Lecture 02: First look at the data

Meelis Kull

meelis.kull@ut.ee

Autumn 2019
✓ 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
Lecture 02

- Types of data
- Types of attributes
- First look at a nominal attribute
- First look at an ordinal attribute
- First look at a numeric attribute
- Distribution of attributes
- Types of histograms
- How to describe probability distributions?
- Some standard probability distributions
- More ways to visualise distributions
- Visualising relations of attributes
Lecture 02

• **Types of data**
  • Types of attributes
  • First look at a nominal attribute
  • First look at an ordinal attribute
  • First look at a numeric attribute

• Distribution of attributes
• Types of histograms
• How to describe probability distributions?
• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
Used materials

• Next 4 slides have been adapted from the course:
  – UIUC CS412: An Introduction to Data Warehousing and Data Mining (Fall 2016) https://wiki.illinois.edu/wiki/display/cs412 taught by Jiawei Han, author of a well-known data mining textbook
Types of data: Records

- Relational records:

<table>
<thead>
<tr>
<th>Person</th>
<th>Surname</th>
<th>First_Name</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Miller</td>
<td>Paul</td>
<td>London</td>
</tr>
<tr>
<td>1</td>
<td>Ortega</td>
<td>Alvaro</td>
<td>Valencia</td>
</tr>
<tr>
<td>2</td>
<td>Huber</td>
<td>Urs</td>
<td>Zurich</td>
</tr>
<tr>
<td>3</td>
<td>Blanc</td>
<td>Gaston</td>
<td>Paris</td>
</tr>
<tr>
<td>4</td>
<td>Bertolini</td>
<td>Fabrizio</td>
<td>Rom</td>
</tr>
</tbody>
</table>

- Data matrix (crosstab):

![Data matrix example]

- Transactions:

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

- Word frequencies in documents:

<table>
<thead>
<tr>
<th>Document</th>
<th>Team</th>
<th>Coach</th>
<th>V</th>
<th>Pla</th>
<th>Ball</th>
<th>Score</th>
<th>Game</th>
<th>U</th>
<th>W</th>
<th>Last</th>
<th>Timeout</th>
<th>Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Types of data: Graphs and networks

- Transportation network:
- Molecular structures:
- World Wide Web:
- Social networks
Types of data: Sequential data

- Temporal data (time-series):

- Video data (sequence of frames):

- Genetic sequence data:
  
  - Human
  - Chimpanzee
  - Macaque

- Transaction sequences (logs of an information system)
Types of data: Spatial and multimedia data

- **Spatial data (maps):**

- **Images:**

- **Video:**

- **Audio data:**
Data objects and attributes

- Data sets are made up of data objects
- A **data object** represents an entity
- Examples:
  - sales database: *customers, store items, sales*
  - medical database: *patients, treatments*
  - university database: *students, professors, courses*
- Data objects are also called: *samples, examples, instances, data points, objects, tuples*
- Data objects are described by **attributes** also called: *properties, features*
- In databases typically:
  - rows → data objects; columns → attributes
Types of data

- **Types of attributes**
- First look at a nominal attribute
- First look at an ordinal attribute
- First look at a numeric attribute
- Distribution of attributes
- Types of histograms
- How to describe probability distributions?
- Some standard probability distributions
- More ways to visualise distributions
- Visualising relations of attributes
Attributes (or dimensions, features, variables)

• A data field, representing a characteristic or feature of a data object  
  – *E.g.*, *customer_ID*, *name*, *address*

• Types of attributes:
  – Qualitative (categorical)
    • Nominal
    • Ordinal
  – Quantitative (numeric)
    • Interval
    • Ratio
## Attribute types

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>=, ≠</td>
</tr>
<tr>
<td>Ordinal</td>
<td>=, ≠, &lt;, &gt;</td>
</tr>
<tr>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>=, ≠, &lt;, &gt;, +, −</td>
</tr>
<tr>
<td>Ratio</td>
<td>=, ≠, &lt;, &gt;, +, −, ·, /</td>
</tr>
</tbody>
</table>
How to determine attribute type?

• Take two arbitrary values
  • If you can meaningfully divide one by the other:  
    – Ratio
  • If you can meaningfully subtract one from the other:  
    – Interval
• If you can compare them (for example, one before the other or bigger than the other):  
  – Ordinal
• Otherwise:  
  – Nominal

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>=, ≠</td>
</tr>
<tr>
<td>Ordinal</td>
<td>=, ≠, &lt;, &gt;</td>
</tr>
<tr>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>=, ≠, &lt;, &gt;, +, −</td>
</tr>
<tr>
<td>Ratio</td>
<td>=, ≠, &lt;, &gt;, +, −, /</td>
</tr>
</tbody>
</table>
What is the type of: Name of a person

- A. Nominal
- B. Ordinal
- C. Interval
- D. Ratio

![Response Counter Diagram](image)

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>$\equiv, \neq$</td>
</tr>
<tr>
<td>Ordinal</td>
<td>$\equiv, \neq, &lt;, &gt;$</td>
</tr>
<tr>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>$\equiv, \neq, &lt;, &gt;, +, -$</td>
</tr>
<tr>
<td>Ratio</td>
<td>$\equiv, \neq, &lt;, &gt;, +, -, /$</td>
</tr>
</tbody>
</table>
Name of a person

• Not quantitative (not ratio, not interval)
  – Cannot divide nor subtract

• Ordinal?
  – Could be considered ordinal with alphabetic order

• Nominal
  – Better to consider nominal, as not very sensible to say “my name is bigger than yours”

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>=, ≠</td>
</tr>
<tr>
<td>Ordinal</td>
<td>=, ≠, &lt;, &gt;</td>
</tr>
<tr>
<td>Numeric</td>
<td>=, ≠, &lt;, &gt;, +, −</td>
</tr>
<tr>
<td>Interval</td>
<td></td>
</tr>
<tr>
<td>Ratio</td>
<td>=, ≠, &lt;, &gt;, +, −, /</td>
</tr>
</tbody>
</table>
What is the type of: Phone number

A. Nominal  ✔
B. Ordinal
C. Interval
D. Ratio
Phone number

• Not quantitative (not ratio, not interval)
  – Not meaningful to divide or subtract

• Ordinal?
  – Could be considered ordinal with standard numeric order, but almost nobody ever needs to sort the phone numbers numerically

• Nominal
  – Better to consider nominal
What is the type of: Salary

A. Nominal
B. Ordinal
C. Interval
D. Ratio

Correct answer: D. Ratio

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>$=$, $\neq$</td>
</tr>
<tr>
<td>Ordinal</td>
<td>$=$, $\neq$, $&lt;$, $&gt;$</td>
</tr>
<tr>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>$=$, $\neq$, $&lt;$, $&gt;$, $+$, $-$</td>
</tr>
<tr>
<td>Ratio</td>
<td>$=$, $\neq$, $&lt;$, $&gt;$, $+$, $-$, $/$</td>
</tr>
</tbody>
</table>
What is the type of:
Date of birth

A. Nominal
B. Ordinal
C. Interval
D. Ratio

Response Counter
What is the type of:
Your final grade (A-F)

A. Nominal

B. Ordinal

☑️ C. Interval

D. Ratio

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>(\equiv, \neq)</td>
</tr>
<tr>
<td>Ordinal</td>
<td>(\equiv, \neq, &lt;, &gt;)</td>
</tr>
<tr>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>(\equiv, \neq, &lt;, &gt;, +, -)</td>
</tr>
<tr>
<td>Ratio</td>
<td>(\equiv, \neq, &lt;, &gt;, +, -, \div)</td>
</tr>
</tbody>
</table>
What is the type of: Temperature

A. Nominal  
B. Ordinal  
C. Interval  
D. Ratio

- A. Nominal
- B. Ordinal
- C. Interval
- D. Ratio

<table>
<thead>
<tr>
<th>Attribute type</th>
<th>Supported operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>=, ≠</td>
</tr>
<tr>
<td>Ordinal</td>
<td>=, ≠, &lt;, &gt;</td>
</tr>
<tr>
<td>Interval</td>
<td>=, ≠, &lt;, &gt;, +, −, /</td>
</tr>
<tr>
<td>Ratio</td>
<td>=, ≠, &lt;, &gt;, +, −, /</td>
</tr>
</tbody>
</table>

Response Counter
## Attribute types

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Description</th>
<th>Examples</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. (=, ≠)</td>
<td>zip codes, employee ID numbers, eye color, gender</td>
<td>mode, entropy, contingency correlation, $\chi^2$ test</td>
</tr>
<tr>
<td>Ordinal</td>
<td>The values of an ordinal attribute provide enough information to order objects. (&lt;, &gt;)</td>
<td>hardness of minerals, {good, better, best}, grades, street numbers</td>
<td>median, percentiles, rank correlation, run tests, sign tests</td>
</tr>
<tr>
<td>Interval</td>
<td>For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, −)</td>
<td>calendar dates, temperature in Celsius or Fahrenheit</td>
<td>mean, standard deviation, Pearson’s correlation, $t$ and $F$ tests</td>
</tr>
<tr>
<td>Numeric (Quantitative)</td>
<td>For ratio variables, both differences and ratios are meaningful. (∗, /)</td>
<td>temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current</td>
<td>geometric mean, harmonic mean, percent variation</td>
</tr>
</tbody>
</table>

**Standard operators**

=, ≠, <, >, +, −, ∗, /
Type in database vs actual attribute type

- If something is stored in the database as numeric, does not imply that it’s true type is quantitative
  - Personal identity number in Estonia is categorical, not numeric
  - Percentage stored as string “58%” is numeric
Lecture 02

✓ Types of data
✓ Types of attributes
  • **First look at a nominal attribute**
  • First look at an ordinal attribute
  • First look at a numeric attribute
  • Distribution of attributes
  • Types of histograms
  • How to describe probability distributions?
  • Some standard probability distributions
  • More ways to visualise distributions
  • Visualising relations of attributes
Example of nominal attributes

- Data from clickers:
Example of nominal attributes

- Data from clickers:

```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()
```

<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>396916</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>39B884</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>6430F1</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>3D762E</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>3D766B</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>
Example of nominal attributes

- Data from clickers:

```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
Example of nominal attributes

- Data from clickers:

```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()
```

Removed Q4, as the poll was closed accidentally too early.
Example of nominal attributes

- Data from clickers:

```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
First look at a nominal attribute

- **Print all values** (or first 10 / 100 / 1000 / …)

```python
In [8]: data['Q1'].values
```
First look at a nominal attribute

- Print all values (or first 10 / 100 / 1000 / …)
  - Can give useful insights
  - In this example we see that:
    - ”A” dominates
    - Values “-” are probably special
    - Possible values seem to be “-”, “A”, “B”, “C”

```
In [8]: data['Q1'].values
```
First look at a nominal attribute

• Print all values (or first 10 / 100 / 1000 / …)

• **Count occurrences** of each possible value (textual histogram)
  
  – More information than from just printing:
    
    • The least frequent options are B and C
    • Exact count of A is 36
    • 3 missing values ("-")

```
In [9]: data['Q1'].value_counts()
Out[9]:
A    36
C    16
B    11
-     3
Name: Q1, dtype: int64
```
First look at a nominal attribute

• Print all values (or first 10 / 100 / 1000 / …)
• Count occurrences of each possible value (textual histogram)

• Visual histogram
  – Sometimes reveals more than textual histogram:
    • Magnitudes are easier to notice visually than textually

```
In [10]: import matplotlib.pyplot as plt
data['Q1'].value_counts().plot(kind='bar')
plt.show()
```
First look at a nominal attribute

- Print all values (or first 10 / 100 / 1000 / ...)
- Count occurrences of each possible value (textual histogram)
- Visual histogram
- “Histogram of the histogram”
  - Counts how many values with 1 occurrence, how many with 2 occurrences, etc
  - Can be used if the histogram is too big
    - E.g., first name of all people in Estonia – too many different names
    - Instead count how many names used once, twice, ...

```python
In [11]: data['Q1'].value_counts().value_counts()
Out[11]:
            3    1
            36   1
            11   1
            16   1
Name: Q1, dtype: int64
```
First look at a nominal attribute

• Print all values (or first 10 / 100 / 1000 / …)
• Count occurrences of each possible value (textual histogram)
• Visual histogram
• “Histogram of the histogram”
  – Counts how many values with 1 occurrence, how many with 2 occurrences, etc
  – Can be used if the histogram is too big

```
In [12]: data['Q2'].value_counts()
Out[12]:
D    28
B    15
A    14
C     5
    2
E    2
Name: Q2, dtype: int64
```
```
In [13]: data['Q2'].value_counts().value_counts()
Out[13]:
2    2
15    1
14    1
5     1
28    1
Name: Q2, dtype: int64
```
First look at a nominal attribute

• Print all values (or first 10 / 100 / 1000 / …)
• Count occurrences of each possible value (textual histogram)
• Visual histogram
• “Histogram of the histogram”
• Find the most frequent value (mode)
  – Needed again if the histogram is too big

In [14]: data['Q1'].mode()
Out[14]:
0       A
    dtype: object

In [15]: data['Q1'].value_counts().nlargest(1)
Out[15]:
A    36
Name: Q1, dtype: int64
First look at a nominal attribute

- Print all values (or first 10 / 100 / 1000 / …)
- Count occurrences of each possible value (textual histogram)
- Visual histogram
- “Histogram of the histogram”
- Find the most frequent value (mode)
  – Needed again if the histogram is too big

```python
In [14]: data['Q1'].mode()
Out[14]:
0    A
dtype: object
```

```python
In [16]: data['Q1'].value_counts().nlargest()
Out[16]:
   A  36
   C  16
   B  11
   -  3
Name: Q1, dtype: int64
```
First look at a nominal attribute

- Print all values (or first 10 / 100 / 1000 / …)
- Count occurrences of each possible value (textual histogram)
- Visual histogram
- “Histogram of the histogram”
- Find the most frequent value (mode)
- More options depending on the goal, e.g.:
  - How many values occur only once?
  - How many values occur at least 10 times?
Data quality in nominal attributes

• Are there any obvious typos?
  – E.g. “Thusday” vs “Thursday”

• Does the case matter?
  – E.g. “A” vs “a”

• Values denoting missingness?
  – “N/A”, “none”, “-”, “”, other possibilities
  – If multiple denoters used, is meaning different?

• Values that do not follow the pattern?
  – E.g., clicker ID with more than 6 characters?

• If documentation specifies all possible values:
  – Check if all values are present in the data
  – If not, then include empty bars in histograms for these values
Nominal attributes in Pandas

• The standard Pandas data type for categorical attributes is **categorical**

• Converting strings into a nominal attribute:

```
In [17]: data['Q1'].astype('category').head()
Out[17]:
396916   A
39B884   A
6430F1   A
3D762E   A
3D766B   A
Name: Q1, dtype: category
Categories (4, object): [-, A, B, C]
```
Nominal attributes in Pandas

• The standard Pandas data type for categorical attributes is **categorical**

• Converting strings into a nominal attribute:

```python
In [18]:
grades = pd.Series(['A','B','E','B'])
grades = grades.astype('category',categories=['A','B','C','D','E'])
grades

Out[18]:
0    A
1    B
2    E
3    B
dtype: category
Categories (5, object): [A, B, C, D, E]
```
Lecture 02

✓ Types of data
✓ Types of attributes
✓ First look at a nominal attribute
  • **First look at an ordinal attribute**
  • First look at a numeric attribute
  • Distribution of attributes
  • Types of histograms
  • How to describe probability distributions?
  • Some standard probability distributions
  • More ways to visualise distributions
  • Visualising relations of attributes
First look at an ordinal attribute

- All the same applies as for nominal attributes
- Need to make sure that the order is retained in histograms
- Additional ways to look at the attribute:
  - Calculate the min, max of the attribute

```
In [19]:
grades = pd.Series(['A','B','E','B'])
grades = grades.astype('category',
                      ordered=True,
                      categories=['F','E','D','C','B','A'])
grades
```

```
Out[19]:
0    A
1    B
2    E
3    B
dtype: category
Categories (6, object): [F < E < D < C < B < A]
```
First look at an ordinal attribute

- All the same applies as for nominal attributes
- Need to make sure that the order is retained in histograms
- Additional ways to look at the attribute:
  - Calculate the min, max of the attribute

```python
In [21]: grades.min()
Out[21]: 'E'

In [22]: grades.max()
Out[22]: 'A'
```
First look at an ordinal attribute

• All the same applies as for nominal attributes
• Need to make sure that the order is retained in histograms
• Additional ways to look at the attribute:
  – Calculate the min, max of the attribute
  – Also can calculate median
    • Not implemented in Pandas for ordinal attributes
First look at an ordinal attribute

• All the same applies as for nominal attributes
• Need to make sure that the order is retained in histograms
• Additional ways to look at the attribute:
  – Calculate the min, max of the attribute
  – Also can calculate median
    • Not implemented in Pandas for ordinal attributes
  – More generally, calculate quantiles
    • Again not implemented in Pandas for ordinal attributes
Quartiles, percentiles, quantiles

• There are 3 quartiles:
  – lower quartile, median, upper quartile

• They divide a sorted data set into 4 equal parts

• Percentiles divide into 100 equal parts
  – Lower quartile is the 25\textsuperscript{th} percentile
  – Median is the 50\textsuperscript{th} percentile

• Deciles divide into 10 equal parts

• Quantiles generalise to any location over sorted data:
  – Lower quartile = 25\textsuperscript{th} percentile = quantile at 0.25
  – 4\textsuperscript{th} decile = quantile at 0.4
3\textsuperscript{rd} quartile is between ...

A. 1\textsuperscript{st} and 2\textsuperscript{nd} decile
B. 2\textsuperscript{nd} and 3\textsuperscript{rd} decile
C. 3\textsuperscript{rd} and 4\textsuperscript{th} decile
D. 4\textsuperscript{th} and 5\textsuperscript{th} decile
E. 5\textsuperscript{th} and 6\textsuperscript{th} decile
F. 6\textsuperscript{th} and 7\textsuperscript{th} decile
G. 7\textsuperscript{th} and 8\textsuperscript{th} decile
H. 8\textsuperscript{th} and 9\textsuperscript{th} decile

\checkmark G. 7\textsuperscript{th} and 8\textsuperscript{th} decile
Quantiles can be calculated for ... 

A. Nominal attributes 
B. Ordinal attributes 
C. Both 
D. Neither 
E. Not sure
Data quality in ordinal attributes

• Same aspects as in nominal attributes

• Additionally:
  – Is the order specified in the documentation? *(meta-data)*
  – If all possible values specified in the meta-data:
    • Check if all values present in the data
    • If not, make sure that the empty bars in histograms are in the right location corresponding to the order
Lecture 02

✓ Types of data
✓ Types of attributes
✓ First look at a nominal attribute
✓ First look at an ordinal attribute
• First look at a numeric attribute
• Distribution of attributes
• Types of histograms
• How to describe probability distributions?
• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
First look at a numeric attribute

• Is it really numeric?
  – E.g. if contains only values 0.0 and 1.0 then perhaps nominal is more appropriate?

• Is it **discrete** (only integers) or **continuous**?
  – E.g. if contains integers 0 to 100 then can have a first look as if it were an ordinal attribute

• Calculate quantiles as for ordinal attributes

• Calculate the mean (arithmetic average)

**Note:** the term discrete in mathematics refers to any countable sets, that is sets which can be indexed by integers – therefore usually in data science, attribute is called discrete if it is either categorical or integer numeric
Describing a numeric attribute

- Mean and other statistics of a numeric attribute:

```python
In [23]: np.random.seed(0)
   x = pd.Series(np.random.rand(20))
   x.head()

Out[23]:
0    0.548814
1    0.715189
2    0.602763
3    0.544883
4    0.423655
dtype: float64

In [24]: x.mean()

Out[24]: 0.5815548245225973

In [25]: x.describe()

Out[25]:
       count   20.000000
      mean   0.581555
       std   0.283083
       min   0.020218
      25%   0.434104
      50%   0.585404
      75%   0.801949
      max   0.963663
dtype: float64
```
First look at a numeric attribute

- Is it really numeric?
- Is it discrete (only integers)?
- Calculate quantiles as for ordinal attributes
- Calculate the mean (arithmetic average)
- Plot a histogram (with default binning)

```
In [26]: x.hist(bins=[i*0.1 for i in range(11)])
plt.show()

In [27]: x.hist(bins=10)

In [28]: x.hist()
```
Data quality in numeric attributes

• Check if some value occurs many times
  – With real numbers often no number occurs more than once
  – If a value is frequent then:
    • Possibly can denote missingness (e.g. value 0), should be treated as N/A (not available, missing)
    • Perhaps can denote an extremal value (more extremal values represented as this value)

• Check for rounding
  – E.g., if all numbers rounded to 2\textsuperscript{nd} digit and one has more digits then it might be a typing error
The mean can be calculated for ...

A. Nominal attributes
B. Ordinal attributes
C. Numeric attributes
D. All of the above
E. None of the above
F. Not sure
Data understanding: Summary after first look at attributes
Summary after first look at attributes

• What to do during the first look:
  – Determine the types of attributes
  – Have a first look at each of the attributes separately
Lecture 02

✓ Types of data
✓ Types of attributes
✓ First look at a nominal attribute
✓ First look at an ordinal attribute
✓ First look at a numeric attribute

• Distribution of attributes
• Types of histograms
• How to describe probability distributions?
• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
You will work on the same dataset in the homework, with minor differences
Data understanding: Exploration

• We know how to make the first look at each attribute separately

• A key question we can now answer: How do the items in this dataset look like?

• Now we have at least 3 answers to this:
How do the items in this dataset look like?

Let us just pick up one as an example:
- Age = 22
- Workclass = Private
- Education = 11th
- Occupation = Other-service
- Capital.gain = 0
- ...
• How do the items in this dataset look like?

• Let us just provide the ranges of attributes
  – Age: \{17,18,19,\ldots,90\}
  – Workclass: \{Federal-gov, Local-gov, Private, \ldots\}
  – Education: \{1^{st}-4^{th},5^{th}-6^{th}, \ldots,\text{Doctorate}\}
  – Occupation: \{Adm-clerical, Exec-managerial,\ldots\}
  – Capital.gain: [0,99999]
  – ...

In the homework, your dataset might have slightly different ranges, but that is intentional
Answer 3

• How do the items in this dataset look like?

• **Let us just provide the histograms**
  
  – Age: <histogram>
  – Workclass: <histogram>
  – Education: <histogram>
  – Occupation: <histogram>
  – Capital.gain: <histogram>
  – ...
Lecture 02

✓ Types of data
✓ Types of attributes
✓ First look at a nominal attribute
✓ First look at an ordinal attribute
✓ First look at a numeric attribute
✓ Distribution of attributes

• Types of histograms
• How to describe probability distributions?
• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
Histograms

• Histograms on discrete data
  – Nominal
  – Ordinal
  – Numeric with few different values
    • E.g. small number of different integers

• Histograms on continuous data
  – Numeric with many different values
Histograms on discrete data

- **Frequency histogram**
  - Frequency = count of items with each value

```python
data['occupation'].value_counts().plot(kind='bar')
```
Histories on discrete data

- Frequency histogram
  - Frequency = count of items with each value

```python
data['age'].value_counts().plot(kind='bar')
```
Histograms on discrete data

• Frequency histogram
  – Frequency = count of items with each value

![Histogram of 'age' values](chart.png)

Unintuitive and not good, as ages ordered by frequency (and hard to read x-axis labels)

data['age'].value_counts().plot(kind='bar')
Histograms on discrete data

- Frequency histogram
  - Frequency = count of items with each value

```python
data['age'] = data['age'].astype('category', ordered=True, categories=[str(x) for x in range(120)])
data['age'].value_counts().sort_index().plot(kind='bar')
```
Histograms on discrete data

- Frequency histogram
  - Frequency = count of items with each value

```python
import pandas as pd
import matplotlib.pyplot as plt

# Assuming 'data' is a DataFrame

# Convert 'age' column to categorical with meaningful order
data['age'] = data['age'].astype('category', ordered=True, categories=[str(x) for x in range(120)])

# Plot frequency histogram by age
plt.figure(figsize=(10, 6))
data['age'].value_counts().sort_index().plot(kind='bar')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Frequency Histogram of Age')
plt.show()
```
Histograms on discrete data

- **Frequency histogram**
  - Frequency = count of items with each value

```python
x_locations = [10*i for i in range(12)]
x_values = [data['age'].cat.categories[i] for i in x_ticks]
plt.xticks(x_locations, x_values)
```
Histograms on discrete data

- **Frequency histogram**
  - Frequency = count of items with each value

```python
data['workclass'].value_counts().plot(kind='bar')
```
Histograms on discrete data

• Relative frequency histogram
  – Relative frequency = proportion (0..1) or percentage (0..100%) of items with each value
  – Heights of bars sum up to 1

```python
counts = data['workclass'].value_counts()
counts = counts/sum(counts)
counts.plot(kind='bar')
```
Which histogram to use on discrete data?

• Depends on the goal

• Frequency histogram
  – Gives **actual counts** in the data

• Relative frequency histogram
  – Gives proportions in the data
  – Interpretable as **probability distribution** of a randomly chosen item
Histograms on continuous data

- Continuous attribute:
  - Usually value is different in each item
  - Need to introduce bins (a.k.a. intervals, ranges)
  - Histogram not informative without bins:

```python
data['salaries'].value_counts().plot(kind='bar')
```
Histograms on continuous data

- Frequency histogram of binned data
  - Frequency = count of items in each bin

```python
data['salaries'].hist()
```
Histograms on continuous data

- Frequency histogram of binned data
  - Frequency = count of items in each bin

```python
data['salaries'].hist(bins=[i*10000 for i in range(15)])
```
Histograms on continuous data

- Relative frequency histogram of binned data
  - Relative frequency = proportion of items in bins
  - Heights of bars add up to 1

```python
hist, bins = np.histogram(data['salaries'], bins=[i*10000 for i in range(15)])
hist = hist / sum(hist)
plt.bar(bins[:-1], hist.astype(np.float32) / hist.sum(), width=bins[1]-bins[0], align='edge')
```
Histograms on continuous data

- **Density histogram of binned data**
  - Density = Y-axis such that areas of bars in the histogram add up to 1
  - Density scale is invariant to the sizes of bins

```python
data['salaries'].hist(bins=[i*10000 for i in range(15)], density=True)
```
Histograms on continuous data

• Density histogram of binned data
  – Density = Y-axis such that areas of bars in the histogram add up to 1
  – Density scale is invariant to the sizes of bins

```python
data['salaries'].hist(bins=[i*10000 for i in range(15)], density=True, alpha=0.5, color='blue')
data['salaries'].hist(bins=[i*1000 for i in range(150)], density=True, alpha=0.5, color='red')```
Histographs on continuous data

• Density histogram of binned data
  – Density = Y-axis such that areas of bars in the histogram add up to 1
  – Density scale is invariant to the sizes of bins

import seaborn as sns
sns.distplot(data['salaries'], kde_kws={'color': 'red', 'linewidth': 5})

Red line is the estimated probability density
Continuous probability distributions

- Represented by the probability density function (pdf)
- Area under the curve is equal to 1
- Areas represent probabilities

\[
\text{Area} = P(a < X < b) = \text{probability that } X \text{ is between } a \text{ and } b
\]
Lecture 02

- Types of data
- Types of attributes
- First look at a nominal attribute
- First look at an ordinal attribute
- First look at a numeric attribute
- Distribution of attributes
- Types of histograms

  • **How to describe probability distributions?**
  • Some standard probability distributions
  • More ways to visualise distributions
  • Visualising relations of attributes