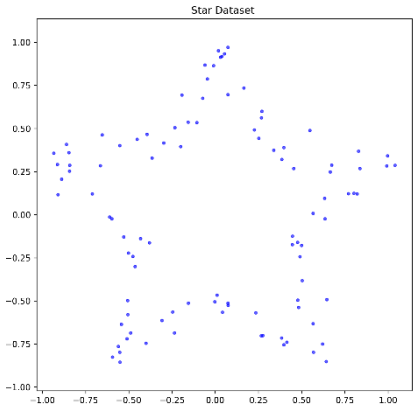
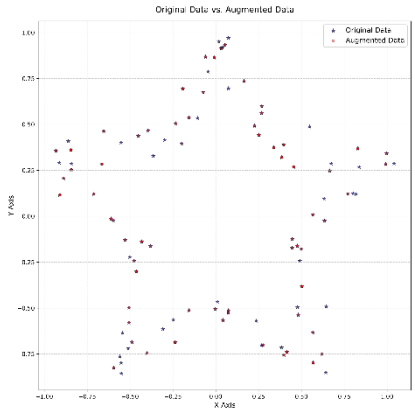
In this document we will generate synthetic data for 6 original dataset

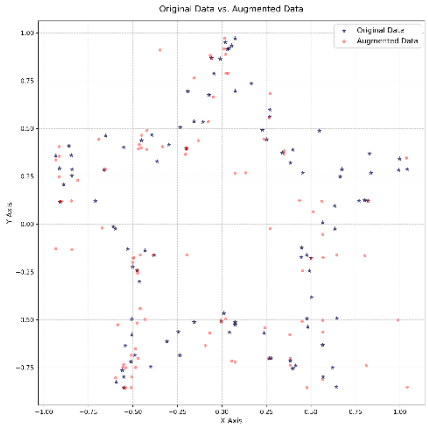
1. Generate original dataset as a star form with 100 points and little noise perturbation (0.05)
    a. Original dataset plot



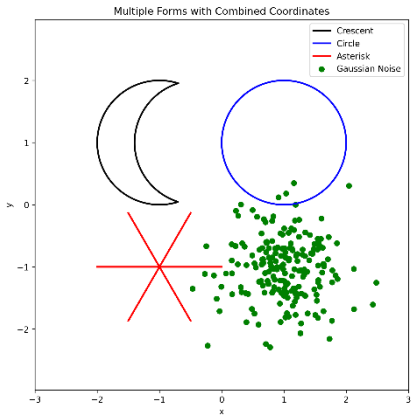
    b. Augmented data with 100 points and noise =0.01 along with original data



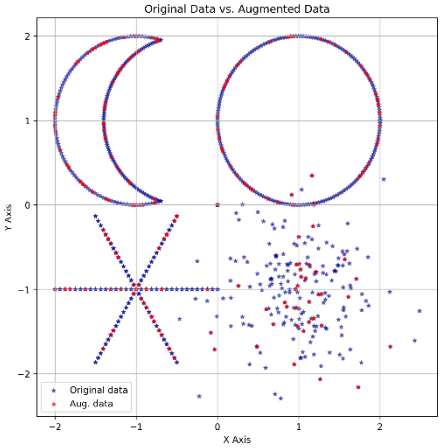
    c. Augmented data with 100 points and noise =5.0 along with original data



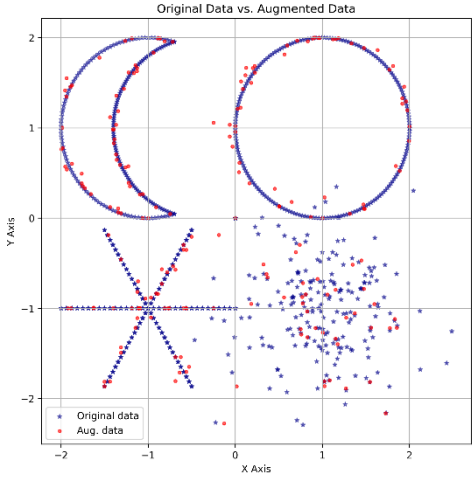
2. Generate multiple forms (crescent, circle, asterisk and Gaussian cloud) each with 200 points
    a. Original dataset plot



    b. Augmented data with 200 points and noise =0.01 along with original data



    c. Augmented data with 200 points and noise =5.0 along with original data

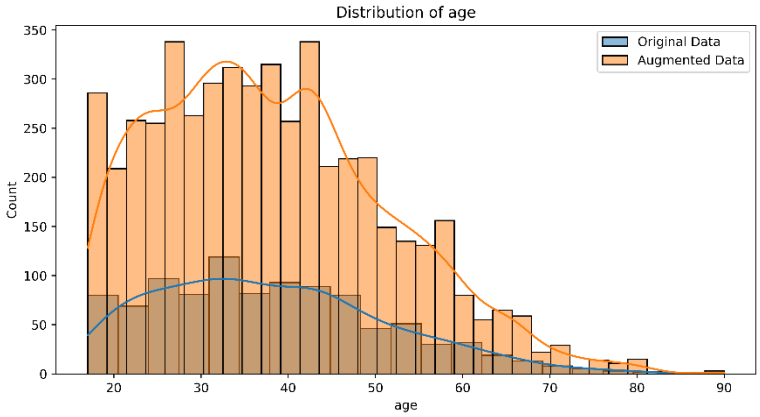


3. The Adult Dataset, also known as the Census Income Dataset, is a widely used dataset in the machine learning and data science communities. It was extracted from the 1994 U.S. Census Bureau database by Ronny Kohavi and Barry Becker and is hosted by the UCI Machine Learning Repository. We will use 6 integer features and 8 categorical that describe various characteristics of an individual, such as age, education, occupation, marital status, etc. Only 1000 records are used . 5000 records are generated with noise =0.01
    a. Automatically determined feature types:
        - Numeric features: ['age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss', 'hours-per-week']
        - Categorical features: ['workclass', 'education', 'marital-status', 'occupation', 'relationship', 'race', 'sex', 'native-country']
    b. Comparing Marginal Distributions

Feature: age

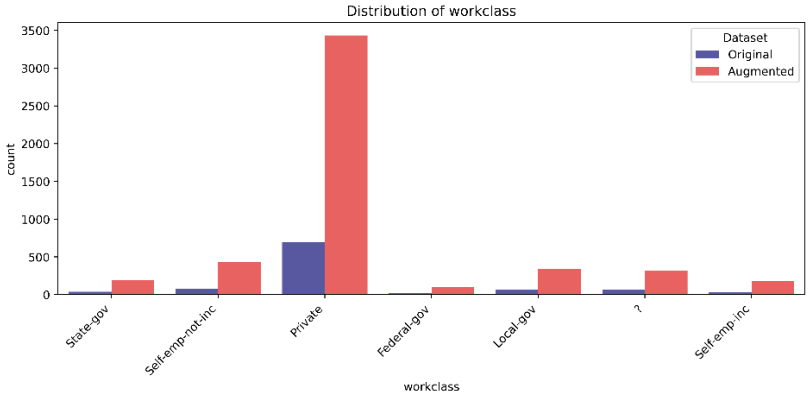
KS Statistic: 0.0084

P-value: 1.0000



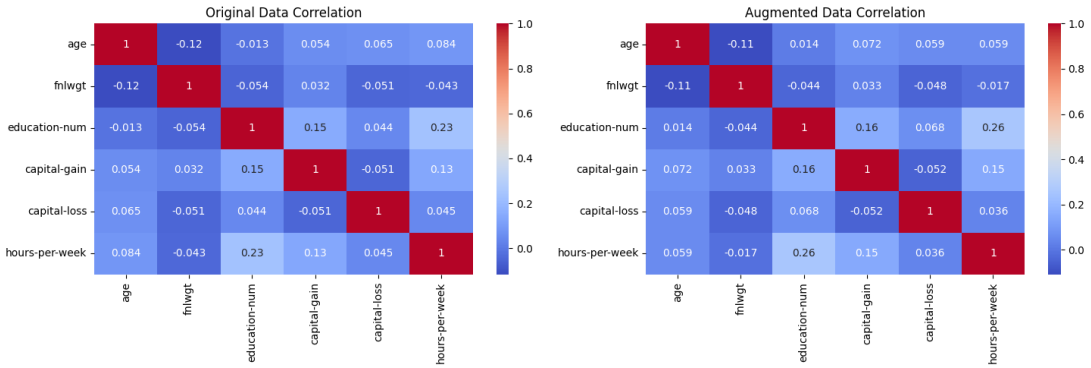
Feature: workclass

Jensen-Shannon Divergence= 0.0107



c. Comparing joint distributions

- Plots correlation heatmaps for the two datasets using only numeric features



- Jensen-Shannon Divergence for Each Feature:

{'age': 0.012009141978358564, 'workclass': 0.010767779748740774, 'fnlwgt': 0.01507315753012395, 'education': 0.015985193560453338, 'education-num': 0.012357914018225663, 'marital-status': 0.013829719726997938, 'occupation': 0.014042447870039021, 'relationship': 0.005534190897702178, 'race': 0.009178185836910418, 'sex': 0.0009033986250309765, 'capital-gain': 0.008308149411012726, 'capital-loss': 0.008086263196793347, 'hours-per-week': 0.014211802977080775, 'native-country': 0.023547239437767143}

Average JS Divergence: 0.0117 (Lower is better)

4. The Ecoli Dataset is a public dataset provided by the UCI Machine Learning Repository. It is used to predict the cellular localization sites of proteins in Escherichia coli (E. coli), a gram-negative bacterium commonly studied in biology. The dataset is a multiclass classification problem, where the goal is to classify proteins into one of several localization sites based on various features derived from the protein sequences. The number of records available is 336 with 7 continuous features . 5000 records are generated with noise =0.01

a. Automatically determined feature types:

Numeric features: ['mcg', 'gvh', 'lip', 'chg', 'aac', 'alm1', 'alm2']

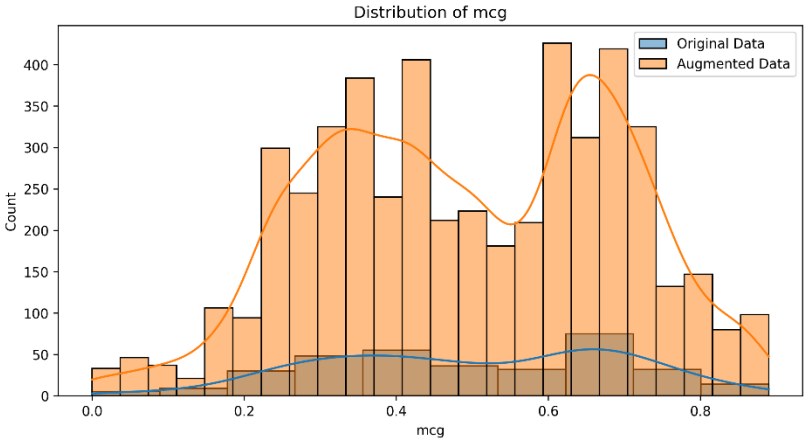
Categorical features: []

b. Comparing Marginal Distributions

Feature: mcg

KS Statistic: 0.0174

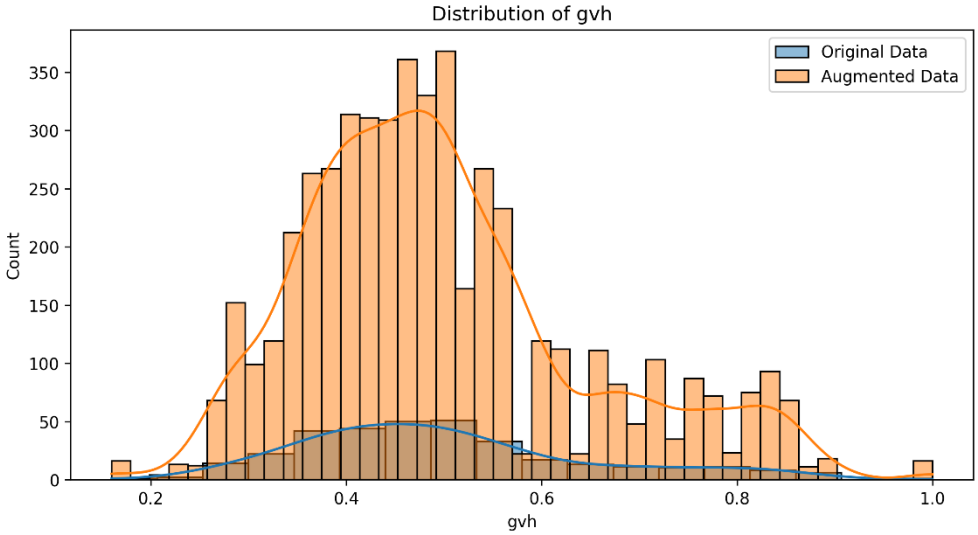
P-value: 1.0000



Feature: gvh

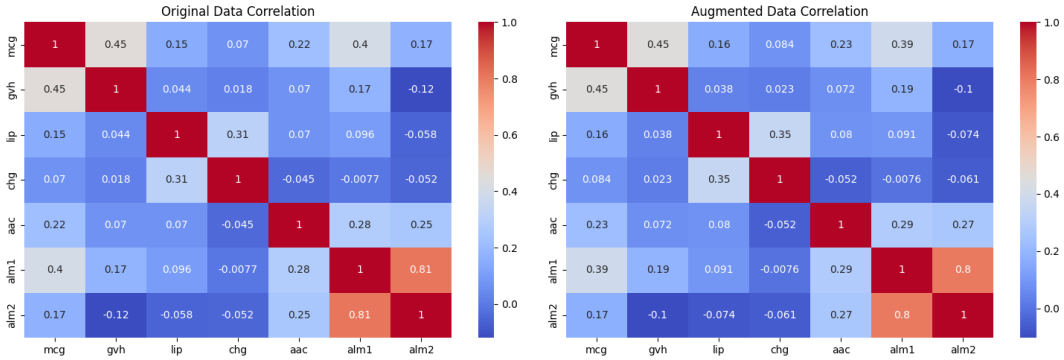
KS Statistic: 0.0073

P-value: 1.0000



c. Comparing joint distributions

- Plots correlation heatmaps for the two datasets using only numeric features



Jensen-Shannon Divergence for Each Feature:

{'mcg': 0.01668222897616381, 'gvh': 0.009229595028560637, 'lip': 0.005774157427971246, 'chg': 0.007253973815855733, 'aac': 0.016942870742773692, 'alm1': 0.017456722728436506, 'alm2': 0.020641212108735277}

Average JS Divergence: 0.0134 (Lower is better)

5. The Forest Fires Dataset is a publicly available dataset hosted by the UCI Machine Learning Repository. It is used for predicting the burned area of forest fires in the Montesinho Natural Park, located in the northeast region of Portuga , based on meteorological and spatial data. We will use 6 integer features and 8 categorical that describe various characteristics of an individual, such as age, education, occupation, marital status, etc The number of records available is 517 with 12 features (10 numeric and 2 categorical) . 5000 records are generated with noise =0.01
    a. Automatically determined feature types:

        Numeric features: ['X', 'Y', 'FFMC', 'DMC', 'DC', 'ISI', 'temp', 'RH', 'wind', 'rain']

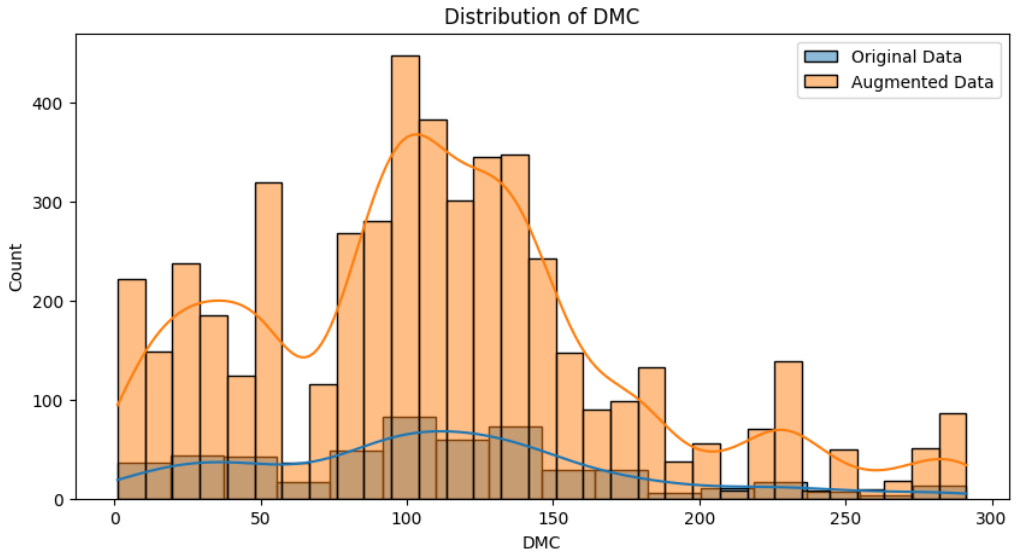
        Categorical features: ['month', 'day']

    b. Comparing Marginal Distributions

        Feature: DMC

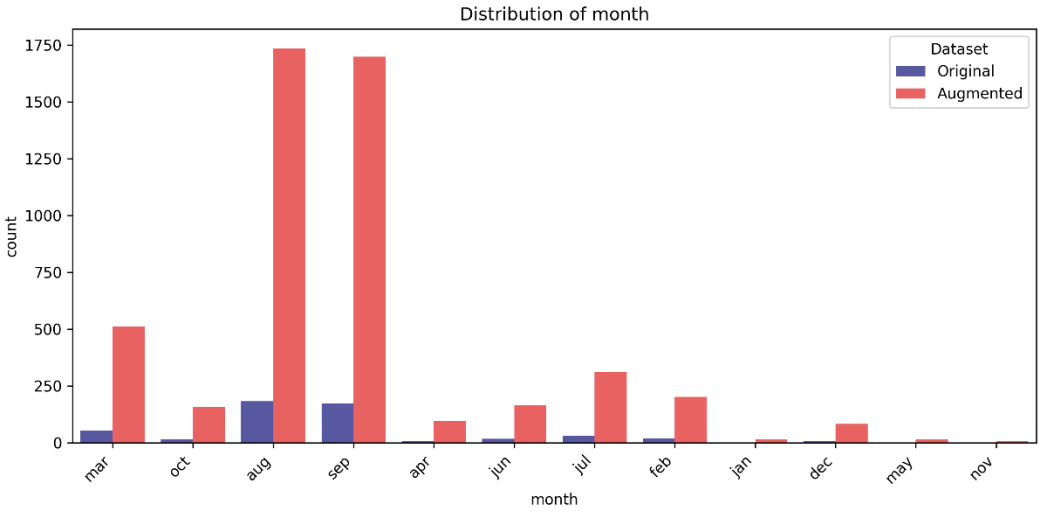
        KS Statistic: 0.0091

        P-value: 1.0000



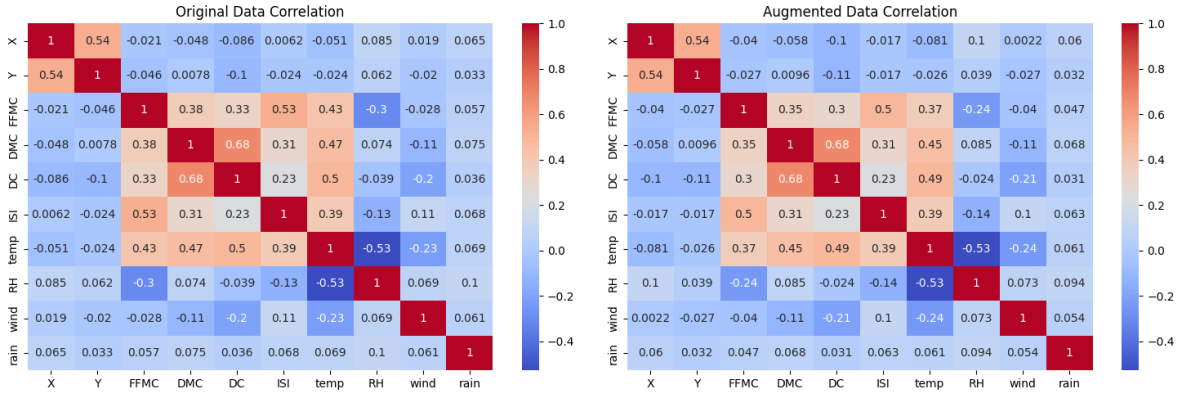
        Feature: Month

        Jensen-Shannon Divergence: 0.0134



c. Comparing joint distributions

Plots correlation heatmaps for the two datasets using only numeric features



Jensen-Shannon Divergence for Each Feature:

{'X': 0.008341369773125055, 'Y': 0.010030899028201867, 'month': 0.013433130643795291, 'day': 0.007384001745135166, 'FFMC': 0.034054381295991705, 'DMC': 0.010506290464019165, 'DC': 0.021674409148876087, 'ISI': 0.009723160911701204, 'temp': 0.017322097523134225, 'RH': 0.016448271552411447, 'wind': 0.015305831325483889, 'rain': 0.007611024252633004}

Average JS Divergence: 0.0143 (Lower is better)

6. The Diagnostic Wisconsin Breast Cancer Dataset, is a widely used dataset in machine learning and data science. The primary task is to classify breast tumors as either benign (non-cancerous) or malignant (cancerous) based on features extracted from digitized images of FNAs. The number of records available is 569 with 29 continuous features. 5000 records are generated with noise =0.01
    a. Automatically determined feature types:

        Numeric features: ['texture1', 'perimeter1', 'area1', 'smoothness1', 'compactness1', 'concavity1', 'concave_points1', 'symmetry1',…']

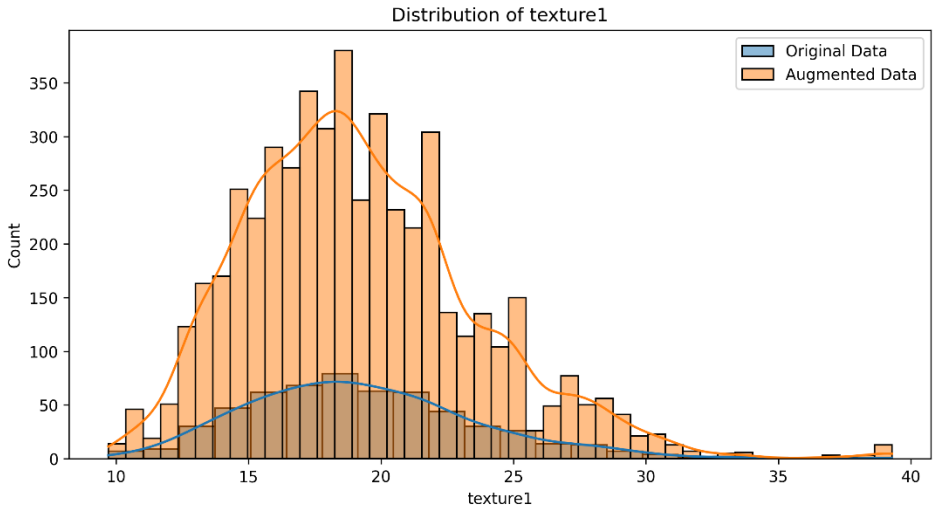
Categorical features: []

b. Comparing Marginal Distributions

Feature: texture1

KS Statistic: 0.0118

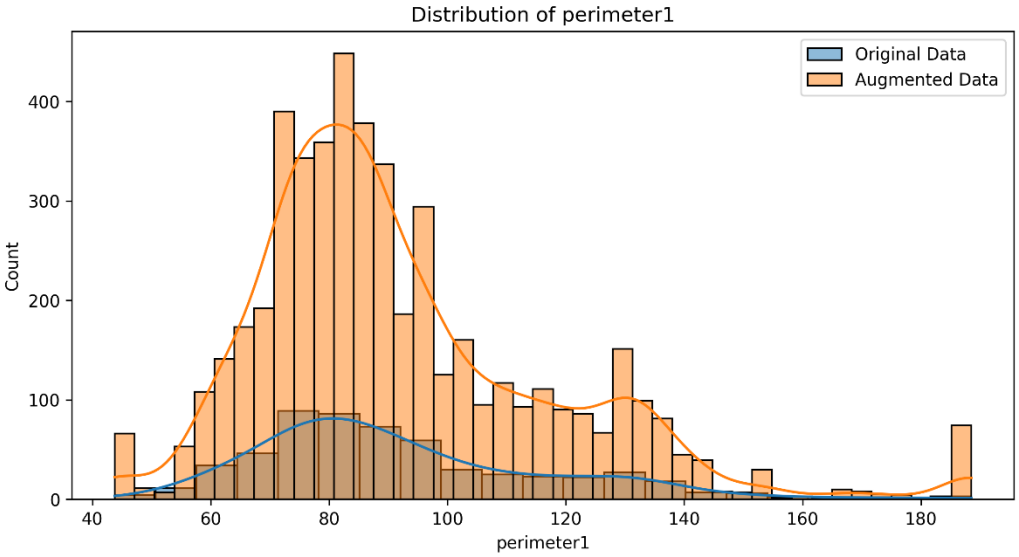
P-value: 1.0000



Feature: perimeter1

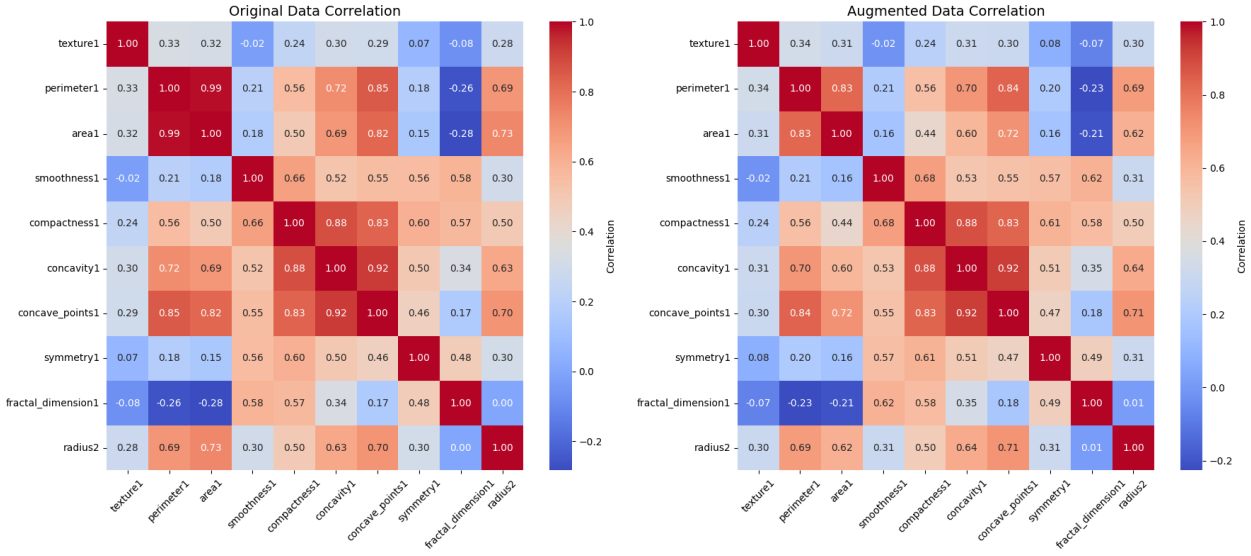
KS Statistic: 0.0126

P-value: 1.0000



c. Comparing joint distributions

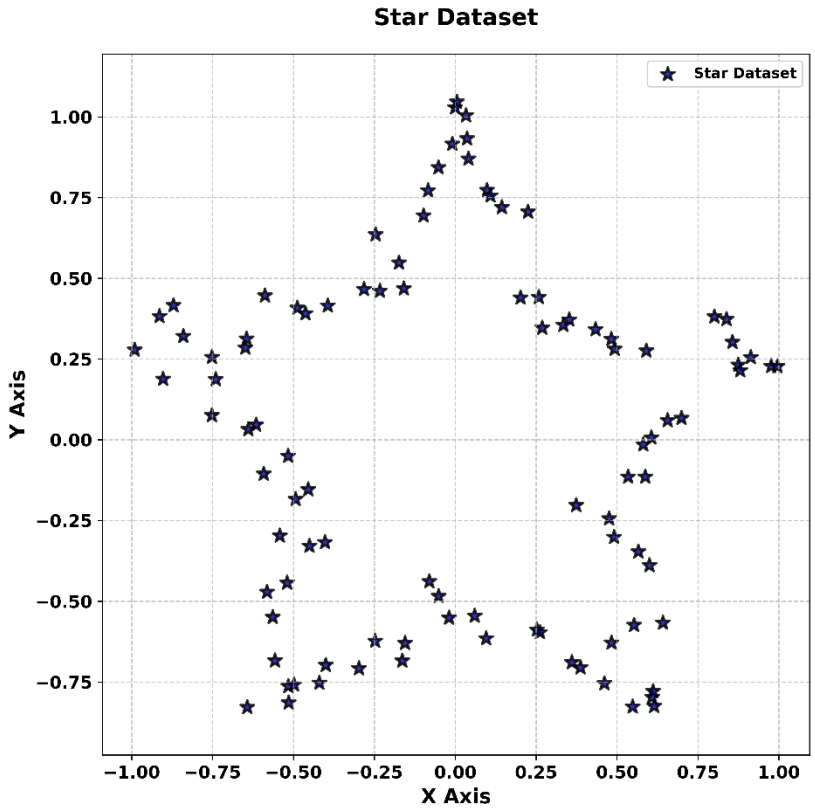
- Plots correlation heatmaps for the two datasets using only numeric features

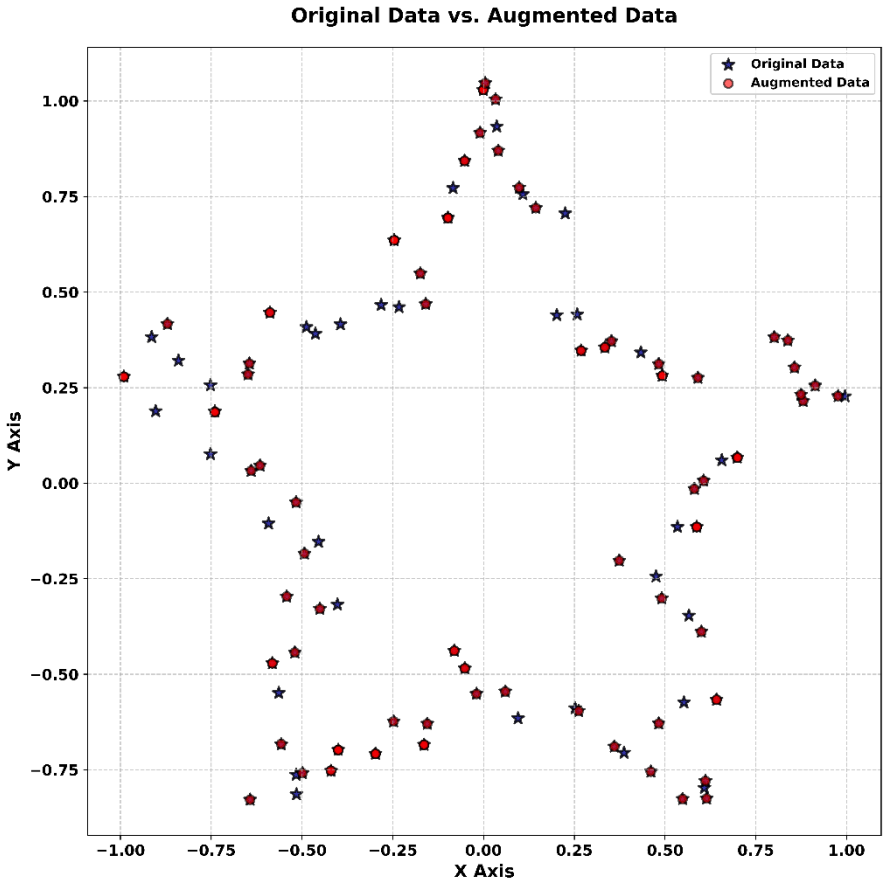


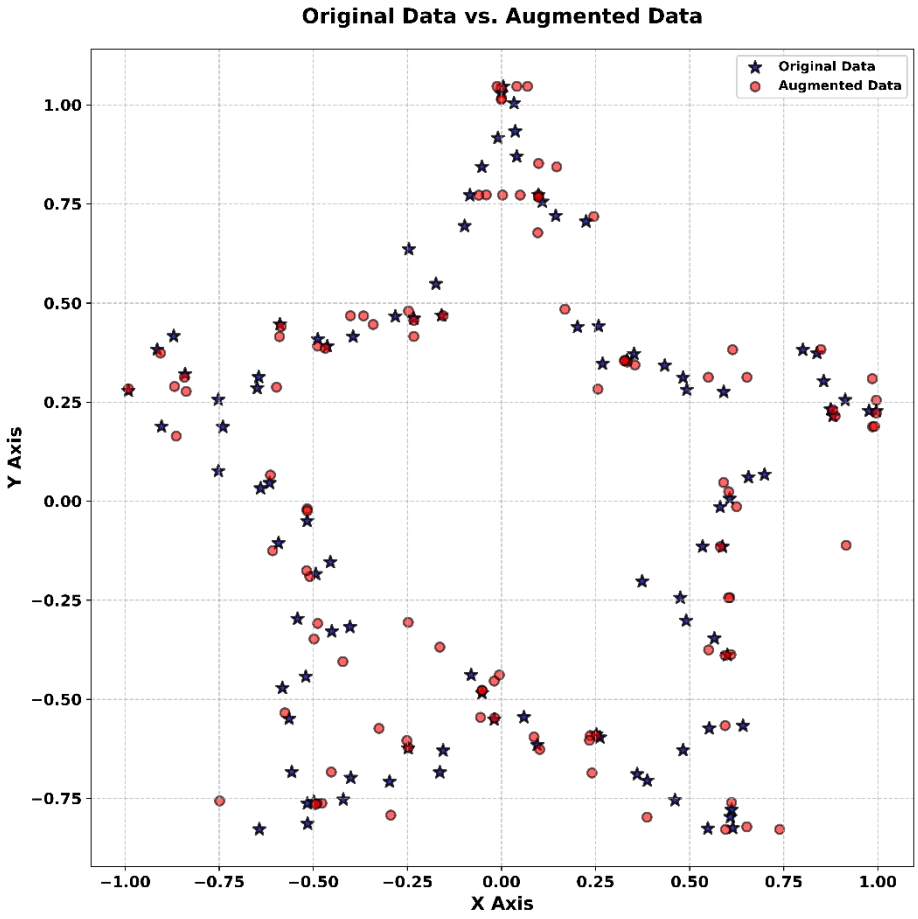
Jensen-Shannon Divergence for Each Feature:

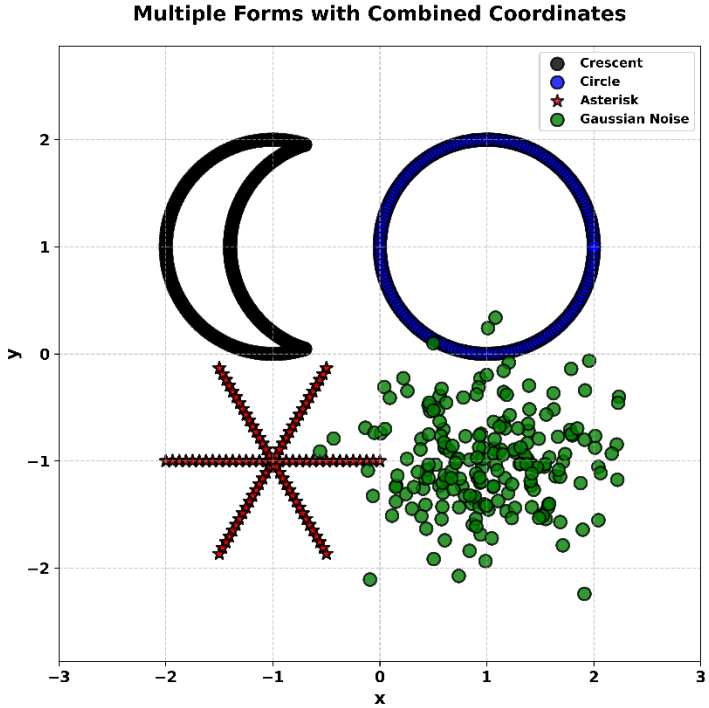
{'texture1': 0.028831283558541997, 'perimeter1': 0.03672664810099611, 'area1': 0.10065016935669929, 'smoothness1': 0.025045012269920838, 'compactness1': 0.020120694205898668, 'concavity1': 0.015580700980194236, 'concave_points1': 0.02134794298122662, 'symmetry1': 0.012536324381881432, 'fractal_dimension1': 0.01522349113042415, 'radius2': 0.019294439613664536, 'texture2': 0.01808996503097313, 'perimeter2': 0.020095642987695752, 'area2': 0.02552554023961594, 'smoothness2': 0.01707724503187744, 'compactness2': 0.02535257237702597, 'concavity2': 0.031621766482832055, 'concave_points2': 0.019972453627675742, 'symmetry2': 0.017411227690286784, 'fractal_dimension2': 0.03653232929050405, 'radius3': 0.015370756309758114, 'texture3': 0.023276908533615373, 'perimeter3': 0.06585064935649561, 'area3': 0.1353978109985651, 'smoothness3': 0.02100317479861306, 'compactness3': 0.014238287091710294, 'concavity3': 0.013924013279411548, 'concave_points3': 0.012277666793283449, 'symmetry3': 0.017981531084874134, 'fractal_dimension3': 0.02103537735393165}

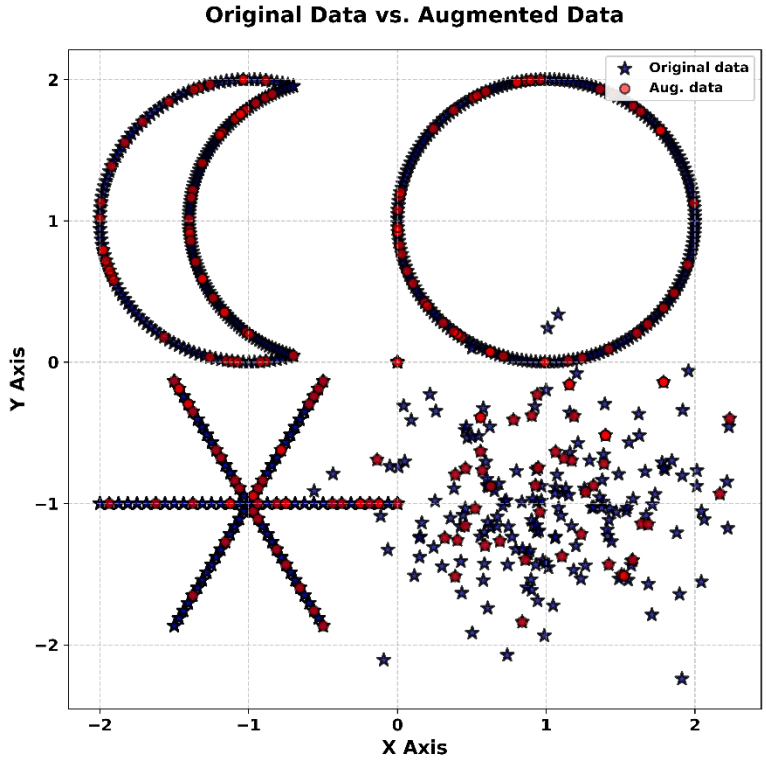
Average JS Divergence: 0.0292 (Lower is better)

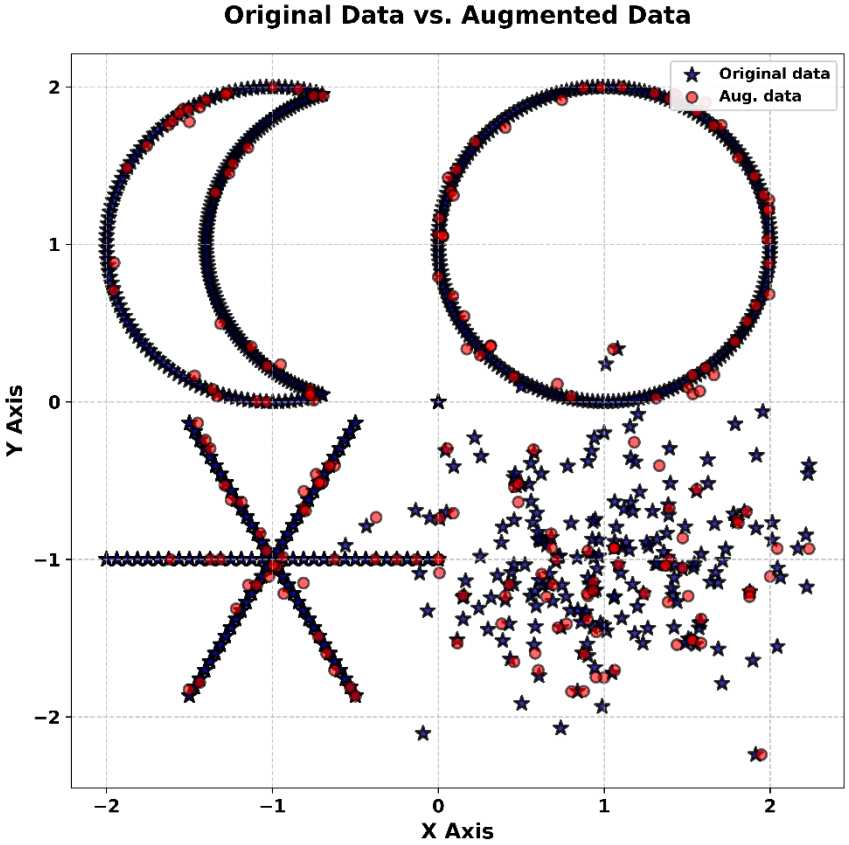

Star Dataset
Star Dataset
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Y Axis
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
X Axis

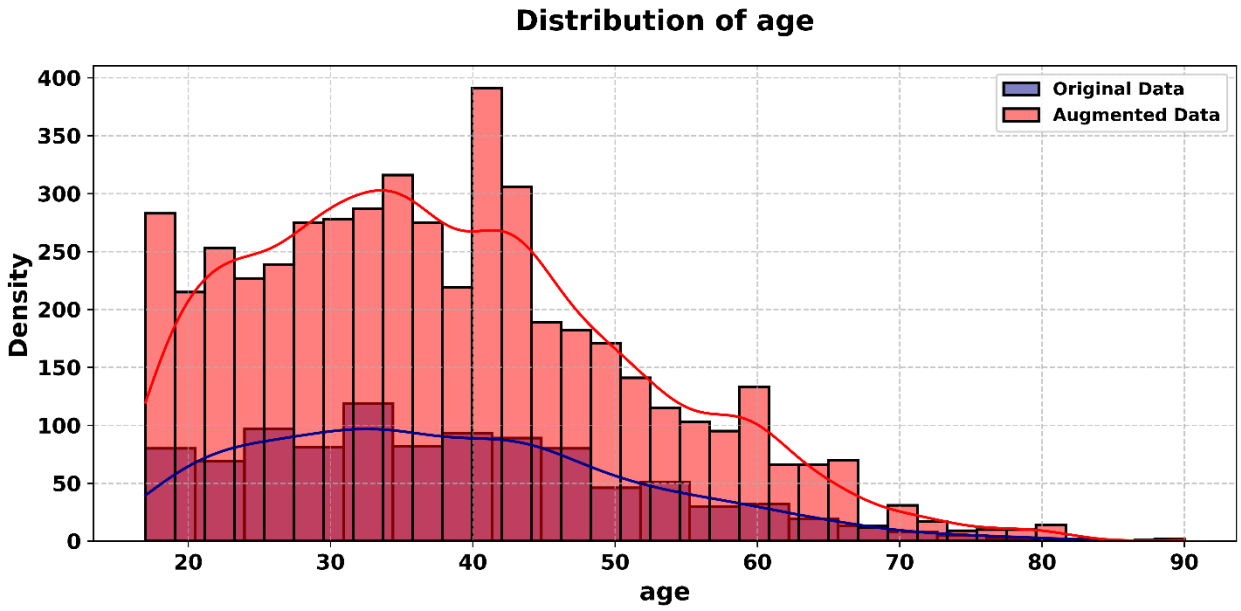

Original Data vs. Augmented Data
Original Data
Augmented Data
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Y Axis
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
X Axis


Original Data vs. Augmented Data
Original Data
Augmented Data
Y Axis
X Axis
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

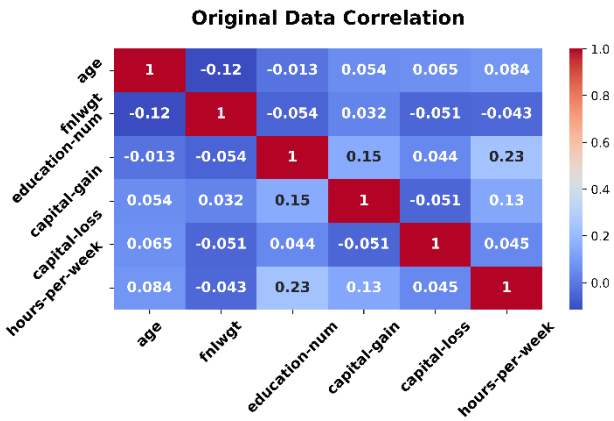

Multiple Forms with Combined Coordinates
Crescent
Circle
Asterisk
Gaussian Noise
y
x
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

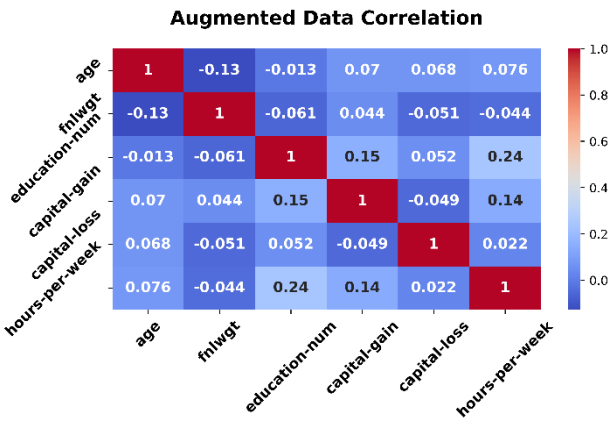

Original Data vs. Augmented Data
Original data
Aug. data
Y Axis
X Axis
2
1
0
−1
−2
−2
−1
0
1
2

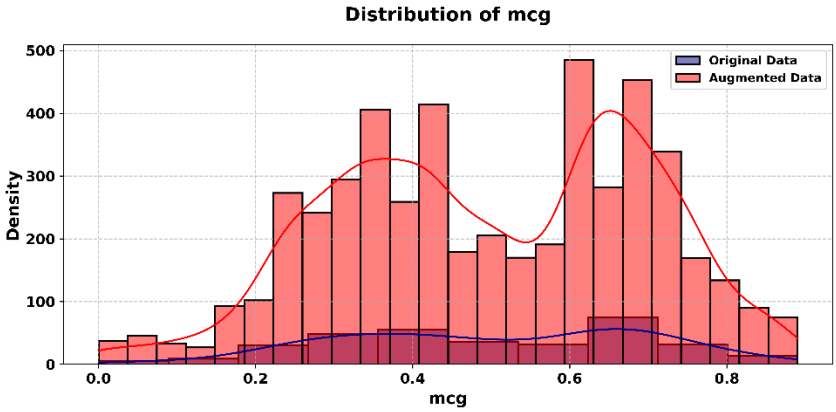

Original Data vs. Augmented Data
Original data
Aug. data
Y Axis
X Axis
2
1
0
−1
−2
−2
−1
0
1
2

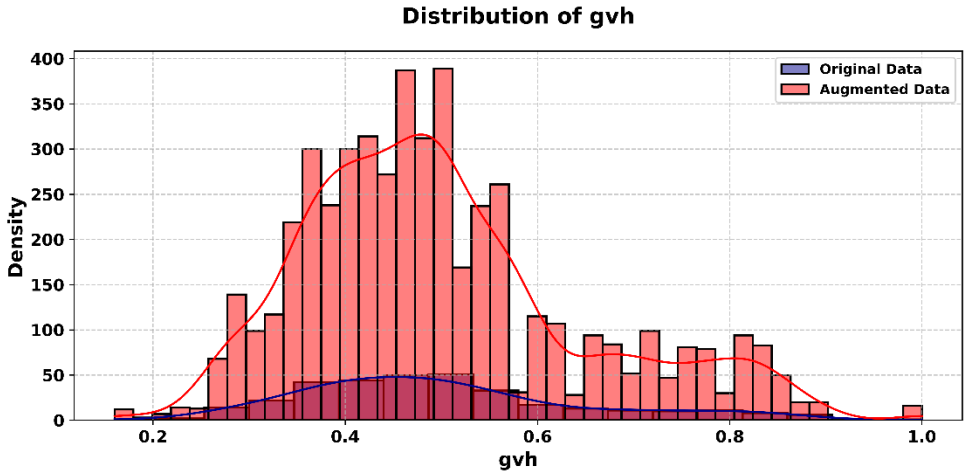

Distribution of age
Original Data
Augmented Data
Density
age
0
50
100
150
200
250
300
350
400
20
30
40
50
60
70
80
90

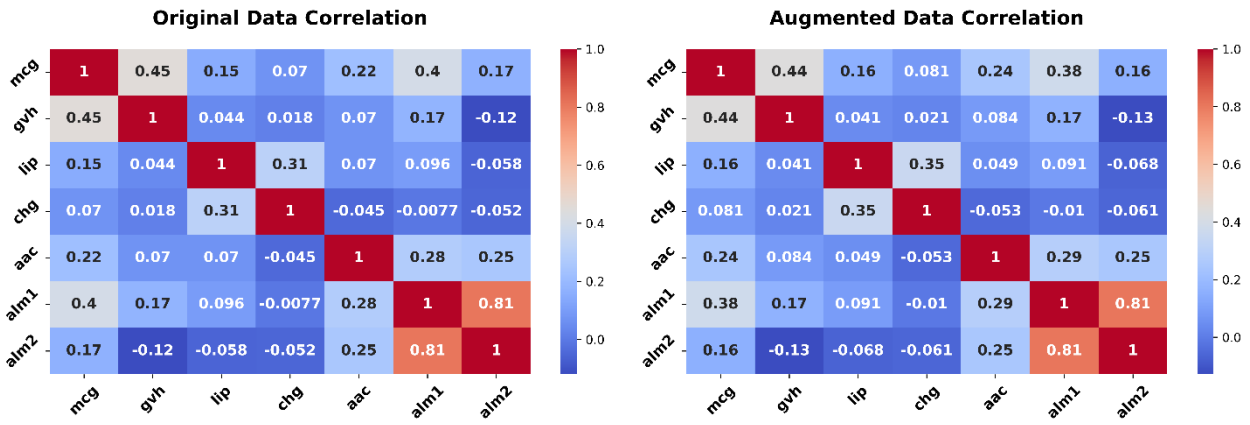

Distribution of workclass
Dataset
Original Data
Augmented Data
Count
workclass
0
500
1000
1500
2000
2500
3000
3500
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
?
Self-emp-inc

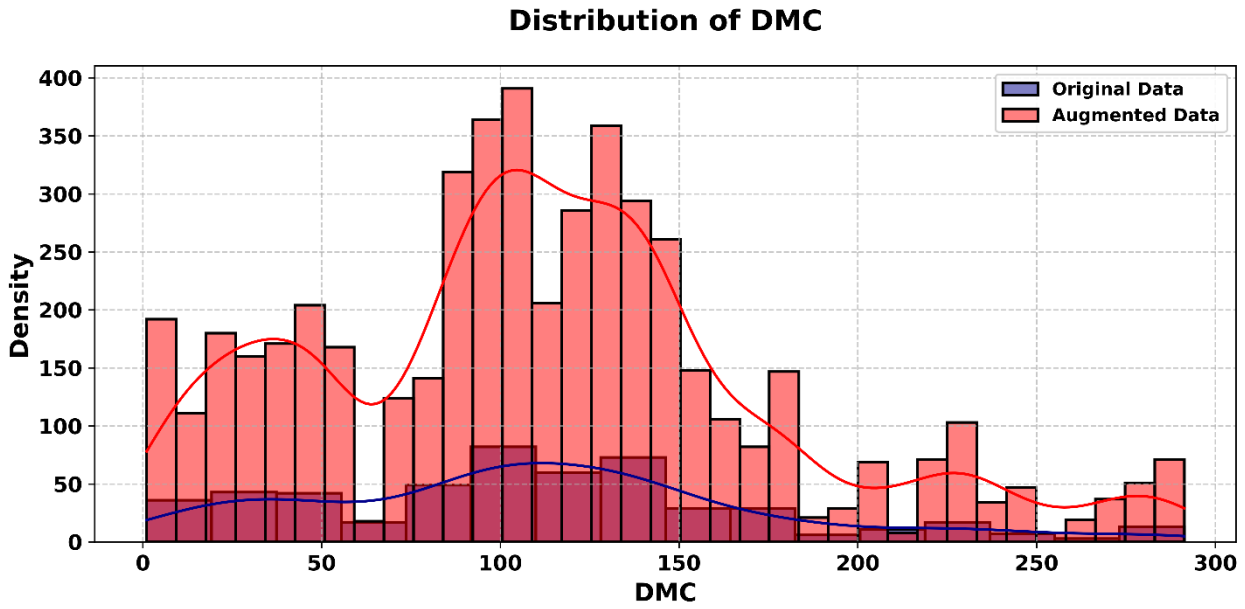

Original Data Correlation
age
fnlwgt
education-num
capital-gain
capital-loss
hours-per-week
1 -0.12 -0.013 0.054 0.065 0.084
-0.12 1 -0.054 0.032 -0.051 -0.043
-0.013 -0.054 1 0.15 0.044 0.23
0.054 0.032 0.15 1 -0.051 0.13
0.065 -0.051 0.044 -0.051 1 0.045
0.084 -0.043 0.23 0.13 0.045 1
1.0
0.8
0.6
0.4
0.2
0.0

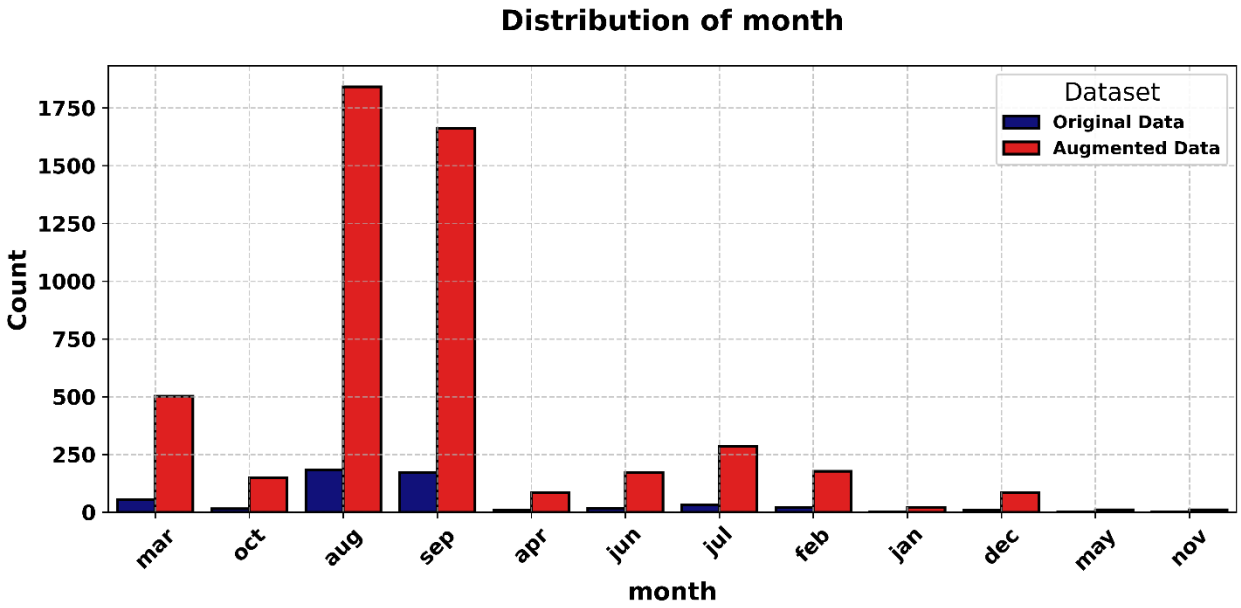

Augmented Data Correlation
age
fnlwgt
education-num
capital-gain
capital-loss
hours-per-week
1 -0.13 -0.013 0.07 0.068 0.076
-0.13 1 -0.061 0.044 -0.051 -0.044
-0.013 -0.061 1 0.15 0.052 0.24
0.07 0.044 0.15 1 -0.049 0.14
0.068 -0.051 0.052 -0.049 1 0.022
0.076 -0.044 0.24 0.14 0.022 1
1.0
0.8
0.6
0.4
0.2
0.0

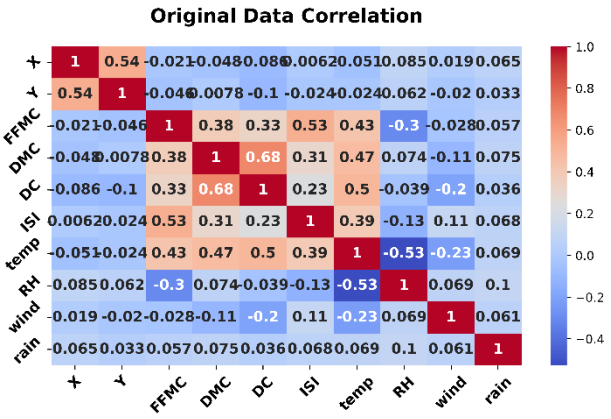

Distribution of mcg
Original Data
Augmented Data
Density
mcg
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8

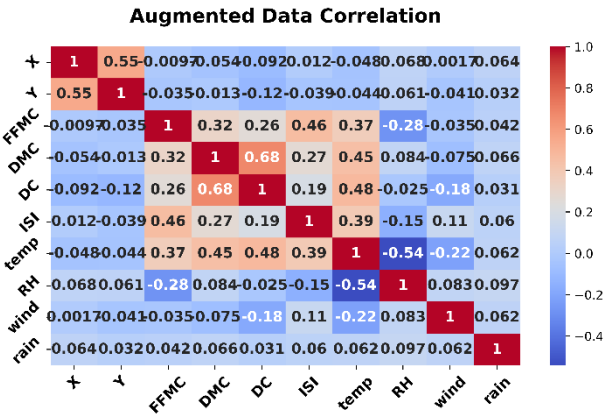

Distribution of gvh
Original Data
Augmented Data
Density
gvh
0
50
100
150
200
250
300
350
400
0.2
0.4
0.6
0.8
1.0


Original Data Correlation
mcg 1 0.45 0.15 0.07 0.22 0.4 0.17
gvh 0.45 1 0.044 0.018 0.07 0.17 -0.12
lip 0.15 0.044 1 0.31 0.07 0.096 -0.058
chg 0.07 0.018 0.31 1 -0.045 -0.0077 -0.052
aac 0.22 0.07 0.07 -0.045 1 0.28 0.25
alm1 0.4 0.17 0.096 -0.0077 0.28 1 0.81
alm2 0.17 -0.12 -0.058 -0.052 0.25 0.81 1
mcg gvh lip chg aac alm1 alm2
1.0 0.8 0.6 0.4 0.2 0.0
Augmented Data Correlation
mcg 1 0.44 0.16 0.081 0.24 0.38 0.16
gvh 0.44 1 0.041 0.021 0.084 0.17 -0.13
lip 0.16 0.041 1 0.35 0.049 0.091 -0.068
chg 0.081 0.021 0.35 1 -0.053 -0.01 -0.061
aac 0.24 0.084 0.049 -0.053 1 0.29 0.25
alm1 0.38 0.17 0.091 -0.01 0.29 1 0.81
alm2 0.16 -0.13 -0.068 -0.061 0.25 0.81 1
mcg gvh lip chg aac alm1 alm2
1.0 0.8 0.6 0.4 0.2 0.0


Distribution of DMC
Original Data
Augmented Data
Density
DMC
0 50 100 150 200 250 300
0 50 100 150 200 250 300 350 400


Distribution of month
Dataset
Original Data
Augmented Data
Count
1750
1500
1250
1000
750
500
250
0
mar
oct
aug
sep
apr
jun
jul
feb
jan
dec
may
nov
month


Original Data Correlation
X
Y
FFMC
DMC
DC
ISI
temp
RH
wind
rain
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Augmented Data Correlation
X
Y
FFMC
DMC
DC
ISI
temp
RH
wind
rain
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4