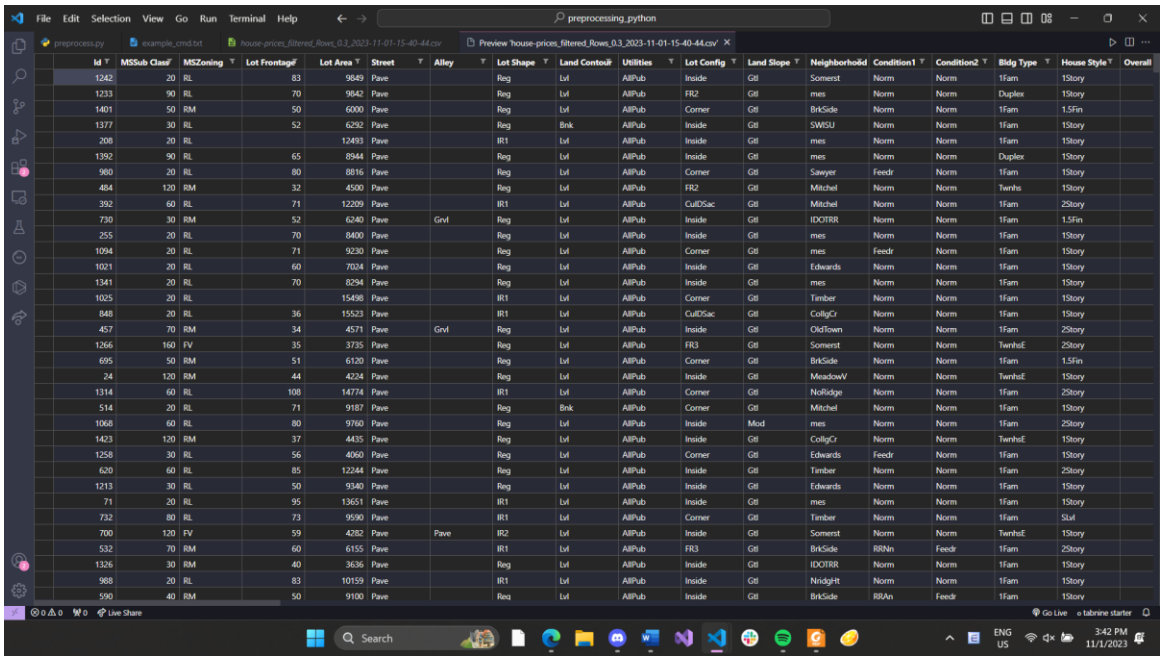
3.4. Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).
Example to delete rows with more than 30% missing values

Prompt: *python preprocess.py house-prices.csv 4 --threshold 0.2*

Output file: 'house-prices_filtered_Rows_0.3_2023-11-01-15-40-44.csv'

```
PS C:\Users\Admin\OneDrive\Desktop\preprocessing_python> python preprocess.py house-prices.csv 4 --threshold 0.2
        Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  ... Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice
0     1242          20       RL         83.0     9849   Pave   NaN      Reg  ...   NaN         NaN       0      6   2007      New       Partial    248328
1     1233          90       RL         70.0     9842   Pave   NaN      Reg  ...   NaN         NaN       0      3   2007       WD        Normal    101800
2     1401          50       RM         50.0     6000   Pave   NaN      Reg  ...   NaN         NaN       0      7   2008       WD        Normal    120000
3     1377          30       RL         52.0     6292   Pave   NaN      Reg  ...   NaN         NaN       0      4   2008       WD        Normal     91000
4      208          20       RL          NaN    12493   Pave   NaN      IR1  ...   GdWo        NaN       0      4   2008       WD        Normal    141000
..     ...         ...      ...          ...      ...    ...   ...      ...  ...   ...         ...     ...    ...    ...      ...           ...       ...
995   1190          60       RL         60.0     7500   Pave   NaN      Reg  ...   NaN         NaN       0      6   2010       WD        Normal    189000
996    192          60       RL          NaN     7472   Pave   NaN      IR1  ...   NaN         NaN       0      6   2007       WD        Normal    184000
997    990          60       FV         65.0     8125   Pave   NaN      Reg  ...   NaN         NaN       0      8   2006      New       Partial    197000
998    982          60       RL         98.0    12203   Pave   NaN      IR1  ...   NaN         NaN       0      7   2009       WD        Normal    336000
999    862         190       RL         75.0    11625   Pave   NaN      Reg  ...   NaN         NaN       0      4   2010       WD        Normal    131500

[995 rows x 81 columns]
```

3.4.a. Cmd display of result file



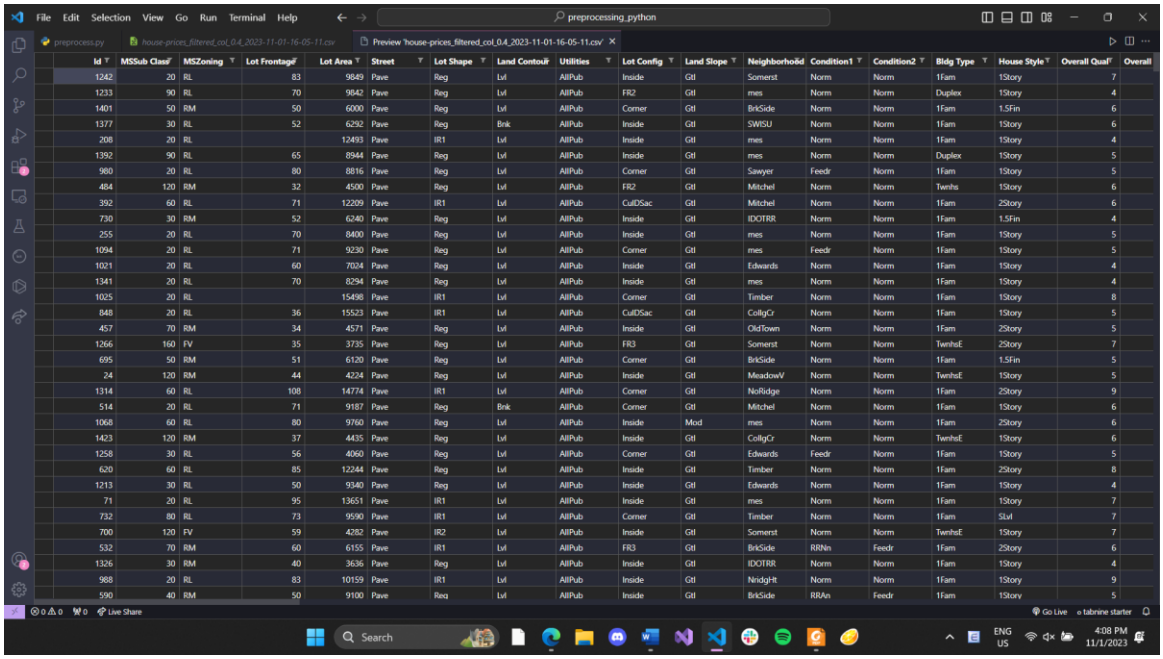
3.4.b Display of result file

3.5. Deleting columns containing more than a particular number of missing values
Example to delete columns with more than 40% missing values

Prompt : *python preprocess.py house-prices.csv 5 --threshold 0.4*

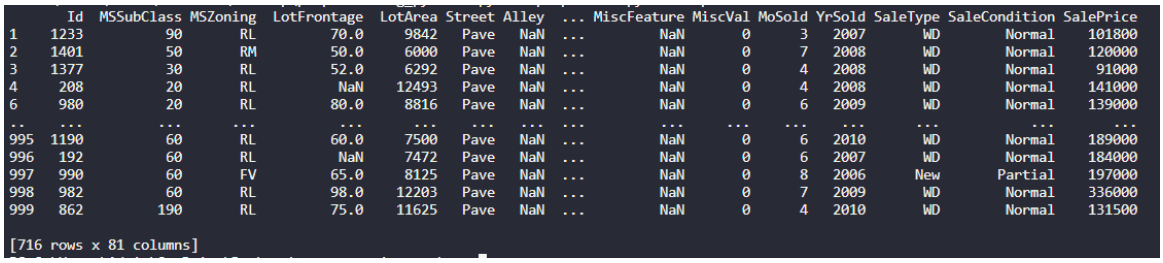
```
PS C:\Users\Admin\OneDrive\Desktop\preprocessing_python> python preprocess.py house-prices.csv 5 --threshold 0.4
     Id  MSSubClass MSZoning  LotFrontage  LotArea Street LotShape LandContour Utilities LotConfig  ... EnclosedPorch 3SsnPorch ScreenPorch PoolArea MiscVal MoSold  YrSold  SaleType  SaleCondition  SalePrice
0  1242          20       RL         83.0     9849   Pave      Reg         Lvl    AllPub    Inside  ...             0         0           0        0       0      6    2007       New        Partial     248328
1  1233          90       RL         70.0     9842   Pave      Reg         Lvl    AllPub       FR2  ...             0         0           0        0       0      3    2007        WD         Normal     101800
2  1401          50       RM         50.0     6000   Pave      Reg         Lvl    AllPub    Corner  ...           112         0           0        0       0      7    2008        WD         Normal     120000
3  1377          30       RL         52.0     6292   Pave      Reg         Bnk    AllPub    Inside  ...             0         0           0        0       0      4    2008        WD         Normal      91000
4   208          20       RL          NaN    12493   Pave      IR1         Lvl    AllPub    Inside  ...             0         0           0        0       0      4    2008        WD         Normal     141000
..  ...         ...      ...          ...      ...    ...      ...         ...       ...       ...  ...           ...       ...         ...      ...     ...    ...     ...       ...            ...        ...
995 1190          60       RL         60.0     7500   Pave      Reg         Lvl    AllPub    Inside  ...             0         0           0        0       0      6    2010        WD         Normal     189000
996  192          60       RL          NaN     7472   Pave      IR1         Lvl    AllPub    CulDSac ...             0         0           0        0       0      6    2007        WD         Normal     184000
997  990          60       FV         65.0     8125   Pave      Reg         Lvl    AllPub    Inside  ...             0         0           0        0       0      8    2006       New        Partial     197000
998  982          60       RL         98.0    12203   Pave      IR1         Lvl    AllPub    Corner  ...             0         0           0        0       0      7    2009        WD         Normal     336000
999  862         190       RL         75.0    11625   Pave      Reg         Lvl    AllPub    Inside  ...             0         0           0        0       0      4    2010        WD         Normal     131500

[1000 rows x 75 columns]
```

3.5.a cmd display of result file



3.5.b display of result file

3.6. Delete duplicate samples.
Prompt: *python preprocess.py house-prices.csv 6*



```
      Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley  ... MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice
1   1233          90       RL         70.0     9842   Pave   NaN  ...         NaN       0      3   2007       WD        Normal    101800
2   1401          50       RM         50.0     6000   Pave   NaN  ...         NaN       0      7   2008       WD        Normal    120000
3   1377          30       RL         52.0     6292   Pave   NaN  ...         NaN       0      4   2008       WD        Normal     91000
4    208          20       RL          NaN    12493   Pave   NaN  ...         NaN       0      4   2008       WD        Normal    141000
6    980          20       RL         80.0     8816   Pave   NaN  ...         NaN       0      6   2009       WD        Normal    139000
..   ...         ...      ...          ...      ...    ...   ...  ...         ...     ...    ...    ...      ...           ...       ...
995 1190          60       RL         60.0     7500   Pave   NaN  ...         NaN       0      6   2010       WD        Normal    189000
996  192          60       RL          NaN     7472   Pave   NaN  ...         NaN       0      6   2007       WD        Normal    184000
997  990          60       FV         65.0     8125   Pave   NaN  ...         NaN       0      8   2006      New       Partial    197000
998  982          60       RL         98.0    12203   Pave   NaN  ...         NaN       0      7   2009       WD        Normal    336000
999  862         190       RL         75.0    11625   Pave   NaN  ...         NaN       0      4   2010       WD        Normal    131500

[716 rows x 81 columns]
```

3.6. display of result file

Omitted 284 duplicated rows.

3.7. Normalize a numeric attribute using min-max and Z-score methods.
3.7.1. Example to scale the 'MSSubClass' column using Min-Max scaling

Prompt : *python preprocess.py house-prices.csv 7 --method min_max --column MSSubClass*

MSSub Class
20
90
50
30
20
90
20
120
60
30
20
20
20
20
20
20
70
160
50
120
60
20
60
120
30
60
30
20
80
120
70
30
20
40

3.7.1.a. Before scaling

MSSub Class
0
0.41
0.18
0.06
0
0.41
0
0.59
0.24
0.06
0
0
0
0
0
0
0.29
0.82
0.18
0.59
0.24
0
0.24
0.59
0.06
0.24
0.06
0
0.35
0.59
0.29
0.06
0
0.12

3.7.1.b. After scaling

3.7.2. Example to scale the 'LotFrontage' column using Z-score standardization

Prompt:*python preprocess.py house-prices.csv 7 --method z-score --column LotFrontage*

Lot Frontage
83
70
50
52
65
80
32
71
52
70
71
60
70
36
34
35
51
44
108
71
80
37
56
85
50
95
73
59
60
40
83
50

3.7.a. before scaling

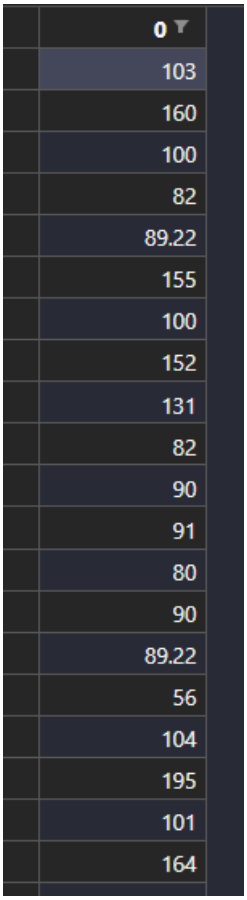
Lot Frontage
0.65
0.04
-0.9
-0.81
-0.2
0.51
-1.75
0.08
-0.81
0.04
0.08
-0.43
0.04
-1.56
-1.66
-1.61
-0.86
-1.19
1.82
0.08
0.51
-1.51
-0.62
0.74
-0.9
1.21
0.18
-0.48
-0.43
-1.37
0.65
-0.9

3.7.b. after scaling

3.8. Performing addition, subtraction, multiplication, and division between two numerical attributes.

Example to perform addition between 'LotFrontage' and 'MSSubClass' columns

Prompt: *python preprocess.py house-prices.csv 8 --operation add --attr1 LotFrontage --attr2 MSSubClass*



0
103
160
100
82
89.22
155
100
152
131
82
90
91
80
90
89.22
56
104
195
101
164

3.8. Results

III. References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

García, S., Luengo, J., & Herrera, F. (2012). Data preprocessing in data mining. In Intelligent data analysis for real-life applications: Theory and practice (pp. 1-37). IGI Global.

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

Patro, S. G. K., & Sahu, K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. Communications of the ACM, 53(6), 59-67.