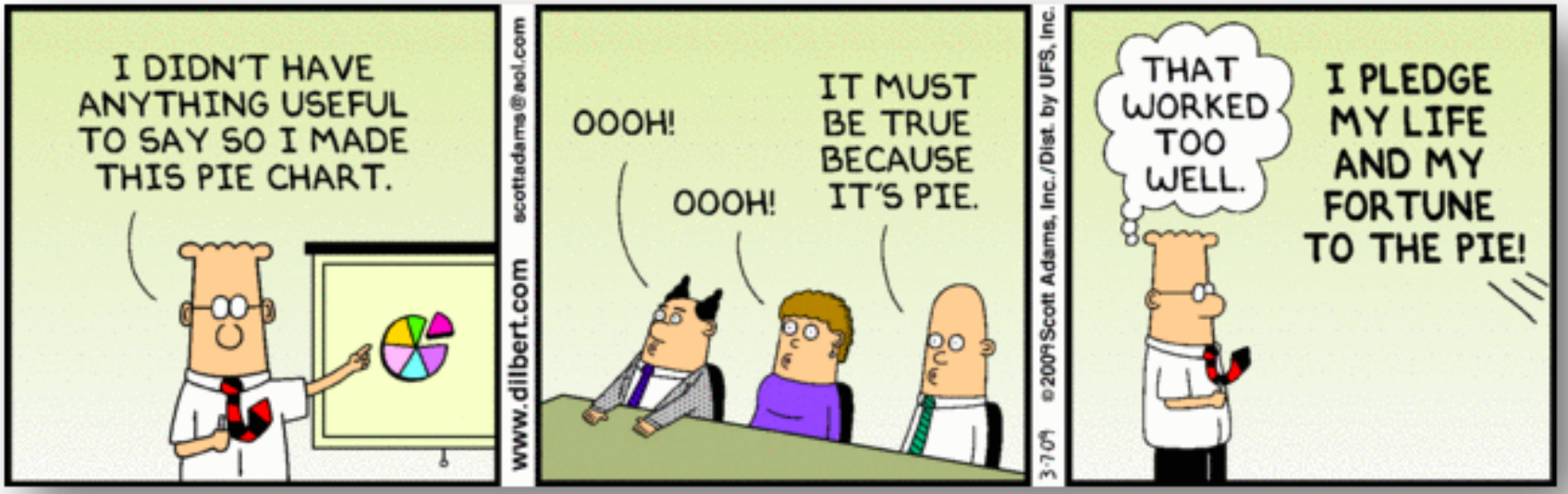
Slides by
Michael Hahsler

DS 1300: Essential Data Preparation, Descriptive Statistics and Visualization



With Examples for rapidminer

Purpose

Import data and "get used" to the data. Use simple statistics and visualization.

Clean the data (e.g., find missing values, outliers, mistakes)

"Make sure the data makes sense."

Find simple relationships between variables.

Prepare data for predictive/prescriptive modeling.

- Install RapidMiner Studio and obtain an educational license (see course website).
- The dataset for the examples can be obtained from http://michael.hahsler.net/SMU/EMIS3309/data/census.csv
- Rapidminer processes for this slide set are available here (save and import process in Rapidminer):
  - http://michael.hahsler.net/SMU/EMIS3309/data/rapidminer/Basic_Statistics_and_Visualizations.rmp
  - http://michael.hahsler.net/SMU/EMIS3309/data/rapidminer/Cleaning_and_preprocessing.rmp

Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms

Gartner Magic Quadrant

Data and Measurements

Scale of Measurement

- Information can be measured on different scales. Depending on the scale, different operations/visualizations are appropriate.

Scale	Examples	Mathematical operators	Advanced operations	Central tendency	
Nominal	Gender, eye Color, Zip code	=, !=	Grouping	Mode	Categorical
Ordinal	hardness of minerals, {good, better, best}, grades	>, <	Sorting	Median	
Interval	temperature in Celsius or Fahrenheit	+, −	Difference	Mean, Variance	Quantitative
Ratio	temperature in Kelvin, monetary quantities, counts, age, mass, length (has a meaningful 0)	×, /	Ratio	Geometric mean, percent variation	

Scale of Measurement

- What is the scale of measurement (nominal, ordinal, or interval/ratio) for the following. What operations are appropriate.
- Grades (letter): A, B, C, D, F
- Grades (for GPA): 4, 3, 2, 1, 0
- Points on a test: 0-100
- Age: 0, 1, 2, … years old
- Age: <20, 21-35, 36-50, 51+
- Waiting time: E.g., 2.5 minutes
- Number of students in classes: E.g., 32
- Percentage of female students in class: E.g., 60%
- Student ID: E.g., 9212354
- Date: March 26, 2018

Importing Data

- *For the examples we use a dataset with census data at the ZIP-code level (Data and processes can be found on the class website).*

should **not** be integer!

Features/Attributes

Observations

	zipcode *integer*	state *polynominal*	population *integer*	housingunits *integer*	landaream... *integer*	waterarea... *integer*	landareami... *real*	waterarea... *real*
1	601	PR	19143	6715	172731389	1082233	66.692	0.418
2	602	PR	42042	15590	80137374	0	30.941	0.000
3	603	PR	55530	21626	78693011	83181	30.384	0.032
4	604	PR	3923	1245	7785336	0	3.006	0.000
5	606	PR	6449	2272	94870047	0	36.630	0.000
6	610	PR	27975	10566	96017843	668032	37.073	0.258
7	612	PR	72730	28754	192519622	2881056	74.332	1.112
8	616	PR	10532	4043	40783913	14091	15.747	0.005
9	617	PR	23370	8725	55340822	1660753	21.367	0.641
10	622	PR	8104	6323	78576914	3070980	30.339	1.186
11	623	PR	38807	16859	103636577	1492416	40.014	0.576
12	624	PR	26719	8735	114880458	649343	44.356	0.251
13	627	PR	35244	12520	120242300	31377	46.426	0.012
14	631	PR	2188	767	23479131	477966	9.065	0.185
15	637	PR	25935	9982	92943730	10072	35.886	0.004
16	638	PR	19634	6823	169524805	428000	65.454	0.165
17	641	PR	35015	12333	284371047	3912858	109.796	1.511
18	646	PR	34099	13096	61037857	268985	23.567	0.104

Categorical
(RM: polynominal)

Quantitative – ratio
(RM: integer or real)


Simple Statistics and Visualization

Single Variable - Quantitative

RapidMiner
- Statistics
- Operator Aggregation

Descriptive Stats

- **5-Number Summary:**
  - min
  - 1st quartile
  - Median
  - Mean
  - 3rd quartile
  - max

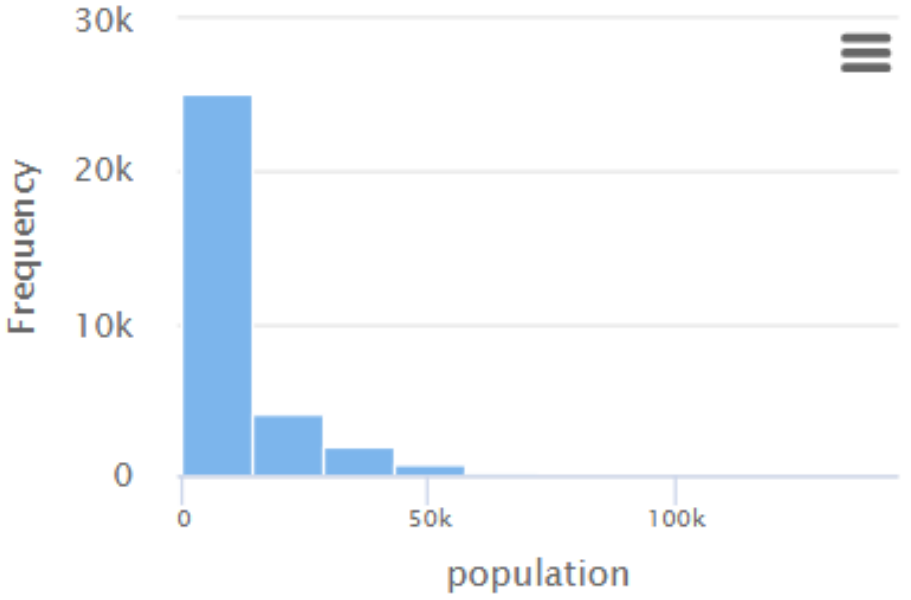
Rapidminer gives you this for population per Zipcode:

			Min	Max	Average
population	Integer	0	0	143987	8901.509

Visualization

- **Histogram**

  to show the distribution



Single Variable - Categorical

Descriptive Stats

RapidMiner
- Statistics

- **Count table**

state	Polynominal	0	 Open visualizations	Least DC (26)	Most TX (1865)	Values TX (1865), PA (1722), CA (1677), NY (1619), ...[48 more] Details...

Visualization

- **Bar chart** for counts
- **Pie chart** (not ideal for more than a few groups)

Two Variables - Quantitative

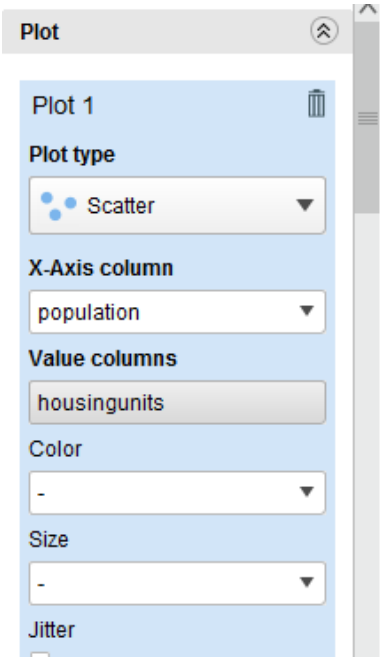
Descriptive Stats

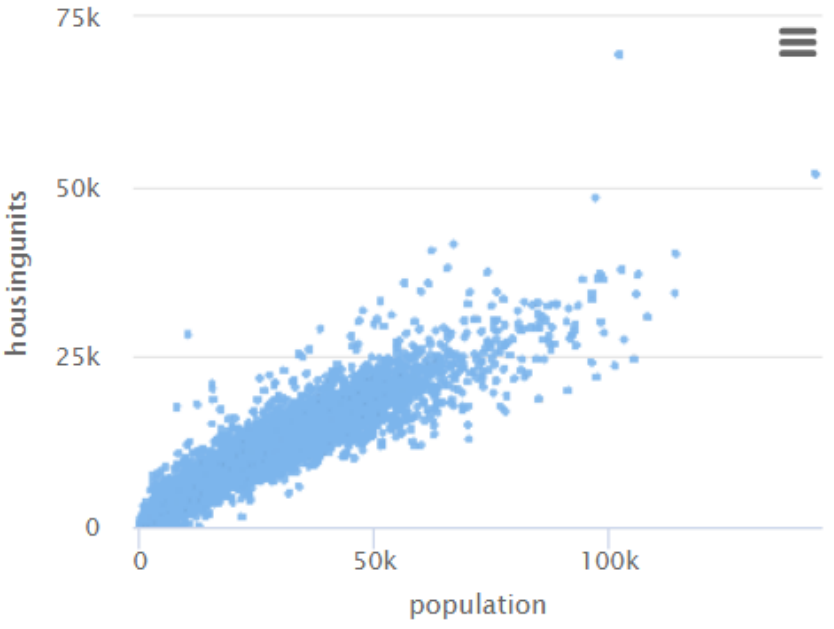
RapidMiner: Use Correlation Matrix node

- **Correlation**
- *Example: population and # of housing units per zipcode have a (Pearson) correlation coefficient of: 0.975*

Visualization

- **Scatterplot**
- Is there a relationship?
- Multivariate Outliers





Two Variables - Categorical

Descriptive Stats

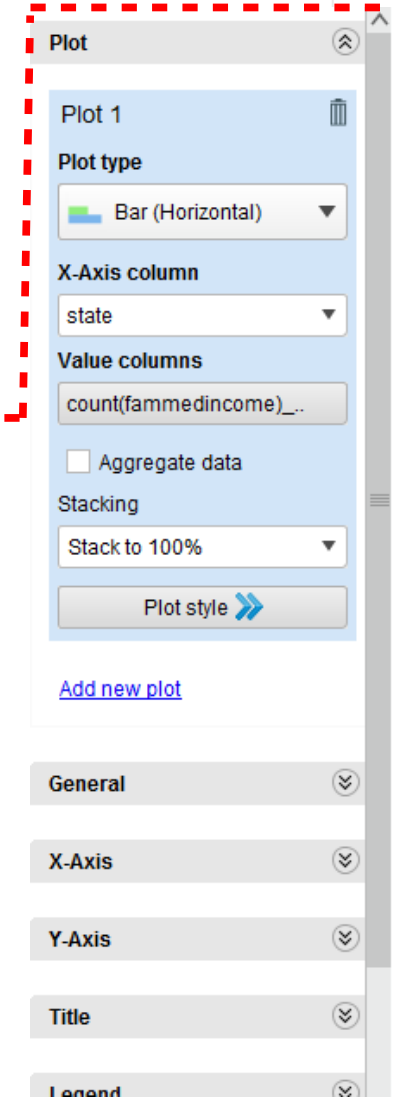
- **Cross-tabulation**
  (i.e., contingency table)

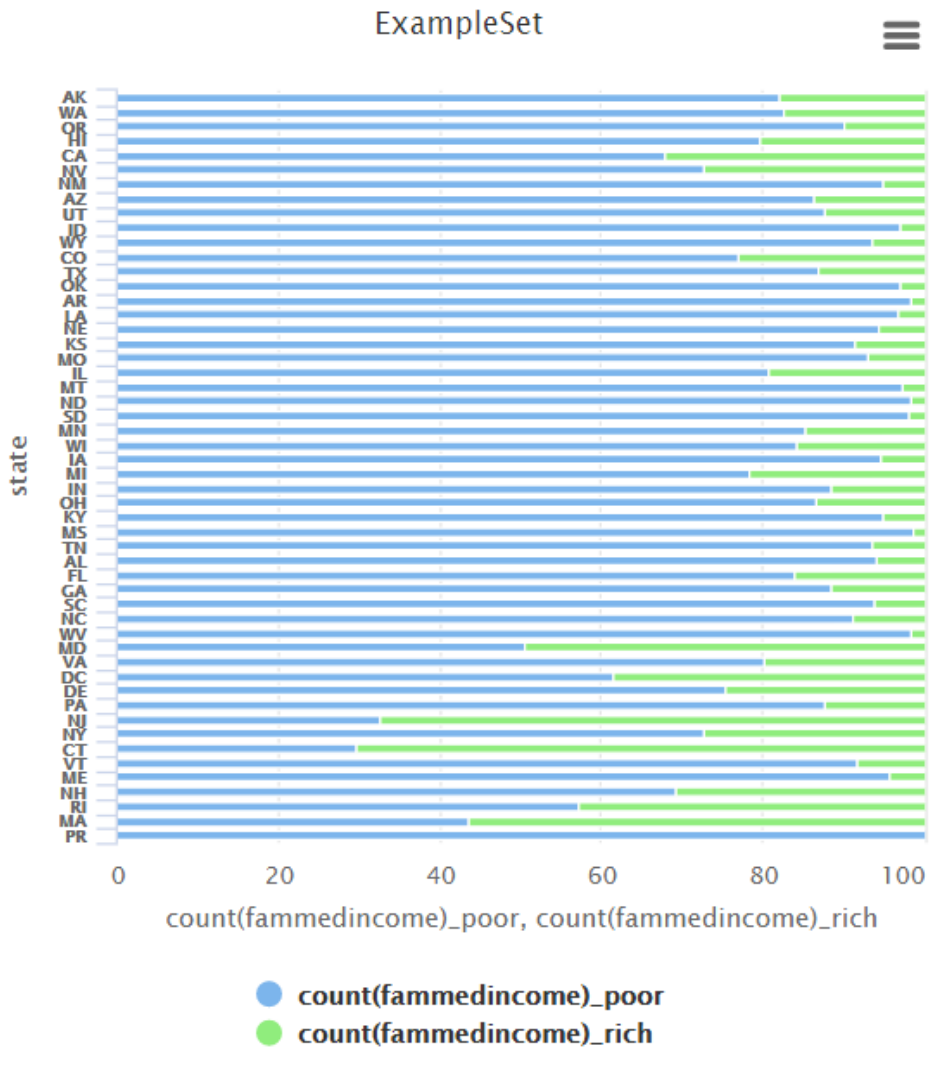
Row No.	state	count(fammedincome)_poor ↓	count(fammedincome)_rich
40	TX	1621	244
10	PA	1508	214
8	NY	1177	442

RapidMiner
Aggregation
Pivot

Visualization

- **Bar chart** (counts)
- **Stacked bar chart**
  (proportion)





Two Variables - Mixed

RapidMiner: Use Aggregate node

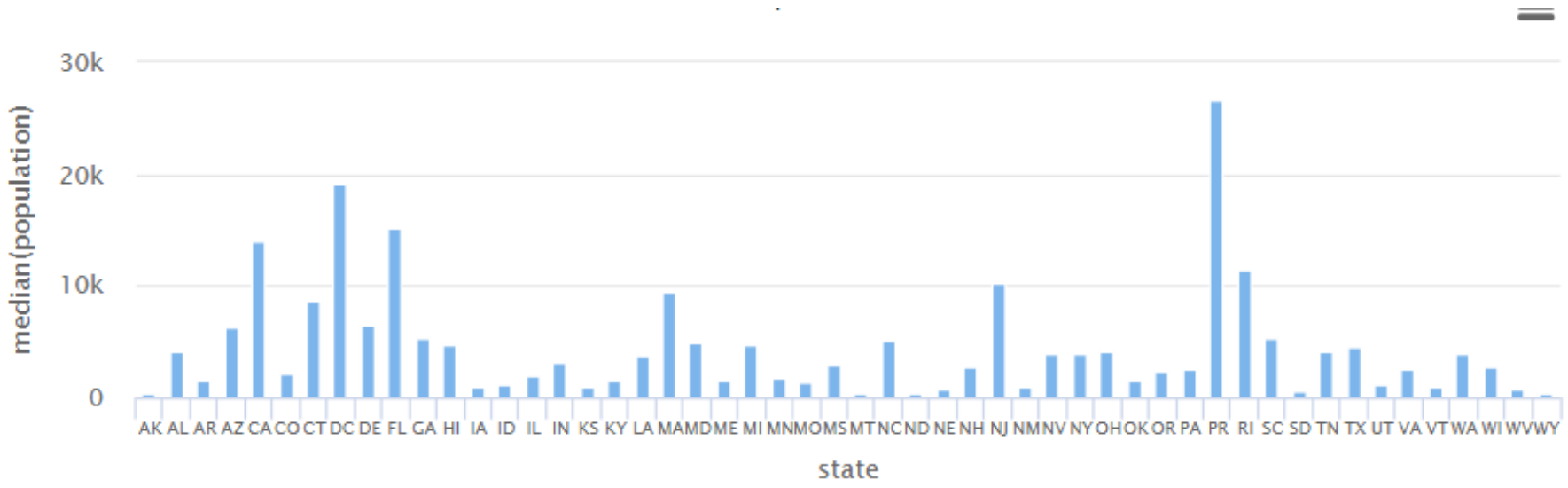
Descriptive Stats

- Compare **5-number statistic** grouped by categorical variable.

Row No.	state ↑	minimum(population)	average(population)	median(population)	maximum(population)
1	AK	10	2593.328	385	37284
2	AL	0	7125.199	4121	46541
3	AR	0	4523.746	1485	59262
4	AZ	32	14177.428	6226	71745
5	CA	0	20193.794	14077	105275
6	CO	0	8923.438	2120	68492

Visualization

- Bar chart for individual statistic to compare groups or box plot



Multiple Variables

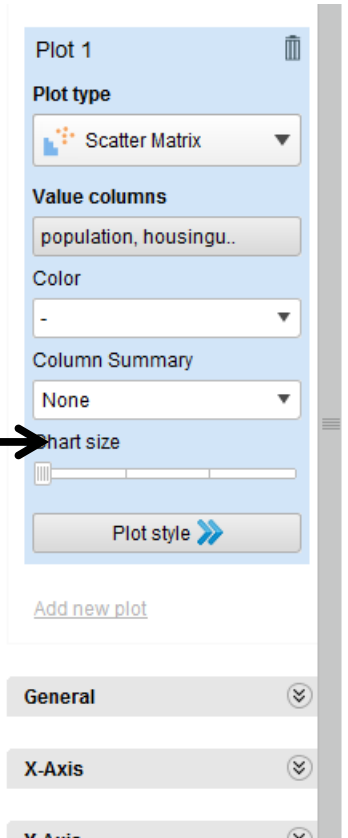
- Are usually broken down into pairwise comparisons.

Attributes	population	housingunits	landareamiles	fammedincome
population	1	0.975	-0.063	0.220
housingunits	0.975	1	-0.057	0.223
landareamiles	-0.063	-0.057	1	-0.123
fammedincome	0.220	0.223	-0.123	1

RapidMiner: Use Correlation Matrix

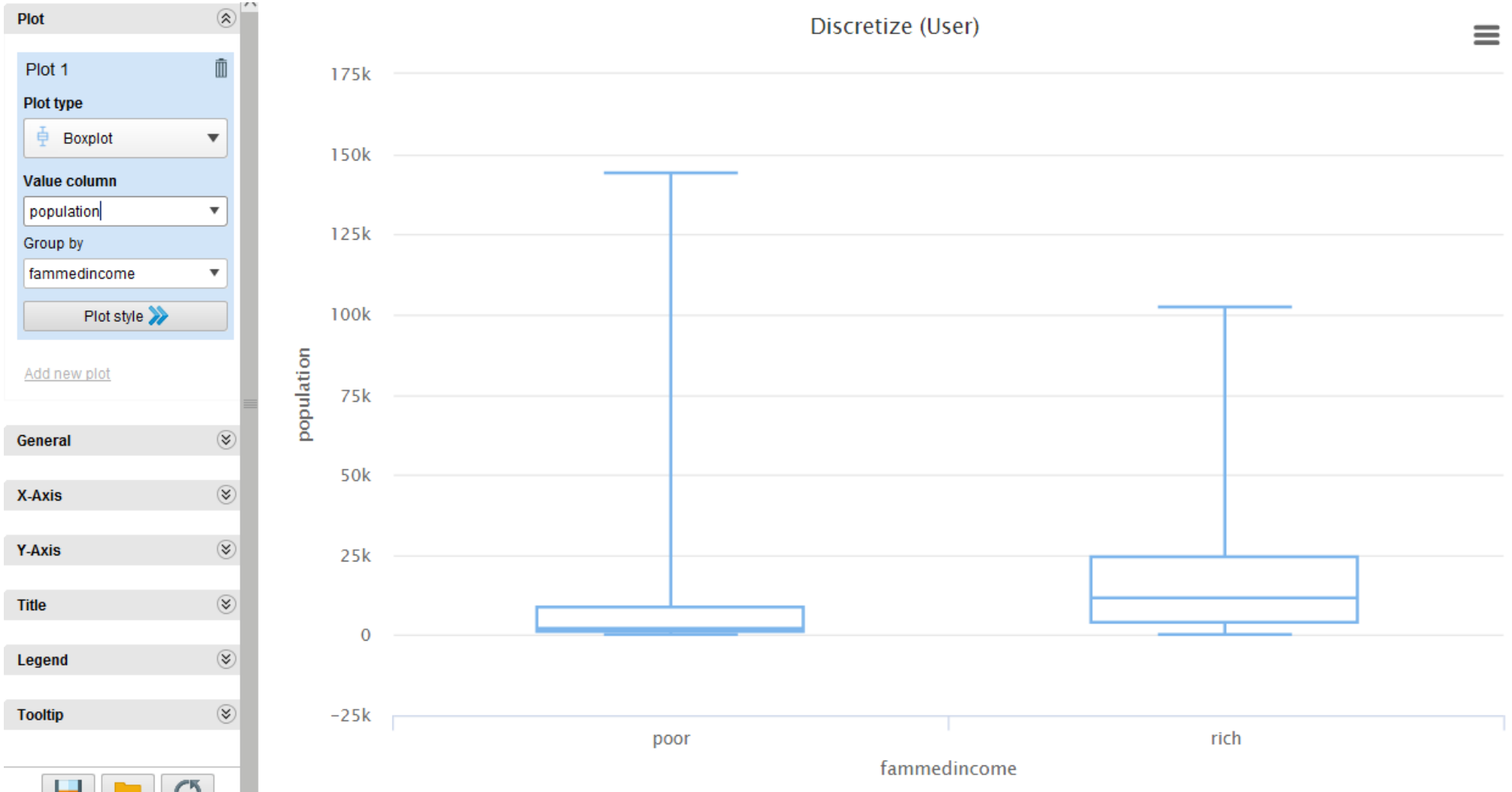
Correlation matrix

Scatterplot matrix



Multiple Variables (cont.)

- Comparing multiple quantitative variables (or comparing a single quantitative variable between groups defined by another categorical variable).
- **Tables with group-wise statistics or Boxplot**



Basic Descriptive Statistics and Data Visualization Cheat Sheet

Single Variable - Explore the distribution

	Statistics	Visualization
Categorical Variable	Counts	Bar chart
Quantitative Variable	5-number summary	Histogram

Two Variables – Compare and explore the relationship

	Statistics	Visualization
Categorical Variables	Contingency table (Cross tabulation)	Grouped bar chart
Quantitative Variables	Correlation	Scatter plot
Mixed Variables	Group-wise statistics (e.g., average)	Box plot Bar chart of group statistics

3+ Variables

Break it down into pairwise statistics or plots.
E.g., Correlation matrix, scatter plot matrix, box plot.

Data Cleaning and Preparation

Data Cleaning

- **Missing values?**
  Is this the result of reading the data?
  Are missing values correctly read in (or are there values like 99, 'N/A' or '.' as text)?
  Do we have to impute the missing values?

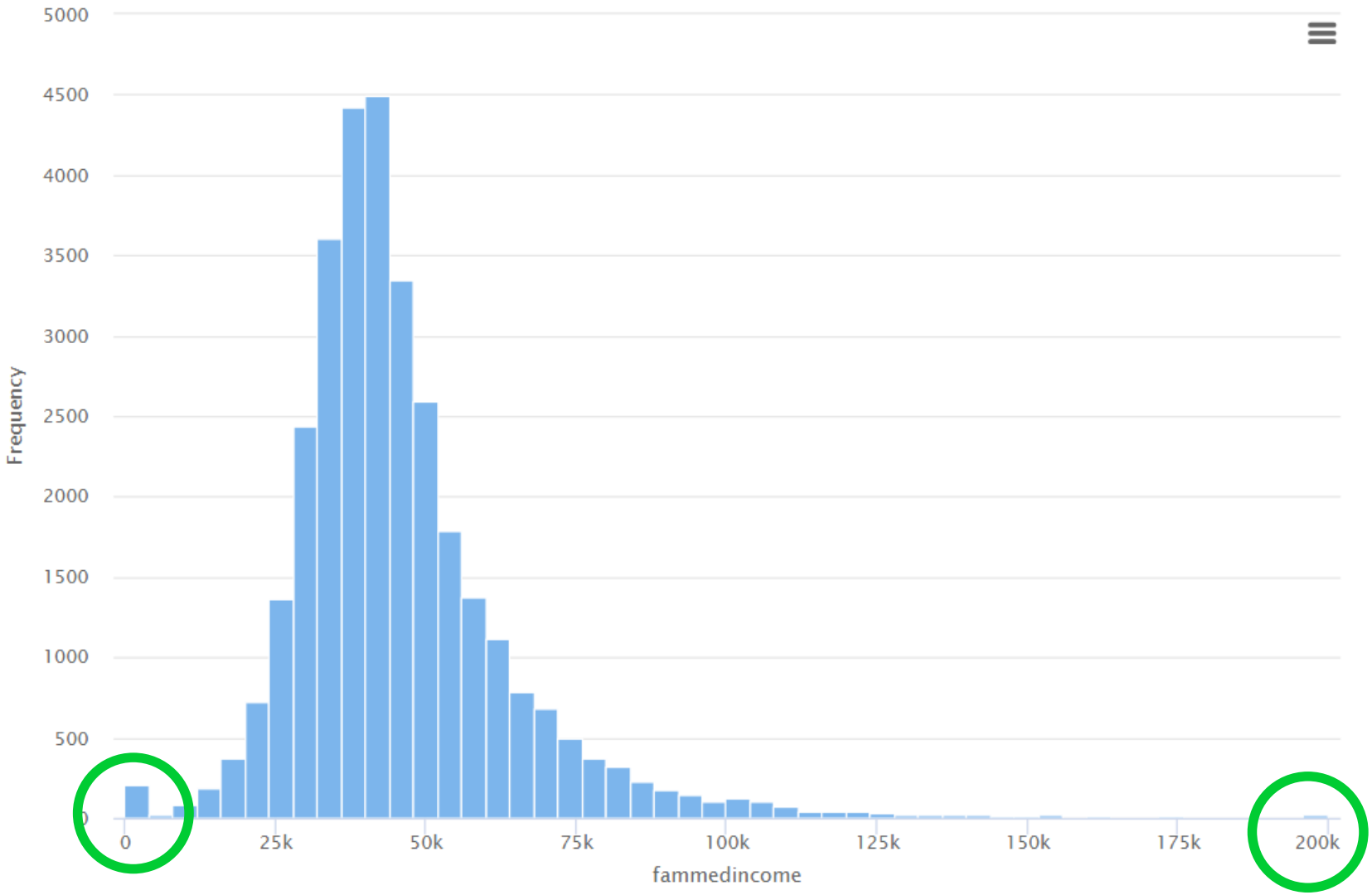
- **Outliers and strange values**
  Identify in histograms and scatter plots. Examples: many zeros, weird visual pattern visible. Might be the result of data collection. Needs investigation and cleaning!

- **Duplicates:** Are these a data problem?

- **Dates**: Make sure that these are read in correctly!

Example: Data Cleaning

Set a higher number of bins.
What do the spikes at 0 and 200,000 for median family income mean?
What should we do?



Data Transformation

- Some data needs to be transformed to be more useful for visualization of for a predictive model.

- **Observations**
  - Sampling/Filtering examples
  - Grouping and aggregation

RapidMiner
Operators in Blending and Cleansing

- **Features**
  - Feature Selection (select attributes)
  - Feature Generation (generate aggregation, etc.)
  - Nomalization to make features comparable (e.g, z-score)
  - Discretization (binning)

Observations

Sampling

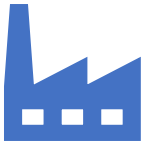
- **Population:** all items of interest (e.g., ZIP code areas in the US)
- **Sample:** a subset of the population. A selection of 500 ZIP codes.
- The purpose of sampling is to **obtain sufficient information to draw valid conclusions about the population.**
- In data science, we often need to sample to reduce the data size.

Grouping and Aggregation

- Many plots (e.g., bar charts) apply grouping and aggregation (e.g., counting the number of ZIP codes per state) automatically.
- Important for comparing groups.

Feature Selection

- Manually select/delete features using **expert knowledge.**
- Delete features of **low quality** (e.g., many missing values)
- Remove features that are **highly correlated** (we only need one)
- For predictive models: Find features that are highly "predictive." E.g., correlated with the variable to be predicted.



Feature Generation

Create better variables. For example:

- Calculate population density from population/area
- Calculate proportions or percentages for comparison. E.g., water to land area
- In a medical setting: Calculate the body mass index (BMI) from height and weight
- For predictive models: Square or multiply values to give larger values more impact.

Feature Normalization

- Make variables with a vastly different range comparable.
  - Normalize between 0 and 1
  - Z-score: Normalize to zero mean and 1 standard deviation
- **Example**: Compare age and income of a person.

Feature Discretization

- Transform a quantitative variable into a qualitative variable.
- **Example**: In a crime data set, change age from a number into a variable that indicates if the perpetrator is younger than 18 (subject to juvenile justice).