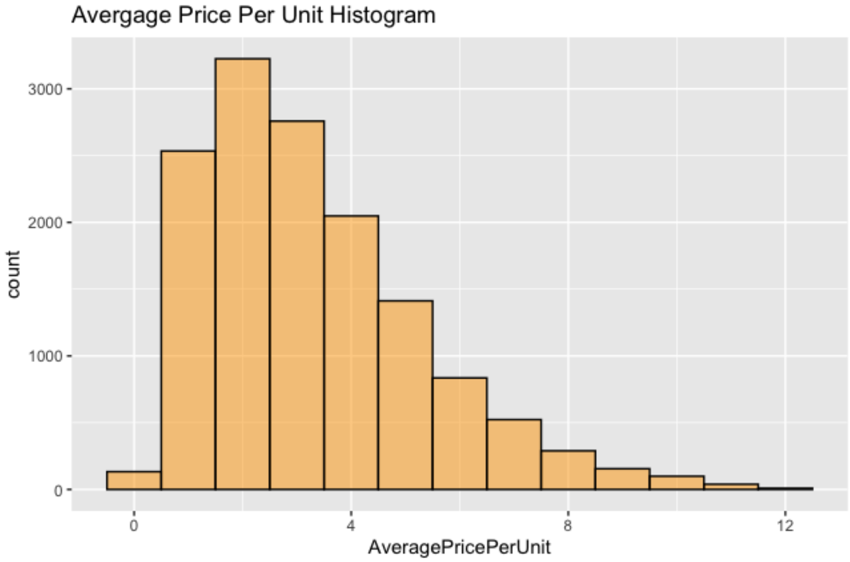
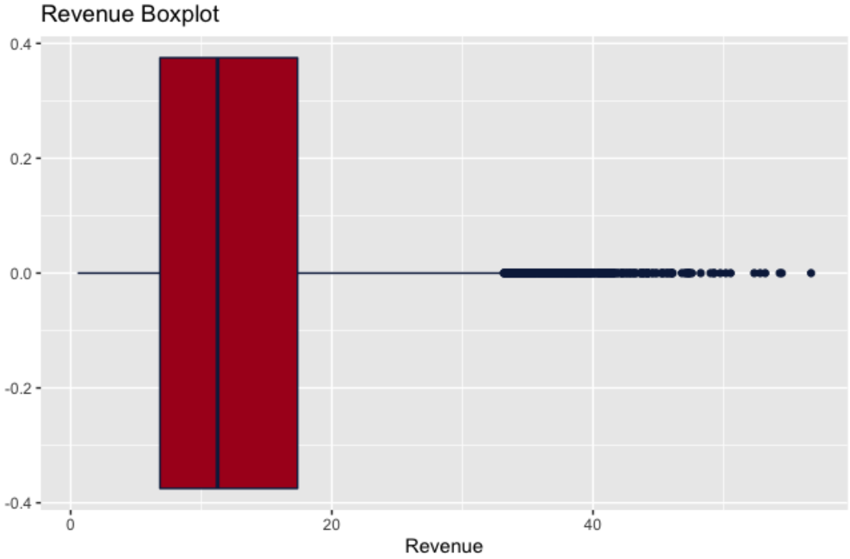
R script code:

```
# Aadhithya Dinesh
# MIS 545 Section 02
# Lab03DineshA.R
# Import and prepare a large dataset of grocery store transactions, assign
# data types, view summary statistics, generate a histogram and a boxplot


# install.packages("tidyverse")
library(tidyverse)

# set the working directory
setwd("~/MIS/Classes/MIS545/Assignments/Lab03")

# read the csv file with column types specified
groceryTransactions1 <- read_csv(file = "GroceryTransactions.csv",
                                 col_types = "iDffffifffffffin",
                                 col_names = TRUE)
# display the tibble
print(groceryTransactions1)

# display the first 20 rows
print(head(groceryTransactions1, n=20))

# display the structure of the tibble
str(groceryTransactions1)

# display the summary of the tibble
print(summary(groceryTransactions1))

# using dplyr summarize function to display mean of Revenue
print(summarize(.data = groceryTransactions1, mean(Revenue)))
# using dplyr summarize function to display median of UnitsSold
print(summarize(.data = groceryTransactions1, median(UnitsSold)))
# using dplyr summarize function to display standard deviation of Revenue
print(summarize(.data = groceryTransactions1, sd(Revenue)))
# using dplyr summarize function to display inter-quartile-range of Units sold
print(summarize(.data = groceryTransactions1, IQR(UnitsSold)))
# using dplyr summarize function to display min of revenue
print(summarize(.data = groceryTransactions1, min(Revenue)))
# using dplyr summarize function to display  max of children
print(summarize(.data = groceryTransactions1, max(Children)))

# creating a new tibble with the below mentioned columns
groceryTransactions2 <- select(.data = groceryTransactions1,
                               PurchaseDate,
                               Homeowner,
                               Children,
                               AnnualIncome,
                               UnitsSold,
                               Revenue)
```

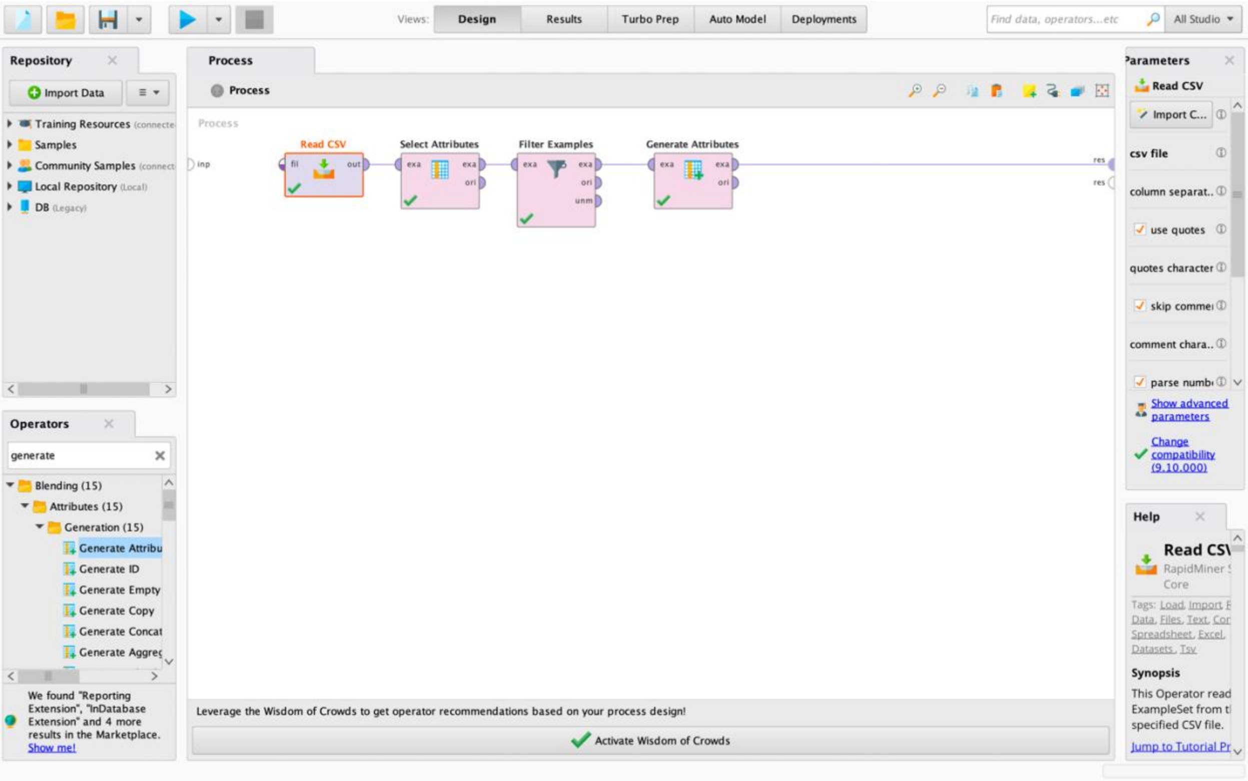
```
# display all features for transactions made by non-homeowners with at least
# 4 children
print(filter(.data = groceryTransactions2,
        Homeowner == 'N' & Children >=4))

# display all of the records and features that were either made by customers
# in the $150K + annual income category OR had more than 6 units sold
print(filter(.data = groceryTransactions2,
        AnnualIncome == '150k +' | UnitsSold > 6))

revnue <- select(.data = groceryTransactions1,
          Revenue)

# display the average transaction revenue grouped by annual income level
print(select(.data = groceryTransactions1,
    Revenue,
    AnnualIncome)%>%
    group_by(AnnualIncome)%>%
    summarize(averageRevnue = mean(Revenue))%>%
    arrange(desc(averageRevnue)))

# calculate average price per unit as revenue/ units sold
groceryTransactions3 <- groceryTransactions2 %>%
  mutate(AveragePricePerUnit = Revenue/UnitsSold)

print(groceryTransactions3)

histogramAveragePricePerUnit <- ggplot(data = groceryTransactions3,
                        aes(x=AveragePricePerUnit))
# creating the histogram for average price per unit
histogramAveragePricePerUnit + geom_histogram(binwidth = 1,
                            color = "black",
                            fill = "orange",
                            alpha = 0.5
                            ) +
  ggtitle("Avergage Price Per Unit Histogram")

boxplotRevnue <- ggplot(data = groceryTransactions3,
               aes(x=Revenue))
# creating the box plot for revenue
boxplotRevnue + geom_boxplot(color = "#0C234B",
                  fill = "#AB0520")
```

Visualizations:





Rapid Miner Process:



Parameters: Read CSV

- Import C...
- csv file
- column separat..
- ☑ use quotes
- quotes character
- ☑ skip comme
- comment chara..
- ☑ parse numb

Tags: Load, Import, Data, Files, Text, Spreadsheet, Excel, Datasets, Tsv

Synopsis

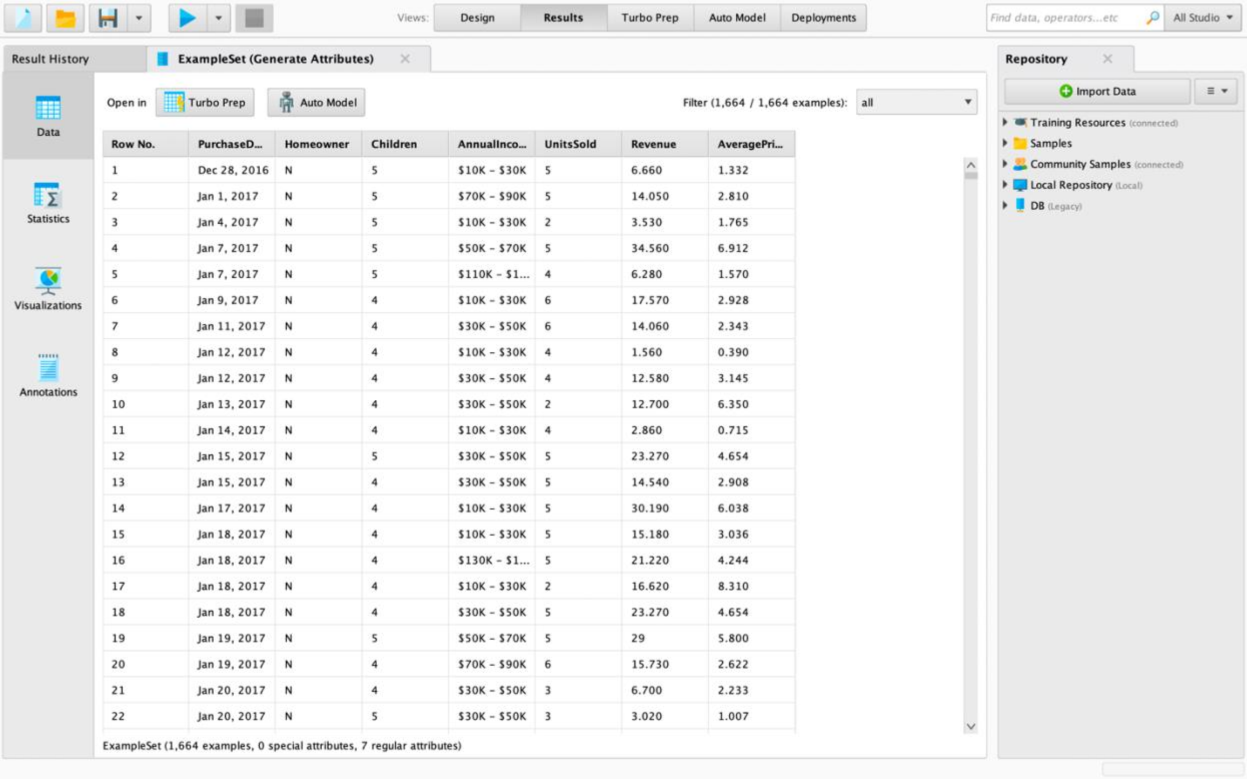
This Operator read ExampleSet from t specified CSV file.

Jump to Tutorial Pr

We found "Reporting Extension", "InDatabase Extension" and 4 more results in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Results:


Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (Generate Attributes)
Data
Statistics
Visualizations
Annotations
Open in
Turbo Prep
Auto Model
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)

Filter (1,664 / 1,664 examples): all

Row No.	PurchaseD...	Homeowner	Children	AnnualInco...	UnitsSold	Revenue	AveragePri...
1	Dec 28, 2016	N	5	$10K - $30K	5	6.660	1.332
2	Jan 1, 2017	N	5	$70K - $90K	5	14.050	2.810
3	Jan 4, 2017	N	5	$10K - $30K	2	3.530	1.765
4	Jan 7, 2017	N	5	$50K - $70K	5	34.560	6.912
5	Jan 7, 2017	N	5	$110K - $1...	4	6.280	1.570
6	Jan 9, 2017	N	4	$10K - $30K	6	17.570	2.928
7	Jan 11, 2017	N	4	$30K - $50K	6	14.060	2.343
8	Jan 12, 2017	N	4	$10K - $30K	4	1.560	0.390
9	Jan 12, 2017	N	4	$30K - $50K	4	12.580	3.145
10	Jan 13, 2017	N	4	$30K - $50K	2	12.700	6.350
11	Jan 14, 2017	N	4	$10K - $30K	4	2.860	0.715
12	Jan 15, 2017	N	5	$30K - $50K	5	23.270	4.654
13	Jan 15, 2017	N	4	$30K - $50K	5	14.540	2.908
14	Jan 17, 2017	N	4	$10K - $30K	5	30.190	6.038
15	Jan 18, 2017	N	4	$10K - $30K	5	15.180	3.036
16	Jan 18, 2017	N	4	$130K - $1...	5	21.220	4.244
17	Jan 18, 2017	N	4	$10K - $30K	2	16.620	8.310
18	Jan 18, 2017	N	4	$30K - $50K	5	23.270	4.654
19	Jan 19, 2017	N	5	$50K - $70K	5	29	5.800
20	Jan 19, 2017	N	4	$70K - $90K	6	15.730	2.622
21	Jan 20, 2017	N	4	$30K - $50K	3	6.700	2.233
22	Jan 20, 2017	N	5	$30K - $50K	3	3.020	1.007

ExampleSet (1,664 examples, 0 special attributes, 7 regular attributes)