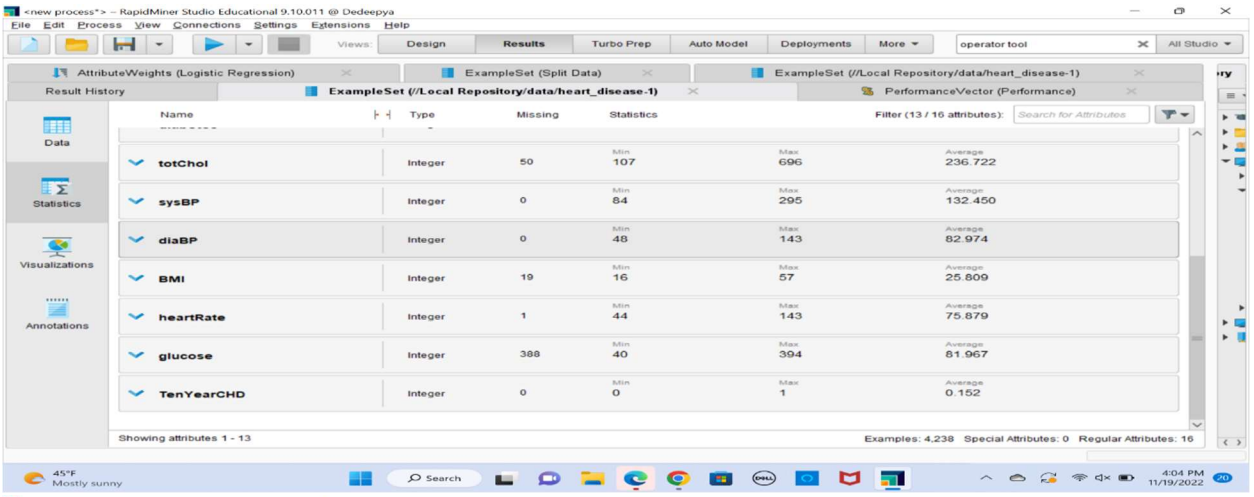
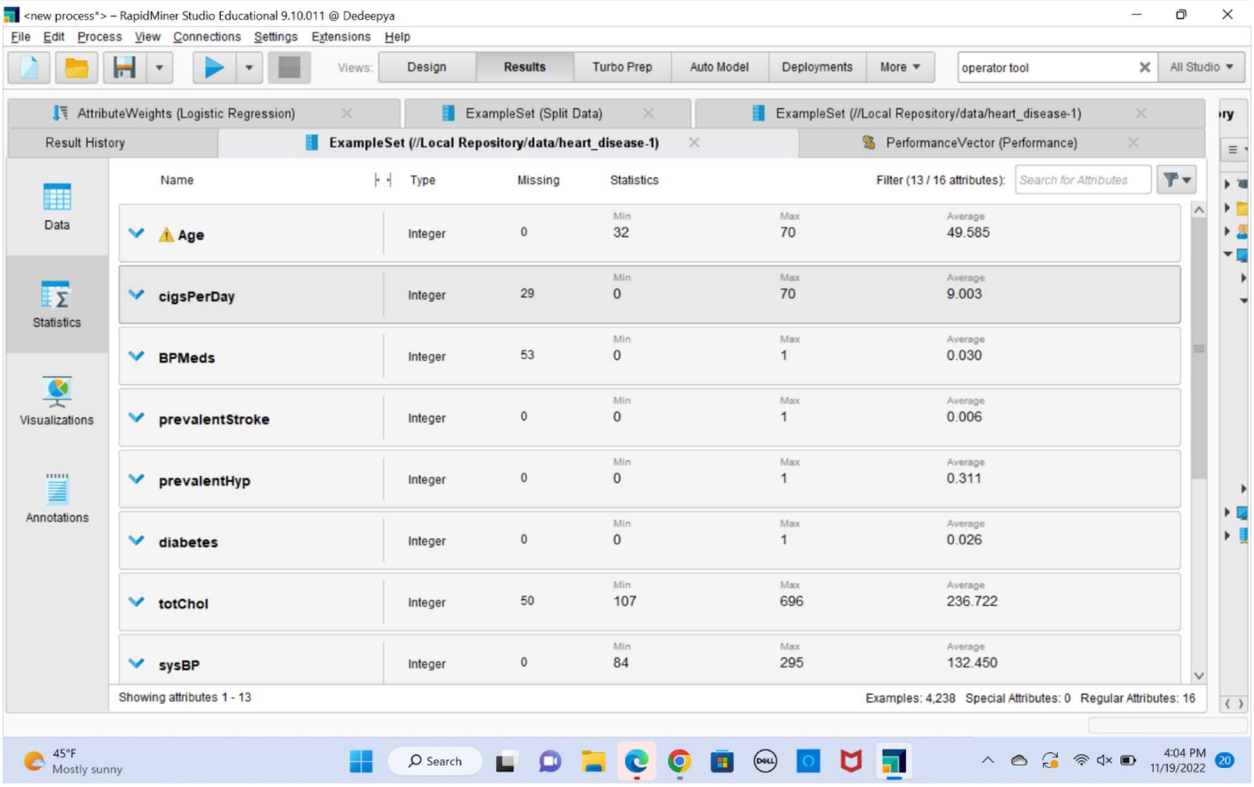
1. Import the dataset “heart_disease.csv” into Rapid Miner, while importing

   A. Convert the below mentioned columns to integers.

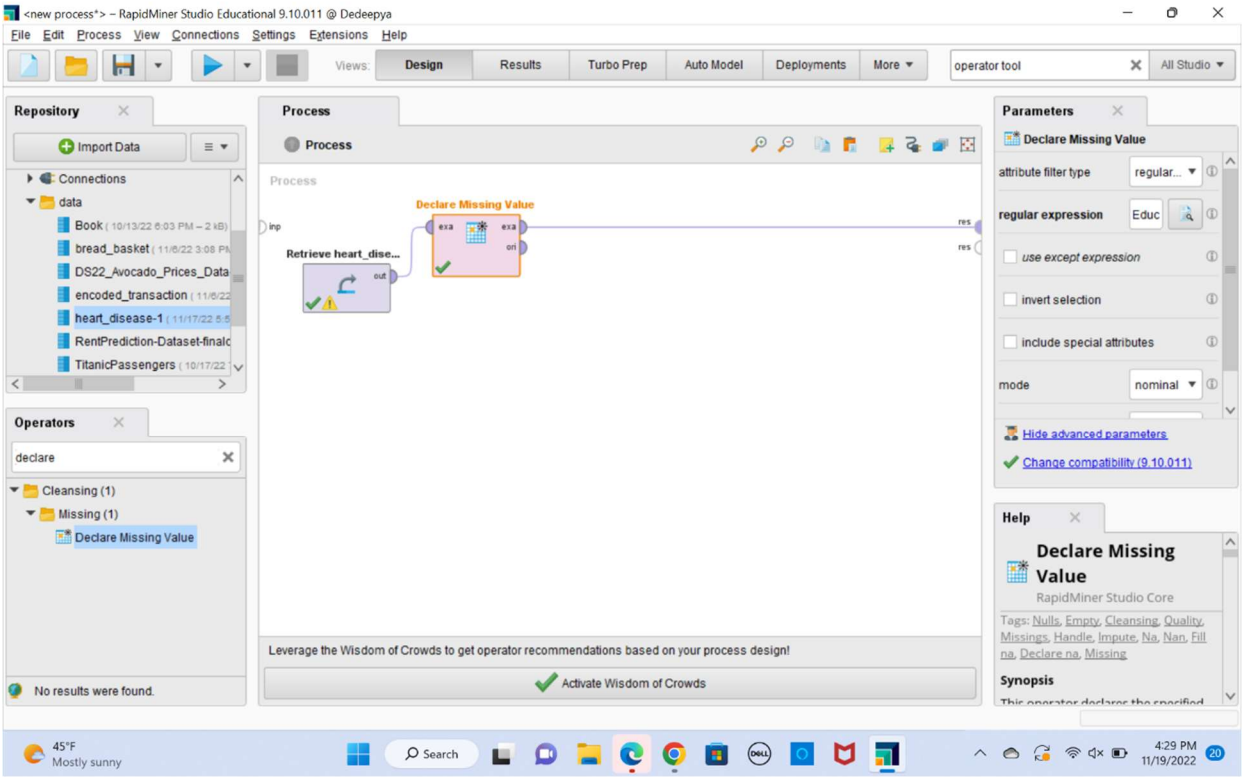
Age	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD

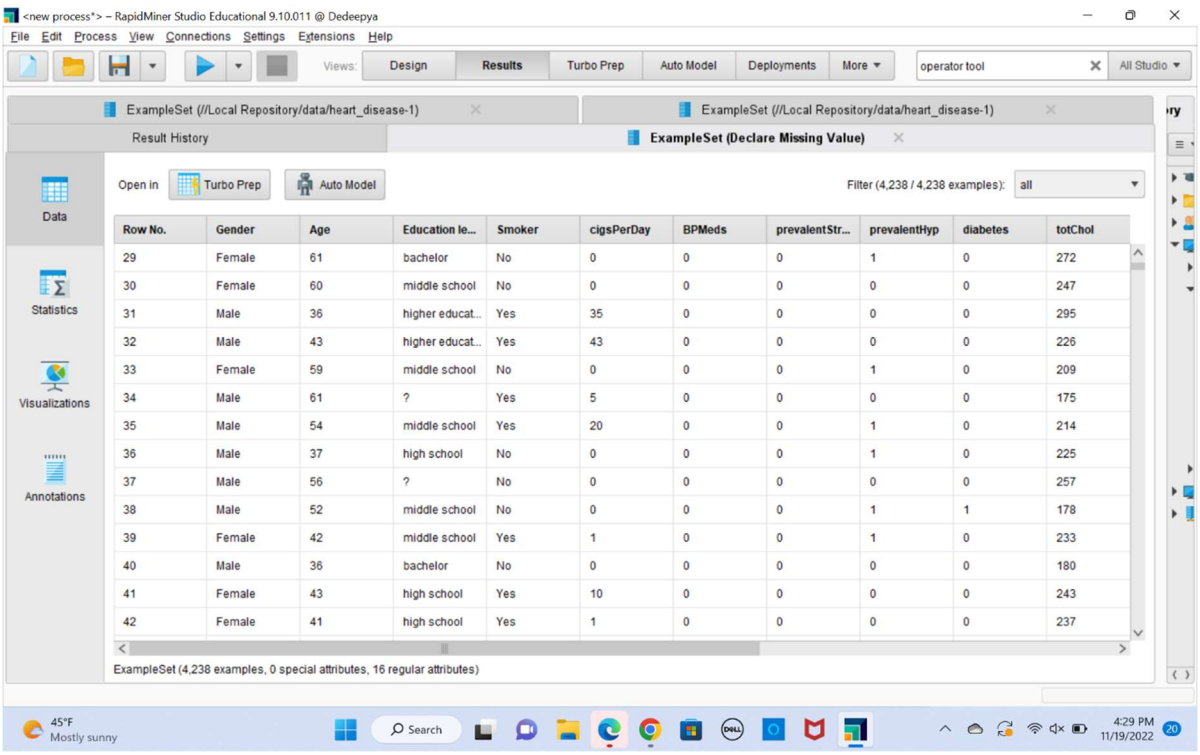
   B. Click on replace errors with missing values check box to avoid parsing error in cigsPerDay column.





C. Using the “Declare missing values operator” identify “N/A” string in Education level column as Missing values (Use nominal mode and enter “N/A” in expression of declare missing values operator).



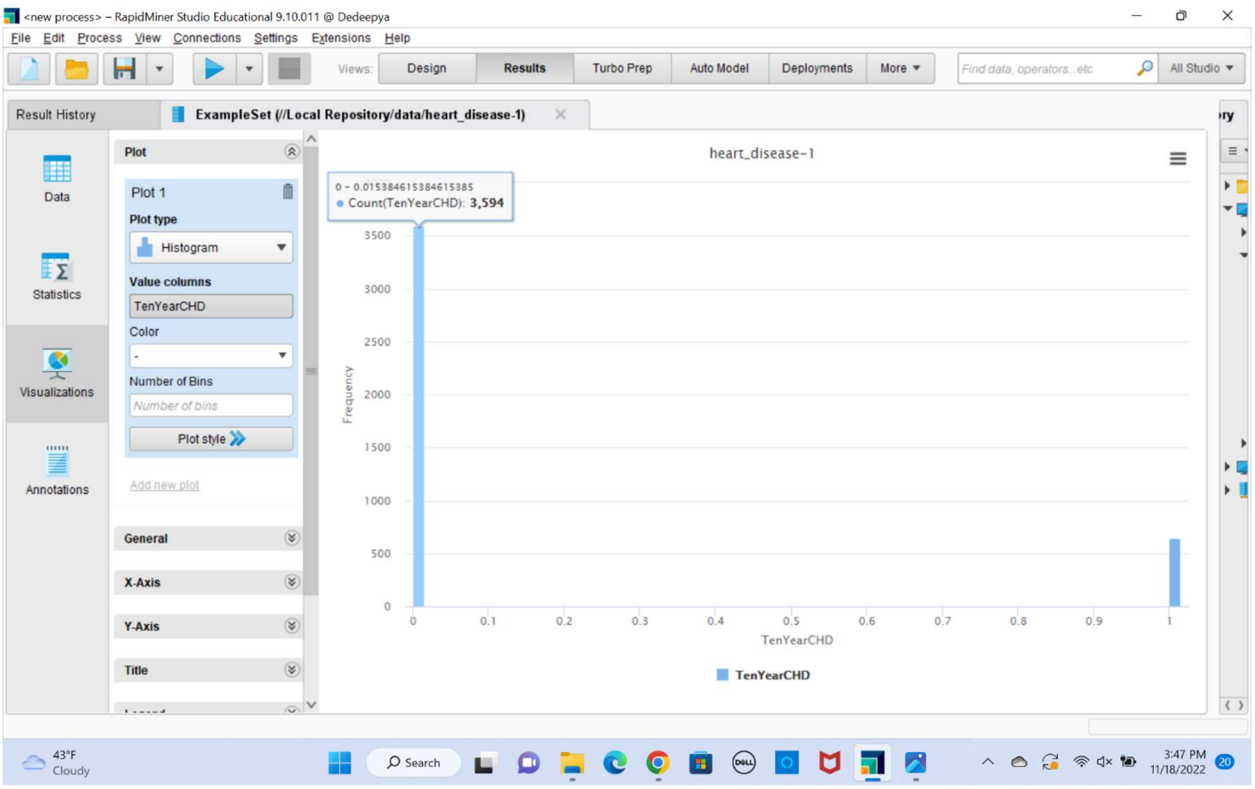


1

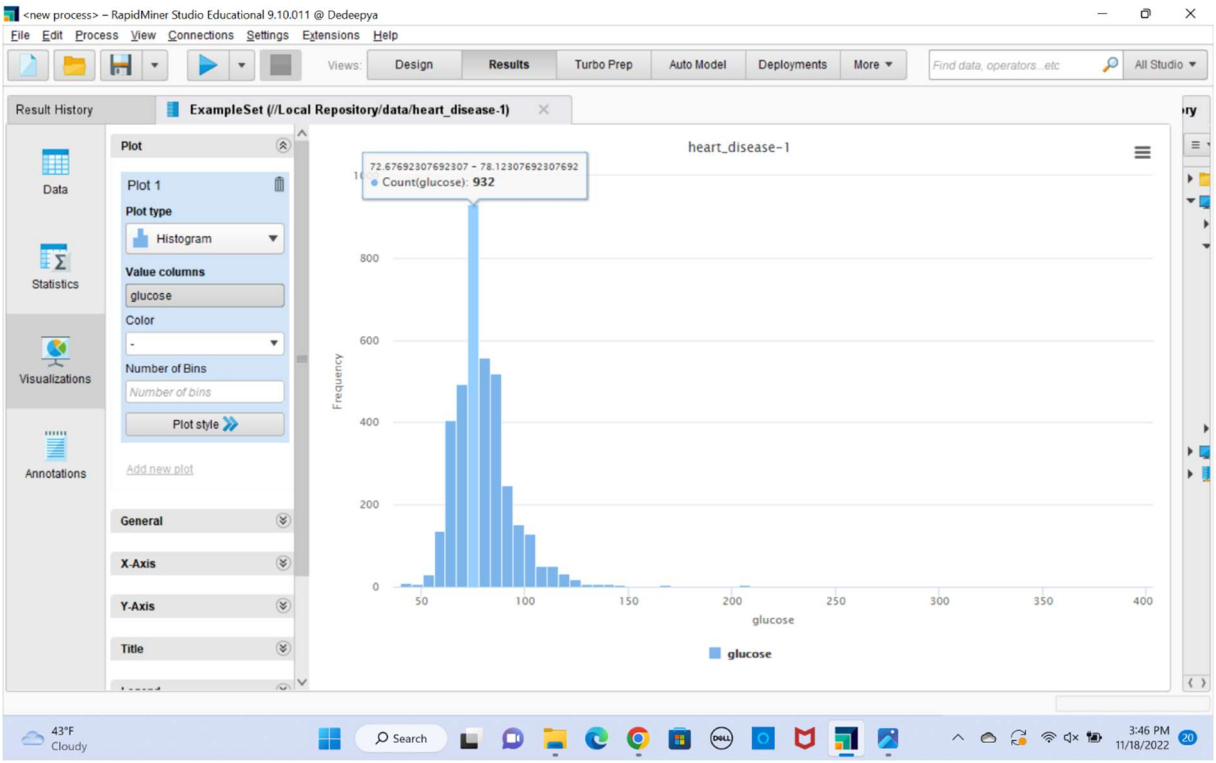
a) Include a histogram of each variable and describe each variable (Use visualization tab after running the dataset in canvas). Need separate histograms for all variables and explain what you see from the histogram plot.

b) Is there any row that contains missing values on all columns?
Which column contains missing values or null values?
How many missing values are there in each column? (Use statistics tab in results view)

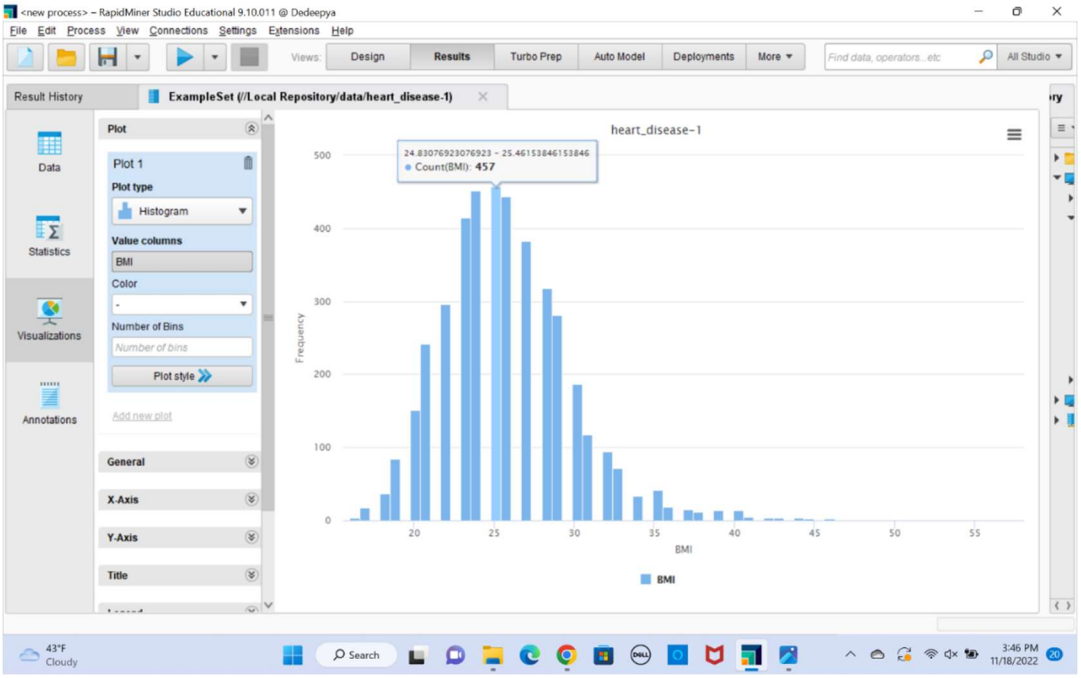
a)TenYearCHD- It has only 2 values 0 and 1 and there are 3594 zeros and 644 ones



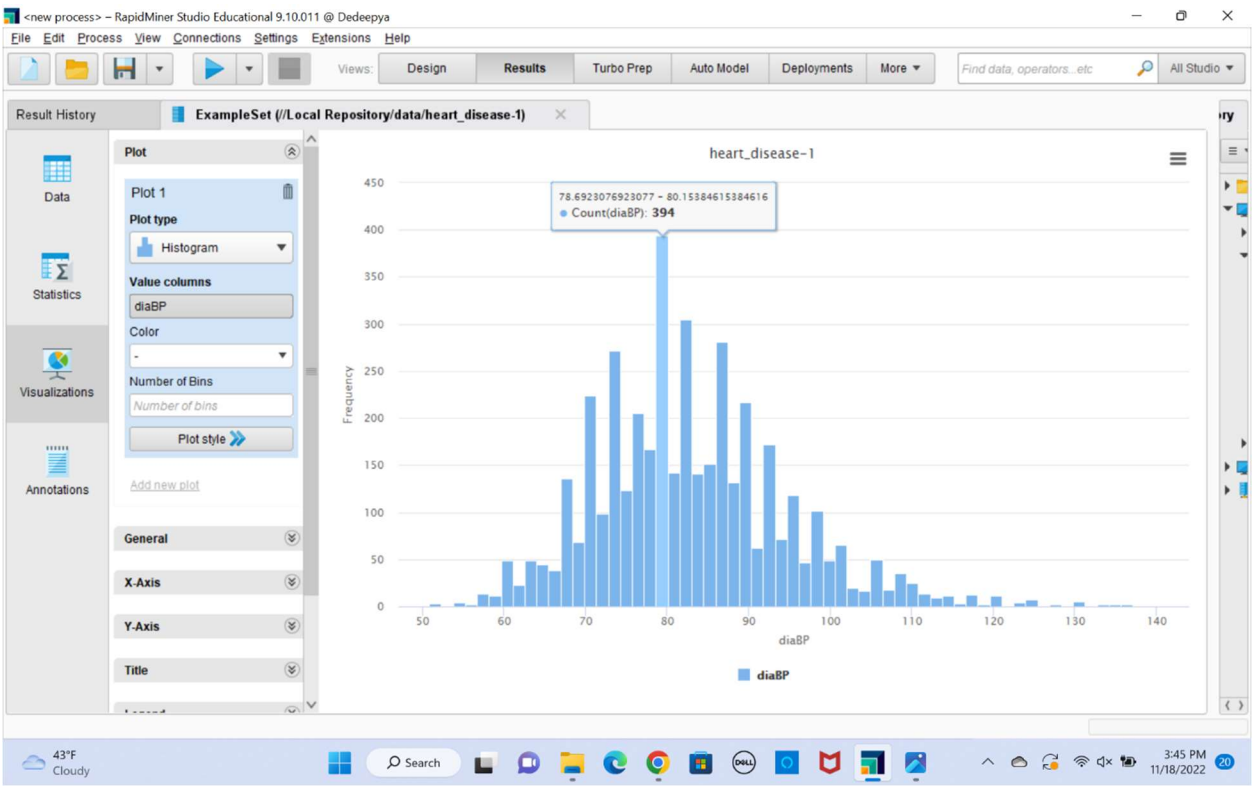
Glucose values ranges from 40 to 208.83, the values between 72.67-78.12 have high frequency of 932.



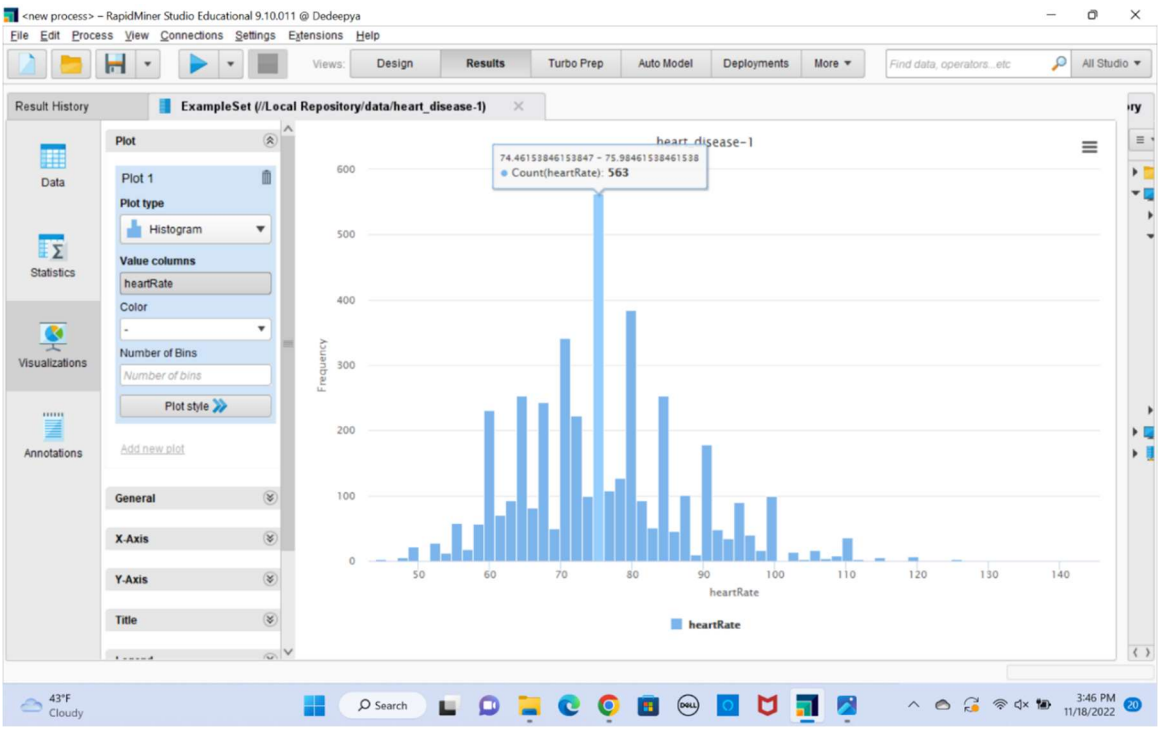
BMI Ranges from 16-46.27 and the values of 24.83-25.46 has high frequency of 457



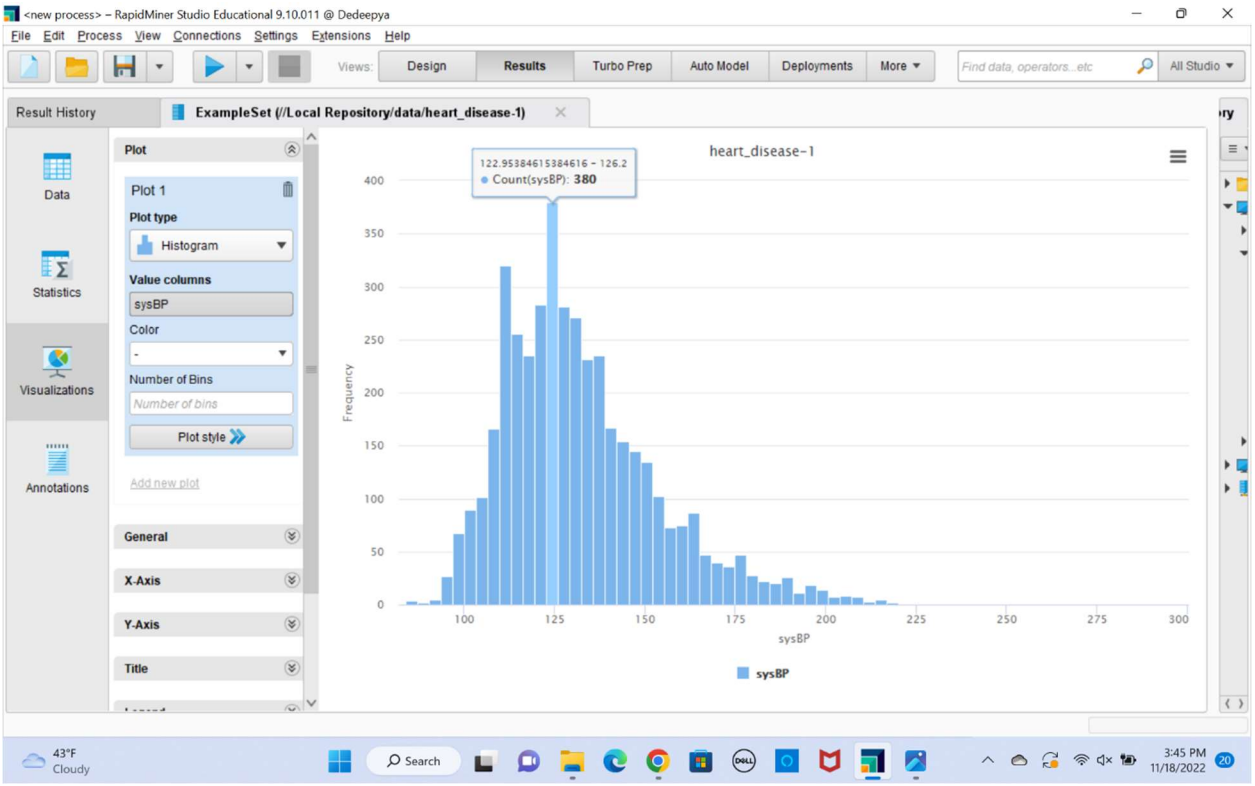
DialBp- range 50.92-137.15 and, values between 78.69-80.15 has high frequency of 394



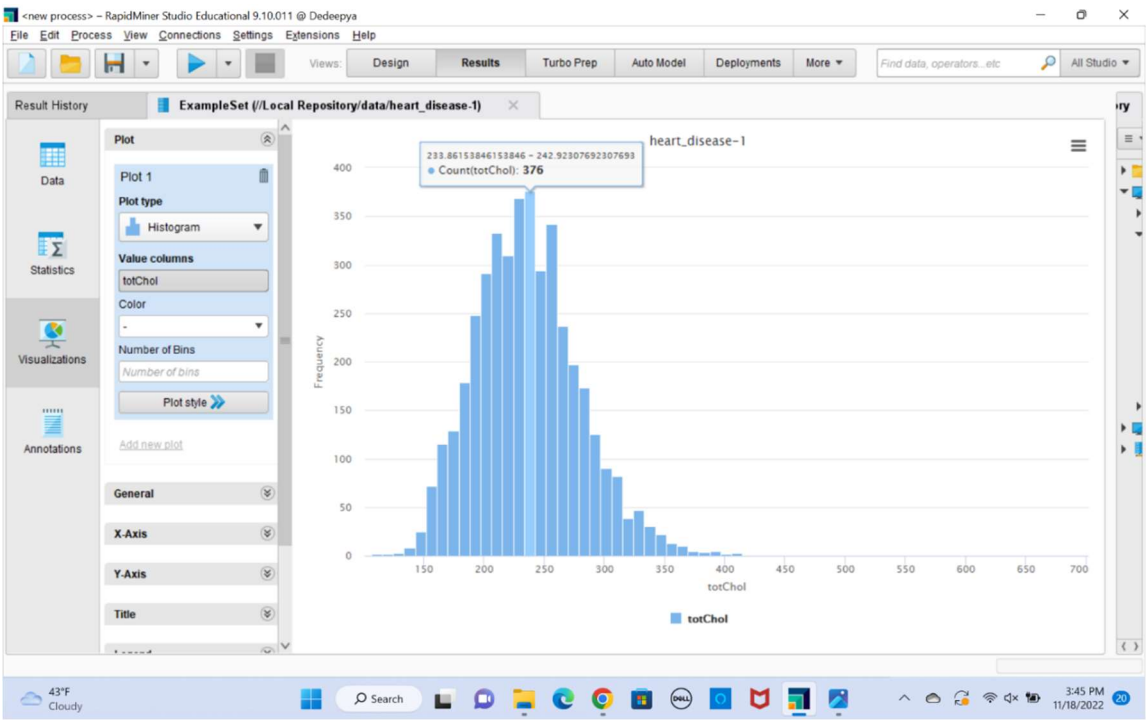
Range of heartrate is 44-126.24, values within range 74.46-75.98 has 563 frequencies



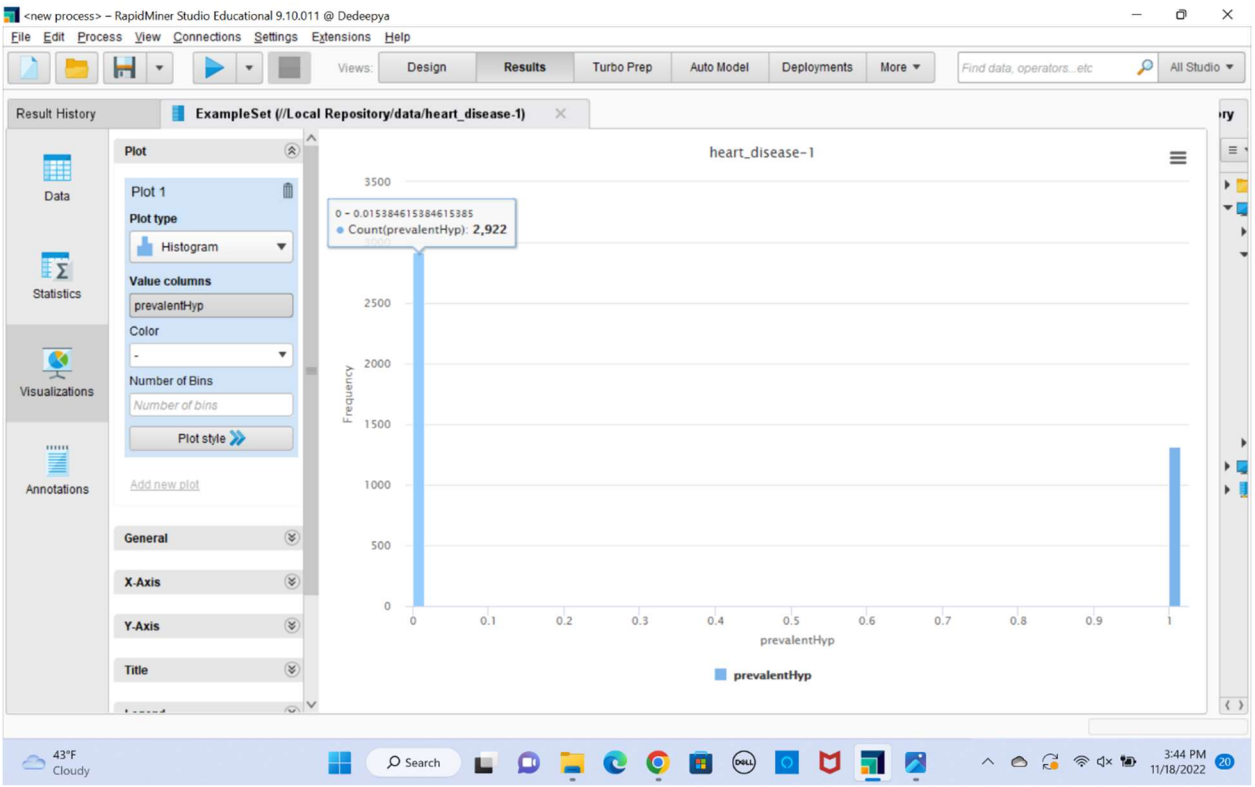
SysBP Range 84-220.33 and values within range 122.95-126.2 has frequency of 380



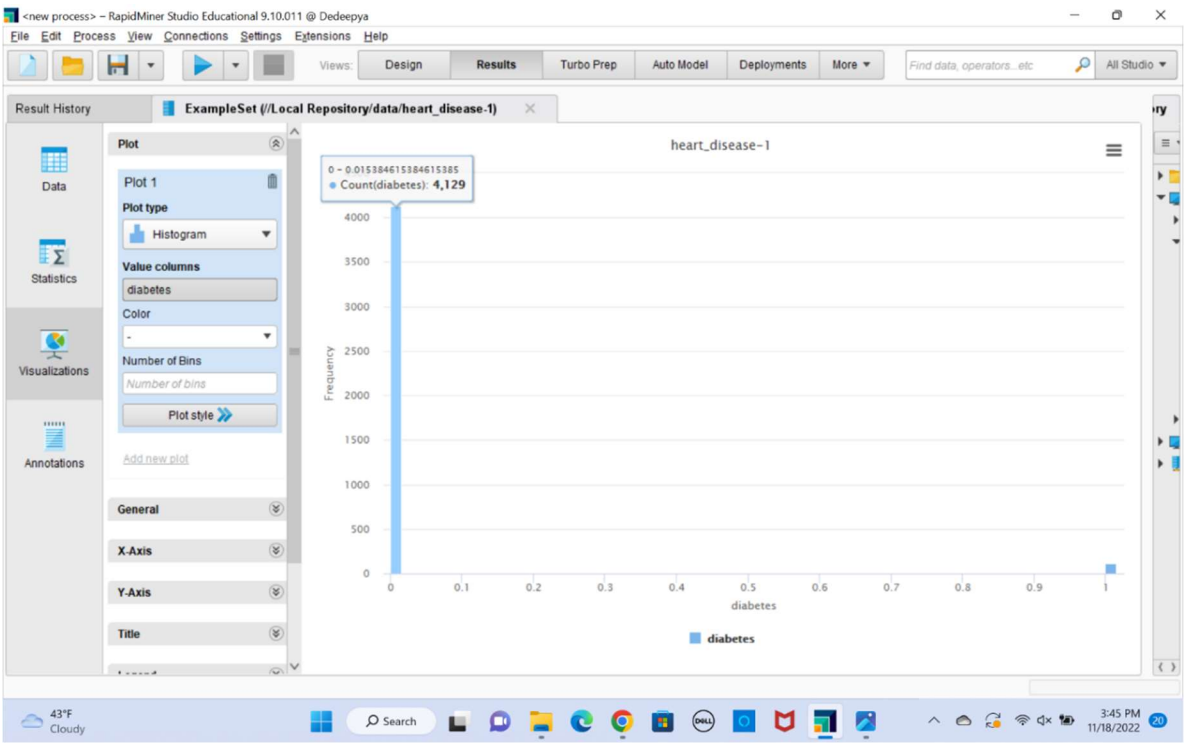
Totchol Range is 107-415, and high frequency of 376 for range 233.86-242.923



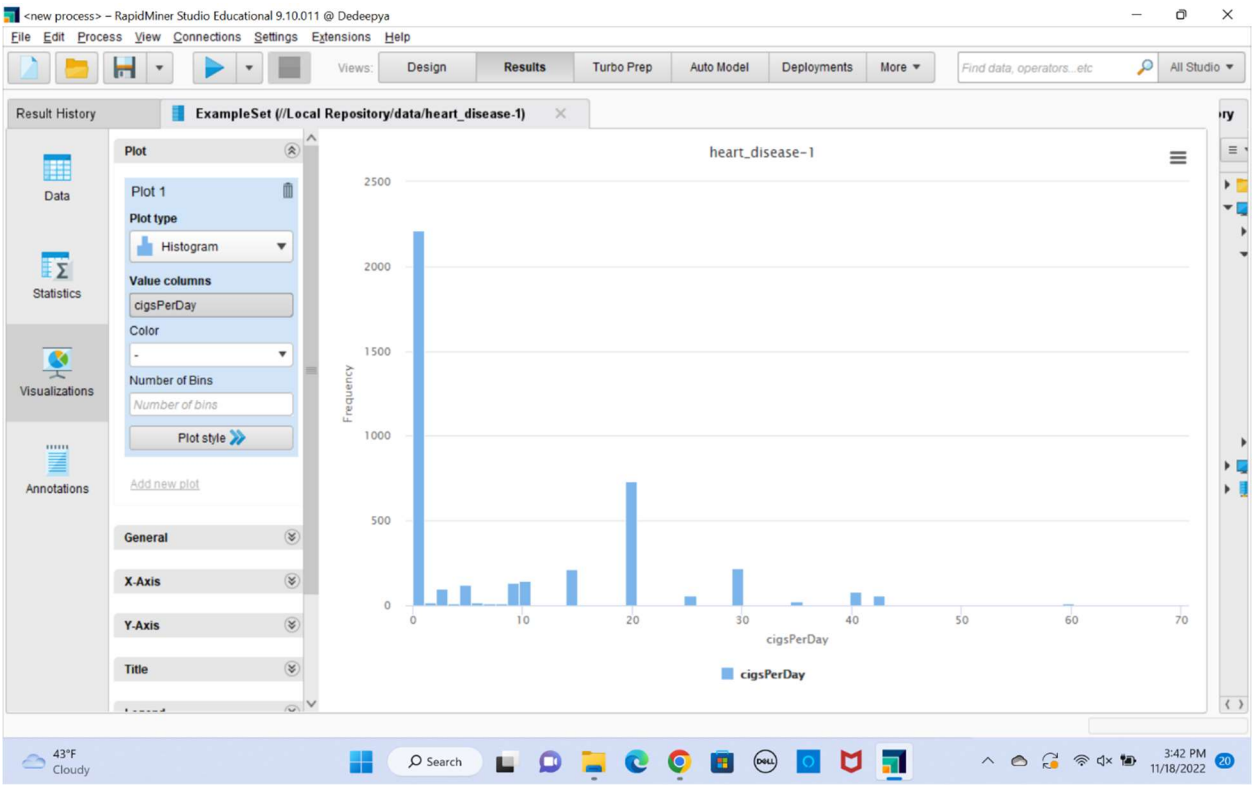
Prevalenthyp has Only 2922 zeros and 1316 ones



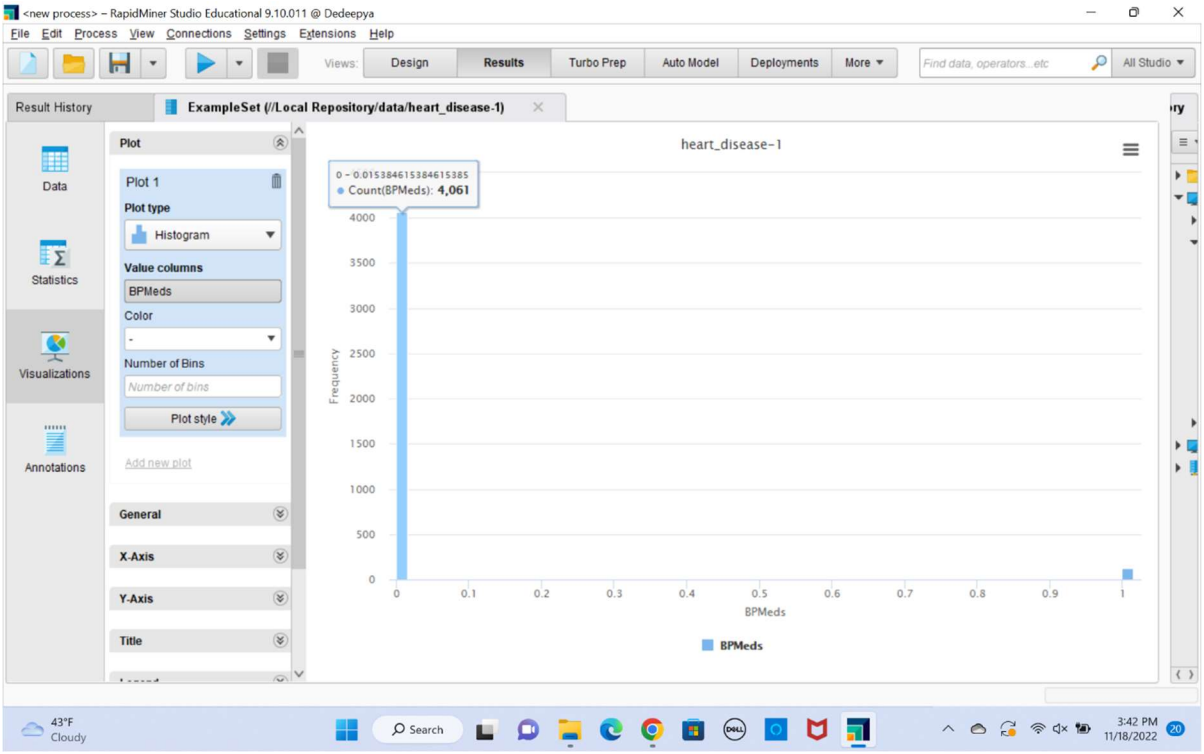
Diabetes has 4129 zeros, 109 ones



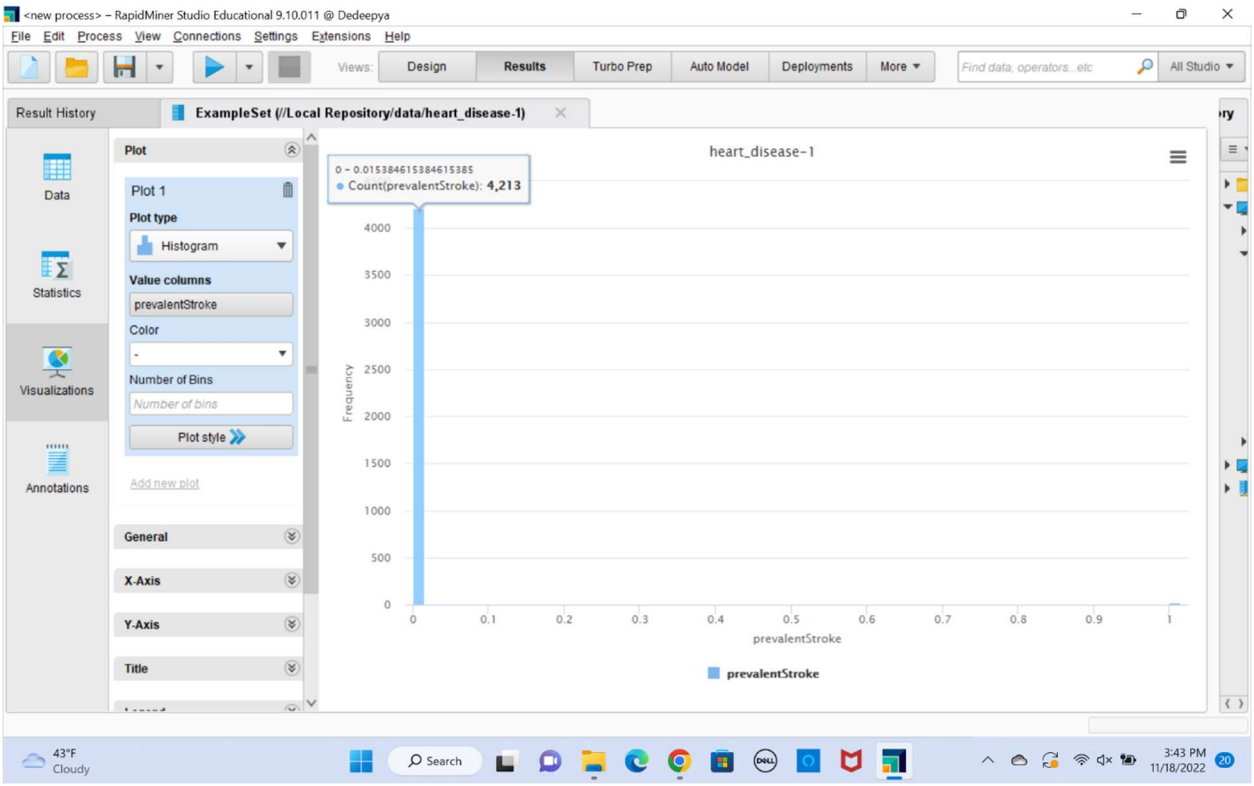
Cigsperday has high frequency of 2211 for range 0-1.076 and values ranges from 0-60.30



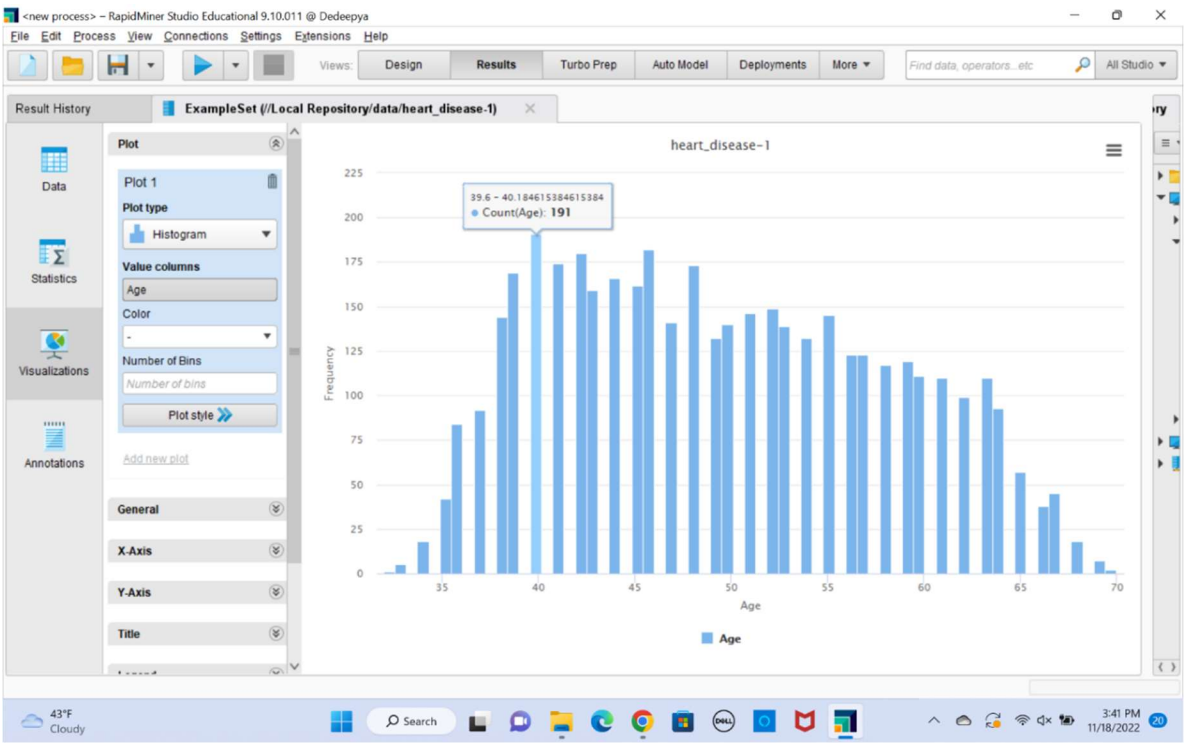
BPmeds has Freq-4061 for 0, and for 1, we have freq-124.



Prevalent stroke has 4213 frequencies for 0nand frequency of 25 for 1

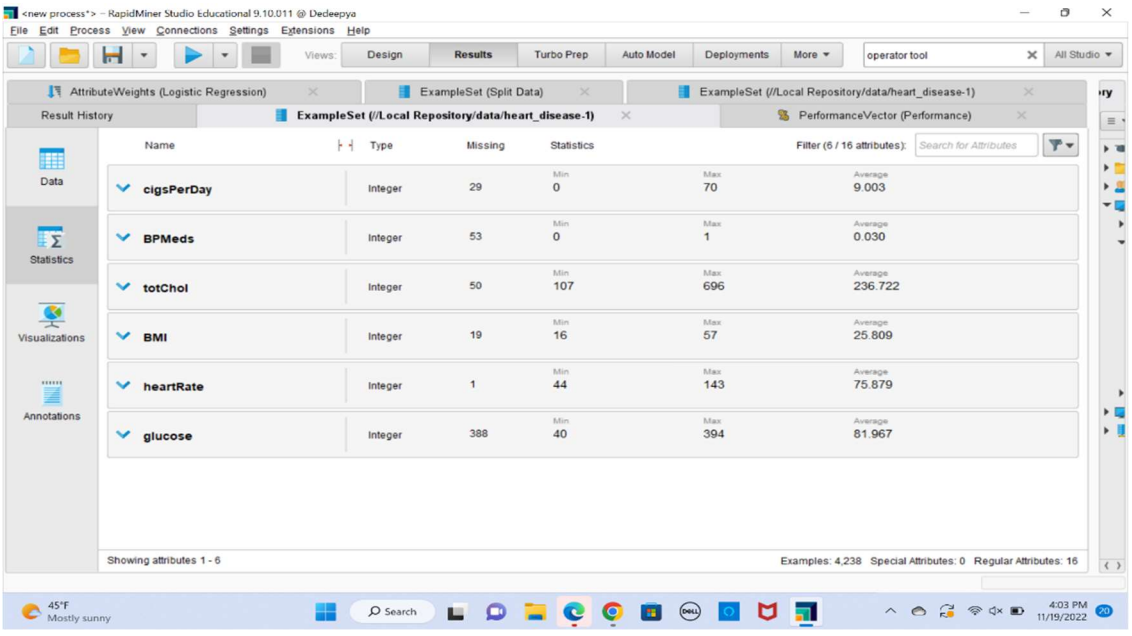


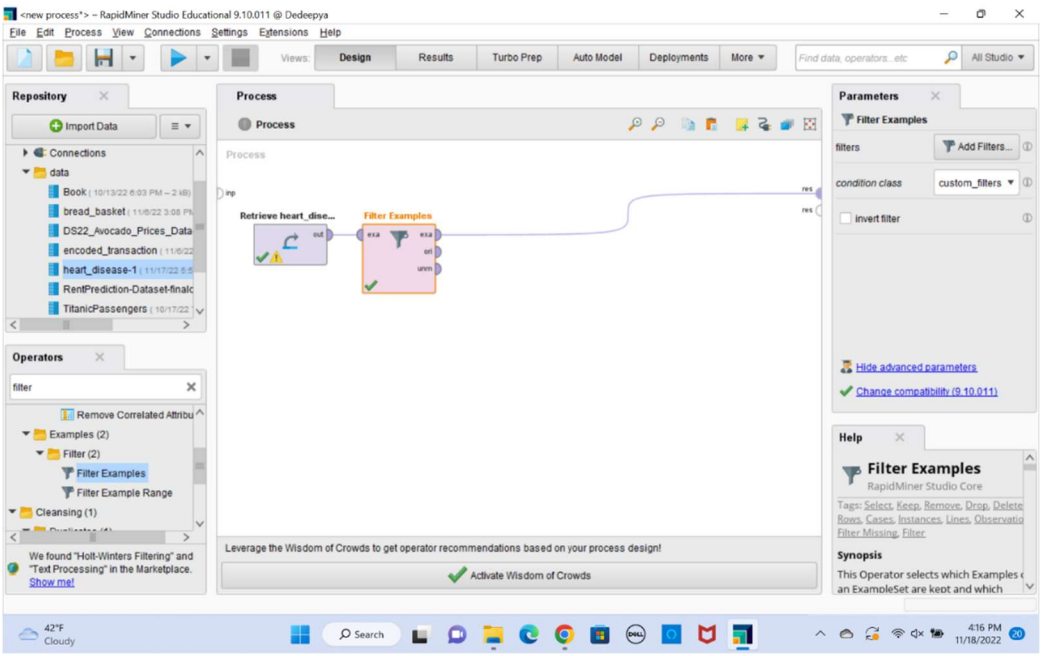
Age has range 32-70 and high frequency of 191 for values between 39.6-40.18

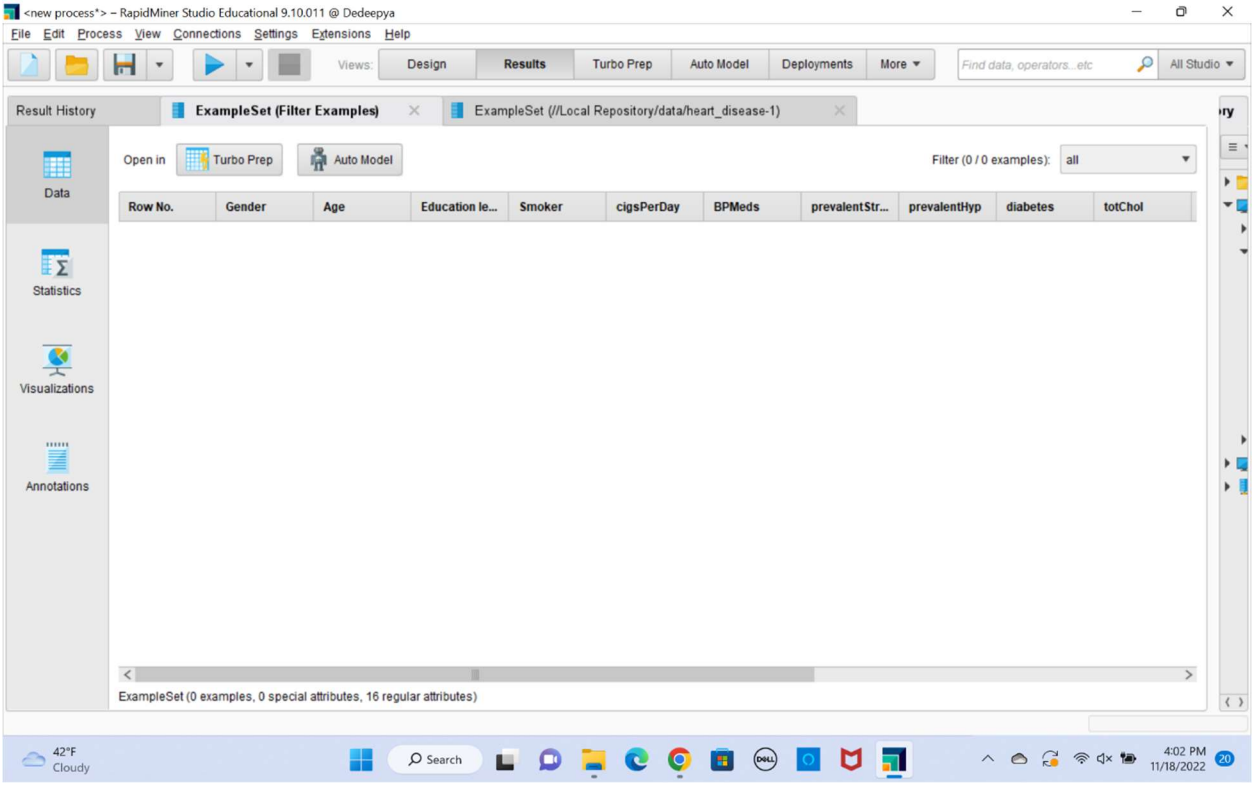


b)There is no record with missing values in all columns.

Cigsperday-29
BPMeds-53
Totchol-50
BMI-19
Heartrate-1
Glucose-388




<new process*> – RapidMiner Studio Educational 9.10.011 @ Dedeepya
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Repository
Import Data
Connections
data
Book
bread_basket
DS22_Avocado_Prices_Data
encoded_transaction
heart_disease-1
RentPrediction-Dataset-finalc
TitanicPassengers
Operators
filter
Remove Correlated Attribu
Examples (2)
Filter (2)
Filter Examples
Filter Example Range
Cleansing (1)
We found "Holt-Winters Filtering" and "Text Processing" in the Marketplace.
Show me!
Process
Process
Retrieve heart_dise...
Filter Examples
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Filter Examples
filters
Add Filters...
condition class
custom_filters
invert filter
Hide advanced parameters
Change compatibility (9.10.011)
Help
Filter Examples
RapidMiner Studio Core
Tags: Select, Keep, Remove, Drop, Delete Rows, Cases, Instances, Lines, Observatio Filter Missing, Filter
Synopsis
This Operator selects which Examples an ExampleSet are kept and which
42°F Cloudy
Search
4:16 PM 11/18/2022


<new process*> – RapidMiner Studio Educational 9.10.011 @ Dedeepya
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Result History
ExampleSet (Filter Examples)
ExampleSet (//Local Repository/data/heart_disease-1)
Data
Statistics
Visualizations
Annotations
Open in
Turbo Prep
Auto Model
Filter (0 / 0 examples):
all
Row No.
Gender
Age
Education le...
Smoker
cigsPerDay
BPMeds
prevalentStr...
prevalentHyp
diabetes
totChol
ExampleSet (0 examples, 0 special attributes, 16 regular attributes)
42°F Cloudy
Search
4:02 PM 11/18/2022

2. For each record that has a missing value in “cigsPerDay” and is identified as a smoker, replaces the missing values with the average of “cigsPerDay” column (use filter examples for filtering the values who are identifies as smokers

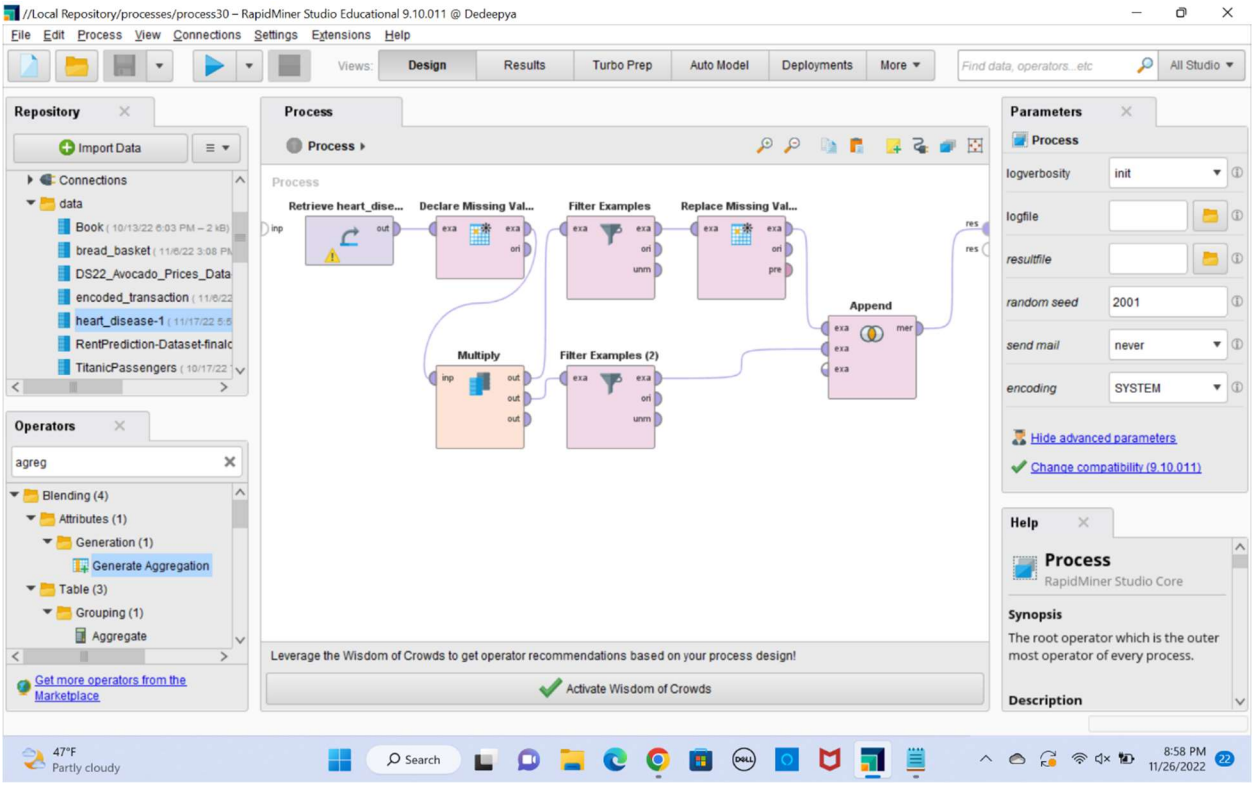
   Use replaces missing values operator and). Answer the following questions

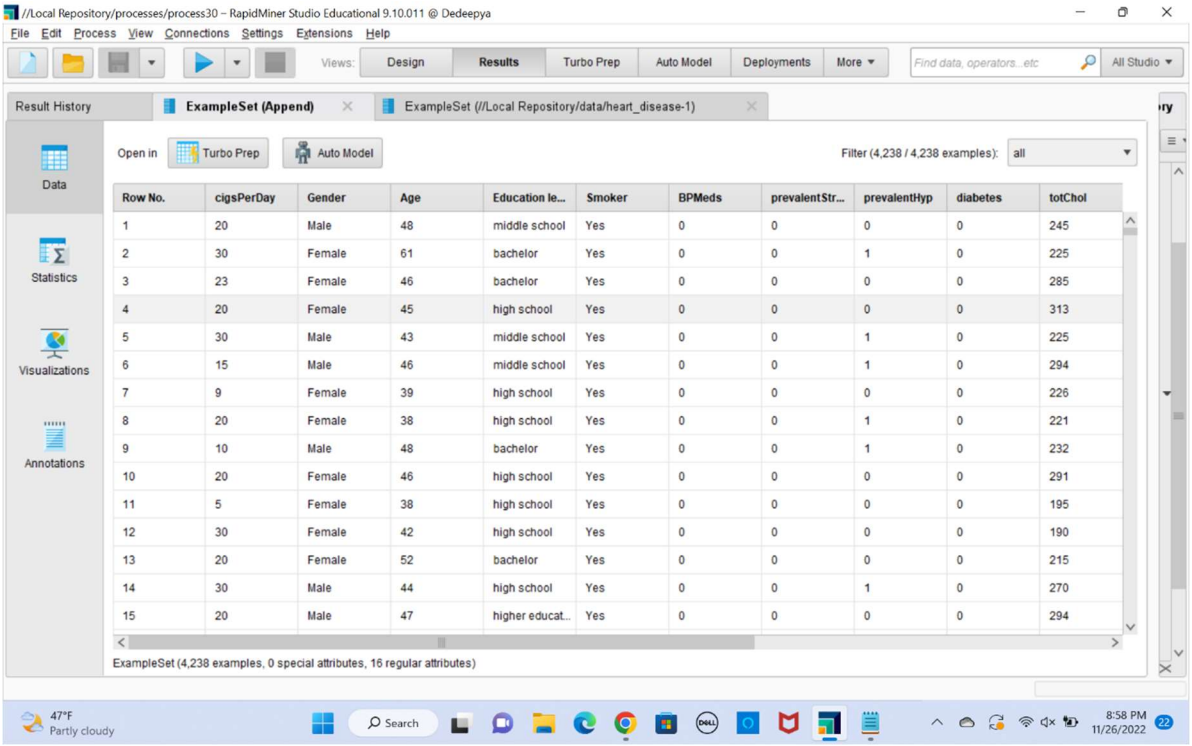
   Use append operator to combine the non-smoker data to missing values replaced data.)

   Include a screenshot of Result set that shows number of data on top right corner of result view using the all filter.

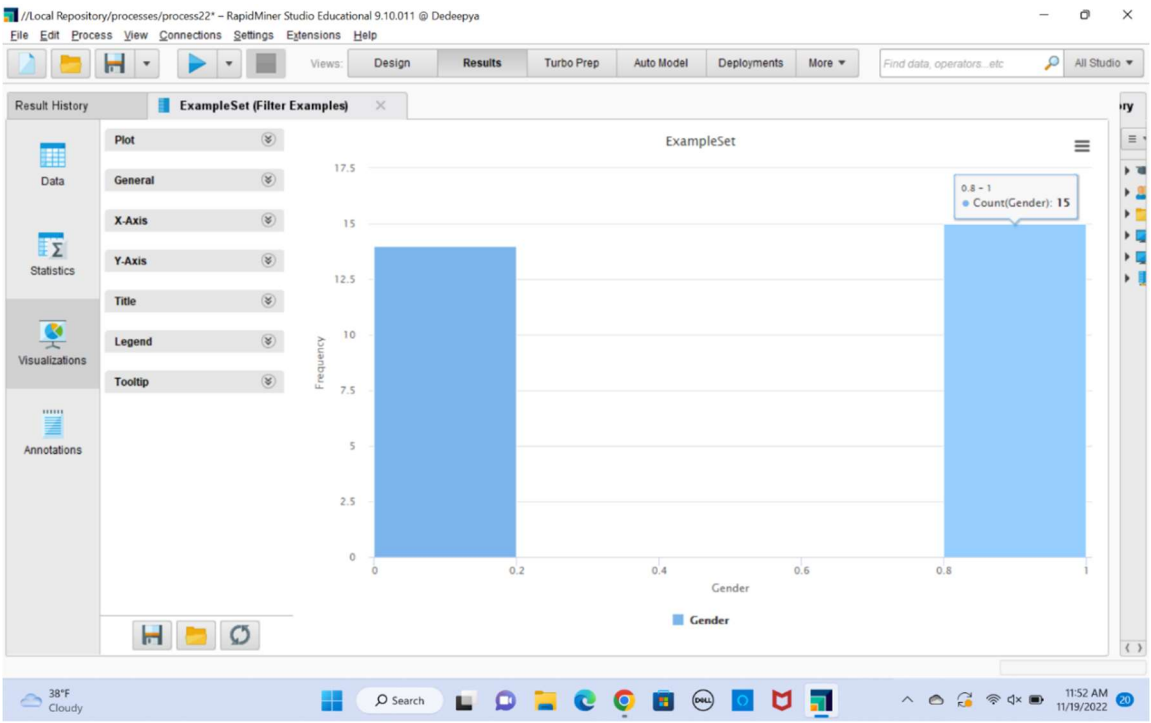
   Answer the following questions

   A. Out of all the records that contain missing values in “cigsPerDay”, how many of them are male, and how many of them are female?
   B. Find the age range of the Male smokers that have missing values in “cigsPerDay”. Do the same thing for Female smokers with missing values in “cigsPerDay”.
   C. What is the average “cigsPerDay” among the Male smokers, whose age is between 30 and 39 years old; 40 and 49 years old; 50 and 59 years old? (Round to the nearest integer)
   D. What is the average “cigsPerDay” among the Female smokers, whose age is between 30 and 39 years old; 40 and 49 years old; 50 and 59 years old; 60 and 69 years old? (Round to the nearest integer)





2A Male-14, female-15

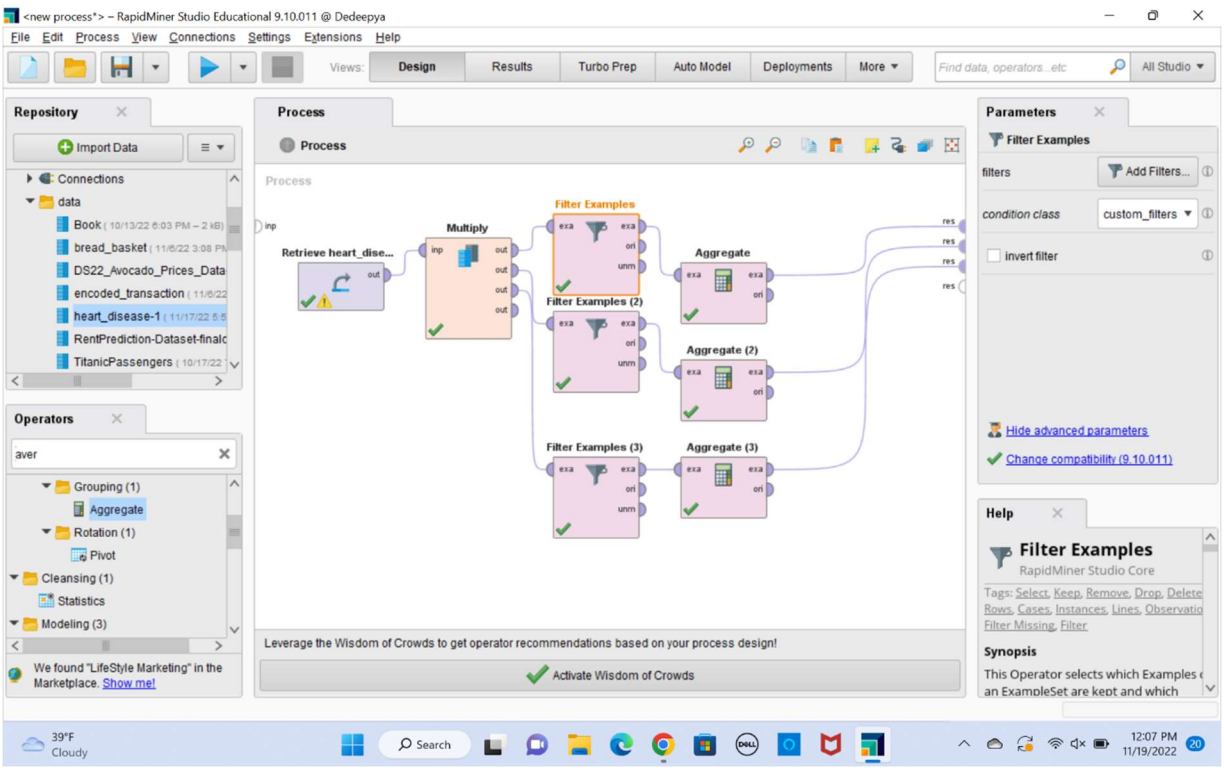


2b

Female age range 37-58, Male Age range - 39-61

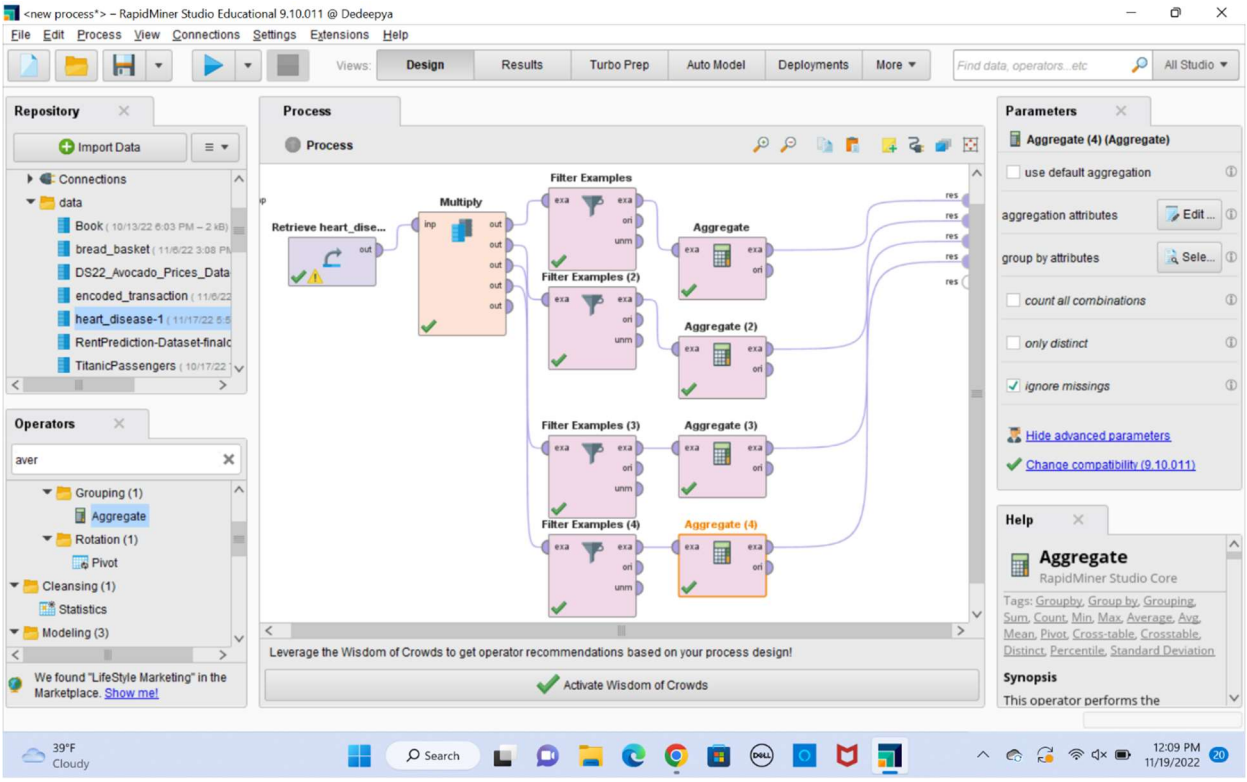
2c

Male Smoker age range	Average Cigerperday
30-39	22
40-49	23
50-59	22



2d.

Female Smoker age range	Average Cigerperday
30-39	15
40-49	14
50-59	14
60-69	14

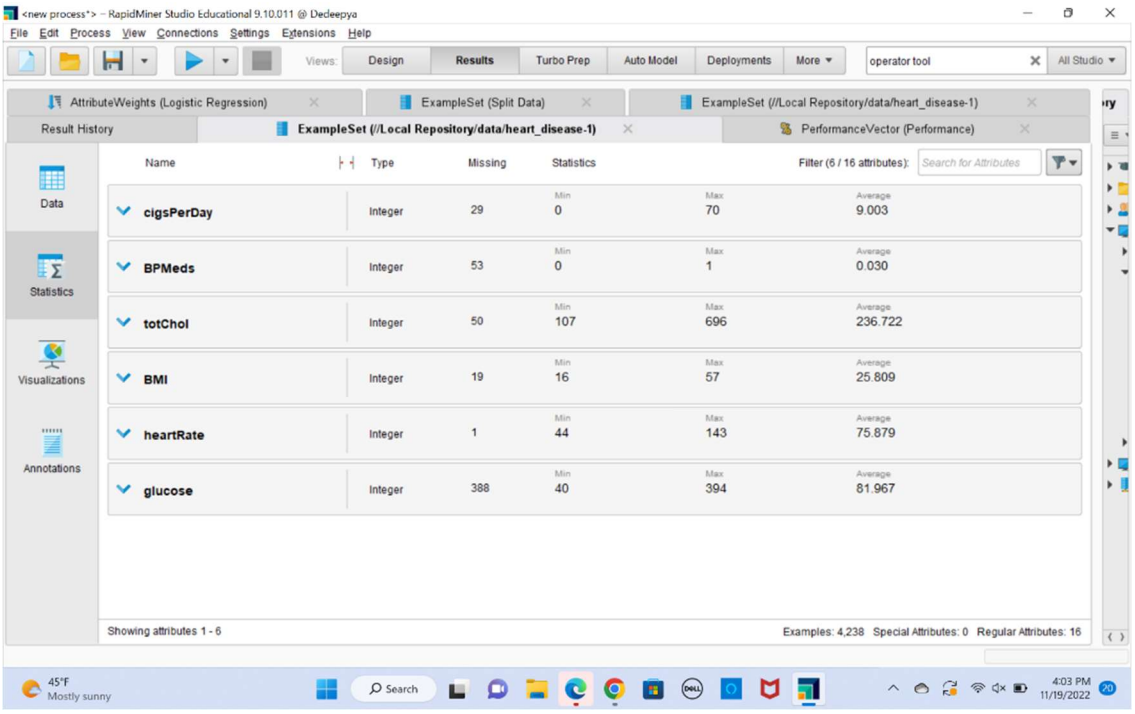

<new process*> – RapidMiner Studio Educational 9.10.011 @ Dedeepya
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Repository
Import Data
Connections
data
Book (10/13/22 6:03 PM – 2 kB)
bread_basket (11/6/22 3:08 PM
DS22_Avocado_Prices_Data
encoded_transaction (11/6/22
heart_disease-1 (11/17/22 5:5
RentPrediction-Dataset-finalc
TitanicPassengers (10/17/22
Operators
aver
Grouping (1)
Aggregate
Rotation (1)
Pivot
Cleansing (1)
Statistics
Modeling (3)
We found "LifeStyle Marketing" in the Marketplace. Show me!
Process
Process
Retrieve heart_dise...
Multiply
Filter Examples
Filter Examples (2)
Filter Examples (3)
Filter Examples (4)
Aggregate
Aggregate (2)
Aggregate (3)
Aggregate (4)
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Aggregate (4) (Aggregate)
use default aggregation
aggregation attributes
Edit ...
group by attributes
Sele...
count all combinations
only distinct
ignore missings
Hide advanced parameters
Change compatibility (9.10.011)
Help
Aggregate
RapidMiner Studio Core
Tags: Groupby, Group by, Grouping, Sum, Count, Min, Max, Average, Avg, Mean, Pivot, Cross-table, Crosstable, Distinct, Percentile, Standard Deviation
Synopsis
This operator performs the
39°F
Cloudy
Search
12:09 PM
11/19/2022

3. Fill in the missing values in the other columns and show proofs that you have completed these steps by including the screenshots of before and after replacing of missing values count in statistics tab: (For all the below questions use the replace missing values operator just like we used in question 2).

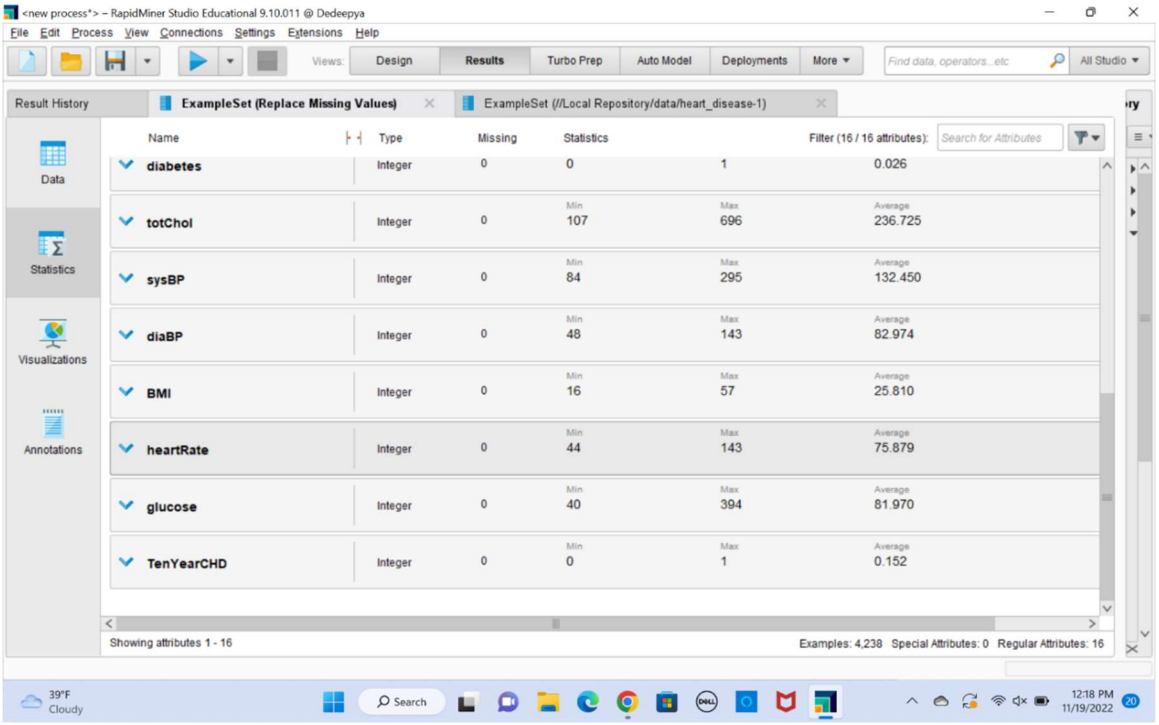
A. For each record that has a missing value in “glucose”, replace the missing values with the average “glucose”.

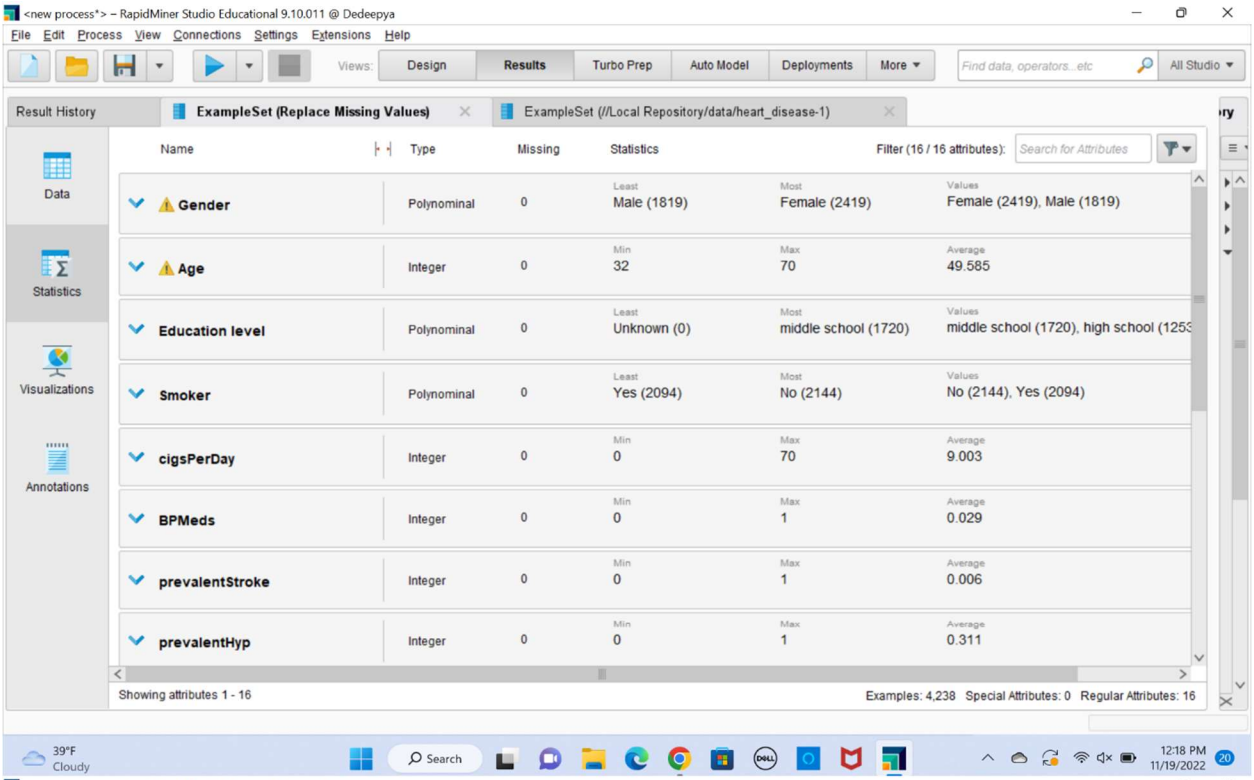
B. For each record that has a missing value in “totChol”, replace the missing values with the average “totChol”.
C. Replace the missing values in “BPMeds” with 0.
D. Replace the missing values in “BMI” with the average BMI.
E. Replace the missing values in “Education level” with the value “Unknown”.
F. Replace the missing values in “heartRate” with the average heart rate value.

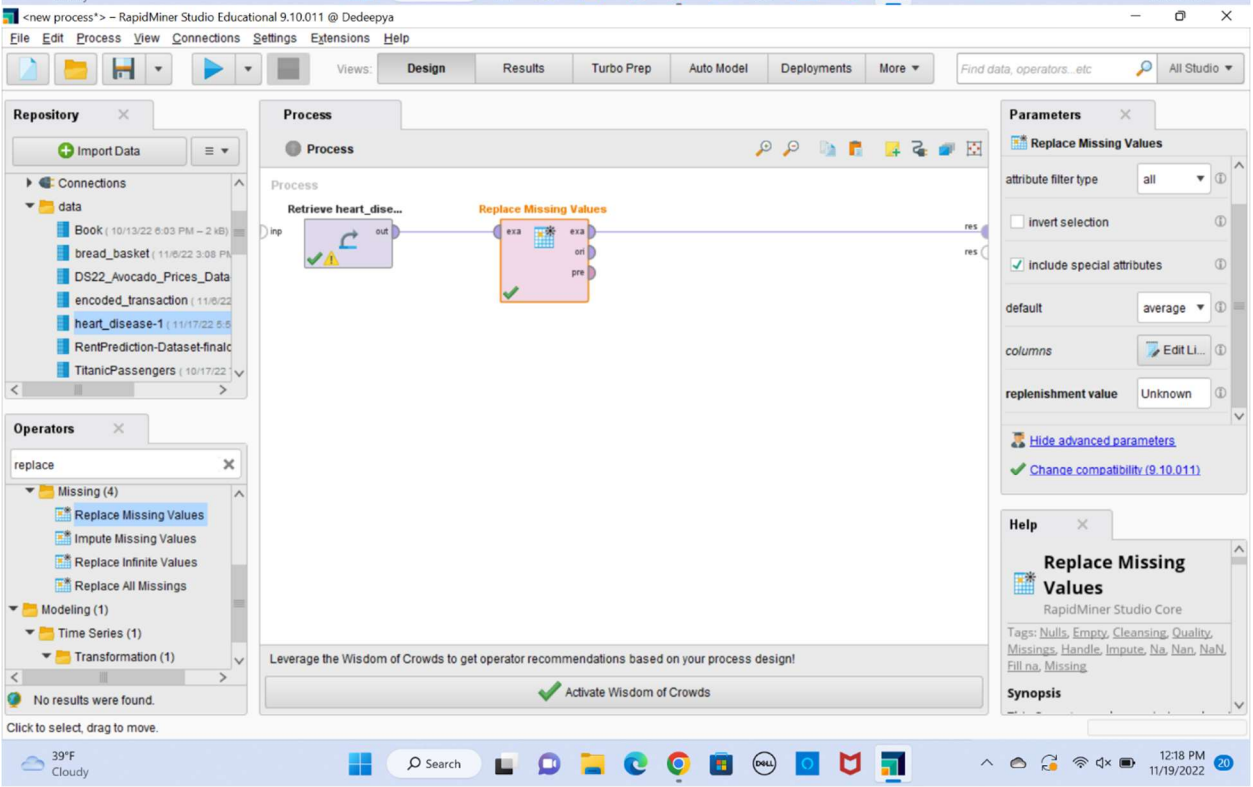
3. Before



After

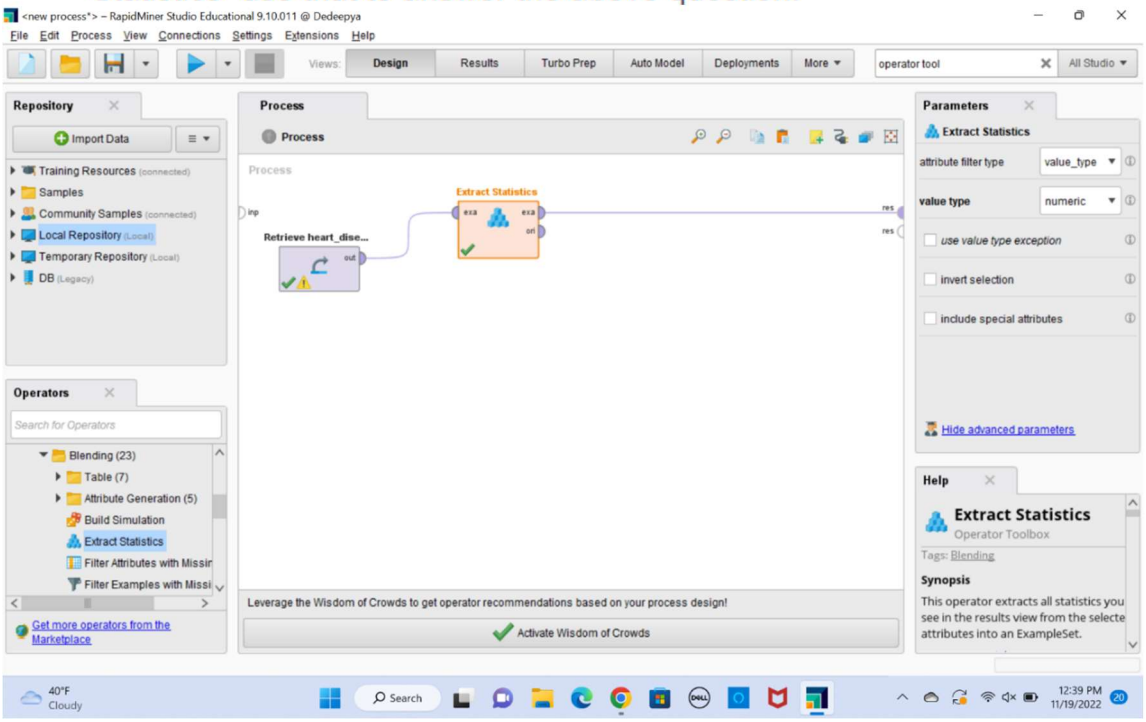

<new process*> – RapidMiner Studio Educational 9.10.011 @ Dedeepya
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Result History
ExampleSet (Replace Missing Values)
ExampleSet (//Local Repository/data/heart_disease-1)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (16 / 16 attributes):
Search for Attributes
diabetes Integer 0 0 1 0.026
totChol Integer 0 Min 107 Max 696 Average 236.725
sysBP Integer 0 Min 84 Max 295 Average 132.450
diaBP Integer 0 Min 48 Max 143 Average 82.974
BMI Integer 0 Min 16 Max 57 Average 25.810
heartRate Integer 0 Min 44 Max 143 Average 75.879
glucose Integer 0 Min 40 Max 394 Average 81.970
TenYearCHD Integer 0 Min 0 Max 1 Average 0.152
Showing attributes 1 - 16
Examples: 4,238 Special Attributes: 0 Regular Attributes: 16
39°F Cloudy
Search
12:18 PM 11/19/2022

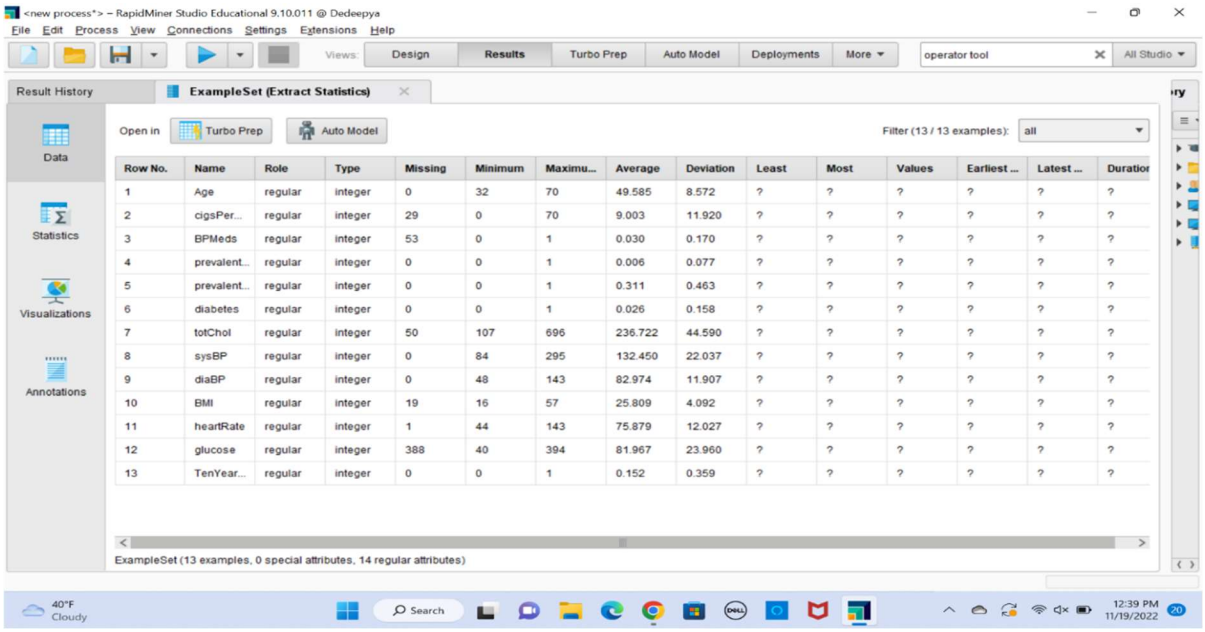

<new process*> – RapidMiner Studio Educational 9.10.011 @ Dedeepya
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Result History
ExampleSet (Replace Missing Values)
ExampleSet (//Local Repository/data/heart_disease-1)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (16 / 16 attributes):
Search for Attributes
Gender Polynominal 0 Least Male (1819) Most Female (2419) Values Female (2419), Male (1819)
Age Integer 0 Min 32 Max 70 Average 49.585
Education level Polynominal 0 Least Unknown (0) Most middle school (1720) Values middle school (1720), high school (1253
Smoker Polynominal 0 Least Yes (2094) Most No (2144) Values No (2144), Yes (2094)
cigsPerDay Integer 0 Min 0 Max 70 Average 9.003
BPMeds Integer 0 Min 0 Max 1 Average 0.029
prevalentStroke Integer 0 Min 0 Max 1 Average 0.006
prevalentHyp Integer 0 Min 0 Max 1 Average 0.311
Showing attributes 1 - 16
Examples: 4,238 Special Attributes: 0 Regular Attributes: 16
39°F Cloudy
Search
12:18 PM 11/19/2022


<new process*> – RapidMiner Studio Educational 9.10.011 @ Dedeepya
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Repository
Import Data
Connections
data
Book (10/13/22 6:03 PM – 2 kB)
bread_basket (11/6/22 3:08 PM
DS22_Avocado_Prices_Data
encoded_transaction (11/6/22
heart_disease-1 (11/17/22 5:5
RentPrediction-Dataset-finalc
TitanicPassengers (10/17/22
Operators
replace
Missing (4)
Replace Missing Values
Impute Missing Values
Replace Infinite Values
Replace All Missings
Modeling (1)
Time Series (1)
Transformation (1)
No results were found.
Process
Process
Process
Retrieve heart_dise...
Replace Missing Values
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Replace Missing Values
attribute filter type
all
invert selection
include special attributes
default
average
columns
Edit Li...
replenishment value
Unknown
Hide advanced parameters
Change compatibility (9.10.011)
Help
Replace Missing Values
RapidMiner Studio Core
Tags: Nulls, Empty, Cleansing, Quality, Missings, Handle, Impute, Na, Nan, NaN, Fill na, Missing
Synopsis
Click to select, drag to move.
39°F Cloudy
Search
12:18 PM 11/19/2022

3

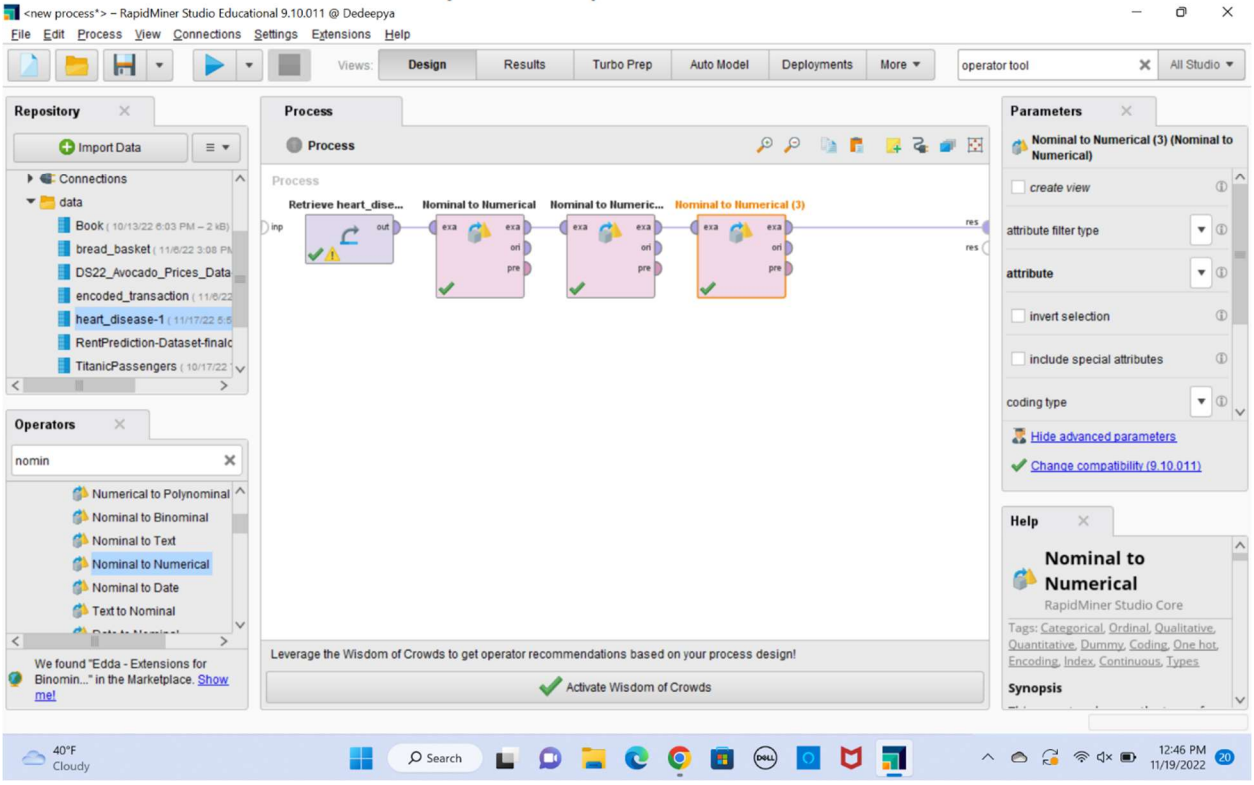
G. What is the mean, median, min, max, and standard deviation of all numerical variables?
Install Operator toolbox extension by click on Extensions tab on top toolbar in rapid miner. This extension contains operator name “Extract Statistics” use that to answer the above question.



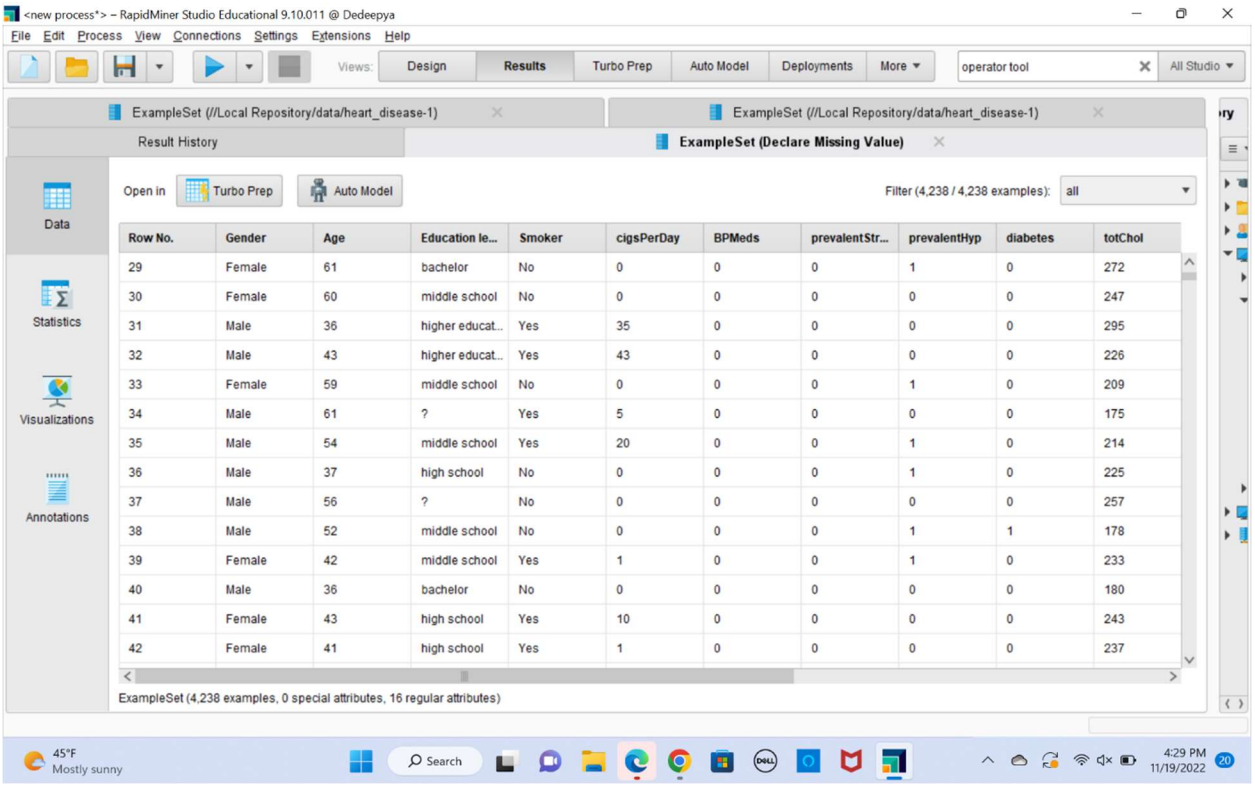


4. Encode categorical variables and include the before and after screenshots of each categorical variable.

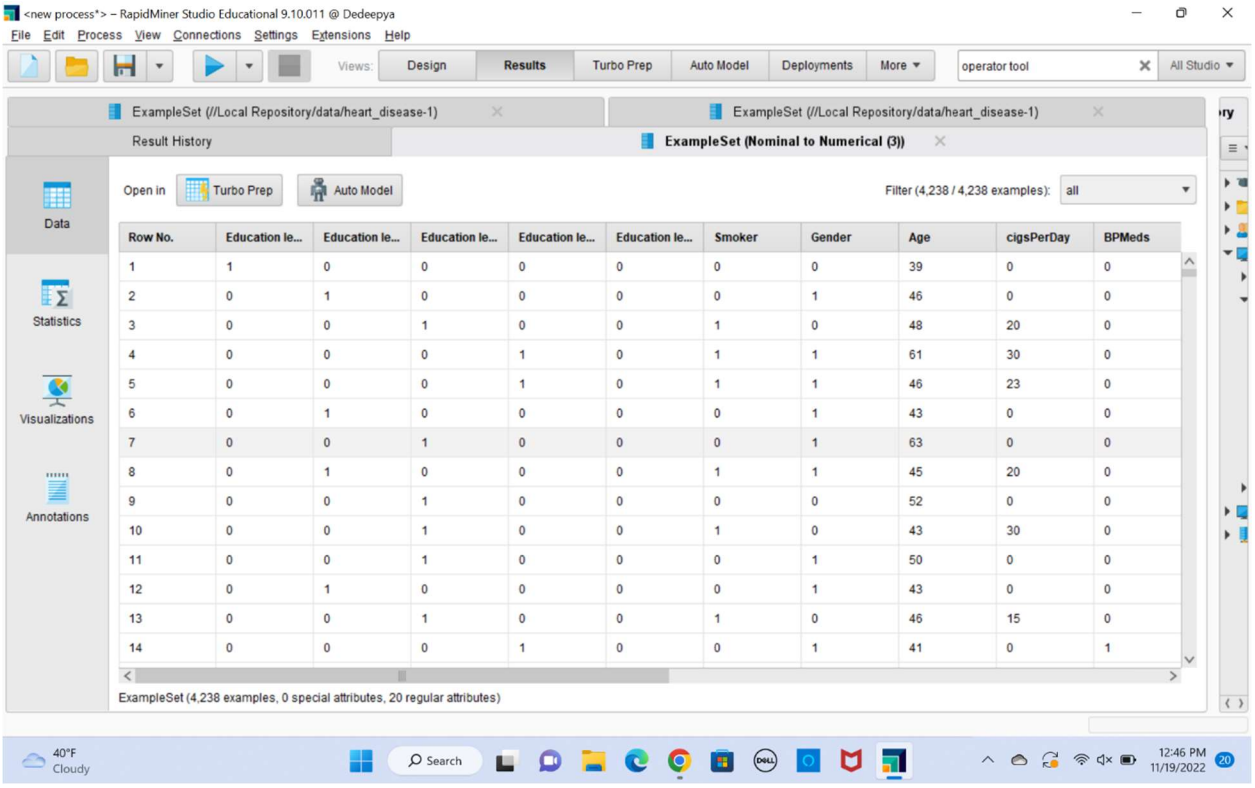
   A. For a categorical variable with more than 2 unique values, create a dummy variable for each unique value. (Use nominal to Numerical operator, this will create one column for each type of value in columns)
   B. For a binary categorical variable, encode it using 0 and 1. (Use the Nominal to numerical operator to replace binary categorical variables to numerical, use Unique integers coding type, it will just convert to integer by relacing one value with 0 and one value with 1, use this only if the column has two unique values).



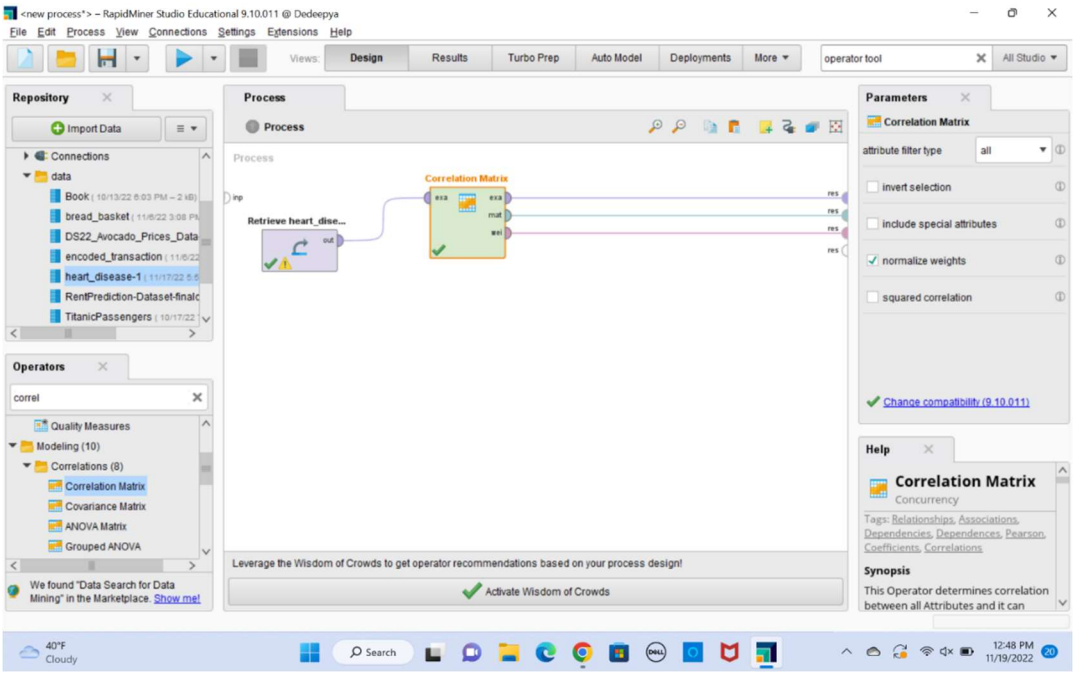
Before-



After-



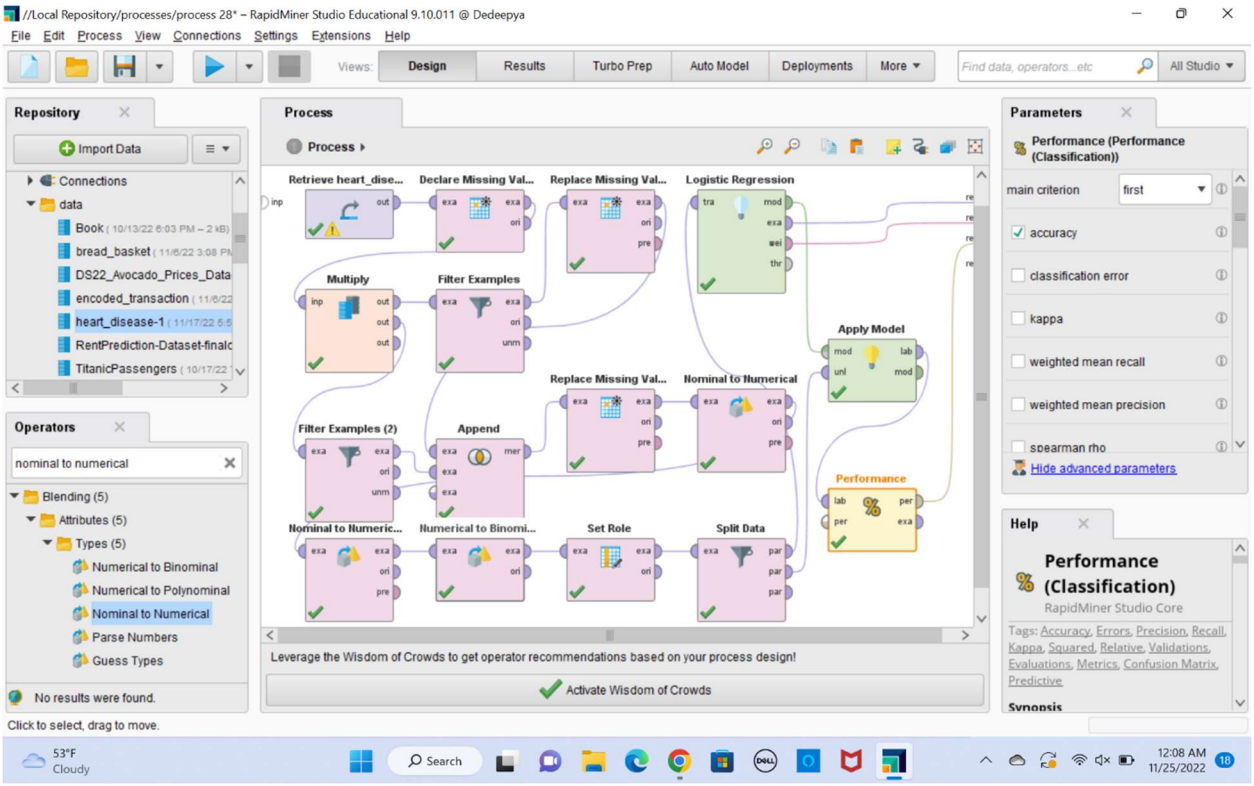
5. Create a correlation matrix and analyze it. Round the correlation values to the 2nd nearest decimal place. List the variables whose correlation with "TenYearCHD" is at least 0.1. List the variables whose correlation with "TenYearCHD" is between -0.1 and -0.01.

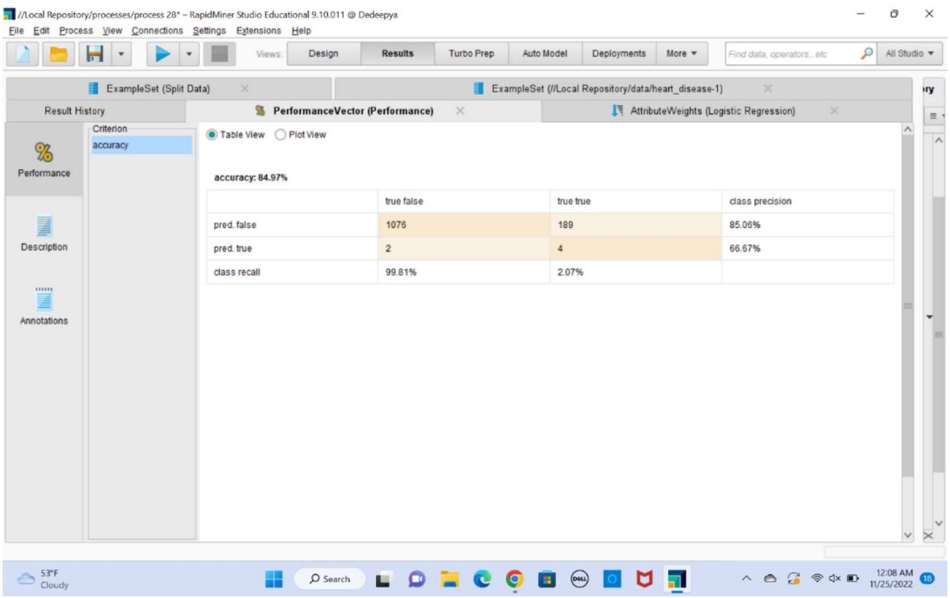


Columns with values of correlation At least 0.1 are - age, prevalenthyp, sysbp, diabp, glucose.

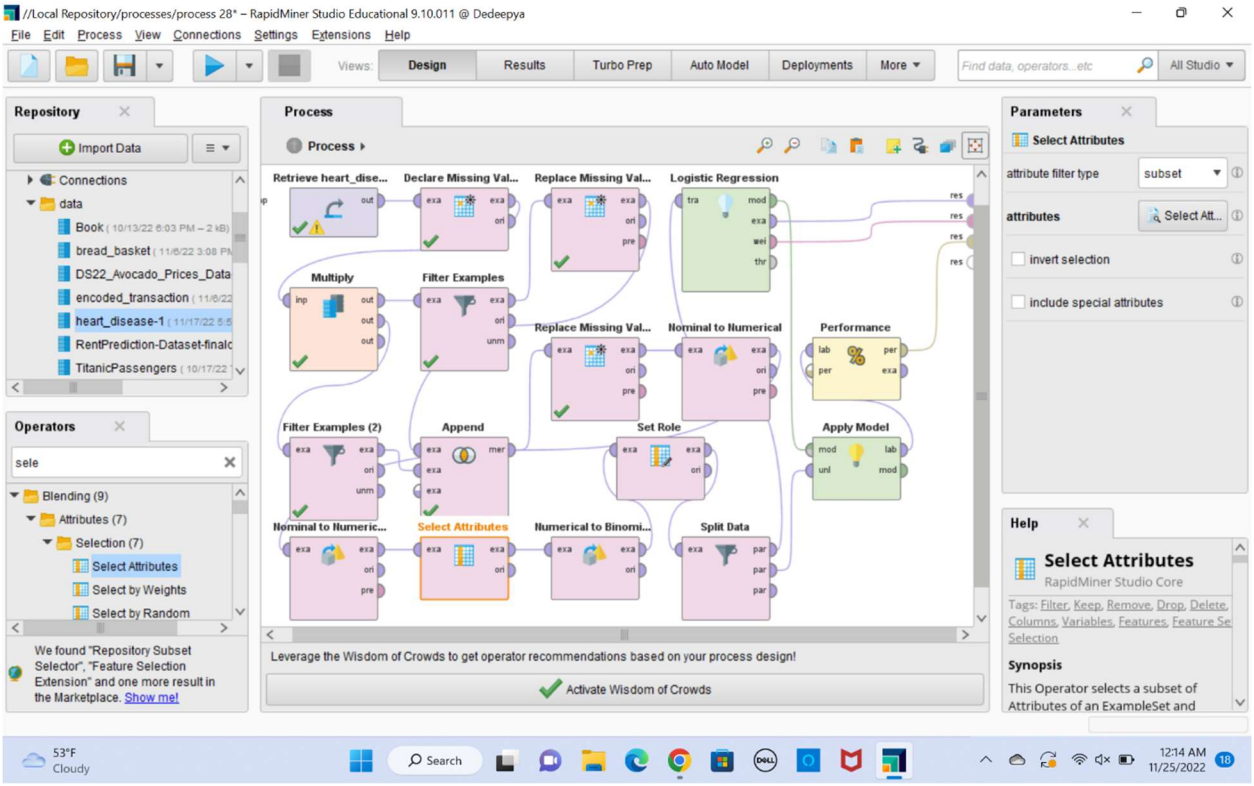
Columns with values of correlation Between -0.1 to -0.01 are – gender,

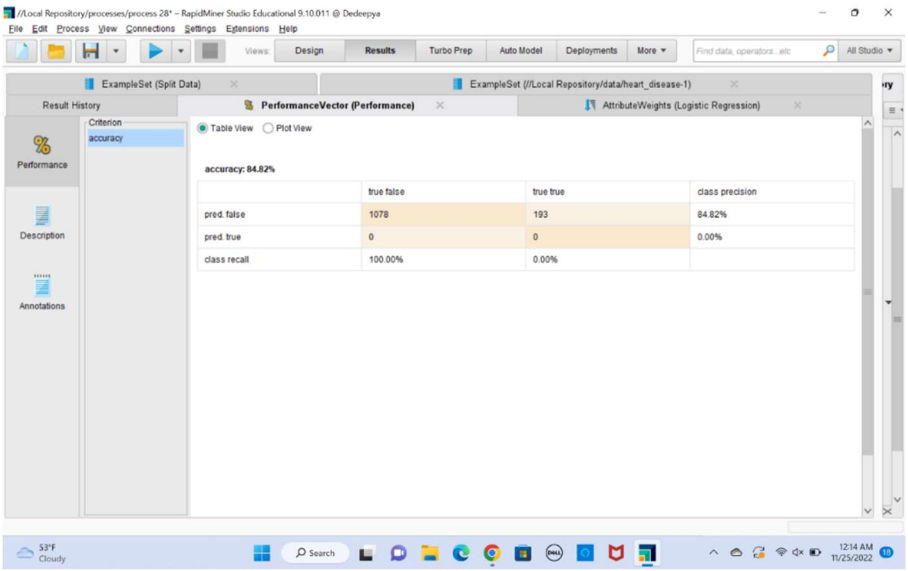
6. Split the data into training and test datasets with a ratio of 70:30 using a random state of 58. Build a Logistic Regression model using all independent variables. What is the test accuracy? How many records were misclassified?





7. Using the same training and test sets above, build a Logistic Regression model using the variables identified in Question 5 and a random state of 58, and evaluate it with the test set. What is the test accuracy? How many records were misclassified? Compared to the model built in Question 6, which model is better? Why?





From both the models we can see that there is high accuracy for model build with all the features instead of models with less features. As we reduced number of features by using correlation matrix the performance of second model decreased.