Lecture 3:
Exploratory Analysis

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

2/23/2016

Overview

1. Exploratory Analysis

2. Transforming Data

Exploratory Analysis: Goals

- Organize and summarize raw data.
- Identify potential problems with your data set.
- Determine if the question you are asking can be answered by the data that you have.
- Prepare the foundation to answer your question.
  - Verify that your hypothesis is worth pursuing.
  - Assess assumptions on which statistical inference is based.
  - Select appropriate modeling tools and techniques.
- Potentially develop new hypotheses.

Exploratory Analysis: Checklist

- Formulate your question.
- Load the data.
- Clean the data (this will be on-going).
- Prod the data.
  - Number of rows and columns.
  - Data types of each column.
  - View top and bottom, look for format and reasonable values.
  - Check spread of numerical variables, and sets of categorical variables.
- Validate with external data source.
- Plot the data.
  - Univariate Distributions: histogram quantitative variables, bar chart of frequency for categorical variables.
  - Bivariate Graphs: Determine explanatory and response variables in your question and plot together to check for a signal of the expected relationship.
  - Check other combinations for potential new hypotheses (bias warning).
- Challenge your assumptions.
  - Do you have the right data to answer your question?
  - Do you need additional data?
  - Do you have the right question?

Exploratory Analysis: Bivariate Plots

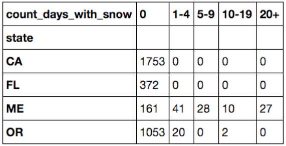
Response

Categorical

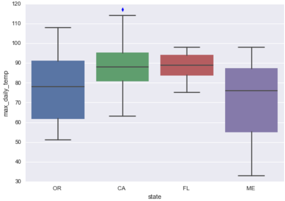
Numerical

Explanatory

Categorical

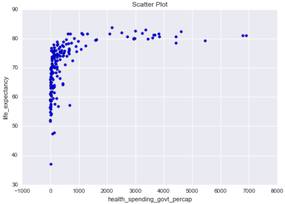


count_days_with_snow	0	1-4	5-9	10-19	20+
state					
CA	1753	0	0	0	0
FL	372	0	0	0	0
ME	161	41	28	10	27
OR	1053	20	0	2	0



Numerical

Bin the quantitative variable to create a categorical variable.



Exercise

Install packages

In terminal:

- pip install seaborn
- pip install statsmodels

IPython notebook: lesson3-common-functions-ipynb

Learn to identify common non-linear relationships.

- Change the parameters for each type of function.

Standardize, Normalize

Transformations to compare variables measured on different scales.

- Ordinal Scale

  Convert the values to a rank order.

- Interval scale

  Standardize to a mean of zero and standard deviation is one.

$$x_i^* = \frac{x_i - \bar{x}}{s_x}$$

- Ratio scale

  Normalize the variable vector, i.e. transform so vector length is equal to 1.

$$x_i^* = \frac{x_i}{\sqrt{\sum_i {x_i}^2}}$$

Normalizing, Rescaling

Other transformations to make disparate data comparable.

- Rescaling

  For some applications, you may want to change the data range, most likely to $[0, 1]$.

  $$x_i^* = \frac{x_i - min(x)}{max(x) - min(x)}$$

- Normalize (another sense of the word)

  If measurement samples are different, divide by the value of another variable.

Exercise

IPython Notebook: lesson3-transforming-data.ipynb

This notebook demonstrates some functionality in pandas using climate data from four metropolitan areas.

IPython Notebook: lesson3-exploratory-analysis.ipynb

Apply some of the principals of cleaning and exploring data discussed up to now.

References

Roger Peng & Elizabeth Matsui (2015)
The Art of Data Science, Leanpub

Recommended Reading

Art of Data Science, Chapter 4