Essential Data Preparation, Descriptive Statistics and Visualizations

with Examples for rapidminer

Slides by Michael Hahsler
Purpose

1) Import data and "get used" to the data.
2) Clean the data (e.g., find missing values, outliers, mistakes)
3) "Make sure the data makes sense."
4) Find simple relationships between variables.
5) 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):
  o http://michael.hahsler.net/SMU/EMIS3309/data/rapidminer/Basic_Statistics_and_Visualizations.rmp
  o http://michael.hahsler.net/SMU/EMIS3309/data/rapidminer/Cleaning_and_preprocessing.rmp
Information can be measured on different scales. Depending on the scale, different operations/visualizations are appropriate.

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

Features/Attributes

<table>
<thead>
<tr>
<th>Observation</th>
<th>Features/Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Categorical</strong></td>
</tr>
<tr>
<td></td>
<td>(RM: polynominal)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Quantitative</strong></td>
</tr>
<tr>
<td></td>
<td>– ratio</td>
</tr>
<tr>
<td></td>
<td>(RM: integer or real)</td>
</tr>
</tbody>
</table>

should not be integer!
Single Variable - Quantitative

5-Number Summary:

- min
- 1st quartile
- Median
- Mean
- 3rd quartile
- max

Rapidminer gives you this for population per Zipcode:

<table>
<thead>
<tr>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integer</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>Min</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>143987</td>
</tr>
<tr>
<td>Average</td>
</tr>
<tr>
<td>8901.509</td>
</tr>
</tbody>
</table>

Histogram
to show the distribution
Single Variable - Categorical

Count table

- **Bar chart** for counts
- **Pie chart** (not ideal for more than a few groups)
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!
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?
Two Variables - Quantitative

Correlation

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

Scatterplot

- Is there a relationship?
- Multivariate Outliers

RapidMiner: Use Correlation Matrix node
Two Variables - Categorical

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

RapidMiner Aggregation Pivot

Grouped bar charts
Two Variables - Mixed

Compare 5-number statistic grouped by categorical variable.

<table>
<thead>
<tr>
<th>Row No.</th>
<th>state</th>
<th>minimum(population)</th>
<th>average(population)</th>
<th>median(population)</th>
<th>maximum(population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AK</td>
<td>10</td>
<td>2593.328</td>
<td>385</td>
<td>37294</td>
</tr>
<tr>
<td>2</td>
<td>AL</td>
<td>0</td>
<td>7125.199</td>
<td>4121</td>
<td>46541</td>
</tr>
<tr>
<td>3</td>
<td>AR</td>
<td>0</td>
<td>4523.746</td>
<td>1485</td>
<td>59252</td>
</tr>
<tr>
<td>4</td>
<td>AZ</td>
<td>32</td>
<td>14177.423</td>
<td>6226</td>
<td>71745</td>
</tr>
<tr>
<td>5</td>
<td>CA</td>
<td>0</td>
<td>20193.794</td>
<td>14077</td>
<td>105275</td>
</tr>
<tr>
<td>6</td>
<td>CO</td>
<td>0</td>
<td>8923.438</td>
<td>2120</td>
<td>68492</td>
</tr>
</tbody>
</table>

RapidMiner: Use Aggregate node

Bar chart for individual statistic to compare groups or box plot
Multiple Variables
Are usually broken down into pairwise comparisons.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>population</th>
<th>housingunits</th>
<th>landareamiles</th>
<th>fammedincome</th>
</tr>
</thead>
<tbody>
<tr>
<td>population</td>
<td>1</td>
<td>0.975</td>
<td>-0.063</td>
<td>0.220</td>
</tr>
<tr>
<td>housingunits</td>
<td>0.975</td>
<td>1</td>
<td>-0.057</td>
<td>0.223</td>
</tr>
<tr>
<td>landareamiles</td>
<td>-0.063</td>
<td>-0.057</td>
<td>1</td>
<td>-0.123</td>
</tr>
<tr>
<td>fammedincome</td>
<td>0.220</td>
<td>0.223</td>
<td>-0.123</td>
<td>1</td>
</tr>
</tbody>
</table>

RapidMiner: Use Correlation 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

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Statistics</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical Variable</td>
<td>Counts</td>
<td>Bar chart</td>
</tr>
<tr>
<td>Quantitative Variable</td>
<td>5-number summary</td>
<td>Histogram</td>
</tr>
</tbody>
</table>

## Two Variables - Explore the relationship

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Statistics</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical Variables</td>
<td>Contingency table (Cross tabulation)</td>
<td>Grouped bar chart</td>
</tr>
<tr>
<td>Quantitative Variables</td>
<td>Correlation</td>
<td>Scatter plot</td>
</tr>
<tr>
<td>Mixed Variables</td>
<td>Group-wise statistics (e.g., average)</td>
<td>Box plot</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bar chart of statistics</td>
</tr>
</tbody>
</table>

## 3+ Variables

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

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