EXPERIMENT N0.8
TITLE: Exploratory data analytics on cloud platform
(Microsoft Azure ML Studio)

Batch: A4 **Roll No.:1914078** **Experiment No.: 8**

Aim: Exploratory data analytics on cloud platform (Azure ML Studio)

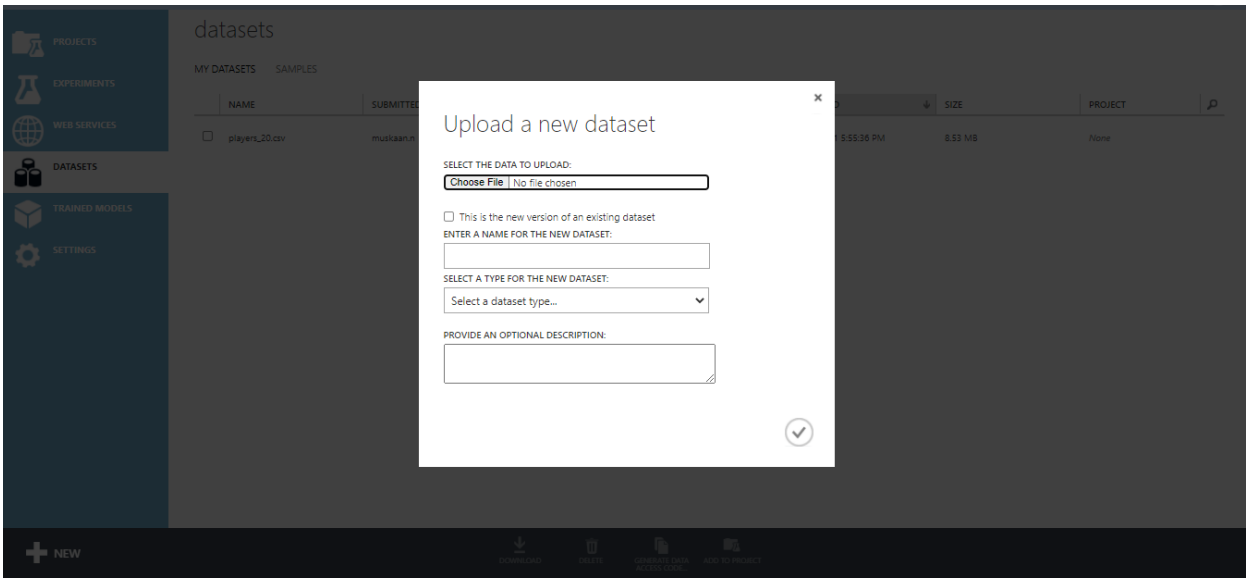
Resources needed: Microsoft Azure Machine learning studio (Classic)

Procedure / Approach /Algorithm / Activity Diagram:

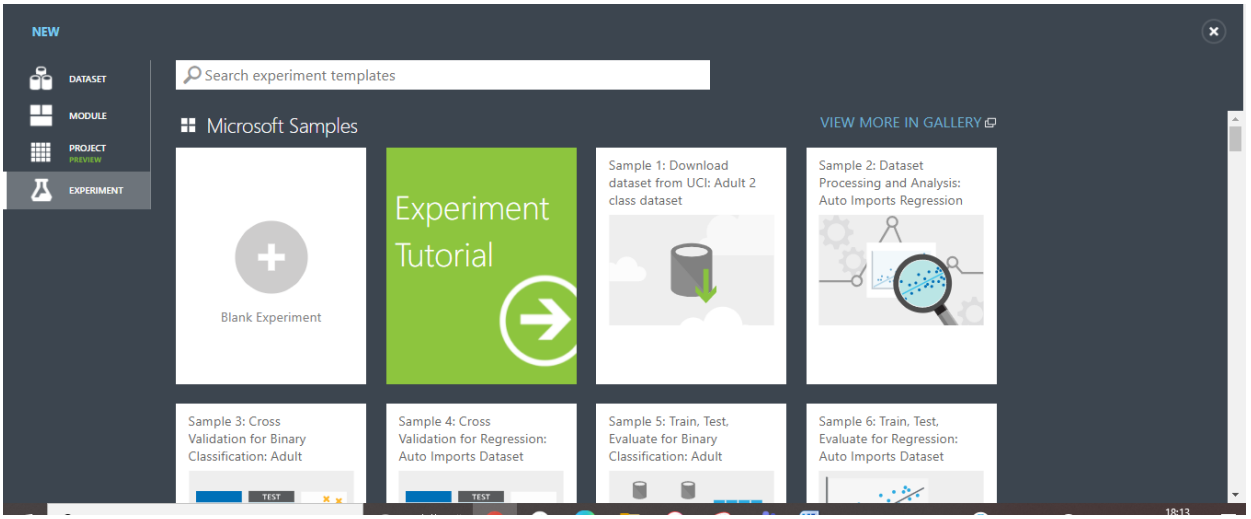
Explore the Microsoft Azure Machine learning studio [1] to perform the exploratory data analytics on your dataset for different purposes such as data normalization, discretization, attribute subset selection, visualization etc.

Results: (Program printout with output / Document printout as per the format)

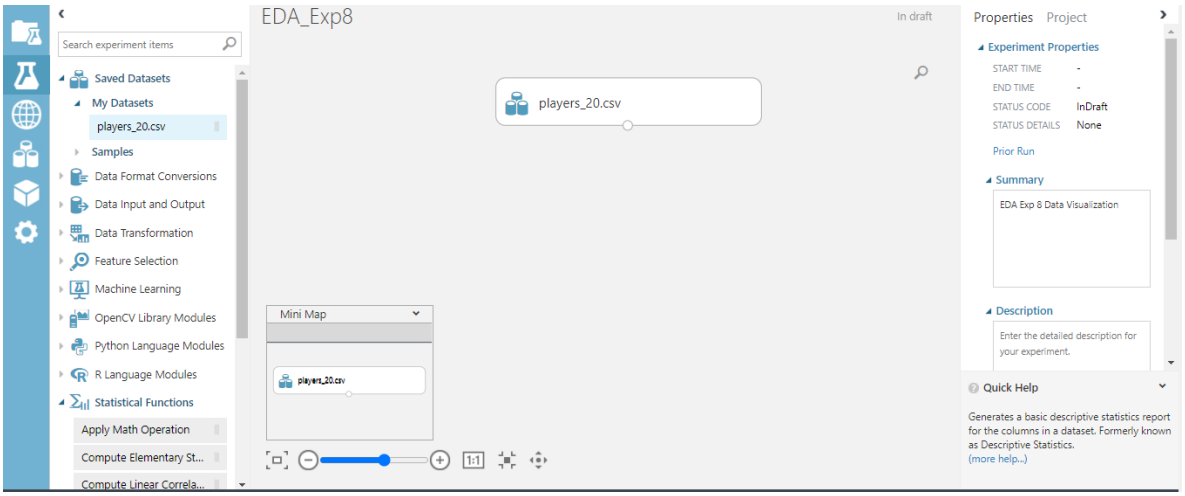
1. Sign In to Azure ML Studio
2. Upload your dataset in the my dataset section



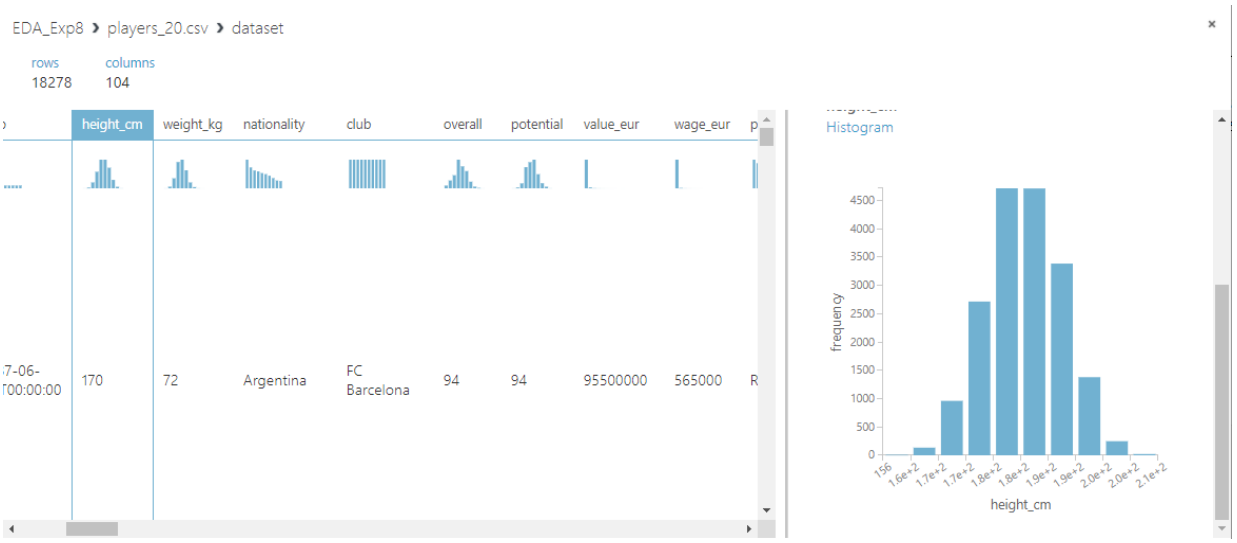
3. Create New Blank Experiment



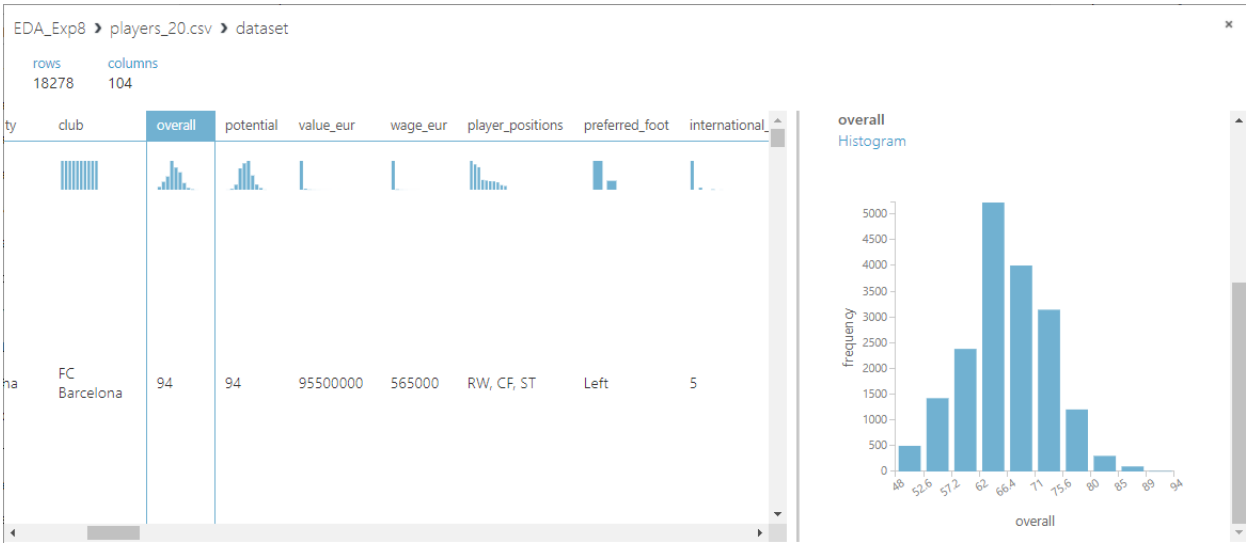
4. Change the name and Add Summary
5. From My Datasets, select and drag your dataset.



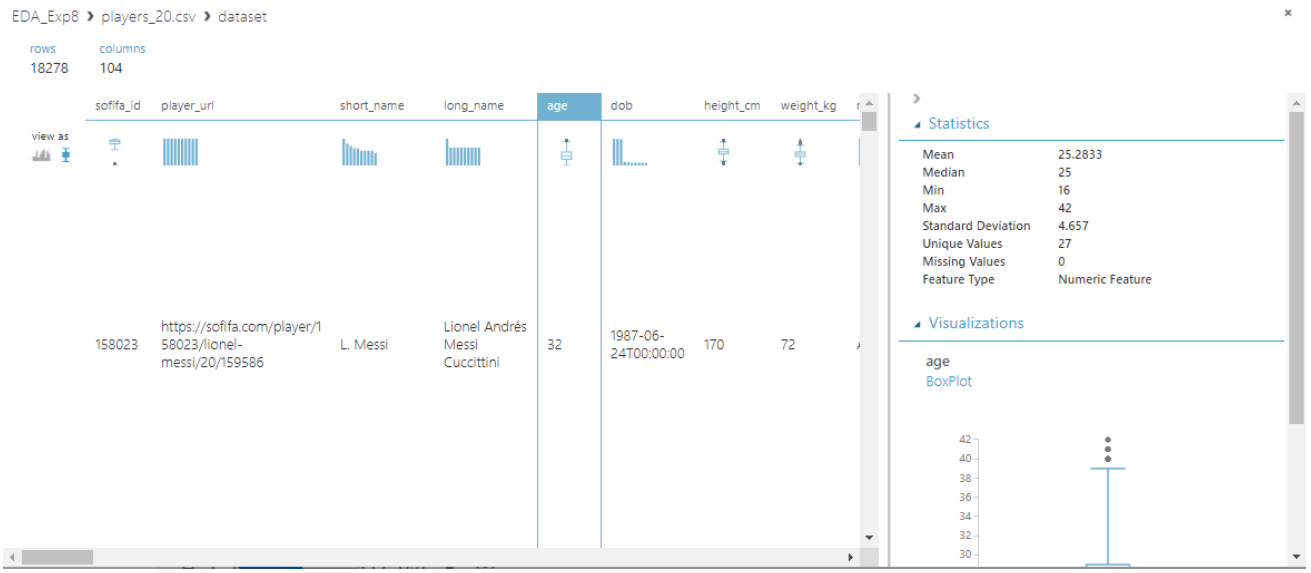
6. Right click and visualize the dataset.
   a. Histogram for player's height

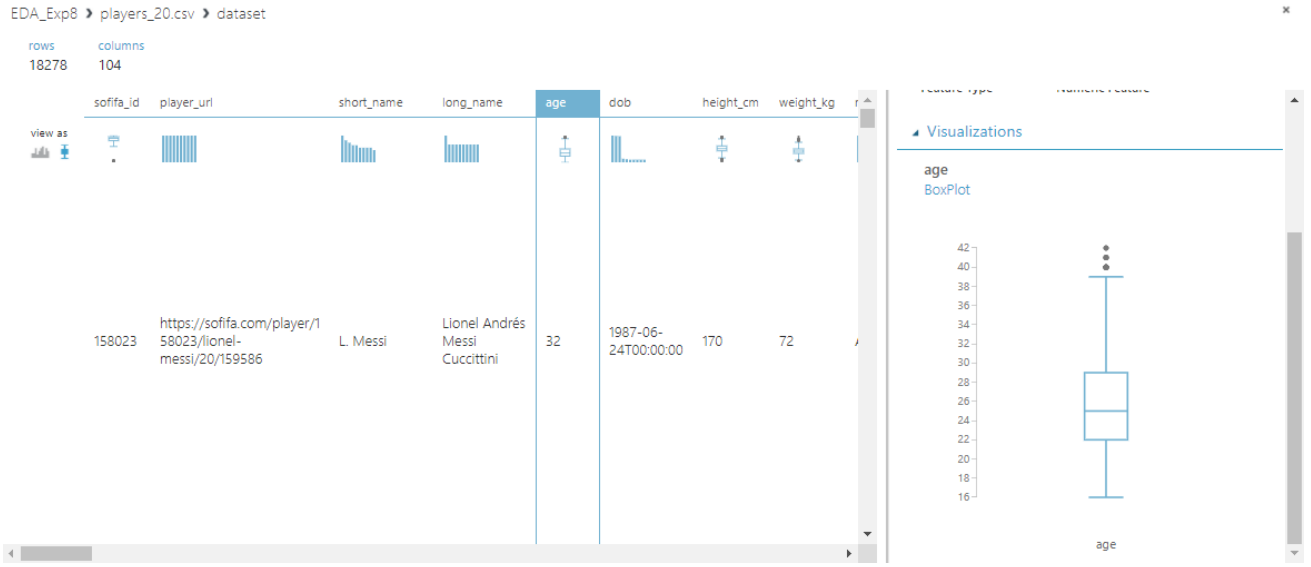


   b. Histogram for player's overall

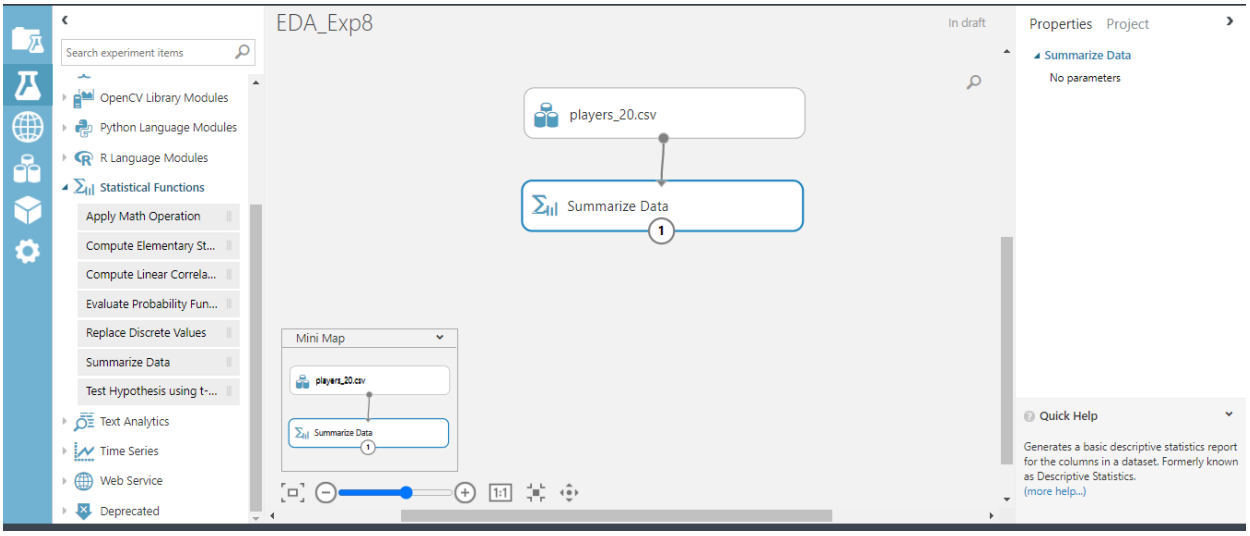


   c. Player's age

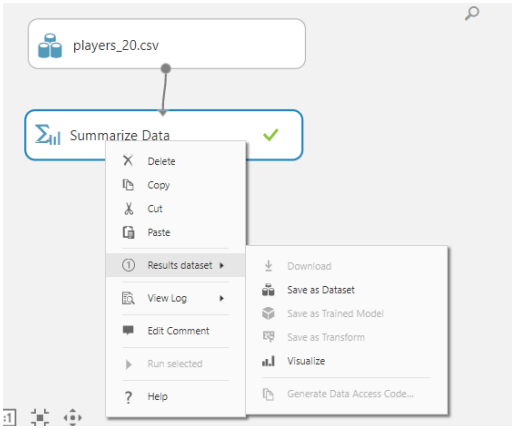


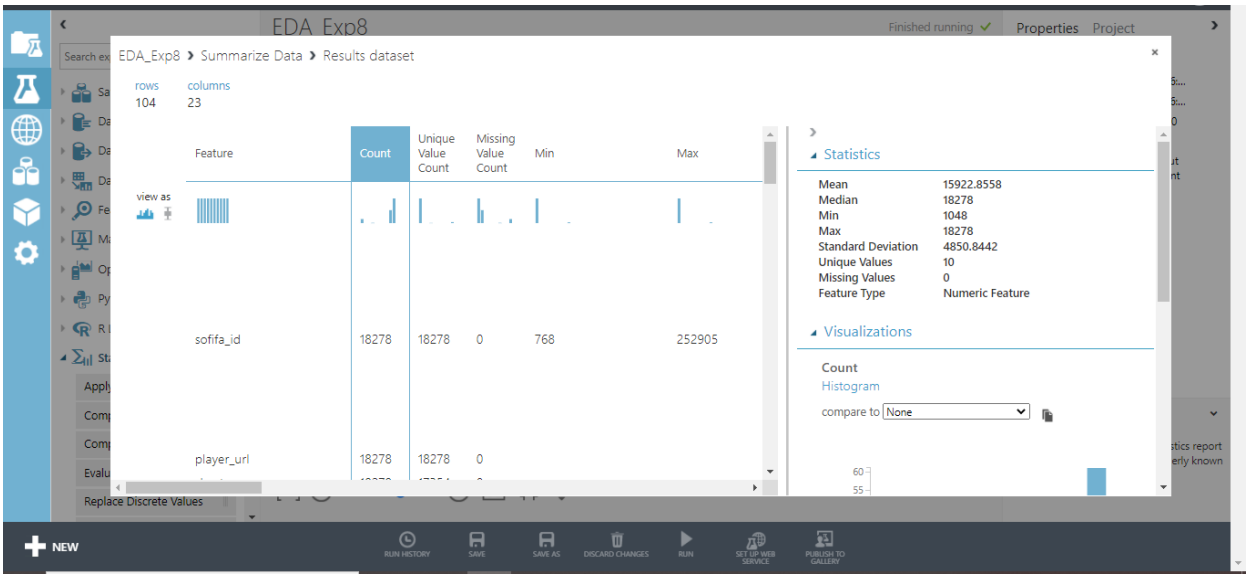


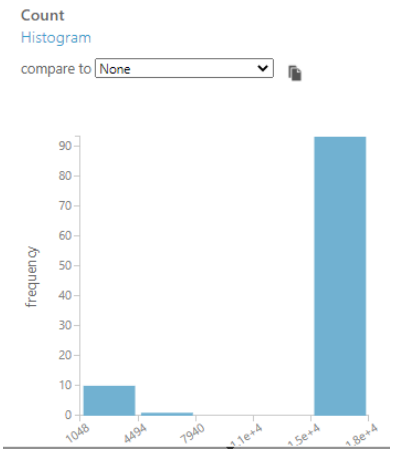
7. From the statistical Functions section select Summarize Data. Connect the output of dataset to the input of Summarize Data



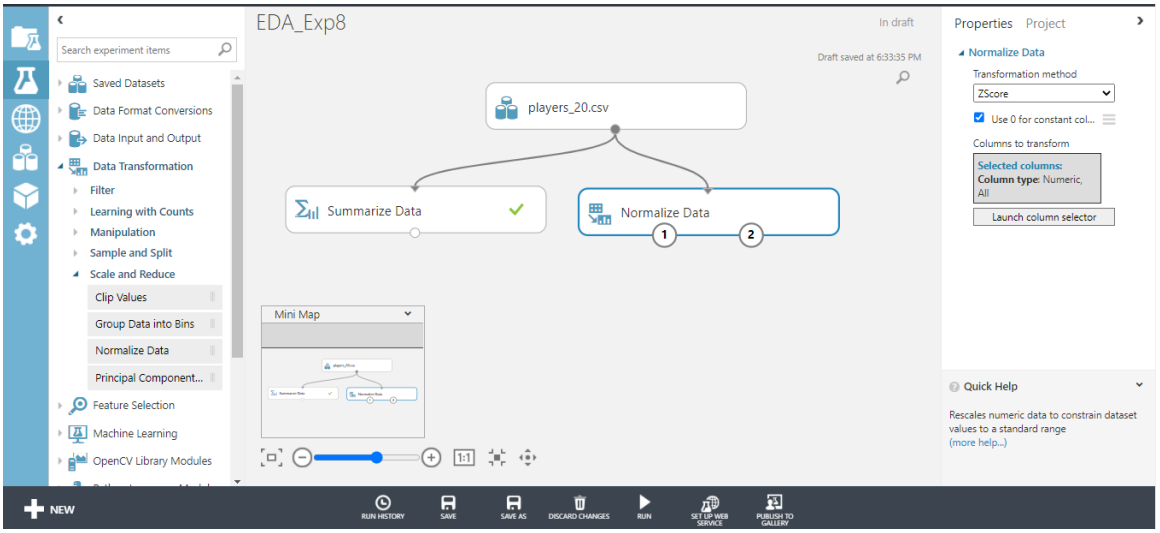
8. Run the simulation and view the Resutls of Summarize Data.


players_20.csv
Summarize Data
Delete
Copy
Cut
Paste
Results dataset
View Log
Edit Comment
Run selected
Help
Download
Save as Dataset
Save as Trained Model
Save as Transform
Visualize
Generate Data Access Code...

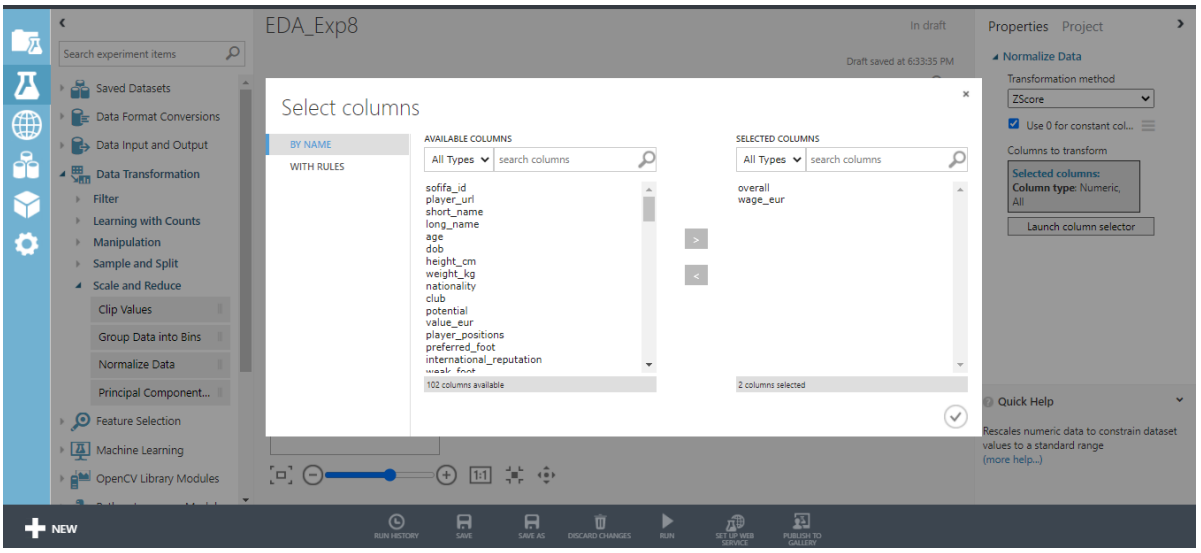

EDA_Exp8
Finished running
Properties
Project
EDA_Exp8 › Summarize Data › Results dataset
rows
104
columns
23
Feature
Count
Unique Value Count
Missing Value Count
Min
Max
view as
sofifa_id
18278
18278
0
768
252905
player_url
18278
18278
0
Statistics
Mean 15922.8558
Median 18278
Min 1048
Max 18278
Standard Deviation 4850.8442
Unique Values 10
Missing Values 0
Feature Type Numeric Feature
Visualizations
Count
Histogram
compare to None
Replace Discrete Values
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY


Count
Histogram
compare to None
frequency
90
80
70
60
50
40
30
20
10
0
1048
4494
7940
1.1e+4
1.5e+4
1.8e+4

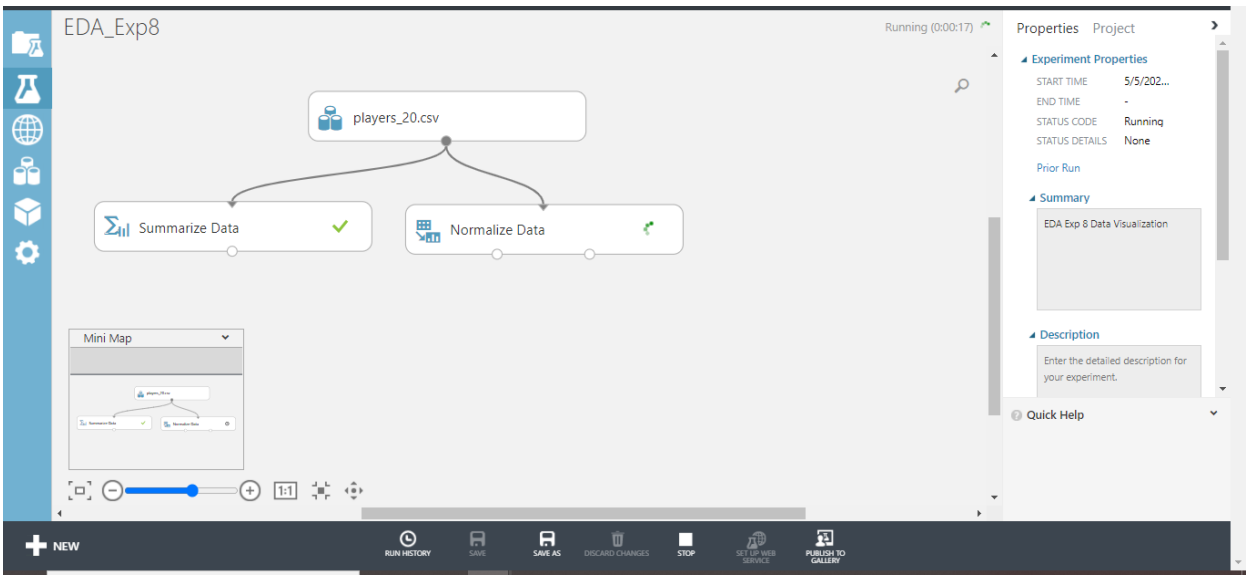
9. From Dataset Transformations select > Scale and reduce > Select Normalize Dataset and Drag it to the workspace. Join the ouput of the Dataset to the input of Normalize Data.



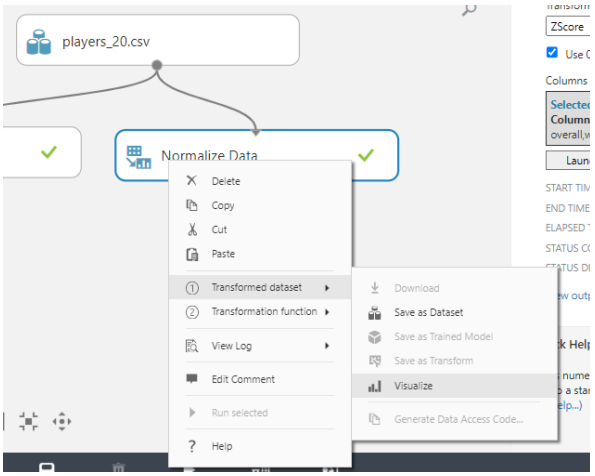
10. Select type of Normalization, I have selected min-max normalization. Launch the column selector and select the columns that you want to be normalized.



11. Run the normalized Dataset

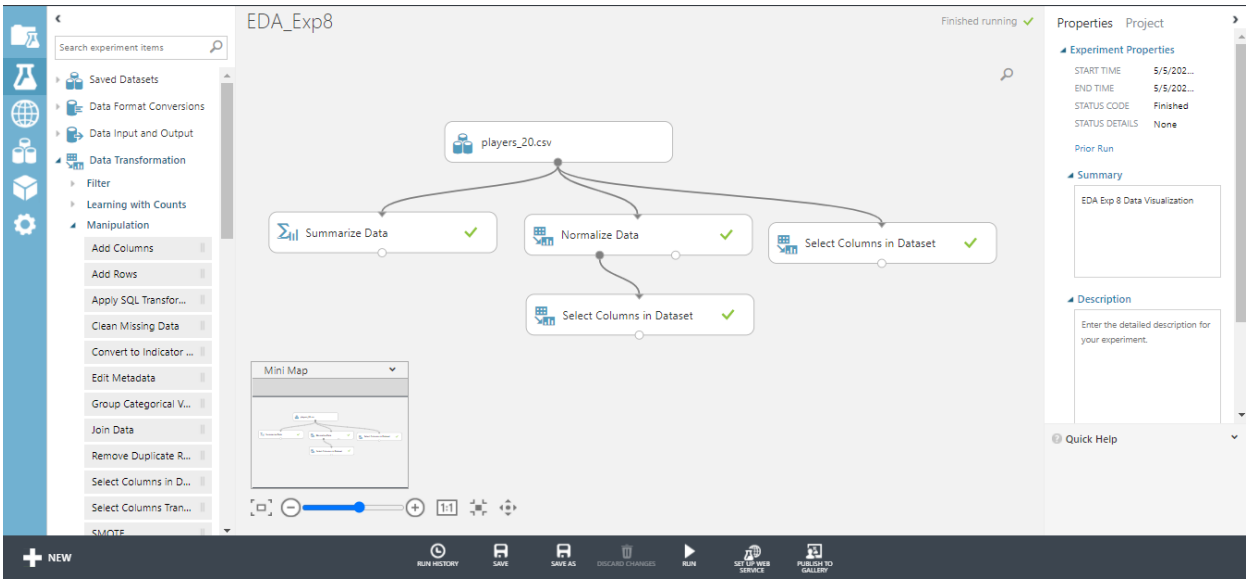


12. Visualize the Normalized Data



13. From Data Transformation > Manipulate > Select and Drag “Select Column from Dataset”. Do this twice.

14. Connect the output of normalized dataset to one “Select Column from Dataset” and output of dataset to other “Select Column from Dataset”. Launch Colum Selector and Select the columns short name, overall and wage.



Then visualize the selected columns

EDA_Exp8 › Select Columns in Dataset › Results dataset

rows 18278 columns 3

short_name	overall	wage_eur
L. Messi	94	565000
Cristiano Ronaldo	93	405000
Neymar Jr	92	290000
J. Oblak	91	125000
E. Hazard	91	470000
K. De Bruyne	91	370000
M. ter Stegen	90	250000
V. van Dijk	90	200000
L. Modrić	90	340000
M. Salah	90	240000

Before Normalization

EDA_Exp8 › Select Columns in Dataset › Results dataset

rows 18278 columns 3

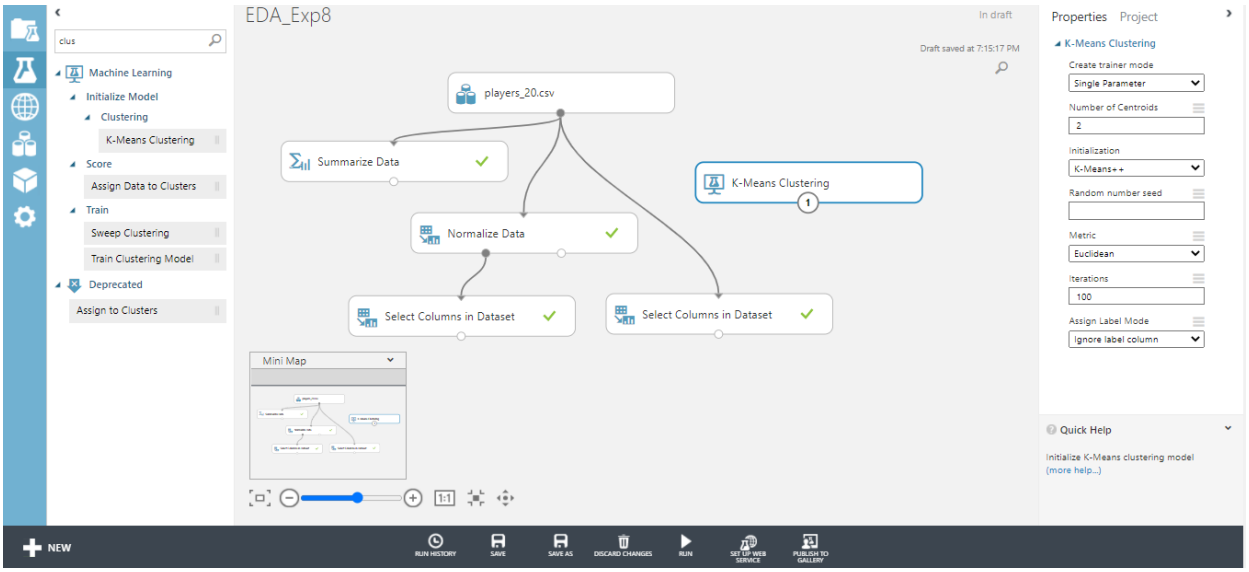
short_name	overall	wage_eur
L. Messi	1	1
Cristiano Ronaldo	0.978261	0.716814
Neymar Jr	0.956522	0.513274
J. Oblak	0.934783	0.221239
E. Hazard	0.934783	0.831858
K. De Bruyne	0.934783	0.654867
M. ter Stegen	0.913043	0.442478
V. van Dijk	0.913043	0.353982
L. Modrić	0.913043	0.60177
M. Salah	0.913043	0.424779

After Normalization

Min-max normalization from Exp 6:

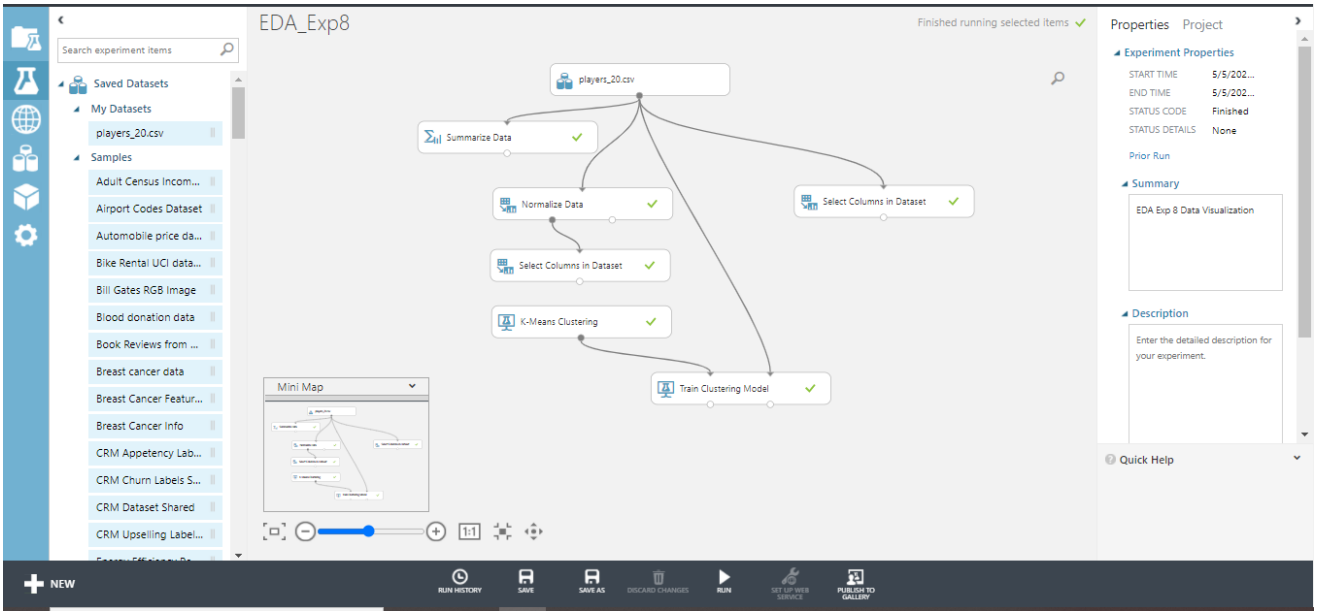
overall	wage_eur
1.0	1.0
0.97826	0.71681
0.95652	0.51327
0.93478	0.22124
0.93478	0.83186
0.93478	0.65487
0.91304	0.44248
0.91304	0.35398
0.91304	0.60177
0.91304	0.42478

15. From Machine Learning > Initialise model > Clustering > Select and drag “K-Means Clustering”. Set the number of clusters as desired.

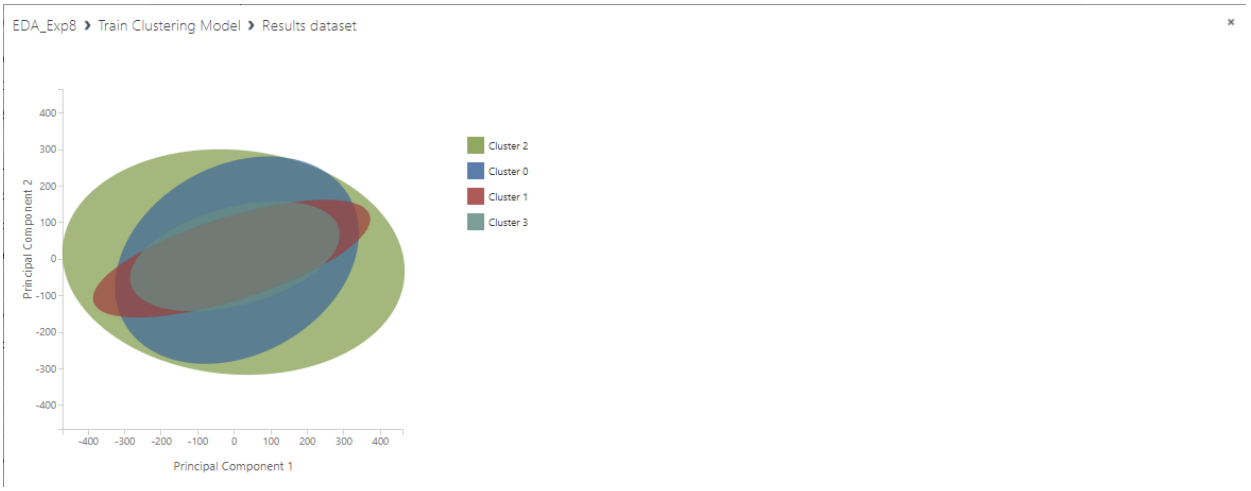


16. Select Train clustering model and drag it to the workspace. Join the output of K-means

Clustering to Input 1 of "Train Clustering Model". Join the output of dataset to input 2 of "Train Clustering Model". Launch Column selector and select columns height_cm and overall.



17. Visualize the data



Questions:

1. What are the different EDA tasks currently available in the Microsoft Azure Machine learning studio?

Ans: Various EDA tasks that can be performed in the Microsoft Azure Machine learning studio are as follows:

1. Create a model
   - Get the data
   - Prepare the data
   - Define features
2. Train the model
   - Choose and apply an algorithm
     a. Classification

- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Networks
- Bayes point machine classification

b. regression

- Bayesian Linear Regression
- Decision Tree
- Decision Forest
- Fast Forest Quantile
- Ordinal
- Neural
- Poisson

3. Train the model
   - Choose and apply an algorithm
   - The following statistics are shown for our model:
     a. Mean Absolute Error (MAE): The average of absolute errors (an error is the difference between the predicted value and the actual value).
     b. Root Mean Squared Error (RMSE): The square root of the average of squared errors of predictions made on the test dataset.
     c. Relative Absolute Error: The average of absolute errors relative to the absolute difference between actual values and the average of all actual values.
     d. Relative Squared Error: The average of squared errors relative to the squared difference between the actual values and the average of all actual values.
     e. Coefficient of Determination: Also known as the R squared value, this is a statistical metric indicating how well a model fits the data.
4. Information Extraction from Receipts: Simple & Complex
5. Develop ML Models to Learn from Past Trends and Forecast Budget
6. Train the Custom Model at Scale with Actual Past Data and Various Data Sources
7. Package the Model and Deploy It for Use by Apps
8. Discretization of Dataset Attributes
9. Normalization of Dataset Attributes
   - Min-Max
   - Z-score
   - Logistic
   - Log Normal
   - Tanh
10. Anomaly Detection

Outcomes: CO4: Comprehend various data visualization techniques and its interpretation

Conclusion:

Different EDA tasks were used, such as Data Normalization (Min-Max Normalization) and Discretization using the K-Means clustering algorithm. We compared the output of Azure's Min-Max Normalization to the Experiment 6 outcome. To visualise data, we used the two data visualisation techniques available on Azure: histograms and boxplots.

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. https://studio.azureml.net/