Lecture 02:
First look at the data

Meelis Kull
meelis.kull@ut.ee
Autumn 2018
Introduction

What is data science?

10 success stories of data science

Data science in Estonia

Terminology: data mining, data science, ...

What can you learn in this course?

Organisational information about this course
The goals of current lecture

• Get a first understanding about how data science projects work
  – I will first present an example mini-project
  – I hope this example makes it easier to understand the standard CRISP-DM about how to run data mining projects

• Learn about first steps when trying to understand the data
Lecture 02

- Demo: Data science mini-project
- CRISP-DM: cross-industrial standard process for data mining
- Data understanding: Types of data
- Data understanding: First look at attributes
  - Types of attributes
  - First look at a nominal attribute
  - First look at a ordinal attribute
  - First look at a numeric attribute
• **Demo: Data science mini-project**

• CRISP-DM: cross-industrial standard process for data mining

• Data understanding: Types of data

• Data understanding: First look at attributes
  – Types of attributes
  – First look at a nominal attribute
  – First look at a ordinal attribute
  – First look at a numeric attribute
Warnings before the demo

• You will see many ugly details about file formats etc
  – This is intended! Similar problems are likely to happen in many data science projects
  – I hope it prepares a bit for when you do your own project

• I will show all the Python code of this mini-project and you are not expected to understand it all
  – It is ugly and not commented
  – I hope it shows you roughly how many lines of code are needed and where the complicated parts are
Demo: Data science mini-project
Data from clickers (1)

• Let’s analyse the data from previous lecture!
• What are my overall goals in this analysis?
  – **Demonstrate a data science mini-project**
  – Where are the problems in understanding?
  – Which questions attract less answers?
  – How many people stopped answering?
  – How many people started later?
  – Identify potential technical problems with clickers
Data from clickers (2)

• What data are available?
  – Public:
    • DS2018_lecture_01_intro_part_1.pdf
    • DS2018_lecture_01_intro_part_2.pdf
  – Private:
    • lecture_01_results_2018.xlsm
    • lecture_01_results_detail_2018.csv
  – No documentation with the data
Is it a good idea to use clickers in data science lectures?

A. Absolutely!
B. Probably good
C. Not sure yet
D. Probably bad
E. Bad idea
Is it a good idea to use clickers in data science lectures?

A. Absolutely!
B. Probably good
C. Not sure yet
D. Probably bad
E. Bad idea

![Bar chart showing the responses.]

Hard to programmatically extract information from PDF.
2. Please tell me about yourself (Multiple Choice)

<table>
<thead>
<tr>
<th>Responses</th>
<th>Percent</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st year master student</td>
<td>21.88%</td>
<td>14</td>
</tr>
<tr>
<td>2nd year master student</td>
<td>23.44%</td>
<td>15</td>
</tr>
<tr>
<td>PhD student</td>
<td>7.81%</td>
<td>5</td>
</tr>
<tr>
<td>Bachelor student</td>
<td>43.75%</td>
<td>28</td>
</tr>
<tr>
<td>Other</td>
<td>3.13%</td>
<td>2</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>100%</strong></td>
<td><strong>64</strong></td>
</tr>
</tbody>
</table>
2. Please tell me about yourself (Multiple Choice)

<table>
<thead>
<tr>
<th>Responses</th>
<th>Percent</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st year master student</td>
<td>21.88%</td>
<td>14</td>
</tr>
<tr>
<td>2nd year master student</td>
<td>23.44%</td>
<td>15</td>
</tr>
<tr>
<td>PhD student</td>
<td>7.81%</td>
<td>5</td>
</tr>
<tr>
<td>Bachelor student</td>
<td>43.75%</td>
<td>28</td>
</tr>
<tr>
<td>Other</td>
<td>3.13%</td>
<td>2</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>100%</strong></td>
<td><strong>64</strong></td>
</tr>
</tbody>
</table>

More information than in PDF: both percents and counts
Results by Participant

*Session does not contain standards*

<table>
<thead>
<tr>
<th>Name</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Id</td>
<td>-</td>
</tr>
</tbody>
</table>

**Responding Device**

396916

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have you used classroom answering systems in any courses yet?</td>
<td>A. Yes, the same clickers!</td>
</tr>
</tbody>
</table>
Results by Participant

Session does not contain standards

<table>
<thead>
<tr>
<th>Name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User Id</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responding Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>396916</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How many clickers were in the classroom?</td>
<td>Name clickers!</td>
</tr>
</tbody>
</table>

**Device ID – always 6-digit?**

**Decimal? Hexadecimal?**
### Results Detail

<table>
<thead>
<tr>
<th>Device ID</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Total Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer Key</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>396916</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>0.00</td>
<td>0.00%</td>
</tr>
<tr>
<td>39B884</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>6430F1</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>3D762E</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>3D766B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>J</td>
<td>A</td>
<td>-</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>643174</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>0.00</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
### Results Detail

<table>
<thead>
<tr>
<th>Device ID</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answer Key</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>396916</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>39B884</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>6430F1</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3D762E</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>3D766B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>643174</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Device ID</td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
<td>Q4</td>
<td>Q5</td>
<td>Q6</td>
</tr>
<tr>
<td>------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Answer Key</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>396916</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>39B88</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>6430F</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>3D762E</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D766B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>643174</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>C</td>
</tr>
</tbody>
</table>
‘Answer Key’ seems to specify which option was considered correct (or ‘-’ if none)
<table>
<thead>
<tr>
<th>Device ID</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer Key</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>396916</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>39B884</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>6430F1</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>3D762E</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td></td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>3D766B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td></td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>643174</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td></td>
<td>-</td>
<td>B</td>
</tr>
</tbody>
</table>

'-' seems to denote that the person did not answer this question.
From lecture_01_results_2018.xlsm

<table>
<thead>
<tr>
<th>Device ID</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer Key</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>

Actually Q4 is the same as Q5 (I looked up the question text and it was identical). It must be the failed attempt where I closed the poll immediately.
Green and red seem to stand for correct and wrong answer, white if none considered correct.
<table>
<thead>
<tr>
<th>Session Name: 05-09-2018, 21-40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date Created: 05.09.2018 16:09:10, Active Participants: 66 of 66</td>
</tr>
<tr>
<td>Average Score: 74,24%, Questions: 10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device ID, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Total Points, Score</td>
</tr>
<tr>
<td>Answer Key, -, -, -, -, A, -, -, -, -, 1.00, 100, 00%</td>
</tr>
<tr>
<td>396916, A, D, A, -, A, B, E, A, A, A, 0.00, 0, 00%</td>
</tr>
<tr>
<td>39B884, A, D, A, -, B, A, E, A, B, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>6430F1, A, B, B, -, A, A, E, A, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>3D762E, A, D, A, -, B, A, E, A, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>3D766B, A, B, B, -, A, A, E, J, A, -, 1.00, 100, 00%</td>
</tr>
<tr>
<td>643174, A, D, B, -, B, C, -, -, A, B, 0.00, 0, 00%</td>
</tr>
<tr>
<td>398C81, A, D, A, -, B, A, E, I, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>3D75F9, A, D, B, -, B, A, E, I, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>64310B, A, B, A, -, B, A, E, A, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>3D7681, C, B, B, -, C, A, E, F, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>643130, A, C, B, -, B, -, E, J, A, A, 0.00, 0, 00%</td>
</tr>
<tr>
<td>3D764A, A, A, C, -, B, A, E, G, A, A, 1.00, 100, 00%</td>
</tr>
<tr>
<td>3968AA, B, A, B, -, B, A, E, A, E, D, 1.00, 100, 00%</td>
</tr>
<tr>
<td>Device ID</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>396916</td>
</tr>
<tr>
<td>39B884</td>
</tr>
<tr>
<td>820F1</td>
</tr>
<tr>
<td>3D6D</td>
</tr>
<tr>
<td>3D71</td>
</tr>
<tr>
<td>3E1C</td>
</tr>
<tr>
<td>643130</td>
</tr>
<tr>
<td>3D764A</td>
</tr>
<tr>
<td>3968AA</td>
</tr>
<tr>
<td>Session Name: 05-09-2018, 21-40</td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td>Date Created: 05.09.2018 16:09:10</td>
</tr>
<tr>
<td>Active Participants: 66 of 66</td>
</tr>
<tr>
<td>Average Score: 74.24%</td>
</tr>
<tr>
<td>Questions: 10</td>
</tr>
</tbody>
</table>

### Results Detail

<table>
<thead>
<tr>
<th>Device ID</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Total Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>396916</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>0.00</td>
<td>0.00%</td>
</tr>
<tr>
<td>39B884</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>6430F1</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>3D762E</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>3D766B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>J</td>
<td>A</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>643174</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>C</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>643130</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>E</td>
<td>J</td>
<td>A</td>
<td>A</td>
<td>0.00</td>
<td>0.00%</td>
</tr>
<tr>
<td>3D764A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
<tr>
<td>3968AA</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>E</td>
<td>D</td>
<td>1.00</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Seems to denote the number of questions that were answered correctly by this person.
Due to locale the percentage 100.00% has been written as 100,00% and this confuses CSV readers, which interpret these as two different columns.
This extra comma is also confusing for some CSV readers.
Data from clickers (3)

• Let’s prepare the data for analysis:

In [1]:
```python
import pandas as pd

data = pd.read_csv('lecture_01_results_detail_2018.csv',
                   skiprows=[0,1,2,3,4,5,7],
                   skipfooter=1, engine='python', index_col=1)
data = data.iloc[:,[1,2,3,5,6,7,8,9,10]]
data.columns = ['Q'+str(i) for i in range(1,10)]
data.head()
```

Out[1]:

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>J</td>
<td>A</td>
<td>-</td>
</tr>
</tbody>
</table>
Let’s prepare the data for analysis:

```python
In [1]: import pandas as pd

data = pd.read_csv('lecture_01_results_detail_2018.csv', skiprows=[0,1,2,3,4,5,7],
                    skipfooter=1, engine='python', index_col=1)
data = data.loc[:,[1,2,3,5,6,7,8,9,10]]
data.columns = ['Q'+str(i) for i in range(1,10)]
data.
```

We will use the Pandas library a lot in this course for handling data matrices.
Data from clickers (3)

• Let’s prepare the data for analysis:

```python
In [1]: import pandas as pd

data = pd.read_csv('lecture_01_results_detail_2018.csv',skiprows=[0,1,2,3,4,5,7],
                      skipfooter=1,engine='python',index_col=1)

data = data.iloc[:,[1,2,3,5,6,7,8,9,10]]
data.columns = ['Q'+str(i) for i in range(1,10)]
data.head()
```

Removed Q4, as the poll was closed accidentally too early.
Data from clickers (3)

• Let’s prepare the data for analysis:

```python
In [1]: import pandas as pd

data = pd.read_csv('lecture_01_results_detail_2018.csv', skiprows=[0, 1, 2, 3, 4, 5, 7],
                   skipfooter=1, engine='python', index_col=1)
data = data.iloc[:, [1, 2, 3, 5, 6, 7, 8, 9, 10]]
data.columns = ['Q' + str(i) for i in range(1, 11)]
data.head()
```

Removed Q4, as the poll was closed accidentally too early.

Renumbering, now old Q5, Q6, … become Q4, Q5, … respectively.
• Let’s analyse the data!
• Goals revisited:
  – Where are the problems in understanding?
    • Proportion of correct answers?
  – Which questions attract less answers?
    • Number of answers per question
  – How many people stopped answering?
    • Clickers that are silent after some point?
  – How many people started later?
    • Clickers that do not answer the first question(s)?
  – Which questions attract more answers?
    • Distribution of answers per question
  – Identify potential technical problems with clickers
    • Distribution of answers per clicker
Number of answers?

• Per person? Per question?

```
In [2]:
n_answered_per_person = data.ne('−').sum(axis=1)
n_answered_per_person.value_counts()

Out[2]:
   9   50
   8   10
   7   3
   6   2
   5   1
dtype: int64

In [3]:
n_answered_per_question = data.ne('−').sum()
n_answered_per_question

Out[3]:
   Q1   63
   Q2   64
   Q3   64
   Q4   65
   Q5   63
   Q6   65
   Q7   63
   Q8   61
   Q9   60
dtype: int64
```
Number of answers?

- Per person? Per question?

In [2]:
```python
n_answered_per_person = data.ne('-').sum(axis=1)
n_answered_per_person.value_counts()
```

Out[2]:
```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
```
dtype: int64

In [3]:
```python
n_answered_per_question = data.ne('-').sum()
n_answered_per_question
```

Out[3]:
```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>63</td>
</tr>
<tr>
<td>Q2</td>
<td>64</td>
</tr>
<tr>
<td>Q3</td>
<td>64</td>
</tr>
<tr>
<td>Q4</td>
<td>65</td>
</tr>
<tr>
<td>Q5</td>
<td>63</td>
</tr>
<tr>
<td>Q6</td>
<td>65</td>
</tr>
<tr>
<td>Q7</td>
<td>63</td>
</tr>
<tr>
<td>Q8</td>
<td>61</td>
</tr>
<tr>
<td>Q9</td>
<td>60</td>
</tr>
</tbody>
</table>
```
dtype: int64

50 people answered all 9 questions
### Number of answers?

- **Per person? Per question?**

| In [2]: n_answered_per_person = data.ne('-').sum(axis=1) n_answered_per_person.value_counts() |
| Out[2]: |
| 9 | 50 |
| 8 | 10 |
| 7 | 3  |
| 6 | 2  |
| 5 | 1  |
| dtype: int64 |

50 people answered all 9 questions

| In [3]: n_answered_per_question = data.ne('-').sum() n_answered_per_question |
| Out[3]: |
| Q1 | 63 |
| Q2 | 64 |
| Q3 | 64 |
| Q4 | 65 |
| Q5 | 63 |
| Q6 | 65 |
| Q7 | 63 |
| Q8 | 61 |
| Q9 | 60 |
| dtype: int64 |

1 person answered only 5. Technical problems? Deliberately? Left early?
Number of answers?

- Per person? Per question?

```python
In [2]:
    n_answered_per_person = data.ne('¬').sum(axis=1)
    n_answered_per_person.value_counts()

Out[2]:
9    50
8    10
7    3
6    2
5    1
dtype: int64

In [3]:
    n_answered_per_question = data.ne('¬').sum()
    n_answered_per_question

Out[3]:
Q1   63
Q2   64
Q3   64
Q4   65
Q5   63
Q6   65
Q7   63
Q8   61
Q9   60
dtype: int64
```

Slightly smaller number of answers. A hard or confusing question?
Number of correct answers?

- Per person? Per question?

```
In [4]:
correct_q5 = data['Q5'].eq('A')*1
correct_q6 = data['Q6'].eq('E')*1
n_correct_per_person = correct_q5 + correct_q6
n_correct_per_person.value_counts()
```

```
Out[4]:
2   41
1   19
0    6
dtype: int64
```

```
In [5]:
n_correct_per_question = {'Q5':sum(correct_q5), 'Q6':sum(correct_q6),
                           'n_people':len(correct_q5)}
n_correct_per_question
```

```
Out[5]:
{'Q5': 49, 'Q6': 52, 'n_people': 66}
```
Number of correct answers?

- Per person? Per question?

In [4]:
correct_q5 = data['Q5'].eq('A')*1
correct_q6 = data['Q6'].eq('E')*1
n_correct_per_person = correct_q5 + correct_q6
n_correct_per_person.value_counts()

Out[4]:
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>41</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>
dtype: int64

41 people got both 2 answers correct

In [5]:
n_correct_per_question = {'Q5':sum(correct_q5), 'Q6':sum(correct_q6),
                          'n_people':len(correct_q5)}
n_correct_per_question

Out[5]:
{'Q5': 49, 'Q6': 52, 'n_people': 66}
Number of correct answers?

- Per person? Per question?

```python
In [4]:
correct_q5 = data[ 'Q5' ].eq('A')*1
correct_q6 = data[ 'Q6' ].eq('E')*1
n_correct_per_person = correct_q5 + correct_q6
n_correct_per_person.value_counts()

Out[4]:
2    41
1    19
0     6
dtype: int64

Out[5]:
{ 'Q5': 49, 'Q6': 52, 'n_people': 66}
```

49 out of 66 people answered Q5 correctly
First / last question answered

• Which question was the first that you answered? The last?

```
In [6]: import numpy as np
first_question_answered = data.ne('-').apply(lambda row:
    min(np.where(row)[0])
    ,axis=1)

first_question_answered.value_counts()
```

```
Out[6]: 0   63
1    2
4    1
dtype: int64
```

```
In [7]: last_question_answered = data.ne('-').apply(lambda row:
    max(np.where(row)[0])
    ,axis=1)

last_question_answered.value_counts()
```

```
Out[7]: 8   60
7    4
6    1
5    1
dtype: int64
```
First / last question answered

- Which question was the first that you answered? The last?

```
In [6]: import numpy as np
   first_question_answered = data.ne(' ').apply(lambda row:
   min(np.where(row)[0])
   ,axis=1)
   first_question_answered.value_counts()

Out[6]:
       0     63
       1      2
       4      1
dtype: int64

In [7]: last_question_answered = data.ne(' ').apply(lambda row:
   max(np.where(row)[0])
   ,axis=1)
   last_question_answered.value_counts()

Out[7]:
      8     60
      7      4
      6      1
      5      1
dtype: int64
```

63 people started answering immediately from Q1
First / last question answered

- Which question was the first that you answered? The last?

In [6]:
```python
import numpy as np
first_question_answered = data.ne('-').apply(lambda row:
    min(np.where(row)[0])
 ,axis=1)

first_question_answered.value_counts()
```

Out[6]:
```
0    63
1     2
4     1
dtype: int64
```

63 people started answering immediately from Q1

In [7]:
```python
last_question_answered = data.ne('X').apply(lambda row:
    max(np.where(row)[0])
 ,axis=1)

last_question_answered.value_counts()
```

Out[7]:
```
8    60
7     4
6     1
5     1
dtype: int64
```

One person started at Q5 – late arrival to the lecture?
First / last question answered

- Which question was the first that you answered? The last?

```python
In [6]:
import numpy as np
first_question_answered = data.ne(' ').apply(lambda row:
                  min(np.where(row)[0])
                  ,axis=1)
first_question_answered.value_counts()

Out[6]:
0   63
1    2
4    1
dtype: int64

In [7]:

Out[7]:
0    8
1    60
2    7
3    4
4    6
5    1
dtype: int64
```

60 people answered the last question Q9
First / last question answered

- Which question was the first that you answered? The last?

```python
In [6]:
import numpy as np
first_question_answered = data.ne('-').apply(lambda row:
    min(np.where(row)[0])
, axis=1)

first_question_answered.value_counts()
```

```
Out[6]:
0    63
1     2
4     1
dtype: int64
```

```
In [7]:

Out[7]:
6  8
7  4
6  1
5  1
dtype: int64
```

60 people answered the last question Q9

One stopped after Q6. Bored? Sleeping? Left?
Data from clickers (5)

- Conclusions
  - Too few questions with correct answers to evaluate understanding properly
  - Good participation in answering
  - No (obvious) technical problems other than one poll closed too early and then repolled
End of Demo
Steps of this data science mini-project

• Understanding the goals

• Understanding the data
  – Which files available? Format?
  – Meanings of fields? Errors?

• Preparing the data
  – Making the table of all answers

• Analysis of data

• Evaluation of results
Steps in data science projects

• It turns out that other data science projects have similar steps!
• Being aware of these steps helps to make the project successful
• This inspired development of the standard: CRISP-DM
Demo: Data science mini-project

- **CRISP-DM**: cross-industrial standard process for data mining
- Data understanding: Types of data
- Data understanding: First look at attributes
  - Types of attributes
  - First look at a nominal attribute
  - First look at a ordinal attribute
  - First look at a numeric attribute
CRISP-DM

• CRoss-Industrial Standard Process for Data Mining

CRISP-DM

• Published in 1999
• Many companies were involved
• Several data science project management tools support it (e.g. IBM SPSS Modeler)
• IBM is the primary corporation that currently embraces the CRISP-DM process model
• In online polls in 2002, 2004, 2007 and 2014 CRISP-DM was the leading methodology used by industry data miners
• In 2015 IBM published ASUM-DM, an advanced process based on CRISP-DM
Why to use CRISP-DM?

CRISP-DM – a Standard Methodology to Ensure a Good Outcome
Posted by William Vorhies on July 26, 2016 at 9:15am

Summary: To ensure quality in your data science group, make sure you’re enforcing a standard methodology. This includes not only traditional data analytic projects but also our most advanced recommenders, text, image, and language processing, deep learning, and AI projects.

About the author: Bill Vorhies is Editorial Director for Data Science Central and has practiced as a data scientist and commercial predictive modeler since 2001.

Source:
CRISP-DM is data mining – what about data science?

• Data science projects have sometimes less specific goals
  – “Here are the data, make money out of it”

• However, thinking in terms of CRISP-DM can still be beneficial
CRISP-DM

- CRoss-Industrial Standard Process for Data Mining

CRISP-DM: Business understanding

• Subtasks:
  – Understand what you want to accomplish
  – Assess the situation (uncover important factors that could influence the outcome of the project)
  – Translate the business goal in a data mining objective
  – Develop a project plan

• In our mini-project to analyse clicker data:
  – List of goals
CRISP-DM: Data understanding

- Subtasks:
  - Collect initial data
  - Describe the data
  - Explore the data
  - Verify data quality

- In our mini-project:
  - Extracted the files
  - Studied the contents of files
  - Found the problems in formatting
  - Found a column with useless data
CRISP-DM: Data preparation

• Subtasks:
  – Select the data
  – Clean the data
  – Construct new data
  – Integrate the data
  – Format the data

• In our mini-project:
  – Choose two files to be used
  – Extract the matrix of answers
CRISP-DM: Modelling

• Subtasks:
  – Select modelling techniques
  – Generate a test design
  – Build the models
  – Assess the models

• In our mini-project:
  – Simple counting
  – No formal modelling
CRISP-DM: Evaluation

• Subtasks:
  – Evaluate the results
  – Review the process
  – Determine the next steps

• In our mini-project:
  – Conclusions from the analysis results
  – Decision to continue using clickers
CRISP-DM: Deployment

• Subtasks:
  – Plan for deployment
  – Plan monitoring and maintenance
  – Produce a final report
  – Conduct a final project review

• In our mini-project:
  – Plan future use of clickers
CRISP-DM

• CRoss-Industrial Standard Process for Data Mining

Which CRISP-DM step comes first?

A. Business understanding
B. Data preparation
C. Data understanding
D. Deployment
E. Evaluation
F. Modelling
Which CRISP-DM step comes 2nd?

A. Business understanding
B. Data preparation
C. Data understanding
D. Deployment
E. Evaluation
F. Modelling

![Bar chart showing the CRISP-DM steps ranked]

- Business understanding: 1
- Data preparation: 8
- Data understanding: 29
- Deployment: 1
- Evaluation: 0
- Modelling: 0
Which CRISP-DM step comes 3rd?

A. Business understanding
B. Data preparation
C. Data understanding
D. Deployment
E. Evaluation
F. Modelling

The correct answer is B. Data preparation.
Which CRISP-DM step comes 4th?

A. Business understanding
B. Data preparation
C. Data understanding
D. Deployment
E. Evaluation
F. Modelling

![Bar chart showing the number of times each step was chosen for the 4th position.]

- Business understanding: 1
- Data preparation: 0
- Data understanding: 0
- Deployment: 1
- Evaluation: 5
- Modelling: 34
Which CRISP-DM step comes 5th?

A. Business understanding
B. Data preparation
C. Data understanding
D. Deployment
E. Evaluation
F. Modelling

The correct answer is E. Evaluation.
Which CRISP-DM step comes 6th?

A. Business understanding
B. Data preparation
C. Data understanding
D. Deployment
E. Evaluation
F. Modelling

![Bar chart showing the distribution of CRISP-DM steps]

- Business understanding: 39
- Data preparation: 2
- Deployment: 0
- Evaluation: 0
- Modelling: 2
Structure of the course

- The following lectures are going to dive into particular steps of CRISP-DM in detail
- Business Understanding is very much domain-dependent
- Therefore we start with Data Understanding
Lecture 02

- Demo: Data science mini-project
- CRISP-DM: cross-industrial standard process for data mining

- **Data understanding**: Types of data
- Data understanding: First look at attributes
  - Types of attributes
  - First look at a nominal attribute
  - First look at a ordinal attribute
  - First look at a numeric attribute