ISP642: BUSINESS INTELLIGENCE

Graduate Tracer Study

DATA UNDERSTANDING AND PREPARATION

PREPARED BY

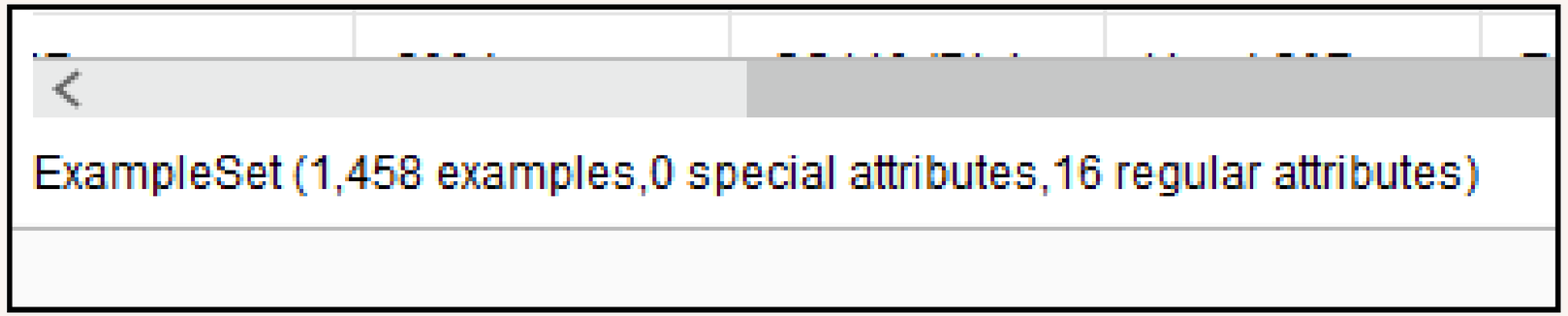
AINA FARZANA ZULKIFLI | 2021817142
NUR EISYATIN RADHIAH ANNUAR | 2021610238
YASMIN NABILA OTHMAN | 2021619506
SITI MUKHLISAH MUSKAMAL | 2022815898
CS2705B

DATA UNDERSTANDING

Data Exploration

1 SIZE OF DATASET

Number of Record: 1458 data
Attributes: 16



Data Exploration

2 DATA TYPES

NO	ATTRIBUTES	DATA TYPE
1.	Timestamp	Interval
2.	Gender	Nominal
3,	Age	Ratio
4.	Graduation Year	Ratio
5.	Program Code	Nominal
6.	Current Occupation	Nominal
7.	Job Category	Nominal
8.	Department	Nominal
9.	Company Name	Nominal
10.	Type of Company	Nominal
11.	Company's nature of business	Nominal
12.	Years of work experience (after graduation)	Ratio

Data Exploration

2 DATA TYPES

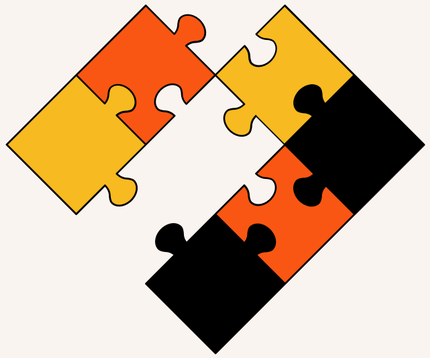
NO	ATTRIBUTES	DATA TYPE
13.	Your Current Salary Range	Nominal
14.	In your opinion, what are the THREE most important skills required from fresh graduates when entering the workforce?	Nominal
15.	In your opinion, what subjects should be added into the existing academic curriculum in order to better prepare our graduates for the industry?	Nominal
16.	What advice would you give to your juniors who are still in the university?	Nominal

Data Exploration

3 STAND OUT CANDIDATES

- Yes, there is a potential attributes which is **Program Code Attributes.**
- From **Program Code attributes**, we can predict the future job based on the historical data.

Data Quality Assessment



1 MISSING VALUES & HOW TO HANDLE THEM?

- 3/16 attributes that have missing values

Handling this problem:

- Use the "Replace Missing Values" operator to replace missing values with the mean, median, or other imputation method.

Name	Type	Missing
Type of Company :	Nominal	0
Company's nature of business :	Nominal	0
Years of work experience (afte...	Integer	0
Your current salary range :	Nominal	0
In your opinion, what are the T...	Nominal	145
In your opinion, what subjects ...	Nominal	338
What advice would you give to ...	Nominal	272

Examples of missing values

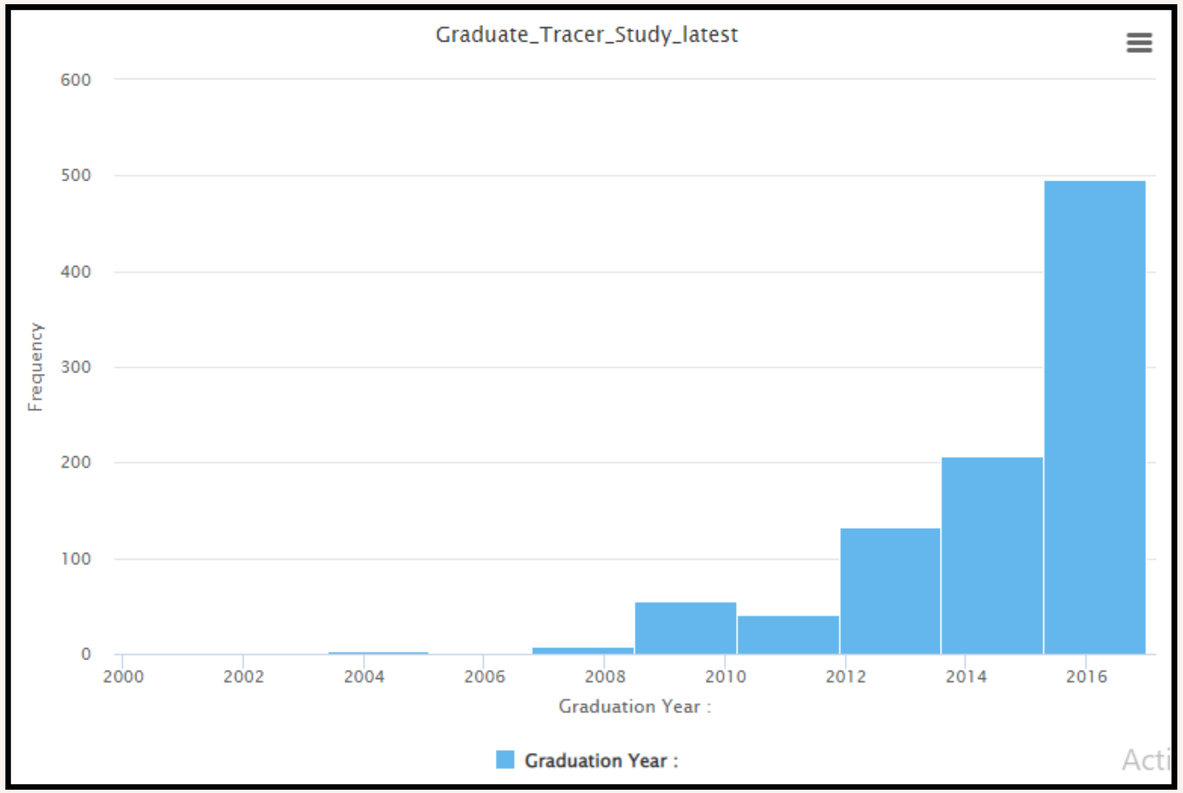
Data Quality Assessment

2 OUTLIERS/ANOMALIES

- There are some outliers found
- For example in Graduation Year attibute, the outlier is 2002 and 2014

Impact:

- Outliers can impact the mean and standard deviation,
- affecting the performance of data mining process
- affecting the accuracy of the prediction result



Data Quality Assessment

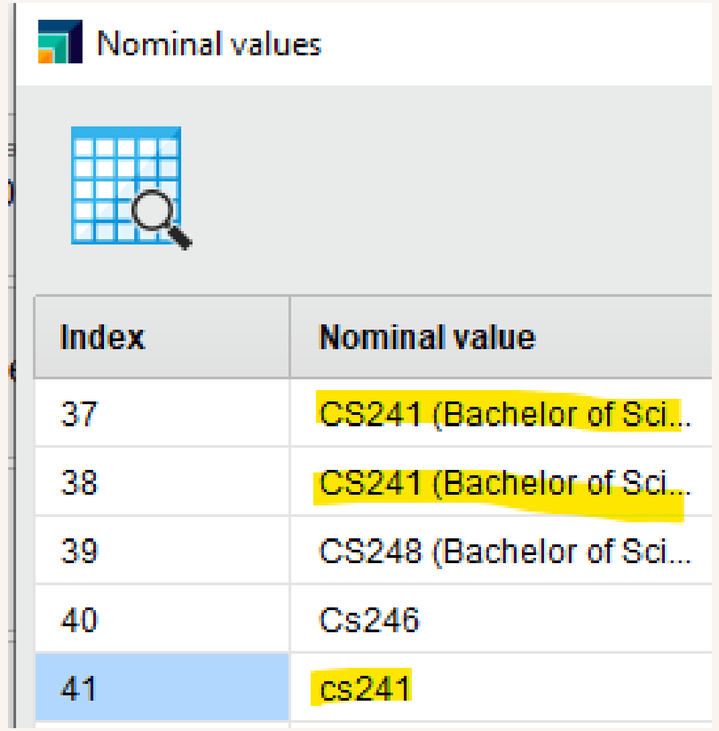
3 DATA ACCURACY AND CONSISTENCY

Analysis:

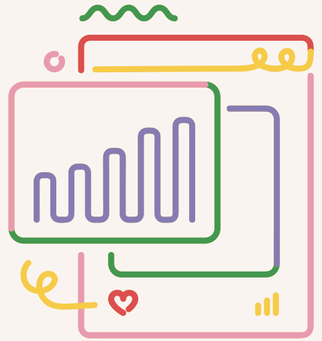
- Check for the consistency across related columns
- Look for disrepancies or unexpected patterns

Is there any discrepancies?

- Yes, there have various data name but have similar meaning
- For example : Program Code Attributes
- CS241 which refer to bachelor of science (hons) statistics have 3 different name/spelled



Statistical Summary



1 STATISTICAL SUMMARY OF KEY ATTRIBUTES

Key attribute : Years of work experience (after graduation)

Row No.	average(Years of wor...	median(Years of w...	mode(Years of wo...	variance(Years of work experience (after graduation...
1	2.637	1	1	7.254

Key attribute : Your current salary

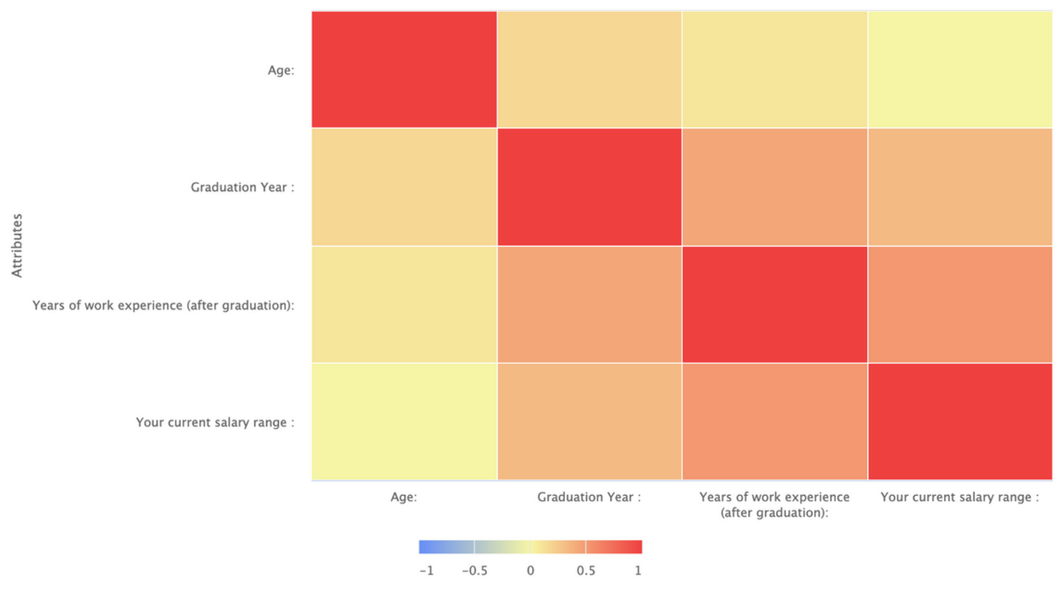
- Salary Range (Data Type : Nominal) is converted to Salary Midpoint (Data Type : Numerical)

Row No.	average(Your current sal...	median(Your current s...	mode(Your current salar...	variance(Your current salary ran...
1	2893.783	2500.500	2500.500	2560688.689

Statistical Summary

2 PAIRS OF ATTRIBUTES THAT SHOW A STRONG CORRELATION

Attributes	Age:	Graduation Year :	Years of work experience ...	Your current salary range :
Age:	1	0.173	0.094	-0.015
Graduation Year :	0.173	1	0.438	0.332
Years of work experien...	0.094	0.438	1	0.518
Your current salary ran...	-0.015	0.332	0.518	1



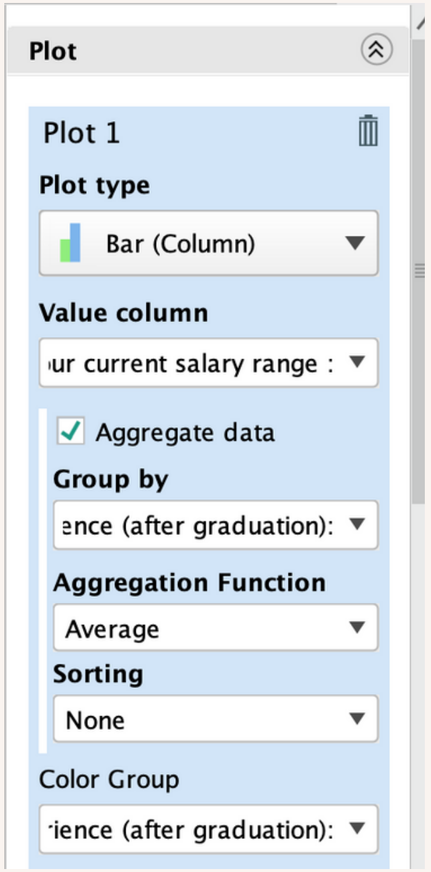
Pair of attributes that show a strong correlation are **'Your Current Salary'** and **'Years of Work Experience (after graduation)'** with correlation coefficient = 0.518.

Data Visualization

1 DATA VISUALIZATION TECHNIQUE

Bar Chart (Column) is used to visualize the **average of salary** received by workers based on their **years of work experience after graduating**

Bar Chart is used because of its clear visualization of data and easy to compare each attributes with another

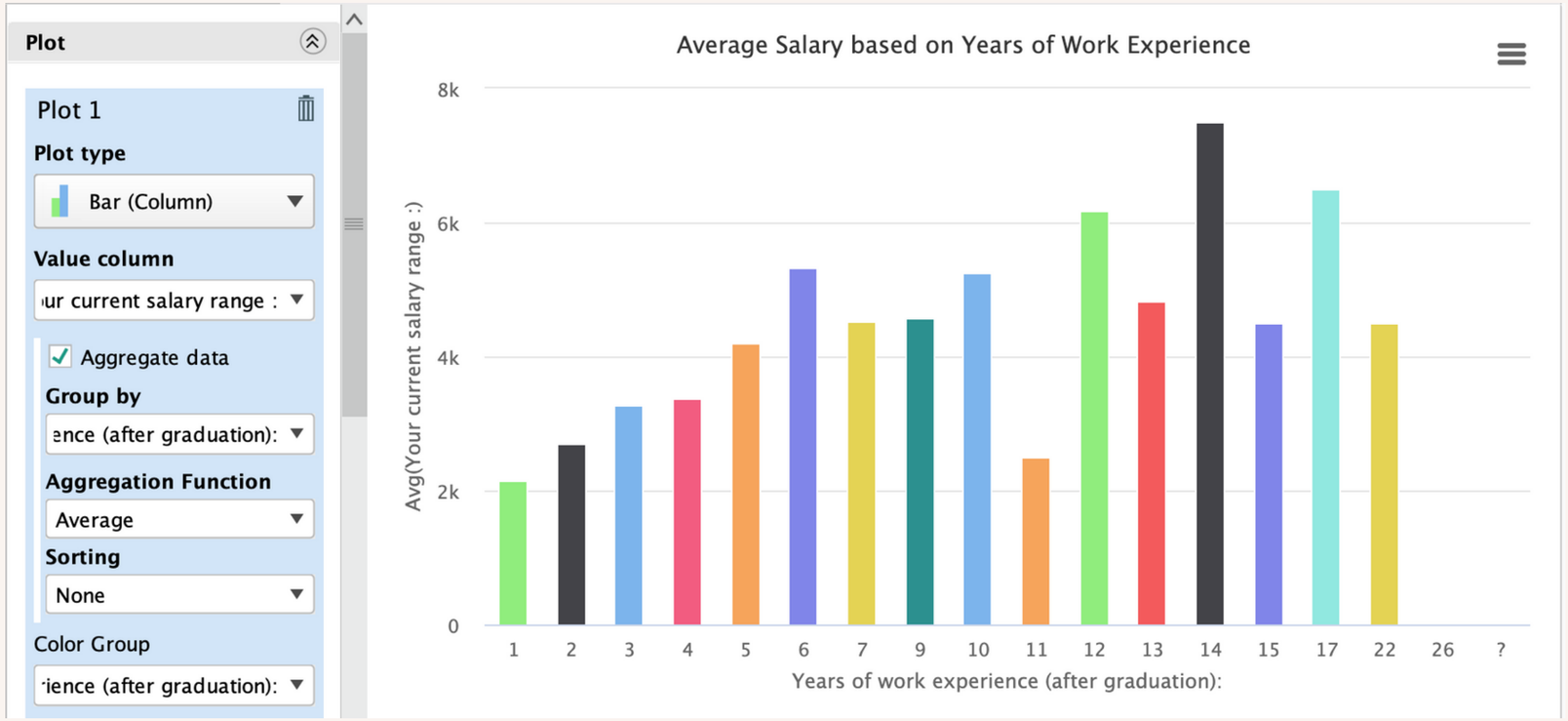


2 RESULT

Value Column is set to be Current Salary to view how much salary is received by workers in Ringgit Malaysia (RM).

Aggregation function is set to Average to find the average of current salary received by the workers

3 RESULT



Based on this data visualization, individuals with **higher years of work experience** tend to receive **higher amount of salary**

However in some cases, higher years of work experience do not guarantee workers to receive higher amount of salary

DATA PREPARATION

Data Cleaning

1 DATA CLEANING PROCESS

Describe the data cleaning process you undertook.

1. Replace Missing Values
   - Use “Replace Missing Values” filter
2. Attribute Selection
   - Excluding irrelevant attributes simplifies the model, making it easier to understand and interpret
   - Use “Select Attributes” filter

2 STRATEGY IN CORRECTING ERROS

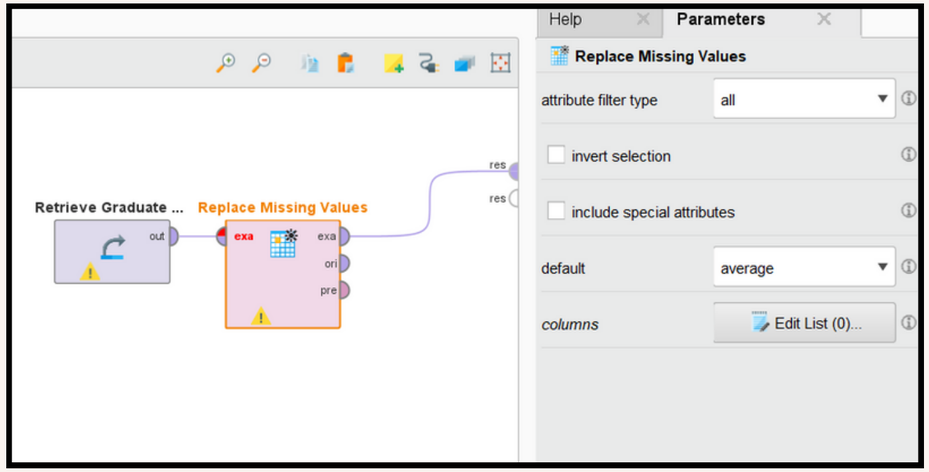
Did you have to correct any errors in the data? What strategy did you use?

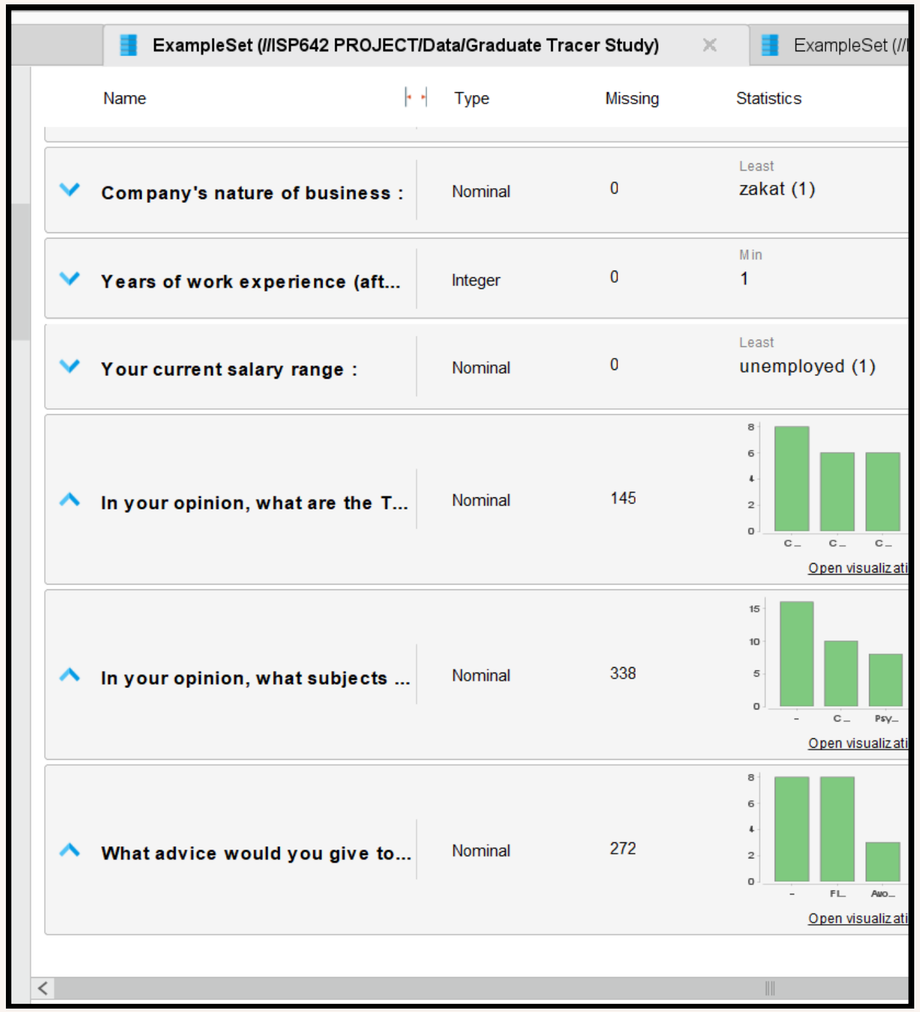
For column “Program Code”, the data may vary, some may own the same meaning but differ because of its capitalization, typo, etc.

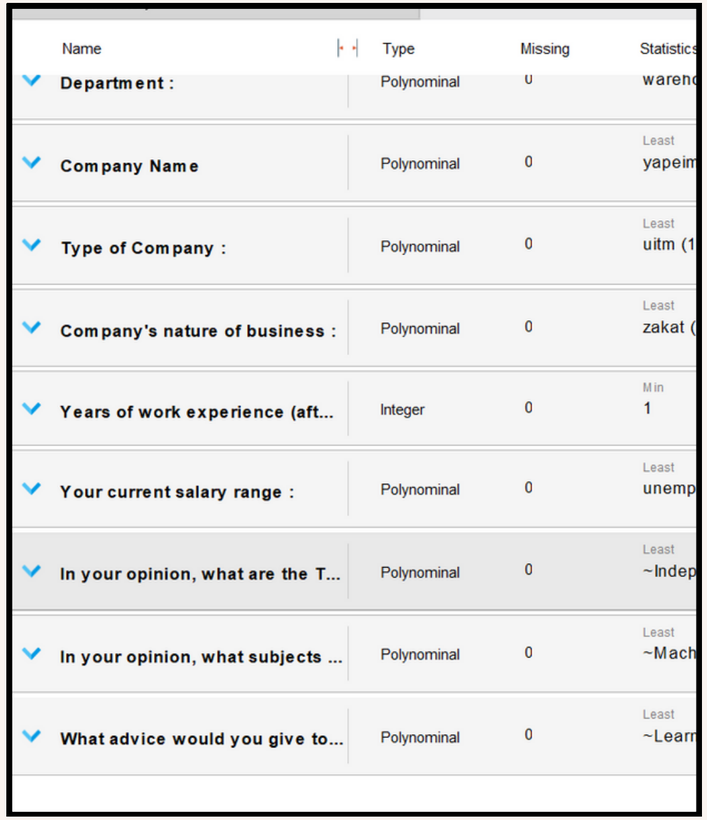
- Find and Replace:
  - Replacing specific words or phrases with others throughout the entire worksheet using Google Sheets.
  - For example, “cs230” and “CS230” are all replaced with “CS230”
  - Allow cleaning of data and ensure consistency

Results (Data Processing)

1 REPLACE MISSING VALUES

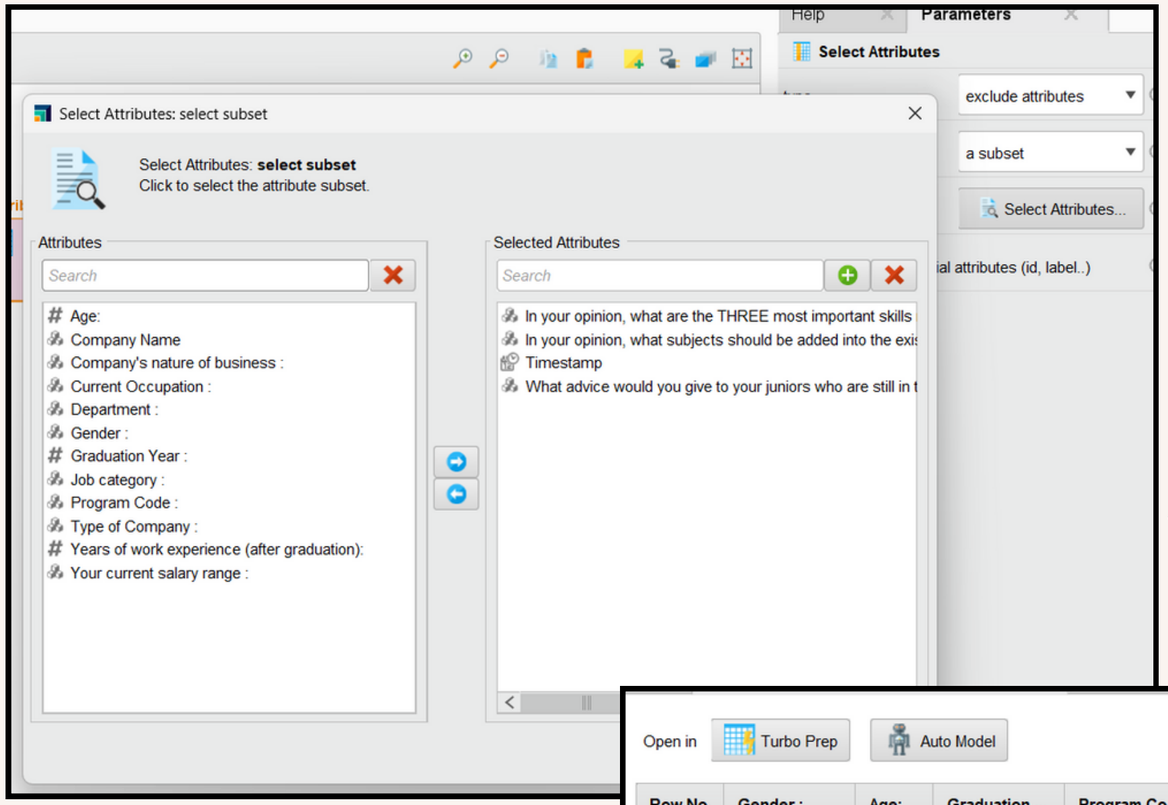


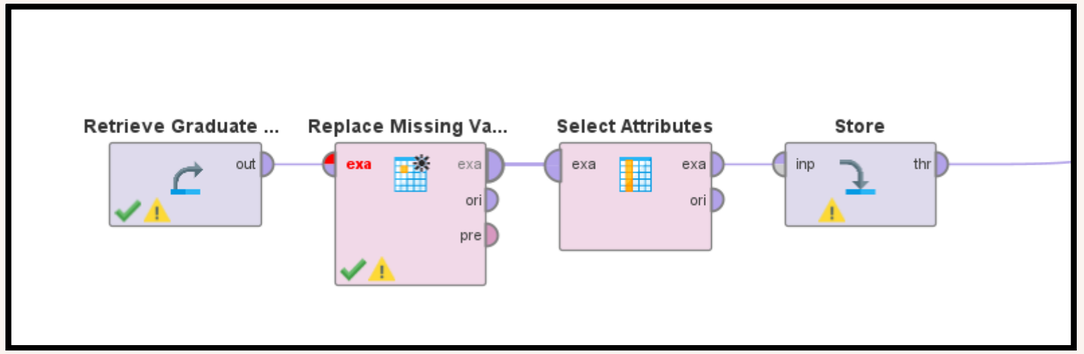




Results (Data Processing)

2 ATTRIBUTE SELECTION





Open in Turbo Prep Auto Model Filter (1,458 / 1,458 examples): all

Row No.	Gender :	Age:	Graduation ...	Program Co...	Current Occ...	Job category :	Department :	Company Na...	Type of Co...	Company's ...	Years of wor...	Your curren...
1	Male	26	2014	CS221	Banker	Executive	Credit Manage...	Agrobank	GLC	Bank for agric...	3	RM 3001 - RM...
2	Female	25	2017	CS224 / CS24...	Purchasing co...	Non Executive	Finanve	Sentoria grouo...	Private Limited...	Hospitality	1	RM 1000 - RM...
3	Male	26	2017	Cs253	PI1M Kg Tana...	Manager	Pelaksana	MSD Digital Int...	Private Limited...	Information Te...	1	RM 1000 - RM...
4	Female	29	2012	Bachelor of Sc...	Senior Test En...	Executive	Engineering	100 Doors Sd...	Private Limited...	Information Te...	5	RM 5001 - RM...
5	Female	25	2016	CS110 (Diplo...	Web developer	Executive	Ecommerce	PETSMORE ...	Private Limited...	Services	2	RM 2001 - RM...
6	Male	26	2016	CS241(Rachel ...	Permanent	Executive	Technical	Statworks (M)	Private Limited	Customer Ser...	2	RM 2001 - RM...

Results (Handling data errors)

2 ATTRIBUTE SELECTION

Row No.	Gender :	Age:	Graduat...	Program Code :
1	Male	26	2014	CS221
2	Female	25	2017	CS224 / CS244 (Bachelor of IT (Hons.) Business Computing)
3	Male	26	2017	Cs253
4	Female	29	2012	Bachelor of Science (Hons) Information Systems Engineering
5	Female	25	2016	CS110 (Diploma in Computer Science)
6	Male	26	2016	CS241(Bachelor of Science (Hons.) Applied Statistics
7	Male	24	2017	CS245
8	Female	24	2017	Bachelor of Science (Hons.) Statistics
9	Female	25	2016	Bachelor of Science (Hons.) Statistic
10	Female	27	2017	CS221
11	Male	30	2012	Cs226
12	Female	25	2015	CS222 Actuarial Science
13	Female	24	2017	CS249 (Bachelor of Science (Hons.) Mathematics)
14	Male	29	2012	CS110 (Diploma in Computer Science)

Gender :	Age:	Graduation Year :	Program Code :	Cu
Male	26	2014	CS221	Ba
Male	26	2017	CS253	PI
Female	25	2016	CS110	W
Male	26	2016	CS241	Pe
Male	24	2017	CS245	Ne
Female	24	2017	CS241	Da
Female	25	2016	CS241	Fir
Female	27	2017	CS221	Le
Male	30	2012	CS226	Ex
Female	25	2015	CS222	Tr
Female	24	2017	CS249	Ca
Male	29	2012	CS110	Pr
Female	24	2016	CS230	Sy
Male	27	2012	CS251	So
Female	27	2013	CS226	SA
Male	28	2012	CS251	So

Data Transformation

1 DATA TRANSFORMATION PROCESS

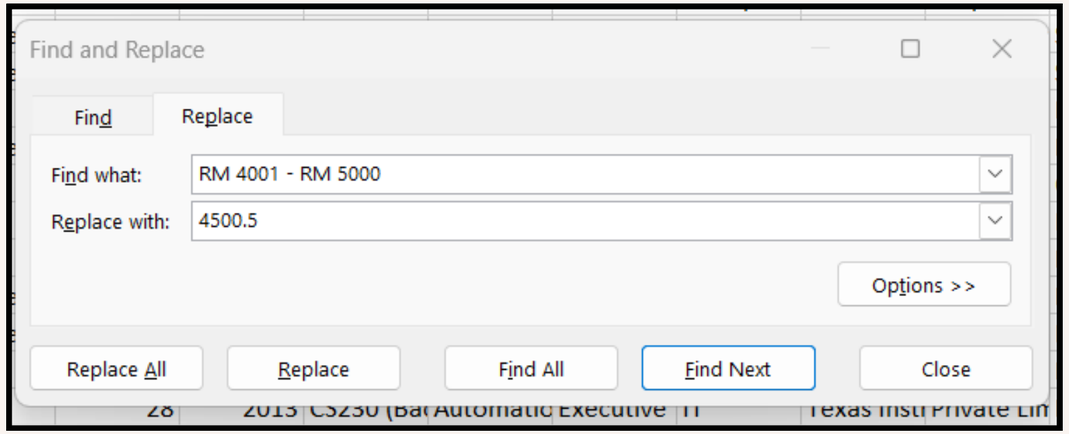
In this dataset, we have transformed the attribute Current Salary Range.

The values of this attribute contains a range where it inconveniences the process to search for the statistical summary.

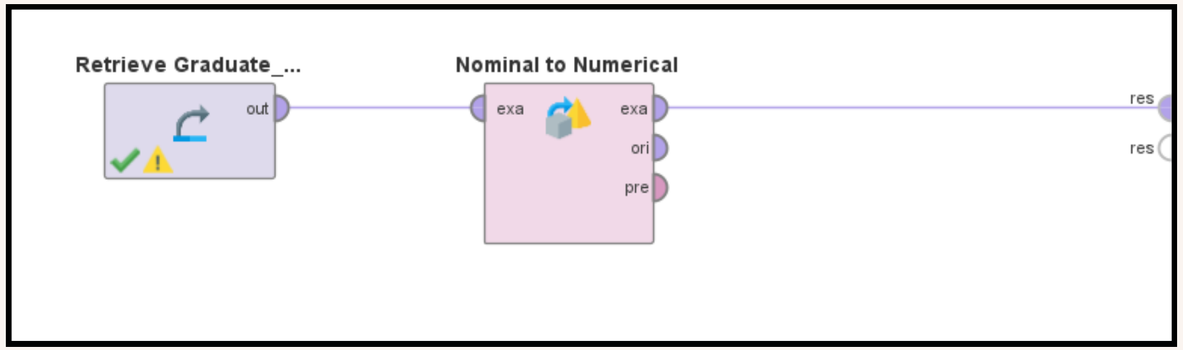
Then, the attribute is converted from nominal to numerical

Process (Data Transformation)

1) Using Excel, using the Find and Replace feature, the “RM “ and “ - “ is removed and the midpoint replaces the range of salary.



2) In RapidMiner, the Numerical to Nominal operator is used to convert the data type of the Salary Range attribute.



2 EXAMPLES (USING EXCEL)

f w	Your current salary range :
3	RM 3001 - RM 4000
5	RM 5001 - RM 10000
2	RM 2001 - RM 3000
2	RM 2001 - RM 3000
1	RM 3001 - RM 4000
1	RM 3001 - RM 4000
1	RM 2001 - RM 3000
1	RM 2001 - RM 3000
6	RM 5001 - RM 10000
2	RM 3001 - RM 4000
1	RM 2001 - RM 3000
4	RM 3001 - RM 4000
3	RM 4001 - RM 5000
2	RM 4001 - RM 5000
4	RM 3001 - RM 4000
5	RM 4001 - RM 5000
4	RM 2001 - RM 3000
2	RM 2001 - RM 3000
5	RM 5001 - RM 10000
5	RM 5001 - RM 10000
3	RM 2001 - RM 3000
7	RM 4001 - RM 5000
3	RM 2001 - RM 3000
7	RM 2001 - RM 3000
3	RM 2001 - RM 3000
2	RM 3001 - RM 4000

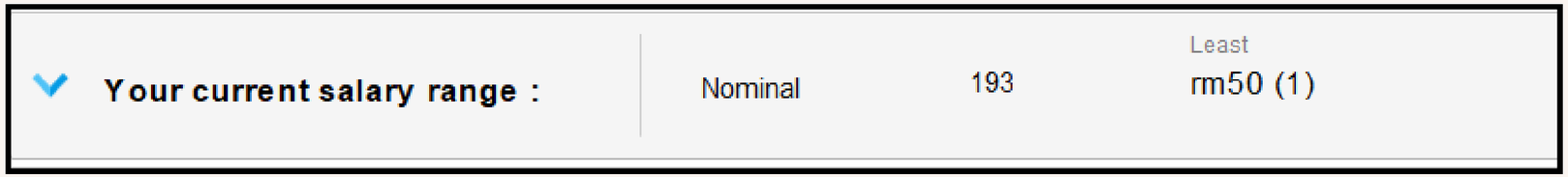
before

Your current salary range :
3500.5
7500.5
2500.5
2500.5
3500.5
3500.5
2500.5
2500.5
7500.5
3500.5
2500.5
3500.5
4500.5
4500.5
3500.5
4500.5
2500.5
2500.5
7500.5
7500.5
2500.5
4500.5
2500.5
2500.5

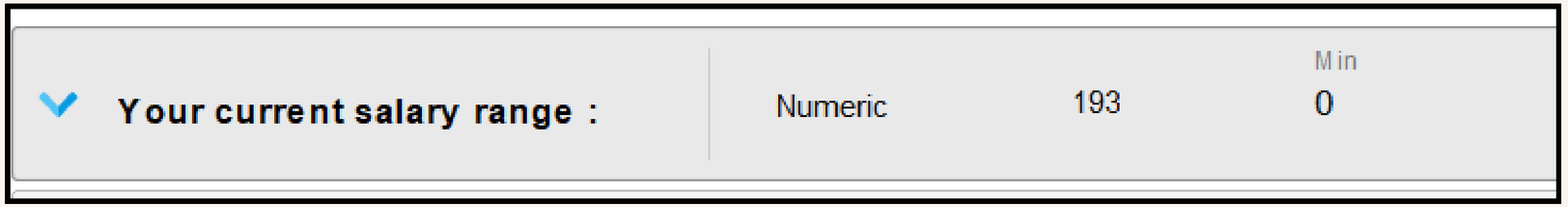
after

2 EXAMPLES (USING RAPIDMINER)

before



after



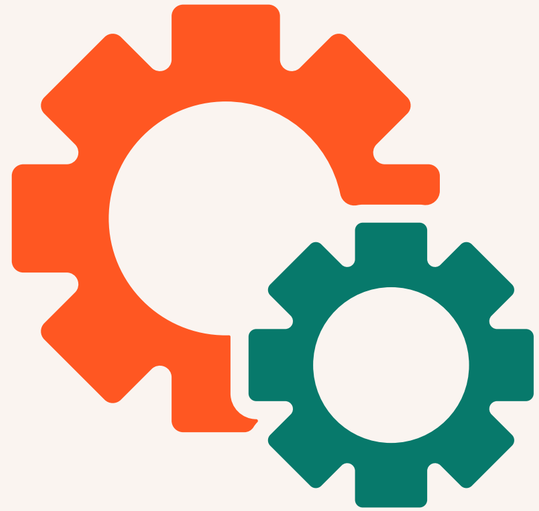
Feature Selection and Engineering

1 DECIDING ATTRIBUTES

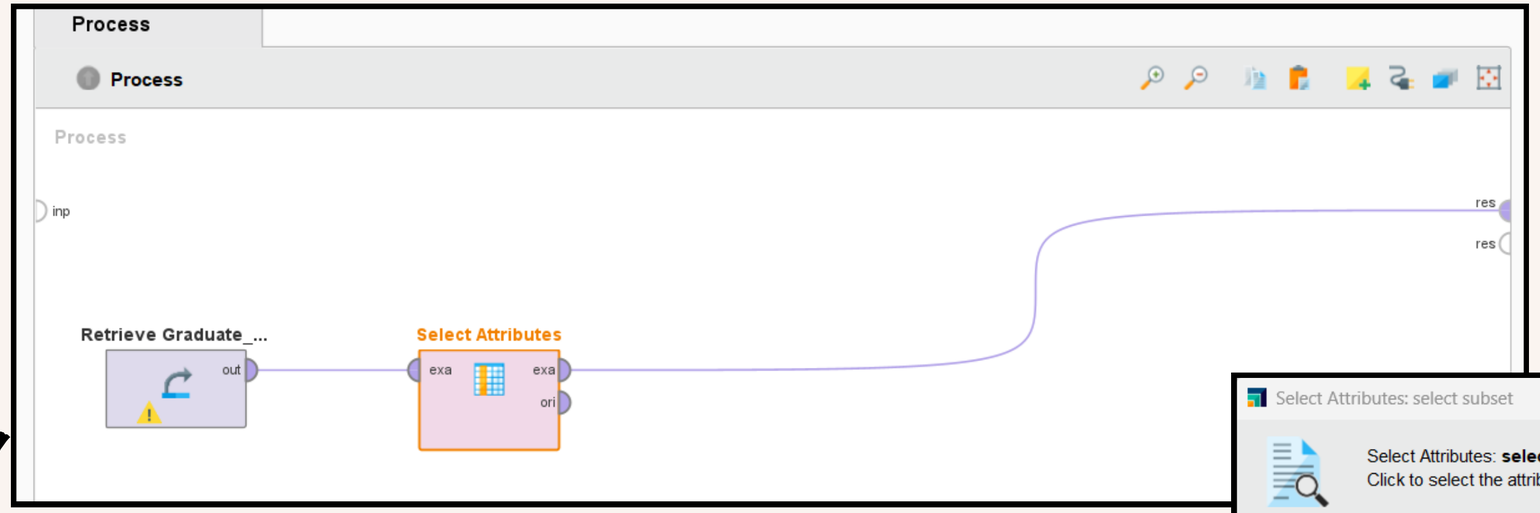
We decided by analyzing the attributes and which showed significant relationship with the target outcome.

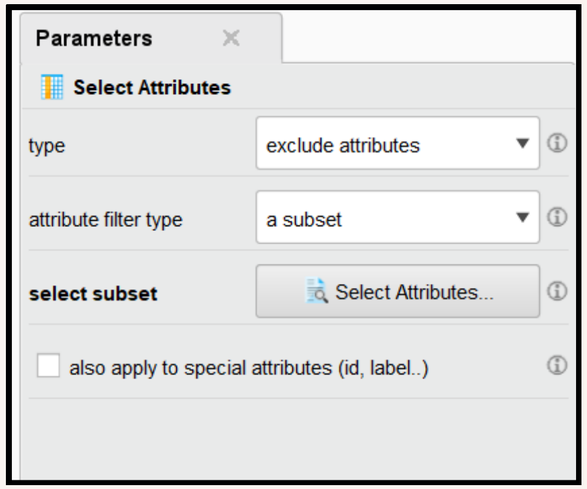
In this project, we have chosen Gender, Age, Program Code, Current Occupation, Job Category, Department, Company Name, Type of Company, Company's nature of business, Years of Working and Current Salary Range.

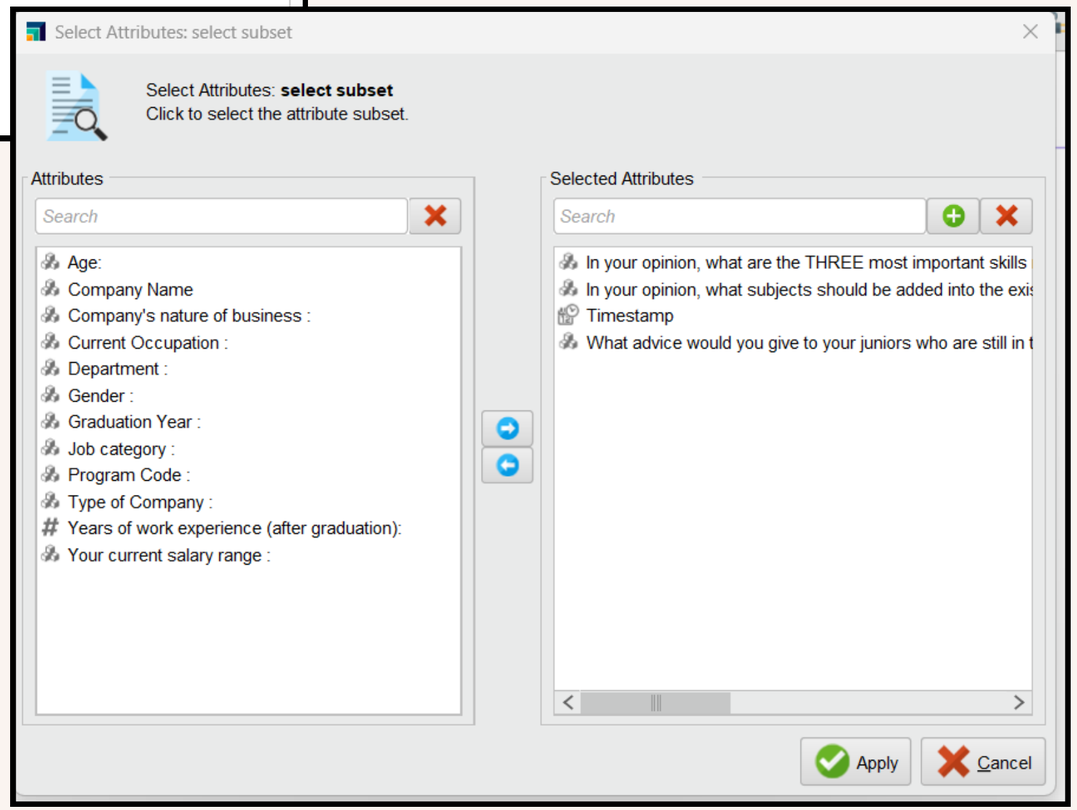
This is done by excluding the attributes using the Select Attributes operator in RapidMiner.



Process (Deciding Attributes)







2 NEW ATTRIBUTES AND WHY

The new attribute created is the Salary Midpoint. This attribute is created by transforming the values of “Your current salary range: “ attribute into the midpoint of the salary range.

The reason is because it is to increase the precision of the data. Other than that, the attribute can be used to perform statistical summaries such as the average and the median.

Thank You.