CptS 475/575: Data Science

Exploratory Data Analysis
CptS 475/575: Data Science

Exploratory Data Analysis

Completed:
• Week 1: What is Data Science?
• Week 2: Intro to R
• Two tutorials on R
• Assignment 1
• Pre-course survey

This week:
• EDA (today)
• The Data Science Process (Friday)
• Assignment 2 (out today, due in a week)
• Tutorial 3 (Python) (today 3:10—4:30, SPRK 339)
Exploratory Data Analysis

*Exploratory Data Analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.*

John Tukey
(1915—2000)
EDA Approach

An approach for data analysis that employs a variety of techniques to

• maximize insight into a data set;
• uncover underlying structure;
• extract important variables;
• detect outliers and anomalies;
• test underlying assumptions;
• develop models;
Confirmatory vs Exploratory Data Analysis

- CDA: concerns itself with modeling and hypothesis
- EDA: there is no hypothesis, no model
- The exploratory aspect means that our understanding of the problem we are solving, or might solve, is changing as we go
Classical vs exploratory data analysis

For classical analysis, the sequence is

Problem => Data => Model => Analysis => Conclusions

For EDA, the sequence is

Problem => Data => Analysis => Model => Conclusions
Basic Tools of EDA

- Plots
- Graphs
- Summary statistics
Plots and graphs

- Plotting distributions of all variables
  - E.g. using box plots
- Plotting time series data
- Transforming variables
- Looking for pairwise relationships between variables using scatter plots
Scatter plot: example
Scatter plot: example
Summary statistics

- Mean, mode, median
- Minimum, maximum
- Standard deviation
- Quantiles
- Identifying outliers
EDA is also a mindset

• Not just a set of tools
• Goal is to understand the data – gain intuition, understand the shape of it, try to connect understanding of the process that generated the data to the data itself
• EDA happens between you and the data itself. It is not about proving anything to anyone else yet.
Philosophy of EDA

• Done even with data in the context of Internet companies
• Additional reasons why EDA is done with data that has been generated from logs
  • Helps with debugging and logging process.
  • E.g. “Patterns” observed in the data could actually be something wrong in the logging process that needs to be fixed
  • In the end, EDA helps you make sure the product is performing as intended

• EDA vs Visualization
  • EDA uses a lot of visualization, but the two are distinct
  • EDA done in the beginning, vis done towards the end
  • With EDA, the graphics are solely done for you to understand what is going on
  • With EDA, you can also use the understanding you get to inform and improve the development of algorithm
## Classical vs exploratory data analysis

<table>
<thead>
<tr>
<th></th>
<th>CDA</th>
<th>EDA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Models</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Focus</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Techniques</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rigor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Treatment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assumptions</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Classical vs exploratory data analysis

<table>
<thead>
<tr>
<th></th>
<th>CDA</th>
<th>EDA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Models</strong></td>
<td>Imposes</td>
<td>Does not impose</td>
</tr>
<tr>
<td><strong>Focus</strong></td>
<td>Model</td>
<td>Data</td>
</tr>
<tr>
<td><strong>Techniques</strong></td>
<td>Quantitative</td>
<td>Graphical</td>
</tr>
<tr>
<td><strong>Rigor</strong></td>
<td>Objective</td>
<td>Subjective</td>
</tr>
<tr>
<td><strong>Data Treatment</strong></td>
<td>Maps to few parameters</td>
<td>Uses all data</td>
</tr>
<tr>
<td><strong>Assumptions</strong></td>
<td>Made (and used to prove)</td>
<td>Not Made</td>
</tr>
</tbody>
</table>
Further readings (side bars)…

• Read about John Tukey:

• Read more about EDA:
  https://en.wikipedia.org/wiki/Exploratory_data_analysis

• Read about Bell Labs:
  https://en.wikipedia.org/wiki/Bell_Labs