Clustering of College Dataset

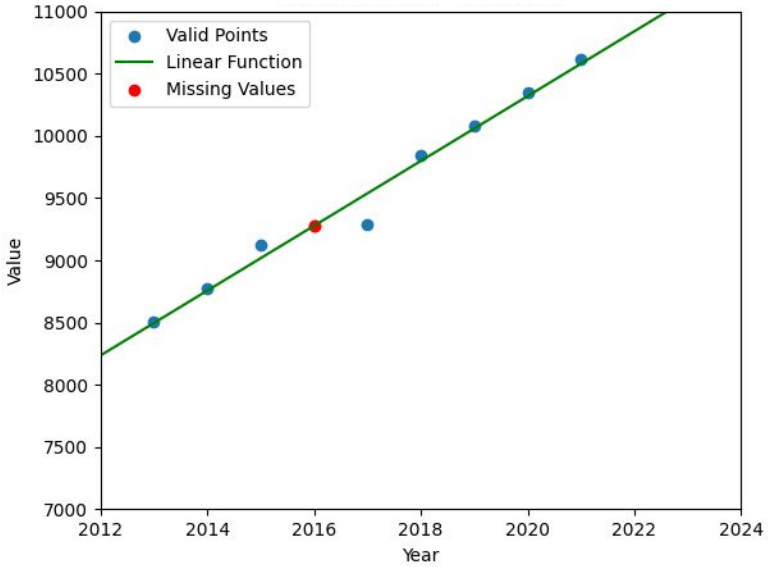
Group 20
Andre Koczka, Igor Lukic, Paul Oberauer, Hans-Peter Wipfler

Data Investigation - Findings

- First column has no name
  - we interpreted it as an enumeration of the data points
- *'Year'* column
  - college information repeats for different years
  - show the cost development over the years
  - colleges report every year
- Missing values
  - for some college id's the whole datapoint is missing
  - missing values in column *'Value'*
  - missing values in column *'Type_2'*
  - missing values in column *'Expense_1'*
  - private colleges do not differentiate between in-state and out-of-state students
- Invalid values
  - *'Year'* is 9999.0
  - *'Value'* is 9999999.0
- Data points with ID above 3547 are redundant

Preprocessing Python

- Data observation: unique values and values counts for each feature
- Replaced wrong inputs like 999999 with nan
- Replacement of missing years (depending on neighbors)
  - except one case is between 2017 and 2018 -> check how often College is listed in other years
- Replacement of state abbreviations (eg. Ar -> Arkansas)
- Conversion of Length from string (2-years) to float (2)
- Replace missing values for Expense_1
  - Fees - Tuition(Expense_2)
  - Room - Board(Expense_2)
- Value estimation with linear regression,
  - take values/years of same College and estimate missing Value
- Removal of empty rows
- Save filtered dataset in new .csv

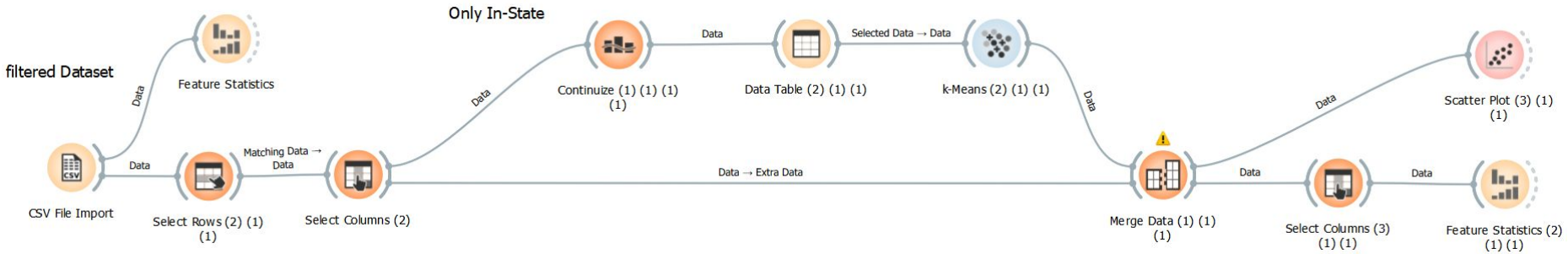


Preprocessing Orange

- Encoding of categorical variables
  - *Expense_2*: ordinal encoding
    - Tuition is more expensive than Board
  - *Type_1*: ordinal encoding
    - Private is more expensive than Public
  - *Type_2*: ordinal encoding
    - Out-of-State is more expensive than In-State
  - *State*: no encoding
    - nominal label
  - *Length*: normalize to interval [0,1]
  - *Value*: normalize to interval [0,1]
- Features for clustering
  - *State, Expense_2, Length, Type_1, Type_2, Value*

Data Processing Pipeline

- Selected college data from *'Year'* 2017
- Remove unnecessary columns
  - IDs, Expense_1, Year
- Second path is used to have non-normalized values in plot
- Clustering done by using k-Means in the more dimensional space
  - Alternatives: DBScan, Hierarchical clustering
  - Alternatives: Dimensionality reduction (PCA, t-SNE)



Clustering

Problem description:

Cluster the colleges based on how much they charge students who live on campus by using the columns that show the average costs for public and private nonprofit institutions

Outcome:

- Analyze how many clusters are found
- Analyze what are the reasons for the clusters
- Analyze how the clusters are distributed
- Analyze effect of pre-processing