Data Engineering Professional Project: Adult Dataset

Citizen Data Science Program – Level 1

Guidelines:

1. Answer each question with a clear and concise explanation of the Operator you used. Explain why you chose that particular operator and how it helps.
2. Outline the steps you followed to arrive at the solution.
3. Provide a screenshot of your process that demonstrated the relevant portion of your answer.
4. Name the document as the {project name - your name} before submitting it.
5. Ensure that you include both your answers word document and RapidMiner process as attachments in a reply to the person who sent you the assessment.
6. Write your name and email at the bottom of this page

Project Description

This project consists of 15 questions related to your process in RapidMiner of Adult dataset. The questions are designed to assess your understanding of the key concepts and your ability to apply them in practical scenarios

You will work on the Adult dataset using RapidMiner. The dataset contains information about person's personal income, including attributes like age, workclass, education, occupation, and whether the income level is more than $50 K per year. You will perform various data engineering tasks to gain insights from this dataset.

Name: Abdullah Mohammad Al Talaq

Email: 30786223@sabiccorp.sabic.com, AlTalaqA@sabic.com

Section 1: Data Import

1. Import the Adult dataset into RapidMiner.

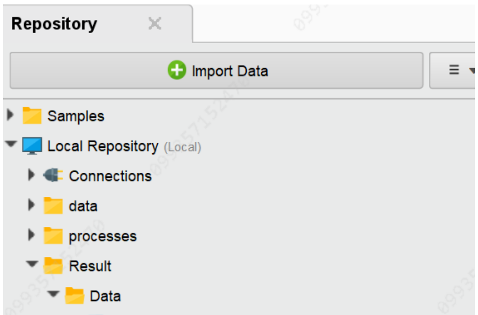


Figure 1 Load Data

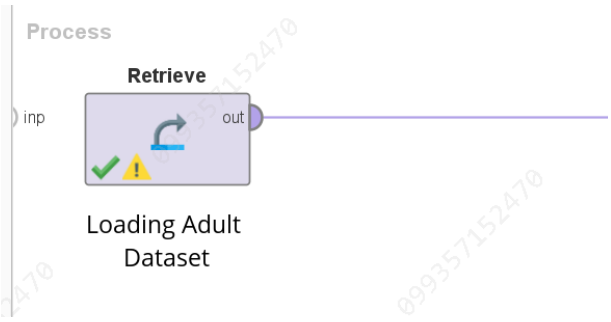
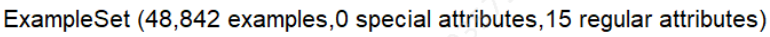


Figure 2 : Box Retrieve Data

By clicking on Import Data, the app will let me to browse the pc and select the data set, after this will show small box Contains the data so I can use operators on it.

2. Provide a brief description of the dataset, including the number of rows, columns, and data types present in the dataset.



ExampleSet (48,842 examples,0 special attributes,15 regular attributes)

Figure 3 : Description Data

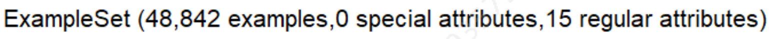
Obviously, the data contains 48,842 record and 15 attributes.

Feature Name	*Type*	*Missing Value*
Age	Int	0
WorkClass	Nominal	2799 (not detected)
Fnlwgt	Int	0
Education	Nominal	0
Education-num	Int	0
Marital-Status	Nominal	0
Occupation	Nominal	2809 (not detected)
Relationship	Nominal	0
Race	Nominal	0
Gender	Nominal	0
Capital-gain	Int	0
Capital-loss	Int	0
Hours-per-week	Int	0
Native country	Nominal	857 (not detected)
income	Nominal	0

Table 1 : data description

Section 2: Data Exploration

1. Determine the total number of individuals in the dataset.

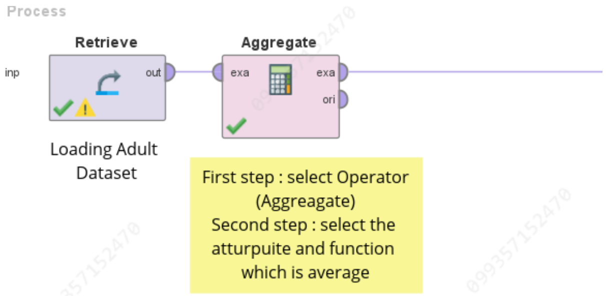


ExampleSet (48,842 examples,0 special attributes,15 regular attributes)

Figure 4 :Description Data

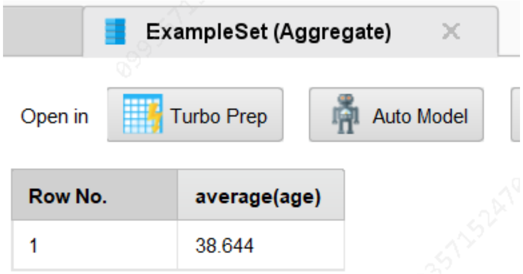
The data contains 48,842 records.

2. Calculate the average age of individuals in the dataset.



First step : select Operator (Aggreagate)
Second step : select the atturpuite and function which is average

Figure 5 : Aggregate operation

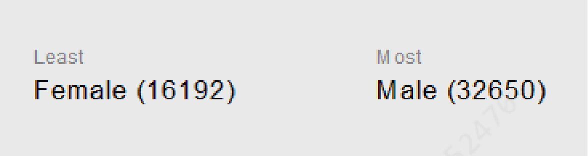


Row No.	average(age)
1	38.644

Figure 6 : Aggregate Result

By using Aggregate operator will give you option to select aggregate function and aggregate attribute

3. Identify the number of male and female in the dataset.



Least
Female (16192)

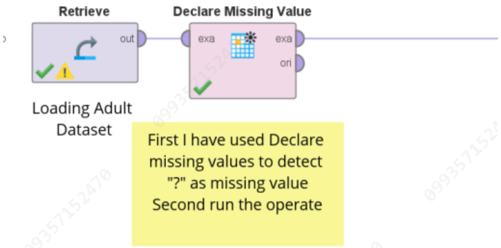
Most
Male (32650)

Figure 7 : Count Result

Selecting the results tab and select statistic option than select the attribute will give you sample details like count of each value.

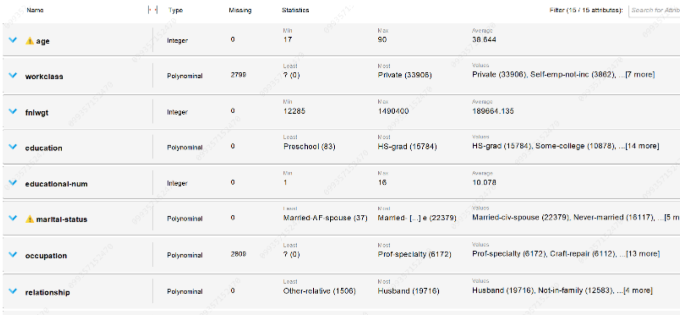
Section 3: Data Cleaning

1. Identify columns with missing values in the dataset if any.



First I have used Declare missing values to detect "?" as missing value Second run the operate

Figure 8 : Operation for declare missing value.



Name	Type	Missing	Statistics			Filter (15 / 15 attributes): Search for Attrib
age	Integer	0	Min 17	Max 90	Average 38.644	
workclass	Polynominal	2799	Least ? (0)	Most Private (33906)	Values Private (33906), Self-emp-not-inc (3862), ...[7 more]	
fnlwgt	Integer	0	Min 12285	Max 1490400	Average 189664.135	
education	Polynominal	0	Least Preschool (83)	Most HS-grad (15784)	Values HS-grad (15784), Some-college (10878), ...[14 more]	
educational-num	Integer	0	Min 1	Max 16	Average 10.078	
marital-status	Polynominal	0	Least Married-AF-spouse (37)	Most Married- [...] e (22379)	Values Married-civ-spouse (22379), Never-married (16117), ...[5 m	
occupation	Polynominal	2809	Least ? (0)	Most Prof-specialty (6172)	Values Prof-specialty (6172), Craft-repair (6112), ...[13 more]	
relationship	Polynominal	0	Least Other-relative (1506)	Most Husband (19716)	Values Husband (19716), Not-in-family (12583), ...[4 more]	

Figure 9 : Result of Detect Null Values

By using Declare missing values operator and declare the parameters like which mode and value to detect as null value (missing value)

2. Remove the unnecessary columns that may not be relevant for analysis and justify your selection.

Removed Education for duplication with Education number.
Removed Marital status for more details can't be used in our situtaion.
Removed Captial Gain and Loss because most of the values are 0.
Removed Native Country.

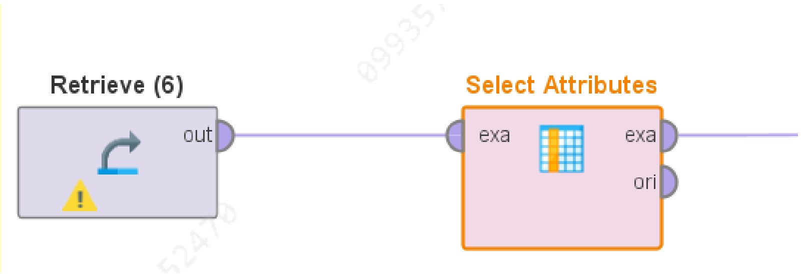
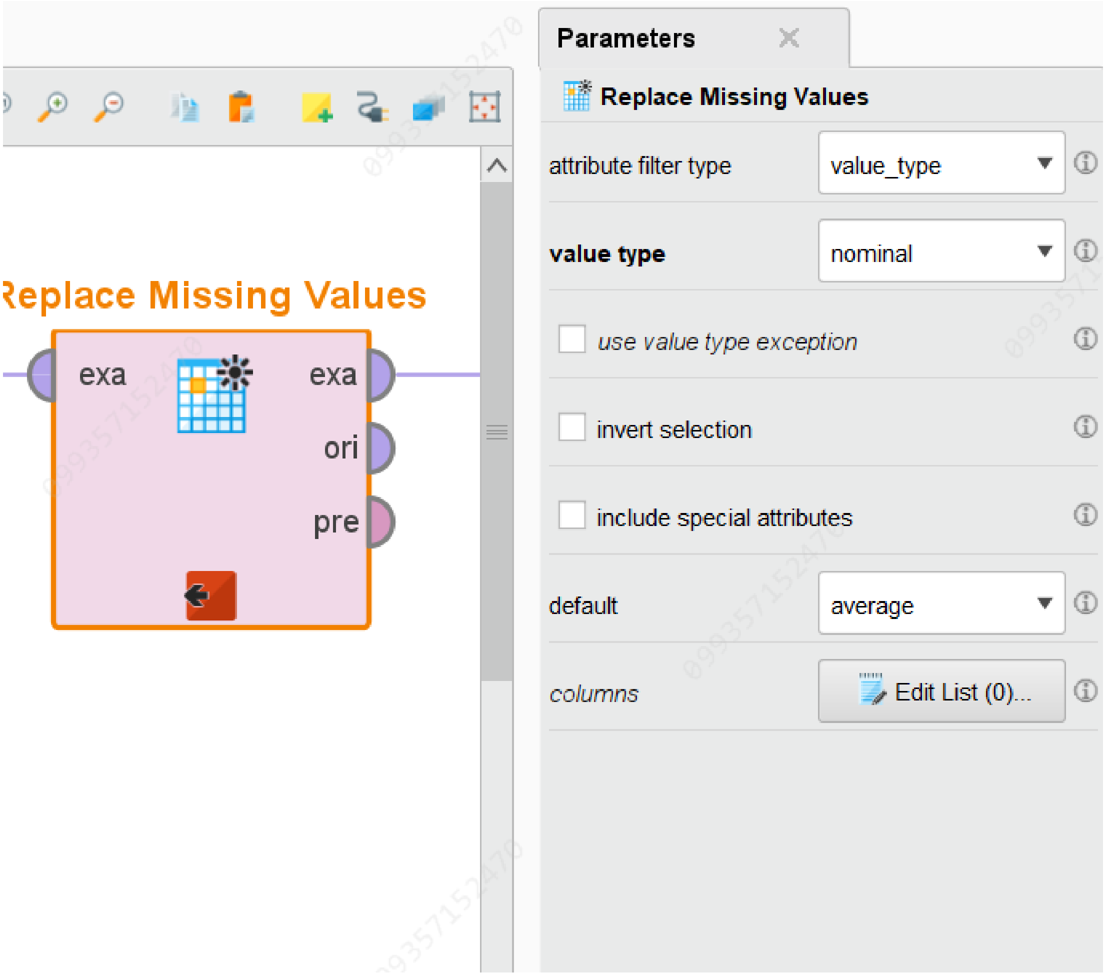


Figure 10 : The operation for deleting selected attributes.

3. Verify if the columns contain values that are not valid or not meaningful and handle them appropriately.



Parameters

Replace Missing Values

attribute filter type: value_type

value type: nominal

□ use value type exception

□ invert selection

□ include special attributes

default: average

columns: Edit List (0)...

Figure 11 : Replace missing value using average.

Using the Replace missing values

Section 4: Data Transformation

1. Generate a new attribute 'net-capital' that represent the difference between the capital-gain & capital-loss for each individual.

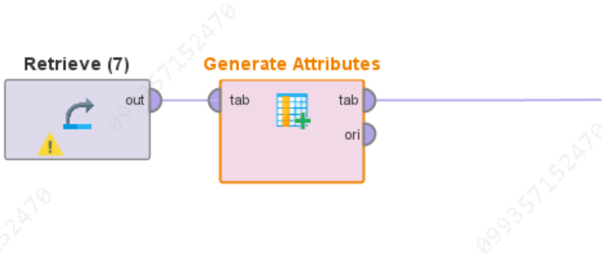
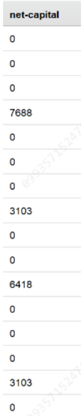


Figure 12 : Generate Attributes

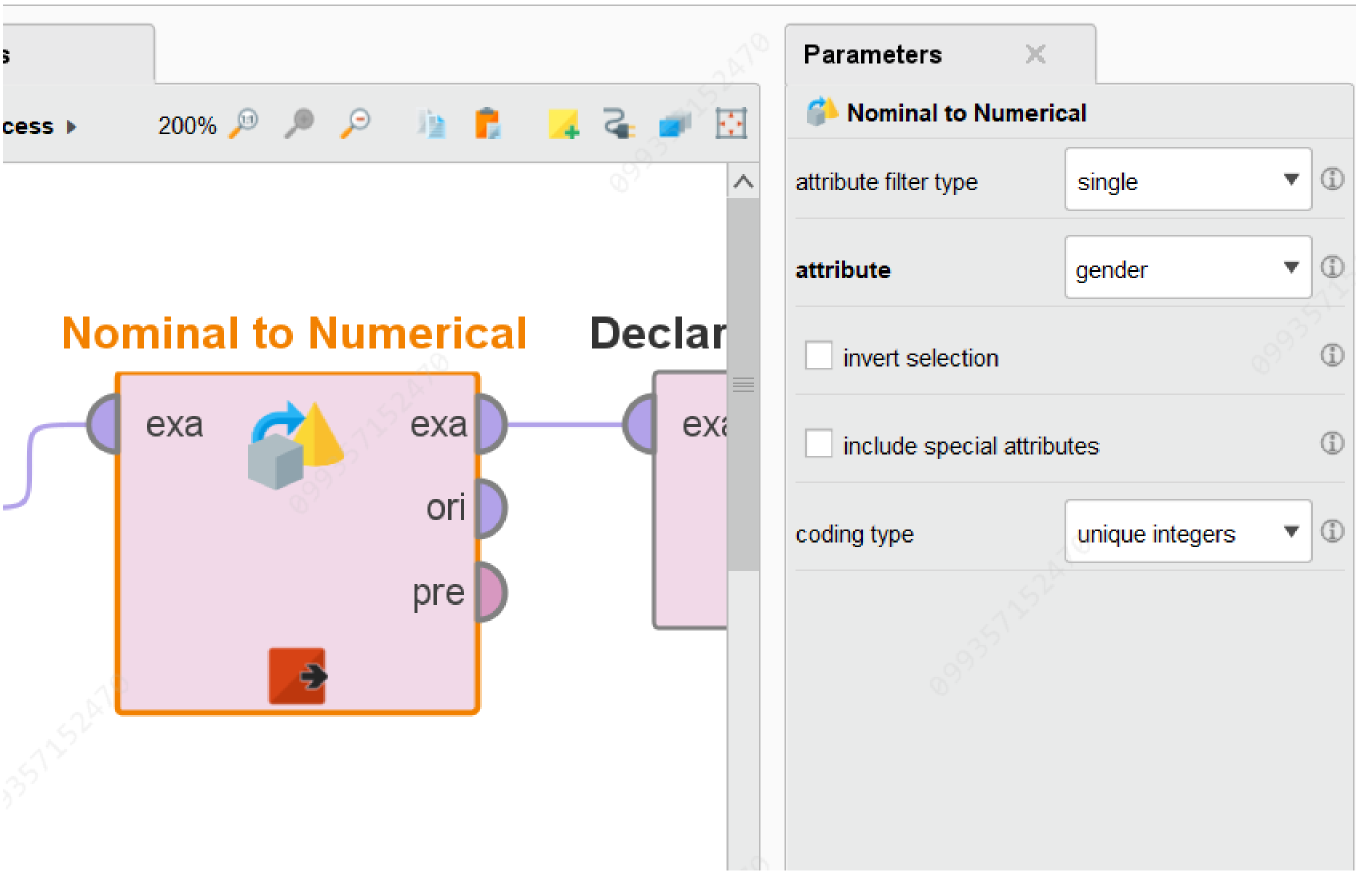


net-capital
0
0
0
7688
0
0
0
3103
0
0
6418
0
0
0
3103
0

Figure 13 : Attributes Result

Generate Attributes operator will allow to crate simple function (argument - argument) and insert the result in new argument.

2. Convert the ‘Gender’ column to a binary variable (0 for male, 1 for female).



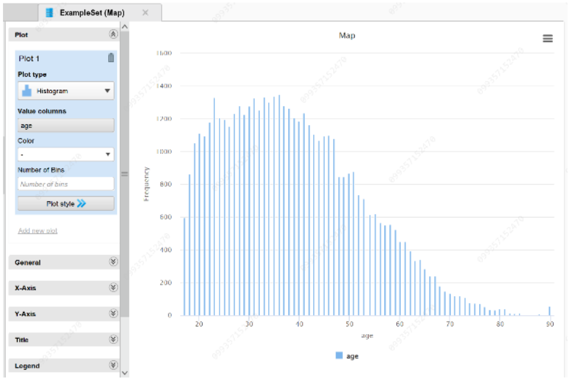
Parameters

Nominal to Numerical

attribute filter type	single
attribute	gender
☐ invert selection	
☐ include special attributes	
coding type	unique integers

Section 5: Data Visualization

1. Create a histogram to visualize the distribution of individual ages.



Plot 1

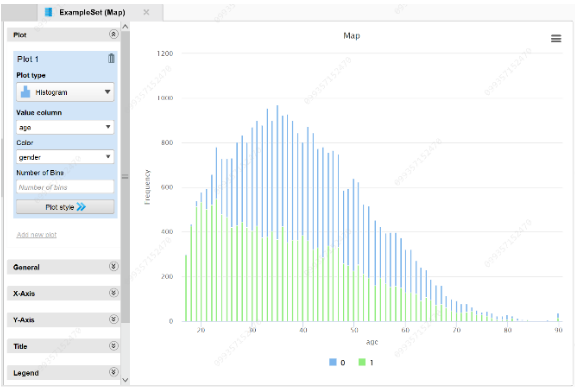
Plot type: Histogram

Value columns: age

Color: -

Number of Bins: Number of bins

Figure 14 : Histogram.



Plot 1

Plot type: Histogram

Value column: age

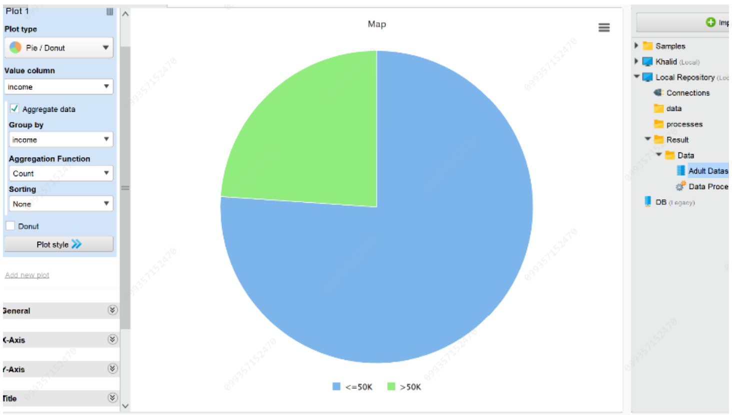
Color: gender

Number of Bins: Number of bins

Figure 15 : Histogram (0 Male ,1 Female)

clicking the results tab after retrieve the data and goes to statistic than select age feature will give you option to visual a histogram graph.

2. Create a pie chart to visualize the number of individuals who earn more than 50K and those who earn less than or equal to 50K.



Plot 1

Plot type: Pie / Donut

Value column: income

☑ Aggregate data

Group by: income

Aggregation Function: Count

Sorting: None

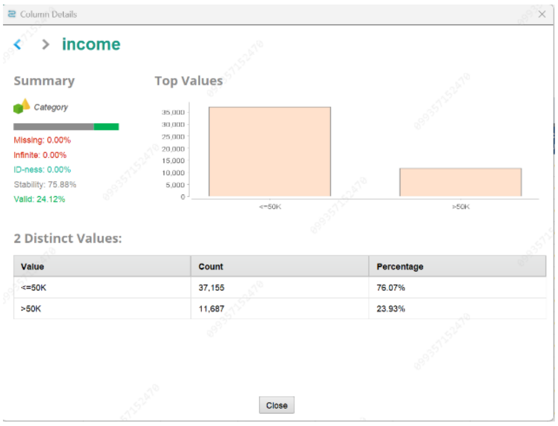
□ Donut

Figure 16 : Pie Chart

By using visualizations option in results and select pie chart graph with the feature income to display it

Section 6: Data Analysis

1. Calculate the percentage of individuals who earn more than 50K.



2 Distinct Values:

Value	Count	Percentage
<=50K	37,155	76.07%
>50K	11,687	23.93%

Figure 17 : Percentage of Income

with turbo prep option you can select the feature by click right which will give you the option show details.

2. Calculate the income rate for individuals who earned less than or equal to 50K of these three workclass (private, state-gov, Self-emp-inc).

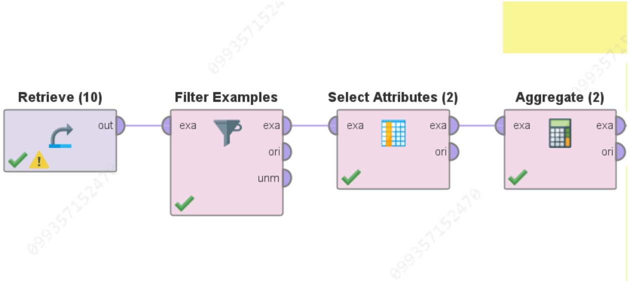
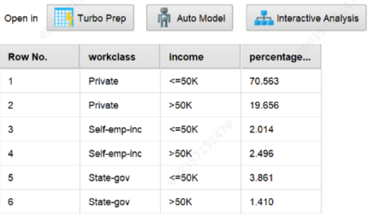


Figure 18 : Calculate Income Operation



Row No.	workclass	income	percentage...
1	Private	<=50K	70.563
2	Private	>50K	19.656
3	Self-emp-inc	<=50K	2.014
4	Self-emp-inc	>50K	2.496
5	State-gov	<=50K	3.861
6	State-gov	>50K	1.410

Figure 19 : Result of the Operation

By Filiter the specifc values of the workclass feature and select age and workclass attribute than aggregate aggrgate feature and select function which is count presentage

#Bonus

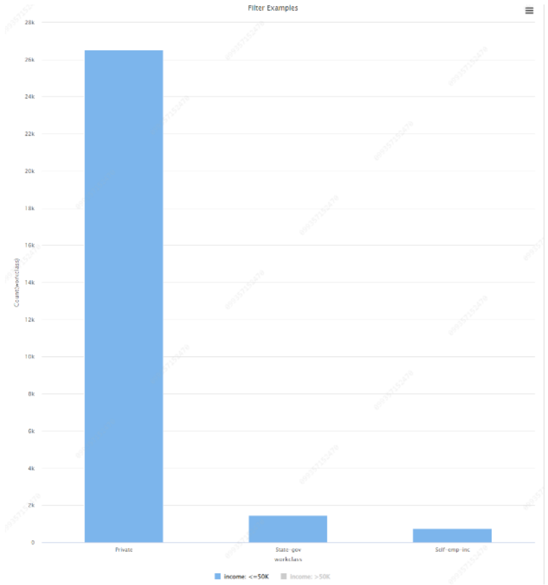


Figure 20 : Bar Graph for (private, state-gov, Self-emp-inc) per Income

This graph represents Income as a bar for (private, state-gov, Self-emp-inc).

Section 7: Conclusion

1. Based on your analysis, provide a concise conclusion discussing any noteworthy findings, relationships, or patterns observed in the data.

In conclusion we get the idea that the dataset about individuals' income in US with some information of them like the education level, relationship status, the place of their jobs and their race. I used the income as a target and see what could hit high impact on the income.

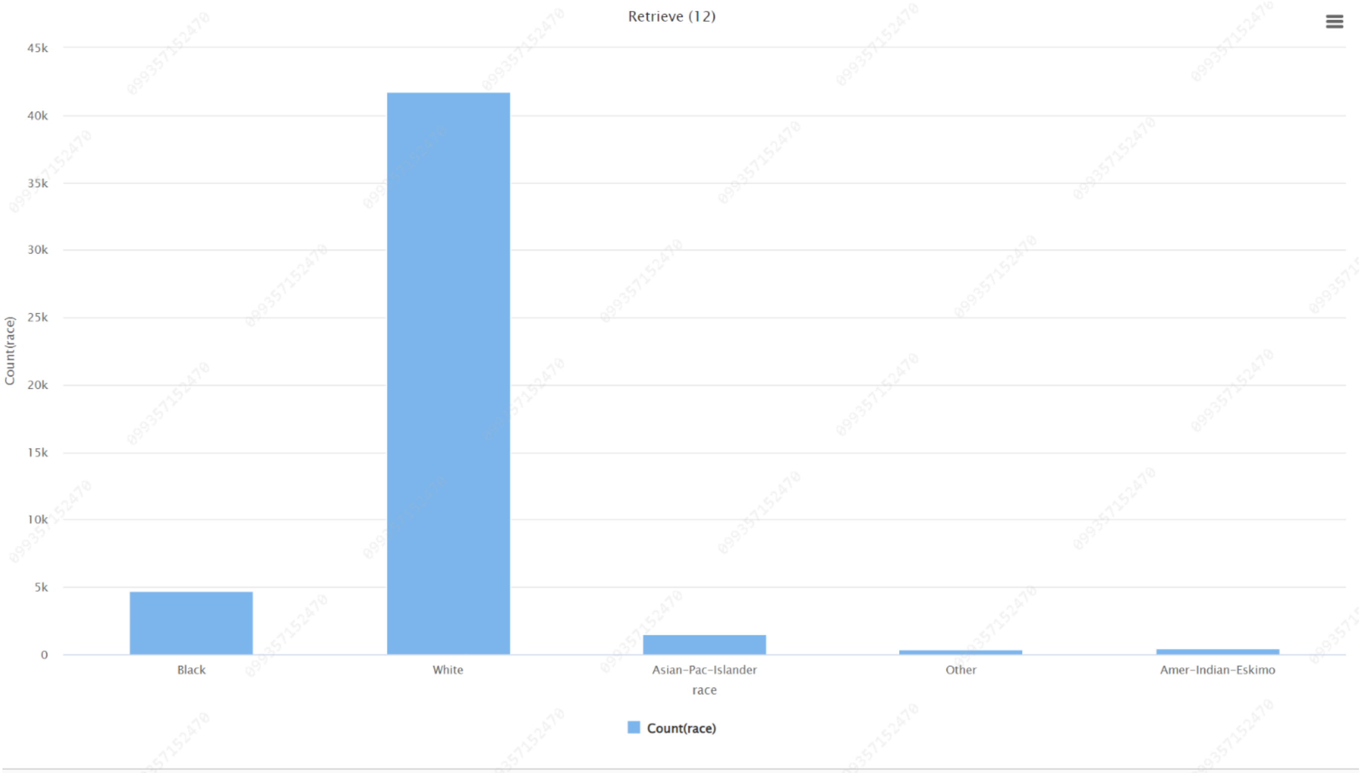


Figure 21 : Race Bar Graph

Firstly, most of the dataset is white people. If we try machine algorithm it might lead to high bias toward white people so we can say it imbalanced data.

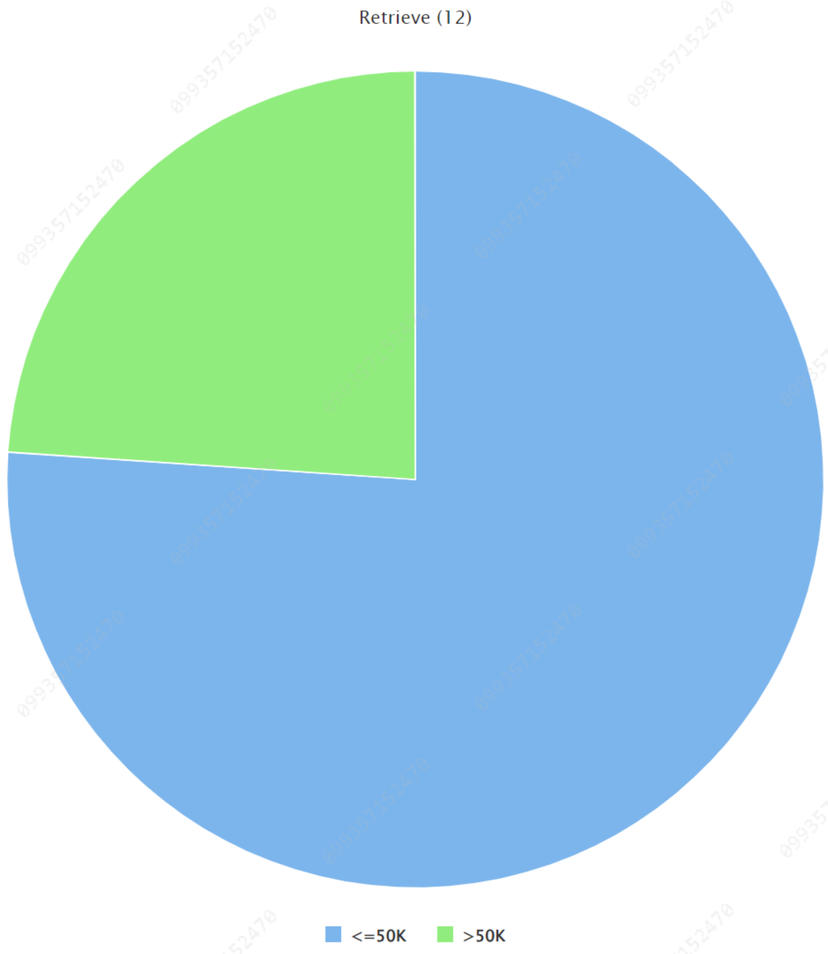


Figure 22 : Pie Chart Graph Income

The pie chart shows the 25% of the individual's income more the 50K while 75% income less or equal 50K

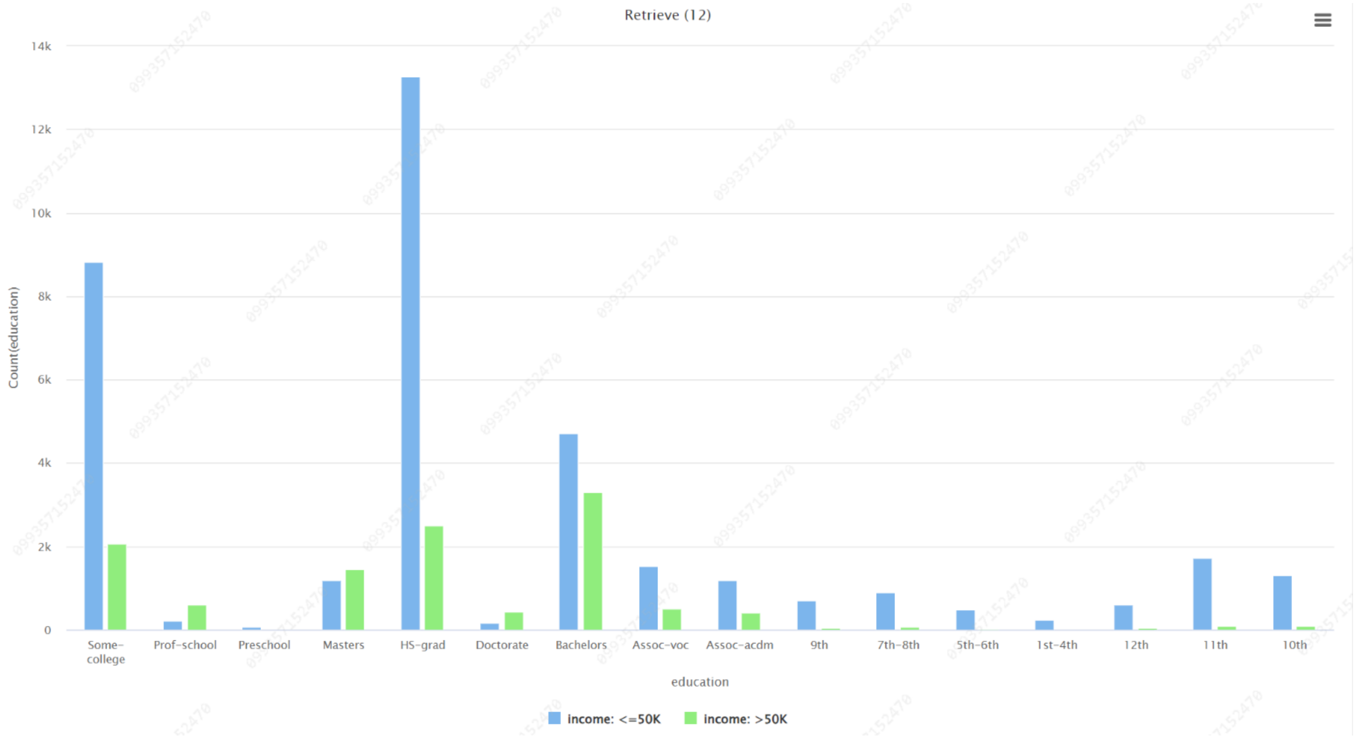


Figure 23 : Education Per income

This graph shows that most of the individuals who has “>50k” their education degree is one of these (Some college, HS-grad, Bachelors) and (Prof-school, Masters, Doctorate) has more people get >50K more then those who get less but other degrees have less possibility of getting high salary.

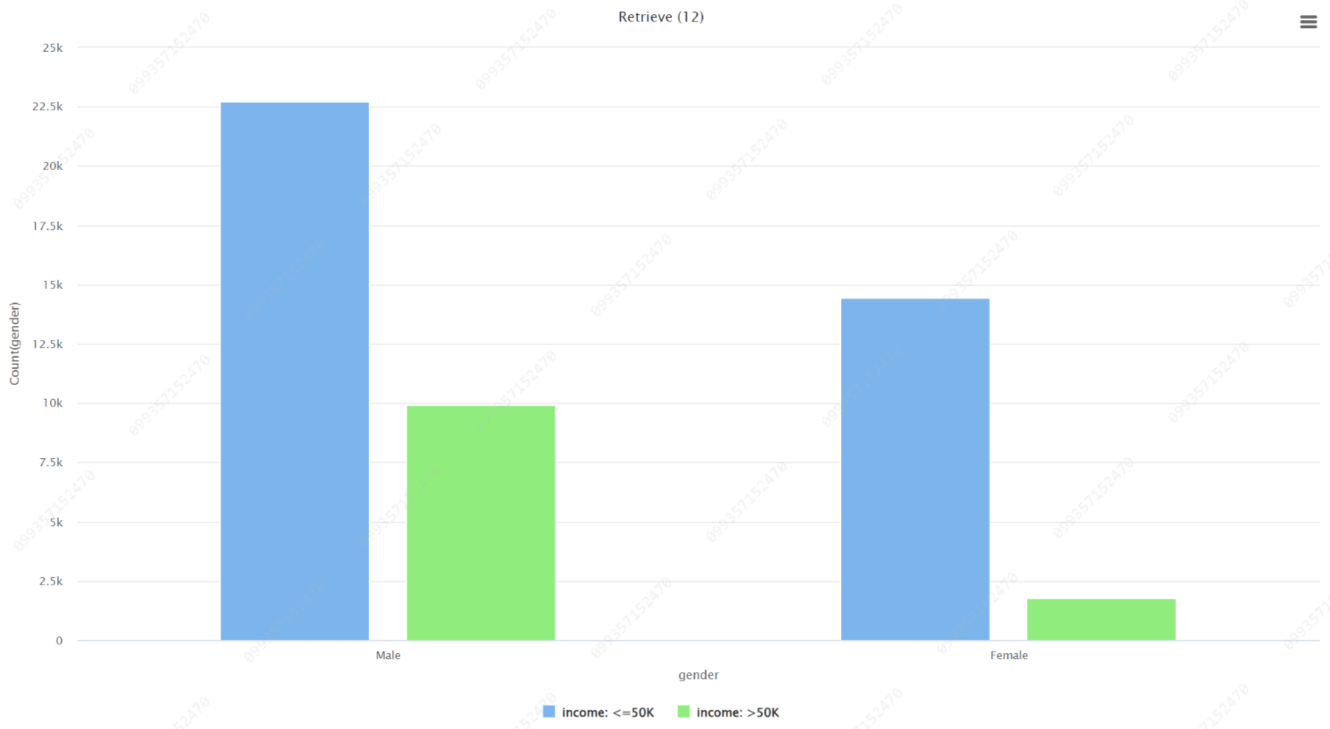


Figure 24 : Gender Income

Both Male and Female most of them have <=50K income, but in male those who get >50k are half of the male while in female they barely reach to 25% which might lead to some people think that female don’t get same salary as male.

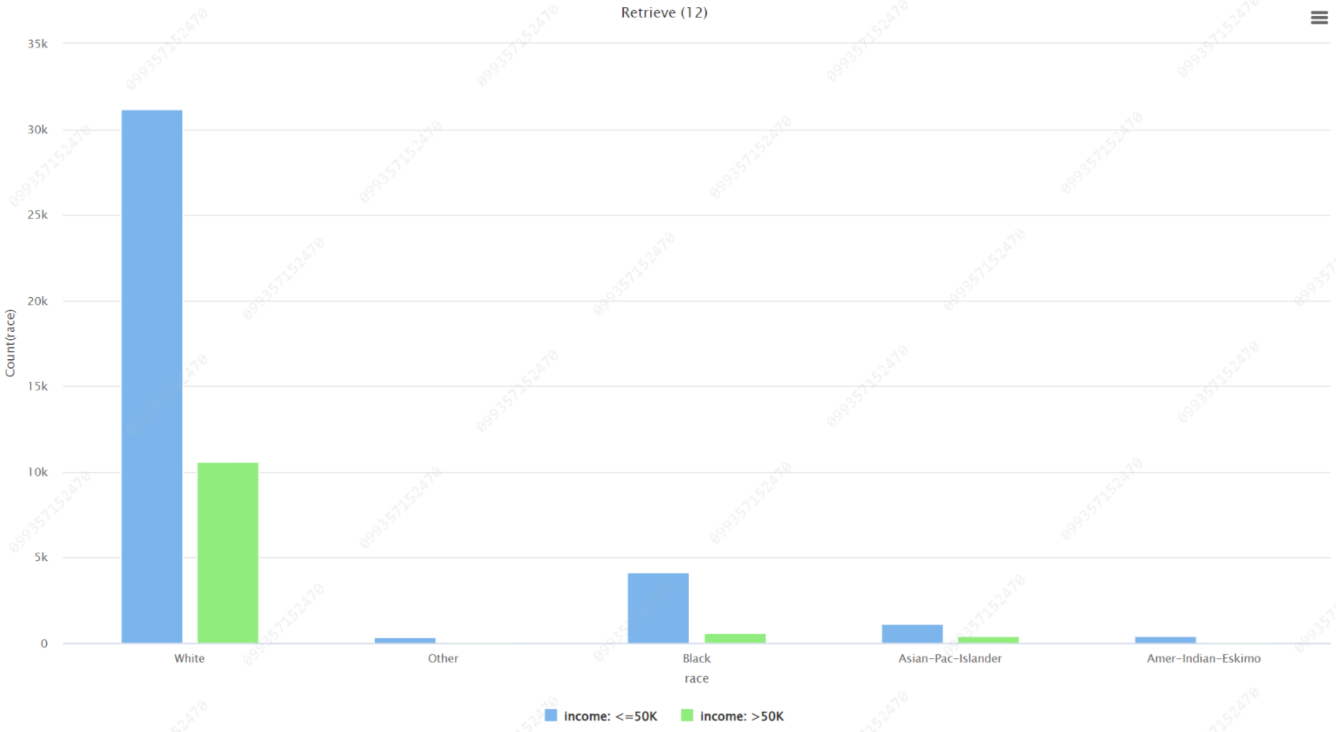


Figure 25 : Race Per Income

This graph shows also how the difference between white race and other races in salary. The income for other races mostly less then 50K but in white there is good percentage for >50k.

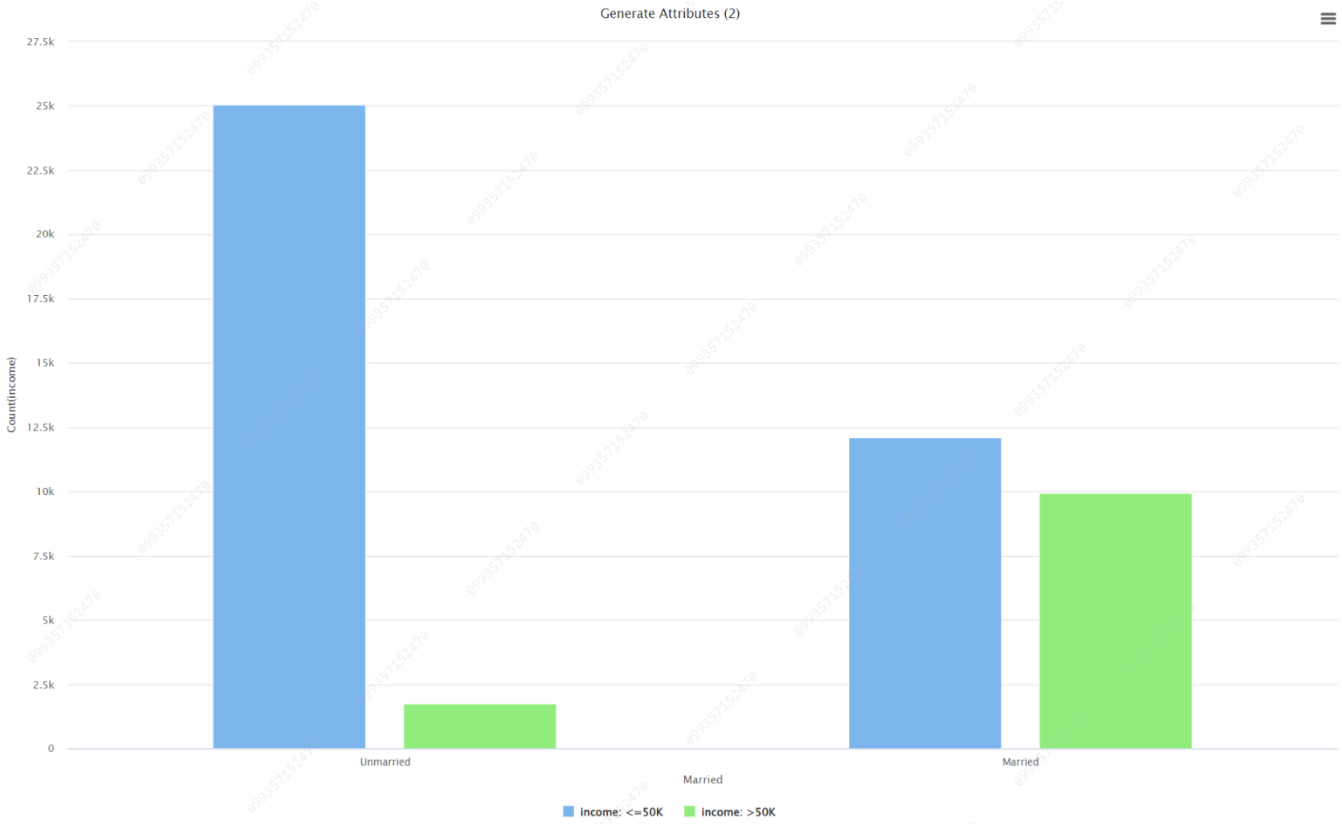


Figure 26 : Relationship per incomes

Making new feature include the Married person either male or female, and we get that marriage can be a reason for getting higher salary.

Optional Bonus:

Perform any additional data engineering or analysis that you find interesting or relevant to the dataset.

Feel free to explore and experiment with the dataset and RapidMiner's capabilities beyond the questions mentioned above.