TEAM F

BIG DATA SYSTEMS FOR MODEL DEVELOPMENT

AGENDA

0. Model Development Approach
1. Automated Data Cleaning
2. Human Assisted Data Cleaning
3. Automated Method Comparison and Choosing
4. Human Assisted Method Picking
5. Automated Dummy Creation and Transformation

MODEL DEVELOPMENT APPROACH

- PYTHON
- 3 TRAINING DATASETS

DEV AND OOTO

USED FOR FRAUD DETECTION
DEV: CSV 82 VARIABLES INCLUDING TARGET, 865 OBS
OOTO: CSV 81 VARIABLES, 2968 OBS
INCLUDES NULL VALUES
BINARY, CATEGORICAL, NUMERIC

MBD_FA2

USED FOR VARIABLE AND RATIO CREATION
CSV WITH 47 VARIABLES INCLUDING TARGET
5951 OBSERVATIONS
INCLUDES NULL VALUES
BINARY, CATEGORICAL AND NUMERIC DATA

DATA

AUTOMATED DATA CLEANING

TASK 1

NAN (NOT A NUMBER)

ID	NUM1	NUM2	NUM3
1	3.56	0.55	
2	2.30	1.60	0.44
3		4.05	3.00
4	0.34	7.00	1.22

SUBSTITUTE NAN FOR 0

ID	NUM1	NUM2	NUM3
1	3.56	0.55	0.00
2	2.30	1.60	0.44
3	0.00	4.05	3.00
4	0.34	7.00	1.22

MISSING VALUES

ID	TOWN
1	MAD
2	BCN
3	NYC
4	BCN
5	
6	NYC
7	BCN

SUBSTITUTE NONE FOR

THE HIGHEST FREQUENCY VALUE

ID	TOWN
1	MAD
2	BCN
3	NYC
4	BCN
5	BCN
6	NYC
7	BCN

OUTLIER DETECTION

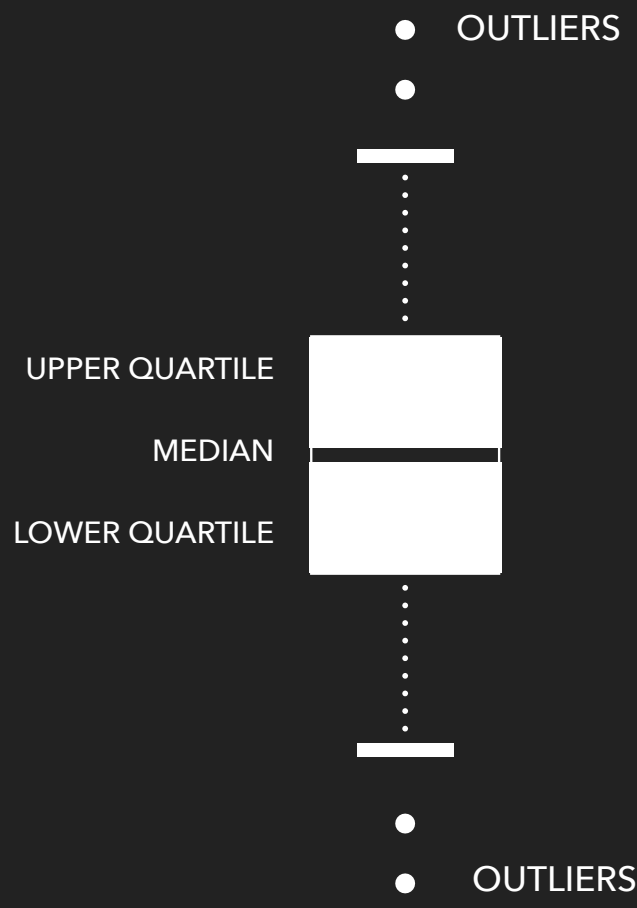
ANYTHING OUTSIDE MEAN +/- 3 * STANDARD DEVIATION WILL BE CONSIDERED AN OUTLIER.

WE CONVERT DATA VALUES TO STANDARD DEVIATIONS FROM THE MEAN:

```
def deviations(x, mean, stddev):

    return math.abs(x - mean) / stddev
```

ABSOLUTE VALUES WHICH ARE HIGHER THAN 3 WILL BE CONSIDERED OUTLIERS



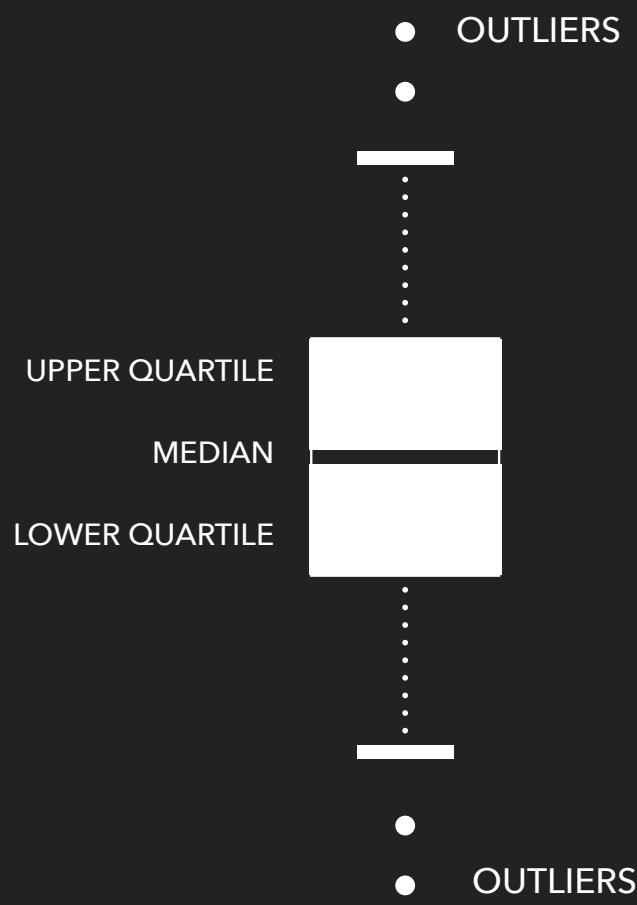
HANDLING OUTLIERS

- IF THE NEXT NON-OUTLIER VALUES IS CLOSE:

TRANSFORM OUTLIERS TO THE NEXT HIGHEST/LOWEST VALUE

- IF THE NEXT NON-OUTLIER VALUE IS FAR:

DROP THE OUTLIERS AND INFORM ABOUT IT



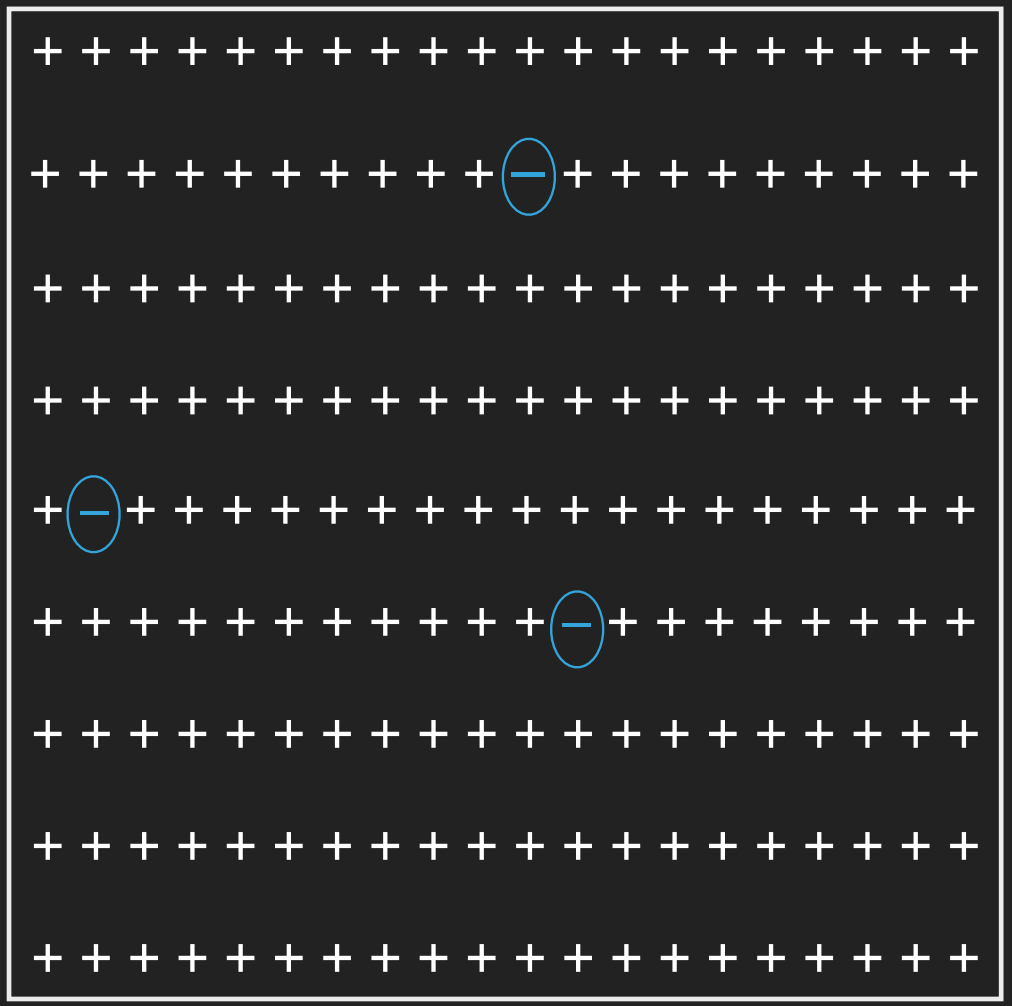
HOW TO DETERMINE WHAT IS FAR?

IF THE OUTLIER IS HIGHER THAN 2 * (MEAN + 3 STDEV)

UNRELIABLE VALUES

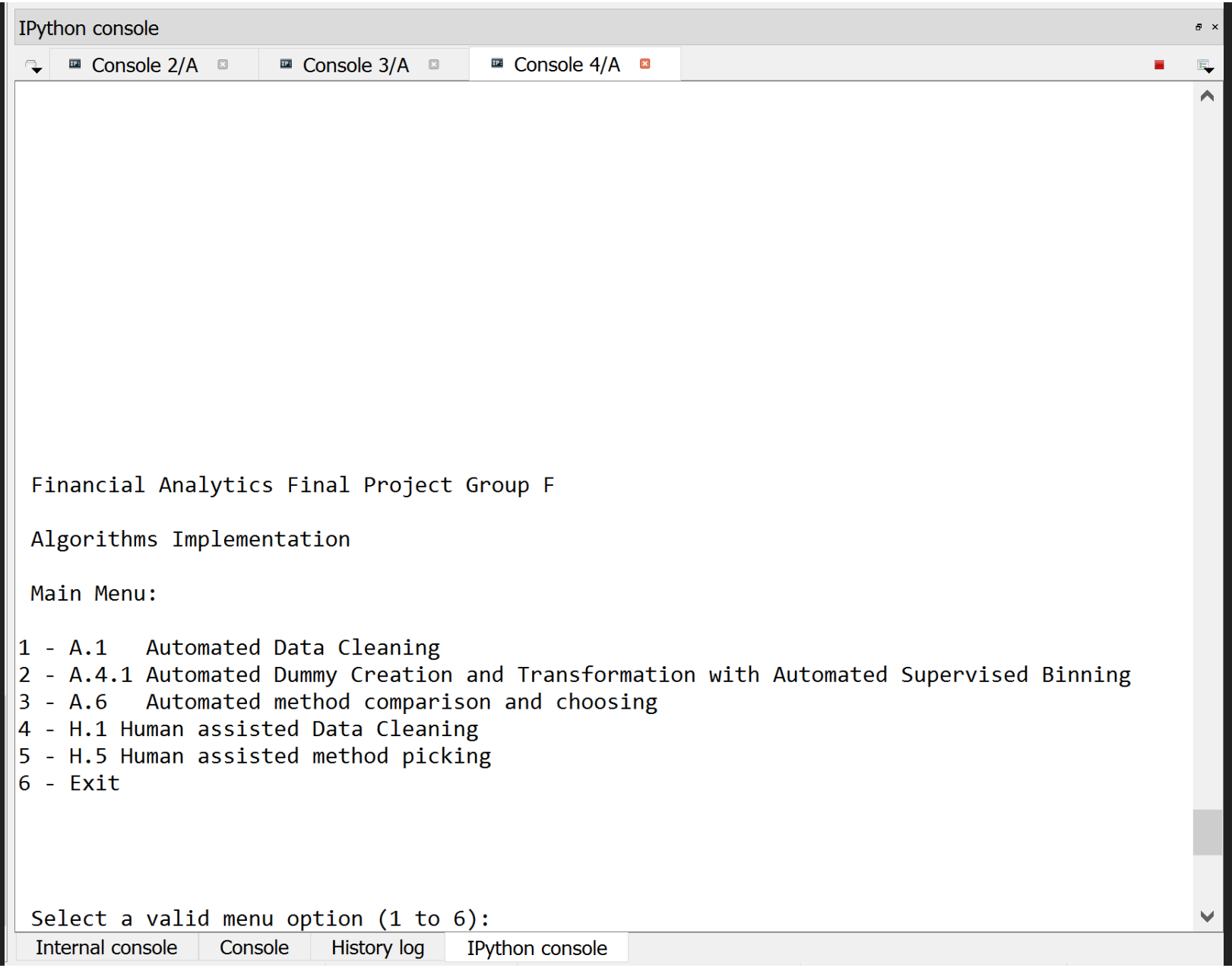
Threshold

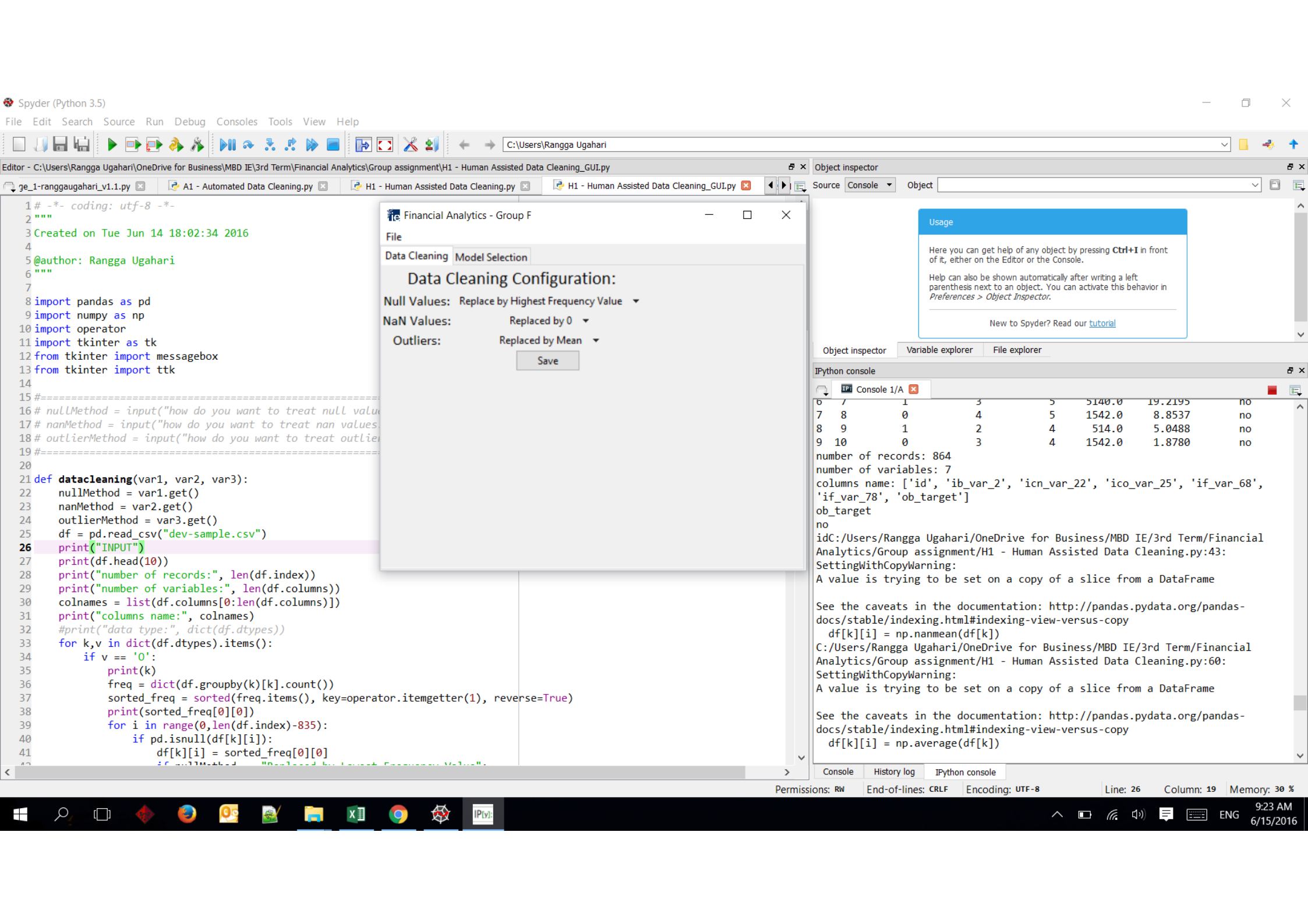
5%

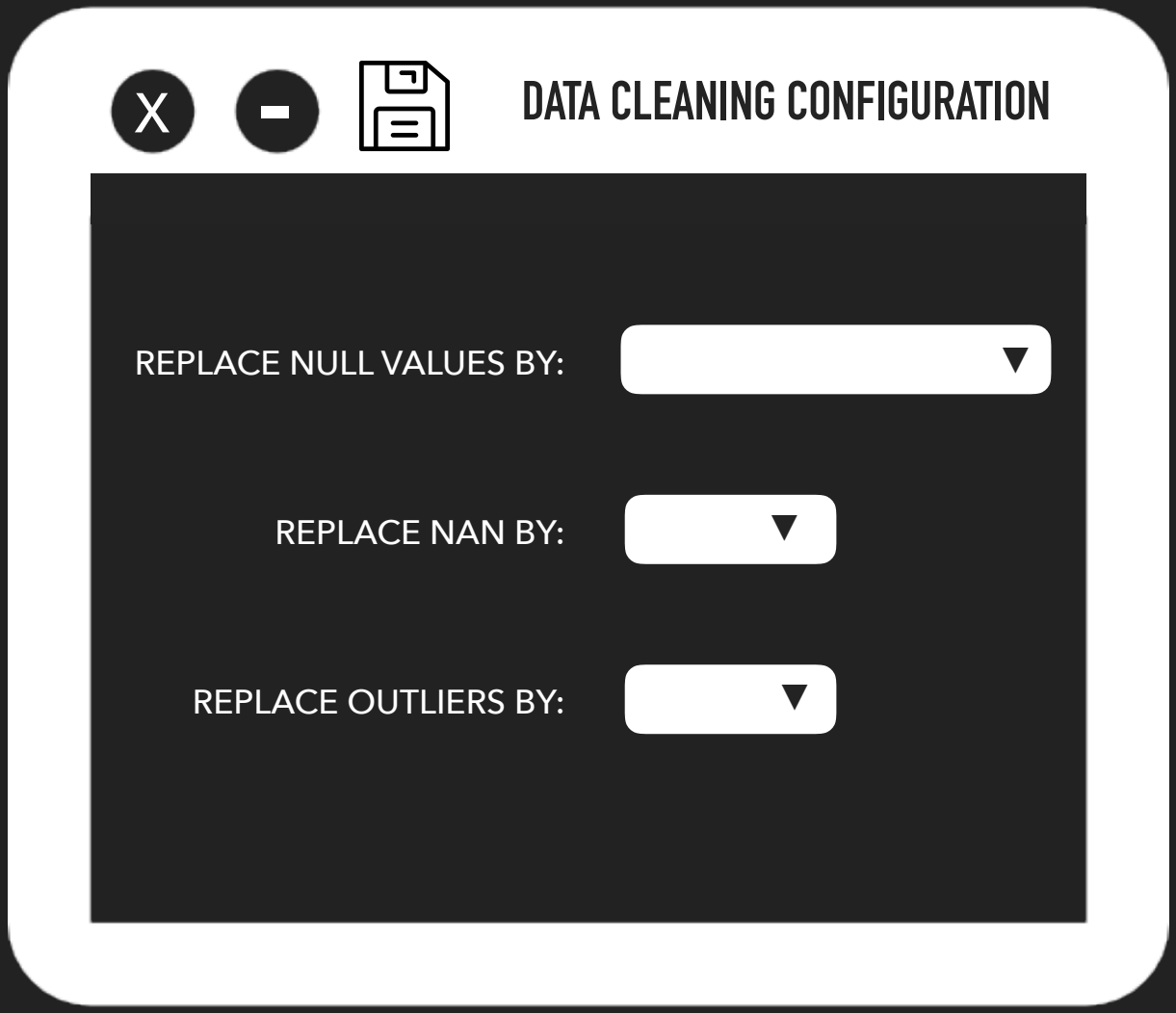


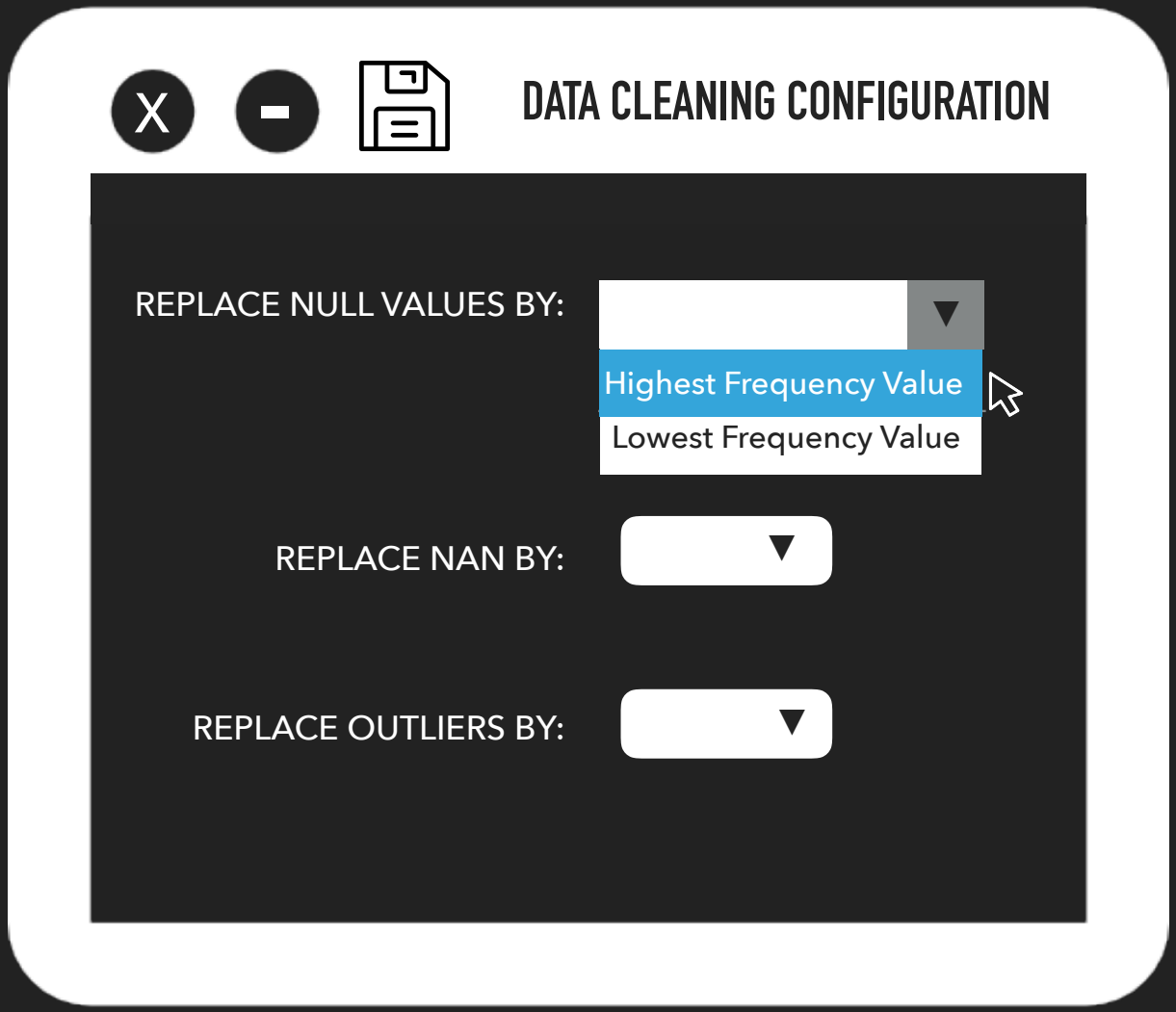
HUMAN ASSISTED DATA CLEANING

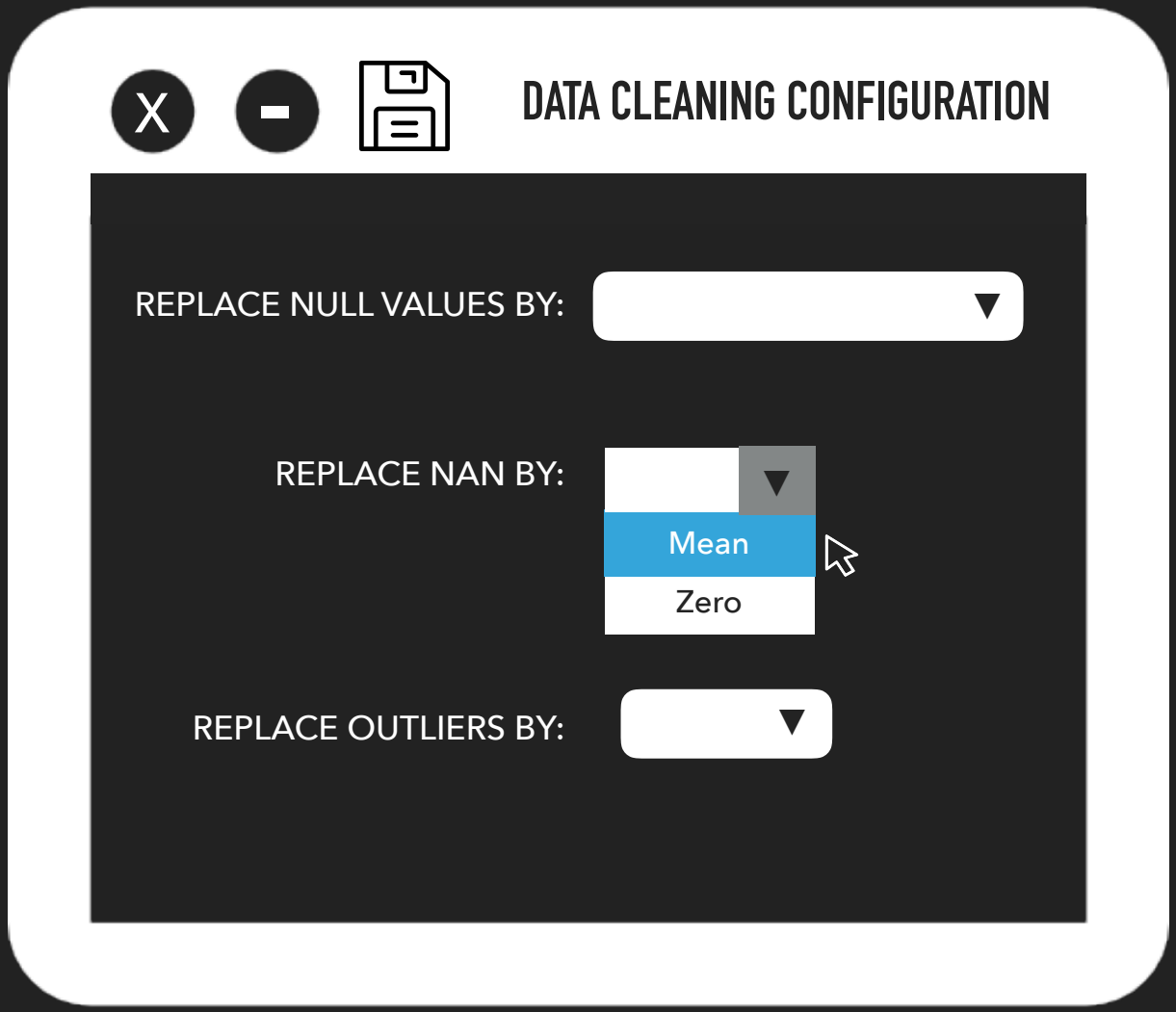
TASK 2

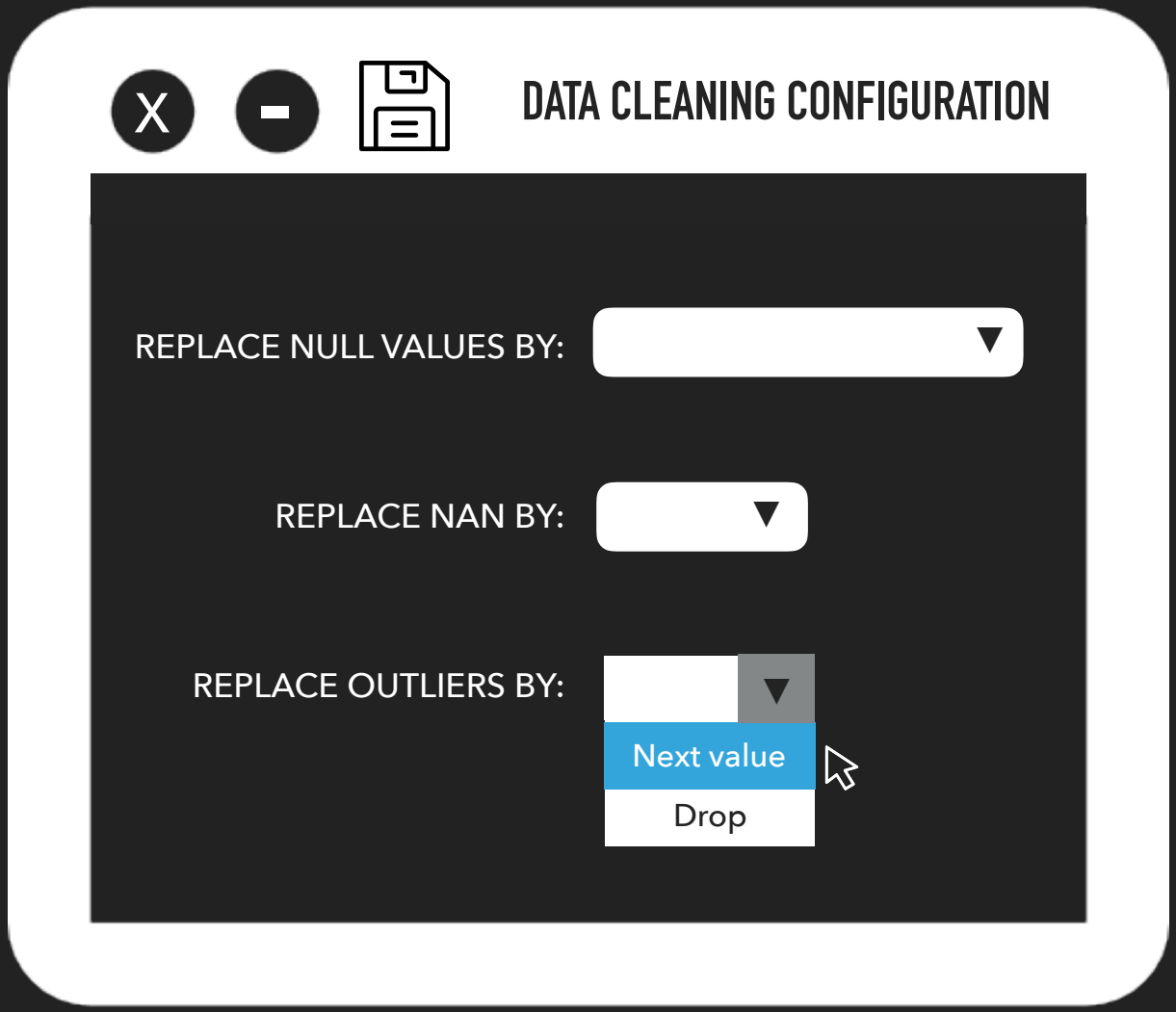

IPython console
Console 2/A
Console 3/A
Console 4/A
Financial Analytics Final Project Group F
Algorithms Implementation
Main Menu:
1 - A.1 Automated Data Cleaning
2 - A.4.1 Automated Dummy Creation and Transformation with Automated Supervised Binning
3 - A.6 Automated method comparison and choosing
4 - H.1 Human assisted Data Cleaning
5 - H.5 Human assisted method picking
6 - Exit
Select a valid menu option (1 to 6):
Internal console
Console
History log
IPython console

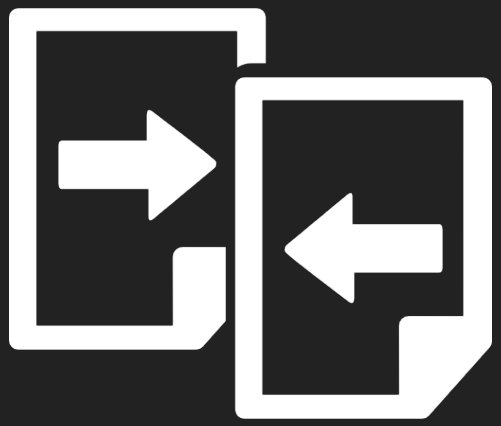

Spyder (Python 3.5)
File Edit Search Source Run Debug Consoles Tools View Help
C:\Users\Rangga Ugahari
Editor - C:\Users\Rangga Ugahari\OneDrive for Business\MBD IE\3rd Term\Financial Analytics\Group assignment\H1 - Human Assisted Data Cleaning_GUI.py
ge_1-ranggaugahari_v1.1.py
A1 - Automated Data Cleaning.py
H1 - Human Assisted Data Cleaning.py
H1 - Human Assisted Data Cleaning_GUI.py
-*- coding: utf-8 -*-
"""
Created on Tue Jun 14 18:02:34 2016
@author: Rangga Ugahari
"""
import pandas as pd
import numpy as np
import operator
import tkinter as tk
from tkinter import messagebox
from tkinter import ttk
# nullMethod = input("how do you want to treat null valu
# nanMethod = input("how do you want to treat nan values
# outlierMethod = input("how do you want to treat outlie
def datacleaning(var1, var2, var3):
nullMethod = var1.get()
nanMethod = var2.get()
outlierMethod = var3.get()
df = pd.read_csv("dev-sample.csv")
print("INPUT")
print(df.head(10))
print("number of records:", len(df.index))
print("number of variables:", len(df.columns))
colnames = list(df.columns[0:len(df.columns)])
print("columns name:", colnames)
#print("data type:", dict(df.dtypes))
for k,v in dict(df.dtypes).items():
if v == 'O':
print(k)
freq = dict(df.groupby(k)[k].count())
sorted_freq = sorted(freq.items(), key=operator.itemgetter(1), reverse=True)
print(sorted_freq[0][0])
for i in range(0,len(df.index)-835):
if pd.isnull(df[k][i]):
df[k][i] = sorted_freq[0][0]
Financial Analytics - Group F
File
Data Cleaning
Model Selection
Data Cleaning Configuration:
Null Values:
Replace by Highest Frequency Value
NaN Values:
Replaced by 0
Outliers:
Replaced by Mean
Save
Object inspector
Source
Console
Object
Usage
Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.
Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Object Inspector.
New to Spyder? Read our tutorial
Object inspector
Variable explorer
File explorer
IPython console
Console 1/A
6 7 1 3 5 5140.0 19.2195 no
7 8 0 4 5 1542.0 8.8537 no
8 9 1 2 4 514.0 5.0488 no
9 10 0 3 4 1542.0 1.8780 no
number of records: 864
number of variables: 7
columns name: ['id', 'ib_var_2', 'icn_var_22', 'ico_var_25', 'if_var_68', 'if_var_78', 'ob_target']
ob_target
no
idC:/Users/Rangga Ugahari/OneDrive for Business/MBD IE/3rd Term/Financial Analytics/Group assignment/H1 - Human Assisted Data Cleaning.py:43:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
df[k][i] = np.nanmean(df[k])
C:/Users/Rangga Ugahari/OneDrive for Business/MBD IE/3rd Term/Financial Analytics/Group assignment/H1 - Human Assisted Data Cleaning.py:60:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
df[k][i] = np.average(df[k])
Console
History log
IPython console
Permissions: RW
End-of-lines: CRLF
Encoding: UTF-8
Line: 26
Column: 19
Memory: 30 %
ENG
9:23 AM
6/15/2016


DATA CLEANING CONFIGURATION
REPLACE NULL VALUES BY:
REPLACE NAN BY:
REPLACE OUTLIERS BY:


DATA CLEANING CONFIGURATION
REPLACE NULL VALUES BY:
Highest Frequency Value
Lowest Frequency Value
REPLACE NAN BY:
REPLACE OUTLIERS BY:


DATA CLEANING CONFIGURATION
REPLACE NULL VALUES BY:
REPLACE NAN BY:
Mean
Zero
REPLACE OUTLIERS BY:


DATA CLEANING CONFIGURATION
REPLACE NULL VALUES BY:
REPLACE NAN BY:
REPLACE OUTLIERS BY:
Next value
Drop



AUTOMATED METHOD COMPARISON AND CHOOSING

TASK 3

MODEL COMPARISON

- SUPPORT VECTOR MACHINE (SVM)
- RANDOM FOREST
- GENERALISED LINEAR MODELS (GLM) - LINEAR REGRESSION

SUPPORT VECTOR MACHINE (SVM)

Non-probabilistic binary linear classifier

Combines aspects of both nearest neighbour classifier and linear regression modeling

RANDOM FOREST

Bagging with random feature selection to add additional diversity to the decision tree models

As the ensemble uses only a small, random portion of the full feature set, random forests can handle extremely large datasets

GENERALISED LINEAR MODELS (GLM) LINEAR REGRESSION

Dependent (Y) and independent variables (X1,X2....XN)

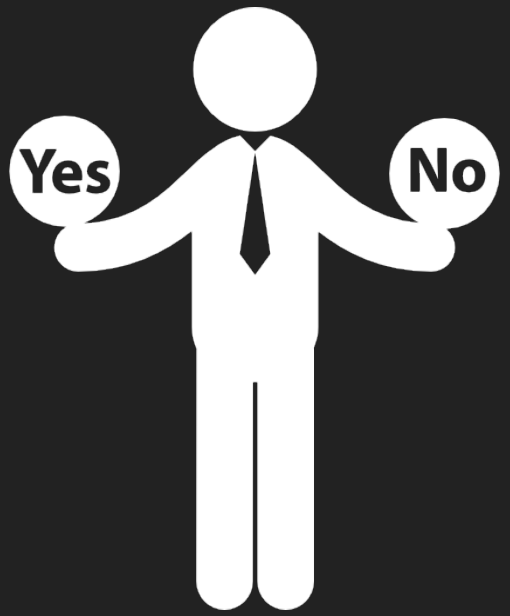
The relationship between the independent and dependent variables follows a straight line

MODEL SELECTION CRITERIA: HIGHEST GINI

Gini = 2*AUC-1

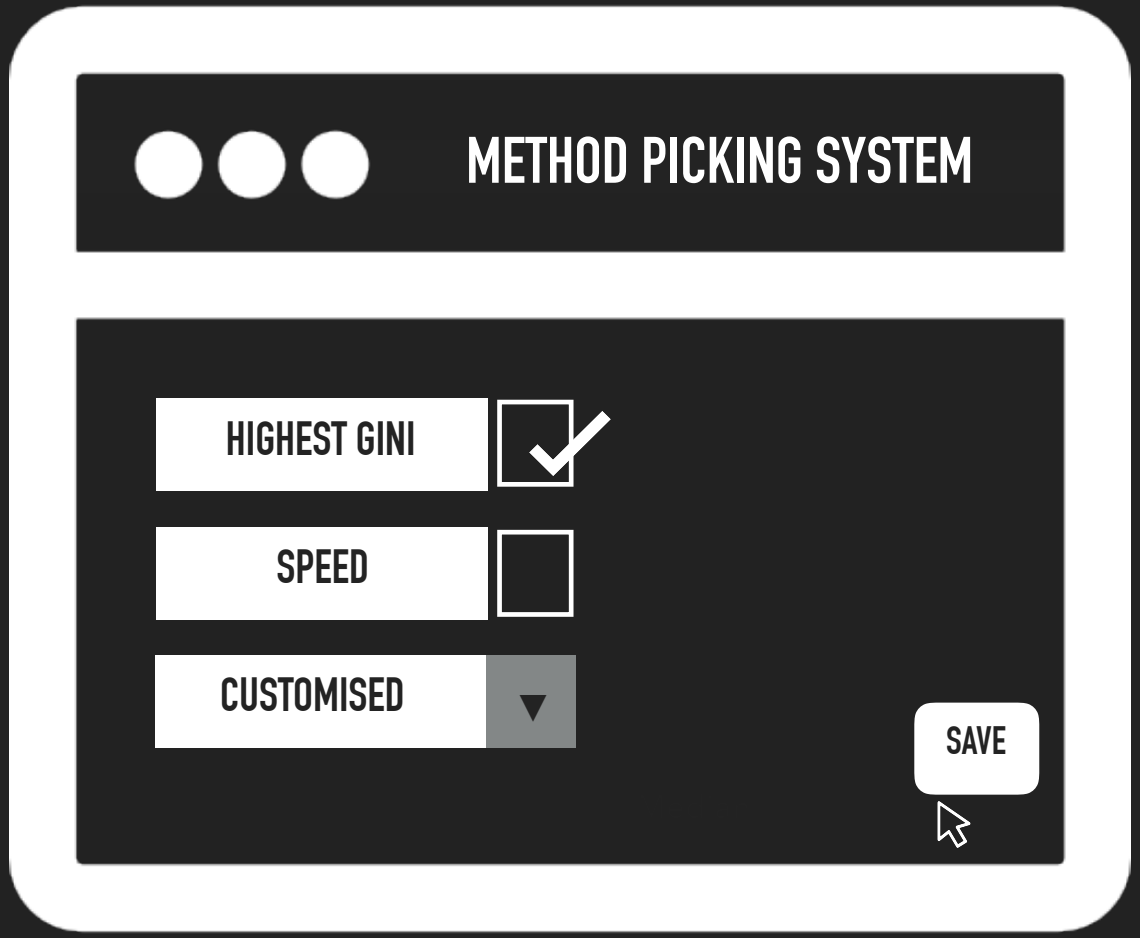
Area Under the ROC Curve

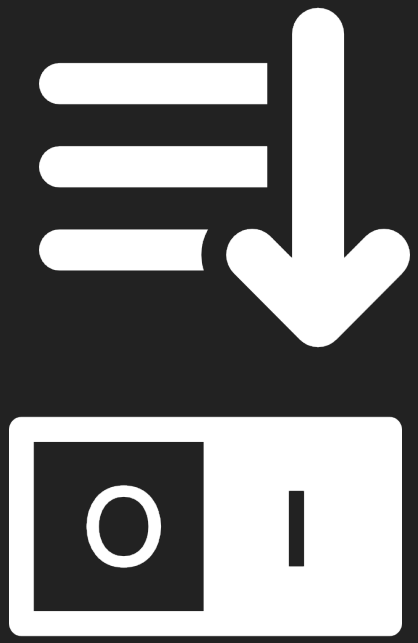
ROC: plot True Positive Rate against False Positive Rate



HUMAN ASSISTED METHOD PICKING

TASK 4


METHOD PICKING SYSTEM
HIGHEST GINI
SPEED
CUSTOMISED
SAVE



AUTOMATED DUMMY CREATION AND TRANSFORMATION

TASK 5

DUMMY CREATION

CATEGORICAL VARIABLE

ID	ANIMAL
1	DOG
2	CAT
3	DOG
4	ELEPHANT
5	CAT
6	DOG

DUMMY VARIABLE

ID	DOG	CAT	ELEPHANT
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0
6	1	0	0

BINNING

CONTINUOUS VARIABLE

ID	AGE
1	20
2	32
3	68
4	33
5	52
6	35

GROUPING

< 25	STUDENT
>25 AND < 60	MATURE
> 60	SENIOR

CATEGORICAL VARIABLE

ID	AGE RANGE
1	STUDENT
2	MATURE
3	SENIOR
4	MATURE
5	MATURE
6	MATURE

BINNING

CONTINUOUS VARIABLE

ID	AGE
1	20
2	32
3	68
4	33
5	52
6	35

CUT POINT?

GROUPING

< 25	STUDENT
>25 AND < 60	MATURE
> 60	SENIOR

CATEGORICAL VARIABLE

ID	AGE RANGE
1	STUDENT
2	MATURE
3	SENIOR
4	MATURE
5	MATURE
6	MATURE

ENTROPY BASED BINNING

Supervised binning

Calculates a value that describes how consistently a potential split will match up with a classifier (Target variable FRAUD)

Refer to the target information when selecting discretisation cut points

Finding the split with the maximal information gain

QUESTIONS?