Chapter 03

Data Exploration

Dr. Steffen Herbold
herbold@cs.uni-goettingen.de

Outline

- Overview
- Summary Statistics
- Visualization for Data Exploration
- Summary

Goal of Data Exploration

- Goal:
  - Understand the basic characteristics of the data

- Examples for characteristics:
  - Structure
  - Size
  - Completeness
  - Relationships

Methods for Data Exploration

- Usually interactive and semi-automated
- Text editors, system calls (head/more/less), etc. to look at raw data directly
  - Helps to understand the structure
- Statistics and visualizations to learn about distributions and relationships
- Exploration should also include meta data
  - Feature names, trace links,etc.

Outline

- Overview
- **Summary Statistics**
- Visualization for Data Exploration
- Summary

Descriptive Statistics

- Summarize data through single value
- Do not predict anything about the data (→ inductive statistics)

- Common statistics covered in this course
  - Central tendency (mean/median/mode)
  - Variability (standard deviation, interquartile range)
  - Range of data (min/max)

- Other important statistics
  - Kurtosis and skewness for the shape of distribuitions
  - More measures for central tendency, e.g., trimmed means, harmonic mean

Central Tendency

- „Typical“ value of the data

- Arithmetic mean
  - $mean(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ with $x = (x_1, \dots, x_n) \in \mathbb{R}^n$

- Median
  - The value that separates the higher half from the data of the lower half

- Mode
  - The value that appears most in the data

Variability

- Measure for the spread of the data
  - Also called dispersion

- Standard deviation
  - Measure for the difference of observation to the arithmetic mean
  - $sd(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - mean(x))^2}{n-1}}$

- Interquartile Range (IQR)
  - Percentile: value below which a given percentage falls
  - Difference between the 75% percentile and the 25% percentile

The median is the 50% percentile

Range of data

- Range for which values are observed
    - Can be infinite!

- Minimum
    - Smallest observed value

- Maximum
    - Largest observed value

- May be strongly distorted by invalid data
    - Makes it also a good tool to discover invalid data

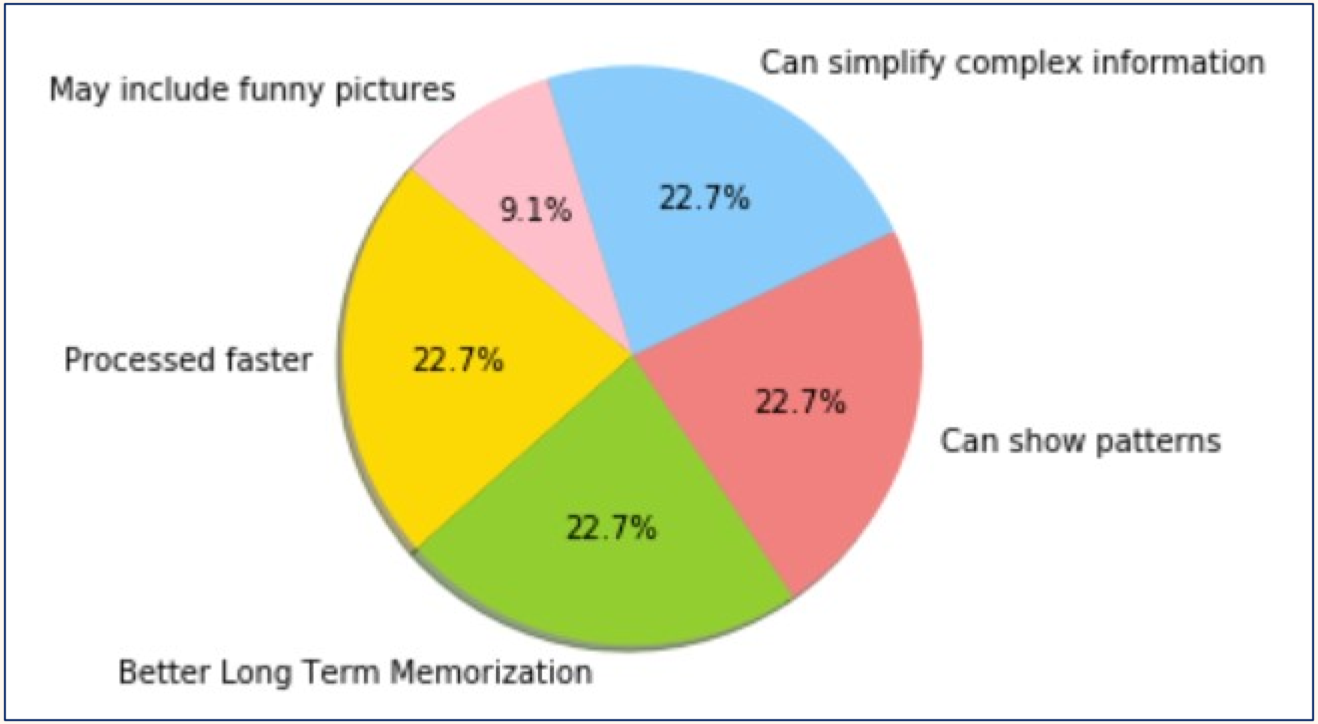
Example

- Random typing on the keypad
  - $x = (1,2,1,1,3,4,5,2,3,4,5,1,3,2,1,6,5,4,9,4,3,6,1,5,6,8,4,6,5,1,3,2,1,6,8,7,6,1,3,1,6,8,4,7,6,4,3,5,4,9,7,4,3,1,4,6,8,7,9, 1,4,6,1,3,8,6,7,4,9,6,5,1,3,6,8,7)$
- central tendency:
  - mean: 4.46052631579
  - median: 4.0
  - mode (count): 1 (14)
- variability
  - sd: 2.41944311488
  - IQR: 3.0
- range
  - min: 1
  - max: 9

Outline

- Overview
- Summary Statistics
- **Visualization for Data Exploration**
- Summary

A Picture Says More than 1000 Words



Numbers are made up and pie charts should actually be avoided

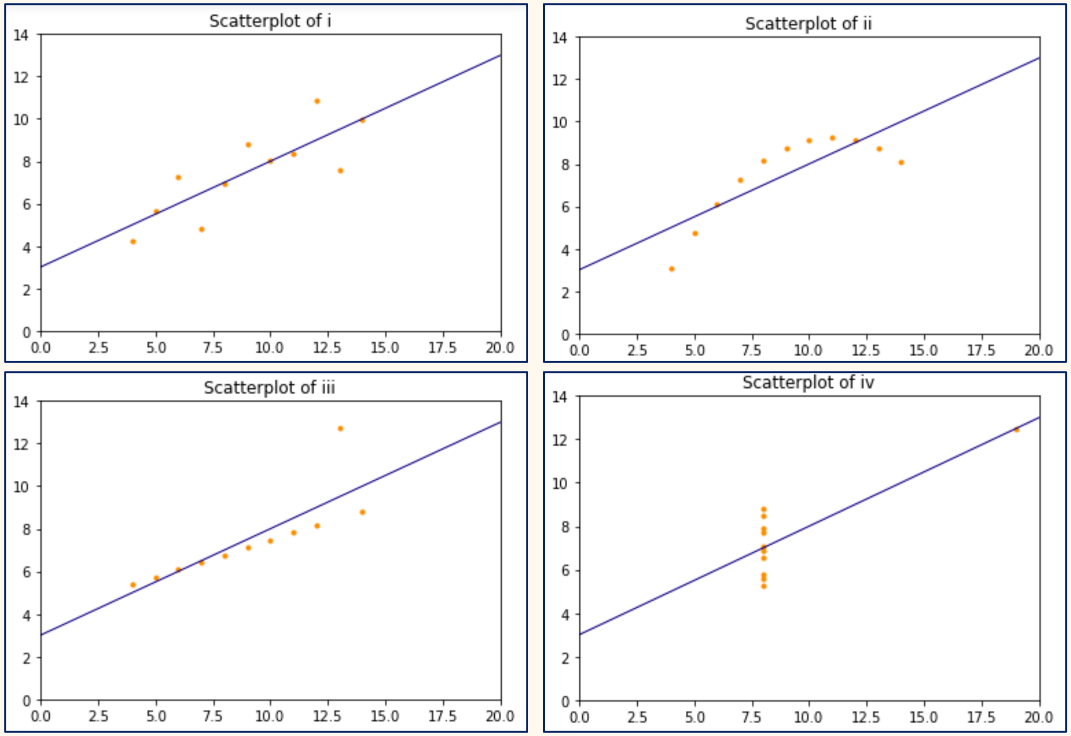
~~Descriptive~~Deceptive Statistics

Have the same
- Mean
- standard deviation
- correlation between x and y
- linear regression

i		ii	
X	y	x	y
10.00	8.04	10.00	9.14
8.00	6.95	8.00	8.14
13.00	7.58	13.00	8.74
9.00	8.81	9.00	8.77
11.00	8.33	11.00	9.26
14.00	9.96	14.00	8.10
6.00	7.24	6.00	6.13
4.00	4.26	4.00	3.10
12.00	10.84	12.00	9.13
7.00	4.82	7.00	7.26
5.00	5.68	5.00	4.74

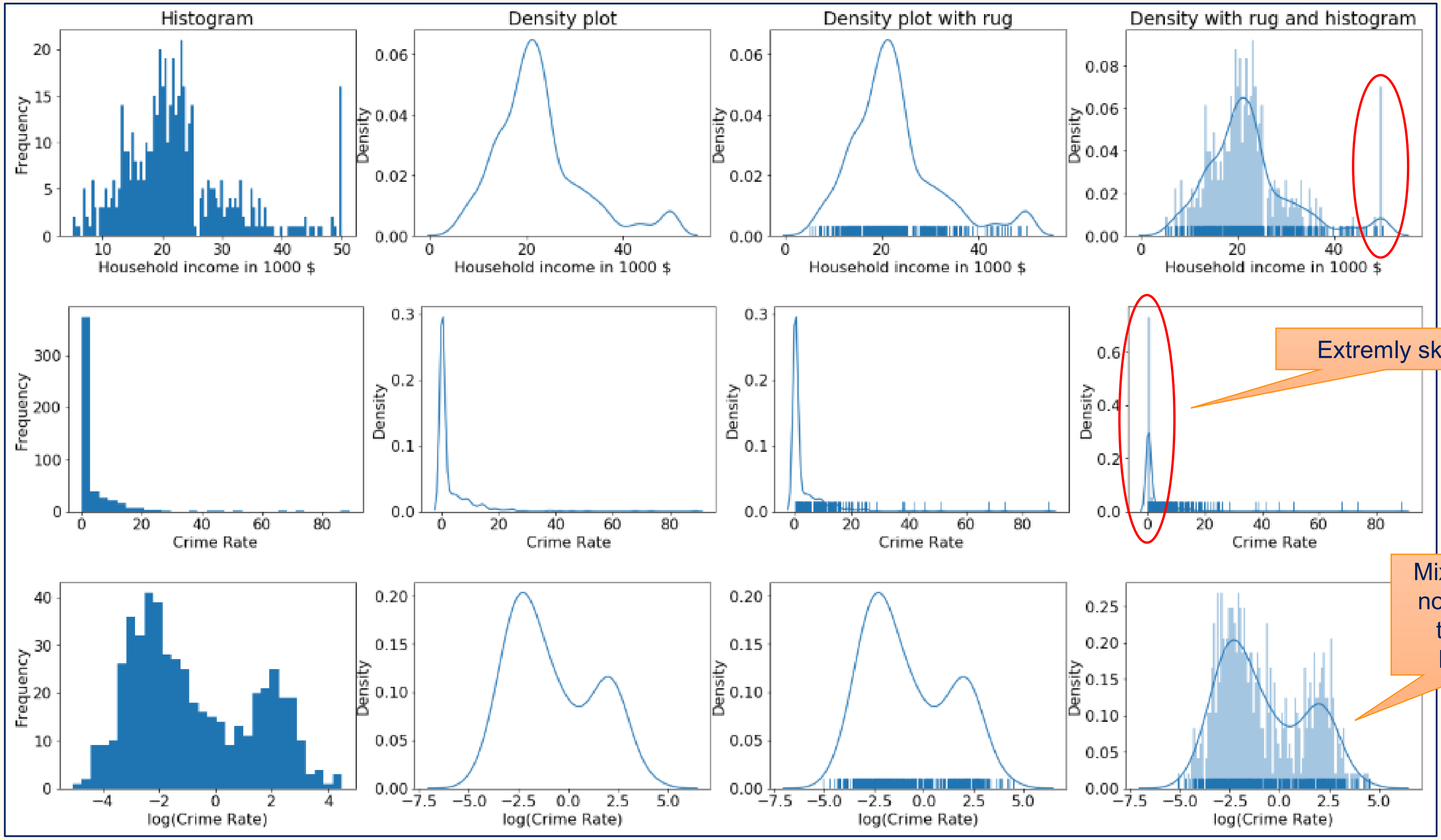
iii		iv	
x	y	x	y
10.00	7.46	8.00	6.58
8.00	6.77	8.00	5.76
13.00	12.74	8.00	7.71
9.00	7.11	8.00	8.84
11.00	7.81	8.00	8.47
14.00	8.84	8.00	7.04
6.00	6.08	8.00	5.25
4.00	5.39	19.00	12.50
12.00	8.15	8.00	5.56
7.00	6.42	8.00	7.91
5.00	5.73	8.00	6.89

Anscombe‘s Quartet



Exploring Single Features

Looks like an artificially high value
→ Groups all higher incomes



Extremly skewed

Mixture of two normals after taking the logarithm

Plots of the Boston house prices data set
http://archive.ics.uci.edu/ml/datasets/Housing

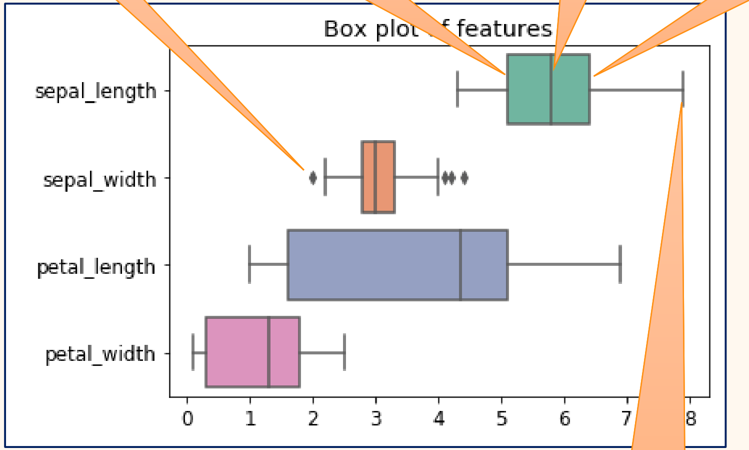
Boxplots

Outlier

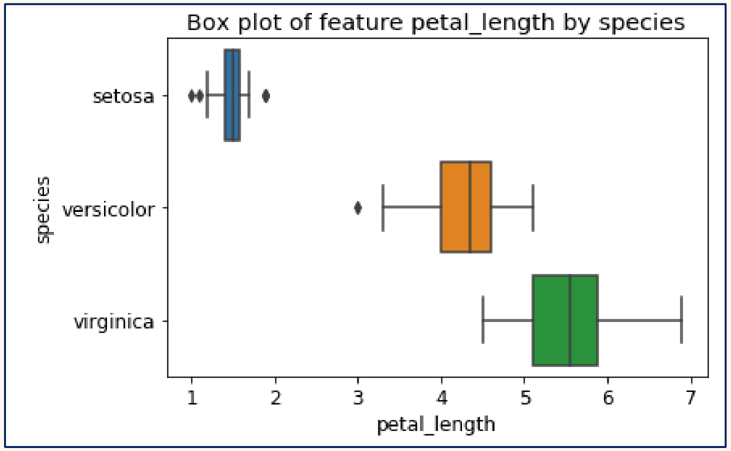
25% percentile

Median

75% percentile

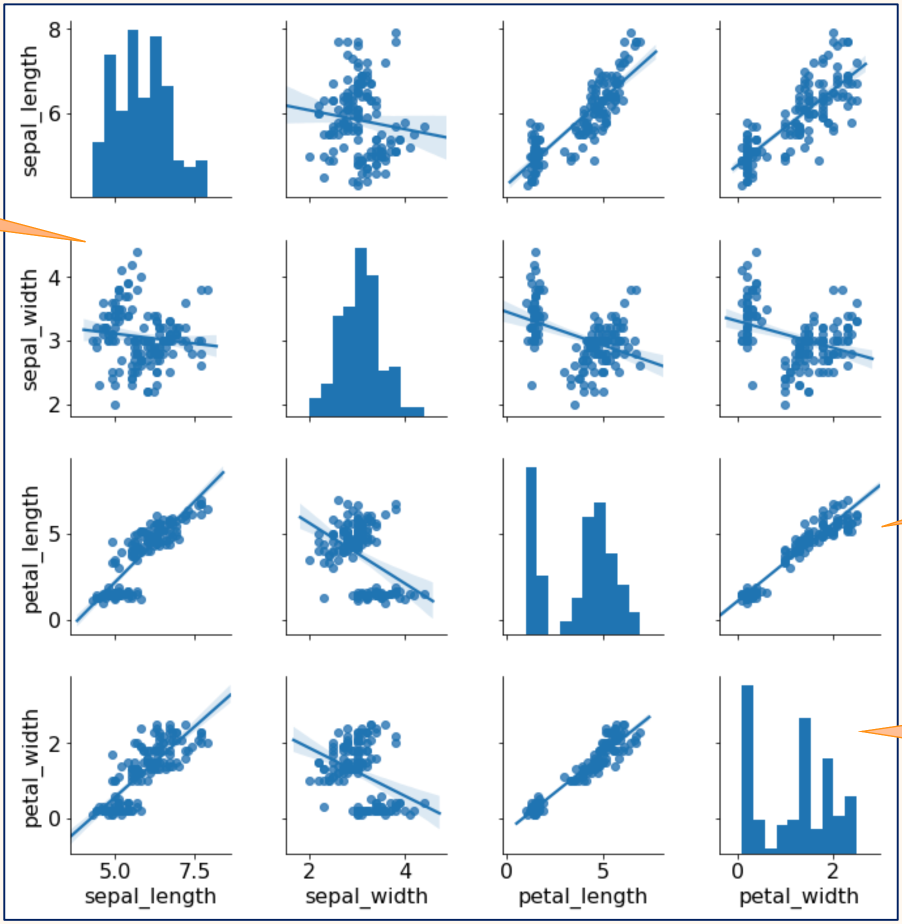


Range of data except outliers



The outlier definition can change. We used „more than 1.5 times the IQR away from the 25%/75% percentile.” You should always check this in the package you use.

Pairwise Scatterplots with Regressions

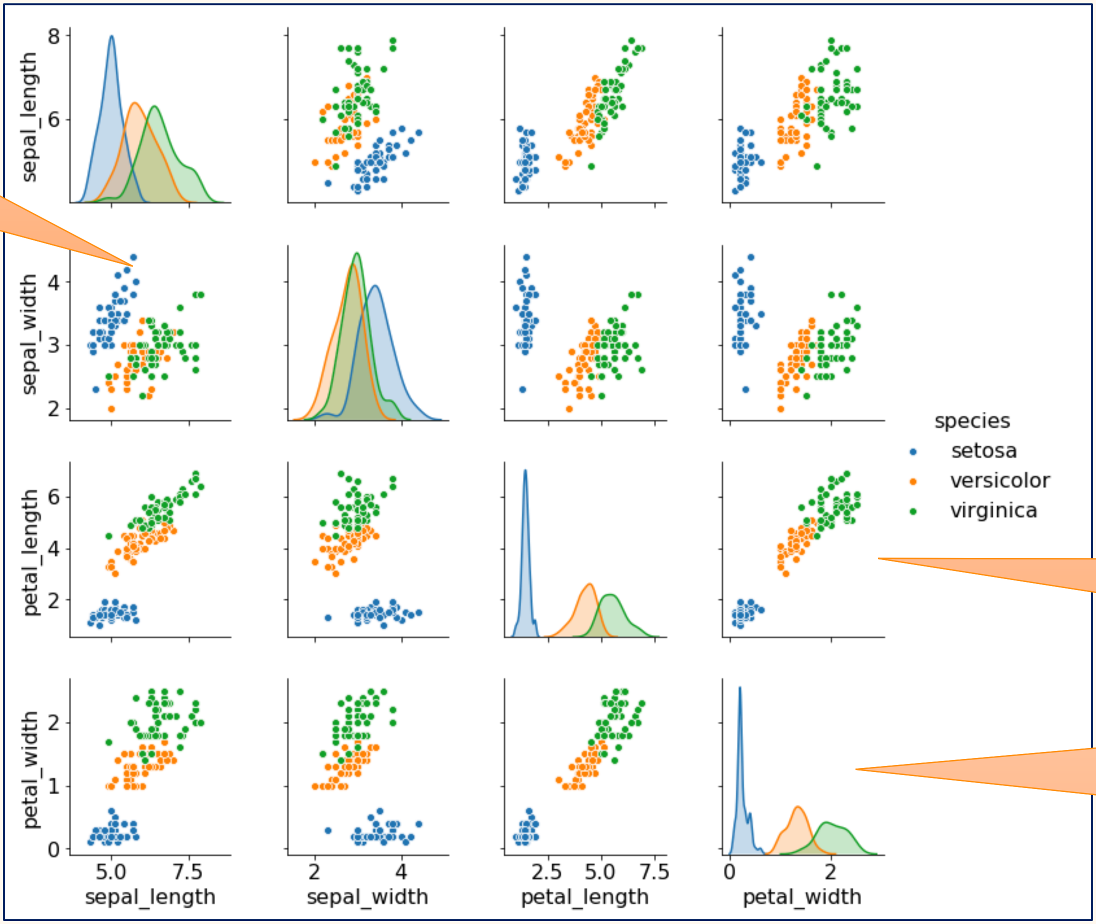


No correlation visible

Strong linear correlation

Histogram of data in the column

Pairwise Plots with Classes



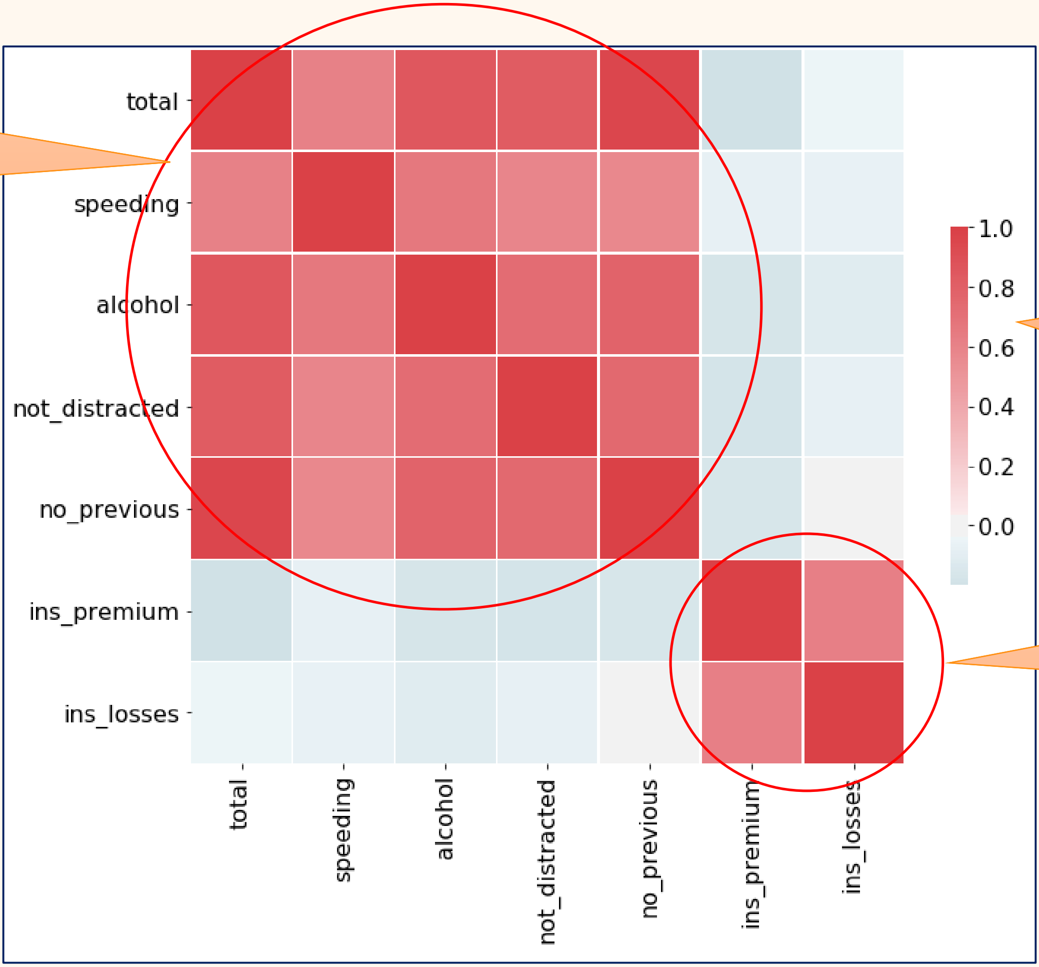
Good separation of blue, but green and orange are overlapping

Good separation of all three classes

Density plots of data in the column separated by classes

Correlation Heatmap

There are different correlation coefficients. We used Pearsons coefficient, which measures linear correlations.

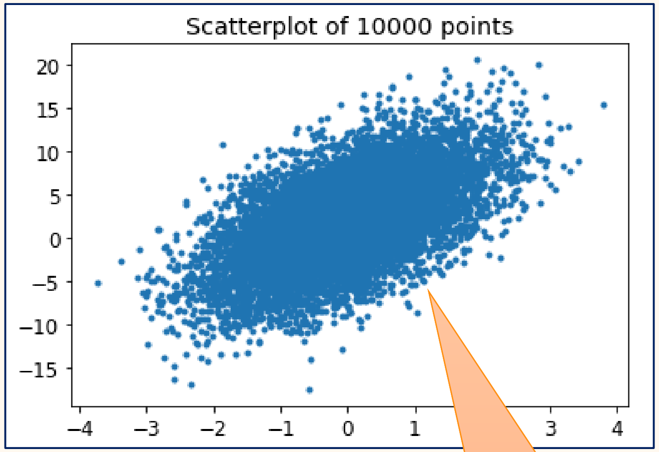


Correlation between reasons for accidents

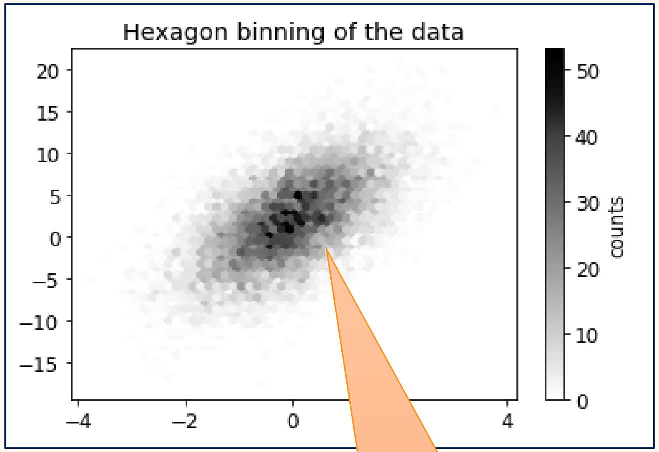
Colors show strength of correlation

Correlation between premiums and losses

Hexbin Plots for Many Instances



Scatterplot of 10000 points

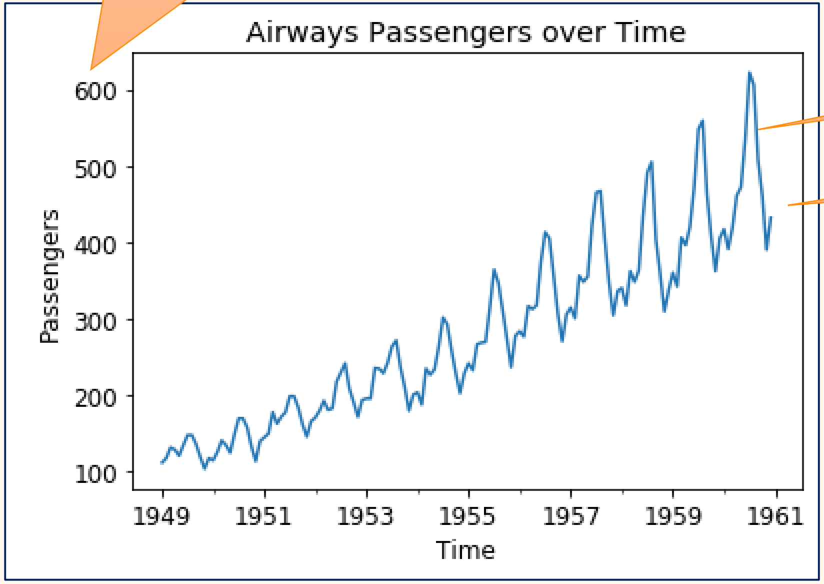


Hexagon binning of the data

Cannot see structure due to amount of data

Hexagonal bins reveal the structure

Line Plots for Timeseries



Range of values

Regular noise pattern → Seasonal?

Linear trend

Outline

- Overview
- Summary Statistics
- Visualization for Data Exploration
- **Summary**

Summary

- Important to understand the data available
- Summary statistics provide a good overview
  - Can be deceptive!
- Visualization is a powerful way to understand data
- Understanding of meta data and how domain expert understand data equally important!