Lecture 03: Exploration of data

Meelis Kull

meelis.kull@ut.ee

Autumn 2018
✓ 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
Lecture 03

- Data understanding: distribution of attributes
- Types of histograms
- How to describe probability distributions?
- Some standard probability distributions
- More ways to visualise distributions
- Visualising relations of attributes
- Are the attributes related?
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 [20]: (grades[2], grades[3], grades[2] > grades[3])
Out[20]: ('E', 'B', True)

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

```python
In [20]: (grades[2], grades[3], grades[2]>grades[3])
Out[20]: ('E', 'B', True)

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

In [22]: grades.max()
Out[22]: 'A'
```

Currently ‘E’>’B’ because comparison is being made alphabetically, as is standard for strings.
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 [ ]:
g2 = grades[[2]]
print(g2)
g3 = grades[[3]]
print(g3)
g3 = g3.rename({3:2})
print(g3)
print(g2 > g3)
```

```
2   E
dtype: category
Categories (6, object): [F < E < D < C < B < A]
Out[3]:
3   B
dtype: category
Categories (6, object): [F < E < D < C < B < A]
Out[4]:
2   B
dtype: category
Categories (6, object): [F < E < D < C < B < A]
Out[5]:
2   False
dtype: bool
```
Lecture 03

• Data understanding: distribution of attributes
• Types of histograms
• How to describe probability distributions?
• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
• Are the attributes related?
You will work on the same dataset in homework 02, with minor differences
Data understanding: Exploration

• We know from previous lecture 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:
Answer 1

• 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
  – ...

Answer 2

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

In homework 02 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 03

✓ Data understanding: distribution of attributes
• Types of histograms
• How to describe probability distributions?
• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
• Are the attributes related?
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
Histgrams on discrete data

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

```python
data['occupation'].value_counts().plot(kind='bar')
```
Histograms 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

```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
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
Histories 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)
```
• 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')
```
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
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

\[
Area = P(a<X<b) = \text{probability that } X \text{ is between } a \text{ and } b
\]
Lecture 03

✓ Data understanding: distribution of attributes
✓ Types of histograms

• How to describe probability distributions?

• Some standard probability distributions
• More ways to visualise distributions
• Visualising relations of attributes
• Are the attributes related?
### Important statistics of distributions

- **Statistic** – measure calculated from all values

- **Mode** of the distribution:
  - The most probable value (or values)
    - The most frequent value if discrete
    - Value with highest density if continuous

- **Median** and other quantiles
  - We have defined earlier

- **Mean** of the distribution:
  - Average value of the attribute
    - Arithmetic average if discrete
    - Expected value if continuous
  - Centre of mass of the distribution
Examples

• Consider attribute with values:
  – 1,2,2,2,3,3,4,7

• Mode:
  – 2 because occurs three times

• Median
  – 2.5 because 2 & 3 are in the middle, \((2+3)/2 = 2.5\)

• Mean:
  – 3.0 because \((1+2+2+2+3+3+4+7)/8 = 24/8 = 3.0\)
Mode of this attribute is ...

A. 0
B. 0.5
C. 1
D. 1.5
E. None of the above
Median of this attribute is ...

A. 0
B. 0.5
C. 1
D. 1.5
E. None of the above
Mean of this attribute is ...

A. 0
B. 0.5
C. 1
D. 1.5
E. None of the above
Standard deviation and variance

• Variance of a distribution is:
  – average squared deviation from the mean

• Standard deviation is square root of variance
  – Quadratic average deviation from the mean

• Example:
  – Values: 1,2,2,2,3,3,4,7
  – Mean: 3.0
  – Variance: \( ((1-3)^2+(2-3)^2+\ldots+(7-3)^2) / 8 = 3.0 \)
  – Standard deviation: \( \sqrt{3.0} = 1.732 \ldots \)
More words to describe probability distributions

- Symmetric
- Skewed / right-skewed / left-skewed
- Heavy-tailed
- Bimodal
- Multi-modal
Simple definitions, not fully correct:

- **Unimodal** – 1 mode
- **Bimodal** – 2 modes
- **Multimodal** – multiple modes

Actually:

- **Bimodal** usually means that the density (pdf) has two local maxima:
- **Multimodal** means pdf has multiple maxima (2 or more)
Symmetric & skewed data

• Symmetric distribution:
  – Symmetric around a vertical axis of symmetry
    • Left and right side are mirror-images of each other

• Left- (or right-)skewed distribution:
  – Non-symmetric and mean below (above) mode
    • Usually used only for unimodal distributions
Lecture 03

- Data understanding: distribution of attributes
- Types of histograms
- How to describe probability distributions?
  - Some standard probability distributions
  - More ways to visualise distributions
  - Visualising relations of attributes
  - Are the attributes related?
Probability distributions in statistics

• Statisticians have names for many different families of probability distributions

• Why need to know some of them?
  – In practice these are used to communicate the distribution without having to visualise it

• We will talk about:
  – Uniform distributions
  – Normal distributions
  – Power law distributions
Discrete uniform distribution

- All options are equally probable
Continuous uniform distribution

• All values in \([a,b]\) are equally probable:
  – The pdf is constant between \(a\) and \(b\), and 0 elsewhere

\[
f(x) = \frac{1}{b-a}\]

\[
\begin{align*}
0 & \quad a & \quad b & \quad x \\
\frac{1}{b-a} & \quad \mathbf{f(x)} & \quad 1 & \quad \mathbf{x}
\end{align*}
\]
Normal distribution

99.7% of the data are within 3 standard deviations of the mean

95% within 2 standard deviations

68% within 1 standard deviation

Represent data dispersion, spread

Represent central tendency
Normal distribution

• $N(\text{mean}, \text{variance})$

• The most common non-uniform continuous distribution
  
  – Why?
  
  – Sum of many independent and identically distributed (i.i.d.) random variables is approximately normally distributed

• Standard normal distribution: $N(0,1)$
More terminology

• Heavy-tailed (or long-tailed) distribution
  – Very high or low values are likely
    • More likely than in case of normal distribution
  – Technical definition:
    • Tails are not exponentially bounded
Power-law distributions

- Alternative names:
  - scale-free / scale-independent
- Distributions on positive real values
- The probability of $M$ times bigger value is $K$ times smaller (power law; $M, K$ - parameters)
- Linearly descending pdf when drawn in log-log-scale (both $x$ and $y$ logarithmic)
- Examples:
  - Number of connections in many real-world graphs
  - Frequencies of words in languages
  - Sizes of craters on the moon
How to describe this distribution?

A. Uniform
B. Unimodal symmetric
C. Unimodal right-skewed
D. Unimodal left-skewed
E. Bimodal symmetric
F. Bimodal asymmetric
G. Multi-modal symmetric
H. Multi-modal asymmetric
I. Normally distributed
J. Power-law distributed
How to describe this distribution?

A. Uniform
B. Unimodal symmetric
C. Unimodal right-skewed
D. Unimodal left-skewed
E. Bimodal symmetric
F. Bimodal asymmetric
G. Multi-modal symmetric
H. Multi-modal asymmetric
I. Normally distributed
J. Power-law distributed
How to describe this distribution?

A. Uniform
B. Unimodal symmetric
C. Unimodal right-skewed
D. Unimodal left-skewed
E. Bimodal symmetric
F. Bimodal asymmetric
G. Multi-modal symmetric
H. Multi-modal asymmetric
I. Normally distributed
J. Power-law distributed
How to describe this distribution?

A. Uniform
B. Unimodal symmetric
C. Unimodal right-skewed
D. Unimodal left-skewed
E. Bimodal symmetric
F. Bimodal asymmetric
G. Multi-modal symmetric
H. Multi-modal asymmetric
I. Normally distributed
J. Power-law distributed
Lecture 03

✓ Data understanding: distribution of attributes
✓ Types of histograms
✓ How to describe probability distributions?
✓ Some standard probability distributions
  • More ways to visualise distributions
  • Visualising relations of attributes
  • Are the attributes related?
Violin plots

- Compactly visualise many distributions
  - Density plots rotated by 90° and mirrored
  - Distribution of salaries for each education level

```python
sns.violinplot(data=data,x='education',y='salaries')
```
Violin plots

• Compactly visualise many distributions
  – Density plots rotated by 90° and mirrored
  – Distribution of salaries for each education level

```python
sns.violinplot(data=data, x='education', y='salaries')
plt.xticks(rotation='vertical')
```
Box plots

- Compact visualise many distributions
  - marked median, upper and lower quartile, and outliers (R ggplot outlier = more than 1.5x inter-quartile range from quartile)

```python
sns.boxplot(data=data, x='education', y='salaries')
plt.xticks(rotation='vertical')
```
Violin and box plots combined

• Violin plots and box plots can also be shown together:

```python
sns.violinplot(data=data, x='education', y='salaries')
sns.boxplot(data=data, x='education', y='salaries', boxprops=dict(alpha=.5))
plt.xticks(rotation='vertical')
```
Scatter plots

- Often too crowded, but sometimes provide extra insights compared to box plots and violin plots

```python
plt.scatter(data['education'], data['salaries'])
plt.xticks(rotation='vertical')
```
Scatter plots

- Often too crowded, but sometimes provide extra insights compared to box plots and violin plots

```python
plt.scatter(data['education'][:100], data['salaries'][:100], s=300)
plt.xticks(rotation='vertical')
```

More useful, here only 100 points
Lecture 03

✓ Data understanding: distribution of attributes
✓ Types of histograms
✓ How to describe probability distributions?
✓ Some standard probability distributions
✓ More ways to visualise distributions
• Visualising relations of attributes
• Are the attributes related?
Visualising relationships between 2 attributes

- 2 categorical attributes:
  - Cross-table (workclass & education)

```python
pd.crosstab(data['education'], data['workclass'])
```
Visualising relationships between 2 attributes

- 2 categorical attributes:
  - Cross-table (workclass & education)
  - Heatmap (workclass & education)

```python
tab = pd.crosstab(data['education'], data['workclass'])
plt.pcolor(tab)
plt.xticks(np.arange(0.5, len(tab.columns), 1), tab.columns, rotation='vertical')
plt.yticks(np.arange(0.5, len(tab.index), 1), tab.index)
```
Visualising relationships between 2 attributes

- 2 categorical attributes:
  - Cross-table (workclass & education)
  - Heatmap (workclass & education)

```
import pandas as pd
import seaborn as sns

# Example data
        'education': ['10th', '11th', '12th', '1st-4th', '5th-6th', '7th-8th', '9th', 'Assoc-acdm', 'Assoc-voc', 'Bachelors', 'Doctorate', 'HS-grad', 'Masters', 'Preschool', 'Prof-school', 'Some-college']}

# Create a cross-tabulation
tab = pd.crosstab(data['education'], data['workclass'])

# Create a heatmap
sns.heatmap(tab)
```
### Visualising relationships between 2 attributes

<table>
<thead>
<tr>
<th>education</th>
<th>workclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>11th</td>
<td>1e+02</td>
</tr>
<tr>
<td>12th</td>
<td>1.2e+02</td>
</tr>
<tr>
<td>1st-4th</td>
<td>40</td>
</tr>
<tr>
<td>5th-6th</td>
<td>12</td>
</tr>
<tr>
<td>7th-8th</td>
<td>30</td>
</tr>
<tr>
<td>9th</td>
<td>72</td>
</tr>
<tr>
<td>Assoc-acdm</td>
<td>51</td>
</tr>
<tr>
<td>Assoc-voc</td>
<td>47</td>
</tr>
<tr>
<td>Bachelors</td>
<td>61</td>
</tr>
<tr>
<td>Doctorate</td>
<td>1.7e+02</td>
</tr>
<tr>
<td>HS-grad</td>
<td>15</td>
</tr>
<tr>
<td>Masters</td>
<td>5.3e+02</td>
</tr>
<tr>
<td>Preschool</td>
<td>48</td>
</tr>
<tr>
<td>Prof-school</td>
<td>5</td>
</tr>
<tr>
<td>Some-college</td>
<td>18</td>
</tr>
</tbody>
</table>

Legend: 0 = Black, 1500 = Deep Blue, 3000 = Green, 4500 = Yellow, 6000 = Orange, 7500 = Red.
Visualising relationships between 2 attributes

- 2 categorical attributes:
  - Cross-table (workclass & education)
  - Heatmap (workclass & education)
  - Cross-table and heatmap combined

```python
tab = pd.crosstab(data['education'], data['workclass'])
sns.heatmap(tab, annot=True, annot_kws={"size": 20})
```
Visualising relationship of 1 categorical, 1 continuous attribute

- Scatter plot, box plot, violin plot
Visualising relationship of 2 continuous attributes

• Scatter plot

```
plt.scatter(data['age'], data['salaries'], s=300)
```
Visualising relationship of 2 continuous attributes

- Scatter plot

```python
plt.scatter(data['capital.gain'], data['salaries'], s=300)
```
Visualising relationship of 2 continuous attributes

• Scatter plot

• Discretise one (or both) of the attributes
  – Discretise = make into categorical
  – For instance, introduce bins

```python
data['capital.gain.discretised'] = pd.cut(data['capital.gain'], [0, 5000, 10000, 20000, 50000, 100000])
sns.boxplot(data=data, x='capital.gain.discretised', y='salaries')
plt.xticks(rotation='vertical')
```
Visualising more than 2 attributes

- Make all pairwise visualisations and organise in a cross-table

```python
df = sns.load_dataset('iris')
sns.pairplot(df)
```
Visualising more than 2 attributes

• Make all pairwise visualisations and organise in a cross-table

• Perform dimensionality reduction
  – Introduced later in the course
  – Projects the data into a 2-dimensional space
  – Then visualise

• Use colours, shapes, etc.
Use colours, shapes, etc

• Three attributes visualised together

```python
sns.scatterplot(data=data, x='capital.gain', y='salaries', hue='workclass', s=300)
plt.legend(loc='upper right')
```
Use colours, shapes, etc

- Three attributes visualised together

Often hard to read when many attributes visualised this way
Lecture 03

✓ Data understanding: distribution of attributes
✓ Types of histograms
✓ How to describe probability distributions?
✓ Some standard probability distributions
✓ More ways to visualise distributions
✓ Visualising relations of attributes

• Are the attributes related?
Are given 2 attributes related?

- There are many statistical tests about this, we will cover some in the next lecture
- Visual inspection can reveal some relations
  - For example, university graduates have higher median salaries than others

```python
sns.boxplot(data=data, x='education', y='salaries')
plt.xticks(rotation='vertical')
```
Are given 2 attributes related?

• There are many statistical tests about this, we will cover some in the next lecture
• Visual inspection can reveal some relations
  – For example, university graduates have higher median salaries than others
• Correlation is a statistic on 2 numeric attributes, quantifying linear relationships
Pearson correlation coefficient

• Mean (actually *sample* mean as explained in next lecture)

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

• Correlation (Pearson correlation coefficient)

\[ r = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \]
Pearson correlation coefficient

- Ranges from -1 to +1:
  - R=-1: perfectly anti-correlated
  - R=+1: perfectly correlated
  - R=0: absolutely uncorrelated
Pearson correlation coefficient

Uncorrelated

Uncorrelated

Uncorrelated
Example of positive correlation (Abalone dataset)
Several sets of (x, y) points, with the Pearson correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
Anscombe’s quartet (question coming up)
The correlations in Anscombe’s quartet are

A. All equal
B. All different
C. Some equal, some different
Anscombe’s quartet: means, variances, correlations equal!

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of ( x ) in each case</td>
<td>11 exact</td>
</tr>
<tr>
<td>Variance of ( x ) in each case</td>
<td>10 exact</td>
</tr>
<tr>
<td>Mean of ( y ) in each case</td>
<td>7.50 (to 2 d.p.)</td>
</tr>
<tr>
<td>Variance of ( y ) in each case</td>
<td>3.75 (to 2 d.p.)</td>
</tr>
<tr>
<td>Correlation between ( x ) and ( y ) in each case</td>
<td>0.816 (to 3 d.p.)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>( y = 3.00 + 0.500x ) (to 2 d.p. and 3 d.p. resp.)</td>
</tr>
</tbody>
</table>
Anscombe’s quartet: means, variances, correlations equal!
Guess the correlation!
Guess the correlation!

- True R: 0.44
- Guessed R: 0.40
- Difference: 0.04
- Streaks: 1
- Mean Error: 0.04

Next:

High Score (0)

Main Menu
CO2 and Noise in Liivi 2-405
Source: Data from prof. Jaak Vilo, 2017

- CO2
- Noise
- Pressure

Source: Data from prof. Jaak Vilo, 2017
CO2 and Noise in Liivi 2-405

Source: Data from prof. Jaak Vilo, 2017

Changed the scale of noise

- CO2
- Pressure
- Noise

Source: Data from prof. Jaak Vilo, 2017
CO2 and Noise in Liivi 2-405

Source: Data from prof. Jaak Vilo, 2017

\[ y = 0.0465x + 19.693 \]

\[ R^2 = 0.65625 \]
CO2 and Noise in Liivi 2-405

Source: Data from prof. Jaak Vilo, 2017

![Graph showing CO2 and Noise correlation]

\[ y = 0.0465x + 19.693 \]

\[ R^2 = 0.65625 \]

**Not the same as correlation!**

Source: Data from prof. Jaak Vilo, 2017
r or $R^2$?

- $r$ – Pearson correlation coefficient
- $R^2$ – coefficient of determination
  - Proportion of variance in the target attribute that is predictable from the source attribute
  - Basically, measures how well the data follow the regression line
  - In simple linear regression $R^2$ is equal to the square of $r$
Stock market correlations (until Brexit)

![Graph showing stock market correlations between Stoxx Europe 50 Index and FTSE 100 Index with a correlation of 0.93.](https://www.valuewalk.com/2016/08/usd-vs-gbp-vs-eur-a-leveling-of-the-playing-field/)

Stock market correlations (after Brexit)

Source:
More correlations

US spending on science, space, and technology correlates with

Source: http://www.tylervigen.com/spurious-correlations
More correlations

US spending on science, space, and technology correlates with
Suicides by hanging, strangulation and suffocation

Source: http://www.tylervigen.com/spurious-correlations
More correlations

Number of people who drowned by falling into a pool correlates with

Swimming pool drownings

1999  2000  2001  2002  2003  2004  2005  2006  2007  2008  2009

80 drownings

100 drownings

120 drownings

140 drownings

Meelis Kull - Autumn 2018 - LTAT.02.002 – Intro to Data Science - Lecture 03
More correlations

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in
More correlations

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Relation ≠ Correlation ≠ Causation

Relation ≠ Correlation ≠ Causation
More correlations

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Relation ≠ Correlation ≠ Causation

Spurious correlations!

Multiple testing problem (we will discuss in the lecture on statistics)
Lecture 03

✓ Data understanding: distribution of attributes
✓ Types of histograms
✓ How to describe probability distributions?
✓ Some standard probability distributions
✓ More ways to visualise distributions
✓ Visualising relations of attributes
✓ Are the attributes related?
Next Lecture 04: Frequent pattern mining
Quotes

• “The most merciful thing in the world... is the inability of the human mind to correlate all its contents.”
  – H. P. Lovecraft

• “Visualization is daydreaming with a purpose.”
  – Bo Bennett

• “Figures never lie, but liars figure.”
  – Anonymous

• Sources: https://www.brainyquote.com
  https://en.wikipedia.org/wiki/Statistical_hypothesis_testing