COMP 499 Introduction to Data Analytics

Lecture 5 — Descriptive Analytics

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Overview of Lecture

Descriptive Analytics is describing your data;
that is, data from past activities

1. Five Numbers
2. Python `pandas` describe()
3. Plots: Bar Chart, Histogram, Box Plot
4. Pareto Diagrams
5. Violin Plot
6. Normalization and Z-scores
7. Comparing Two Attributes
8. Correlation is not Causality

Describing Data

Four Features to Describe Data Sets

Center: the point where about half of the observations are on either side.

Spread: the variability of the data.

Shape: described by symmetry, skewness, number of peaks, etc.

Unusual features: gaps where there are no observations and outliers.

Five Numbers of Robust Statistical Descriptors

Five Number Summary

- maximum
- third quartile Q_3
- median
- first quartile Q_1
- minimum

Descriptors

What Else to Describe?

- number of observations
- number of entries
- number of unique entries
- number of missing entries
- number of outliers
- number of extreme values

Python pandas describe

Describing a numeric `Series`.

```
>>> s = pd.Series([1, 2, 3])
>>> s.describe()
count    3.0
mean     2.0
std      1.0
min      1.0
25%      1.5
50%      2.0
75%      2.5
max      3.0
dtype: float64
```

Describing a categorical `Series`.

```
>>> s = pd.Series(['a', 'a', 'b', 'c'])
>>> s.describe()
count     4
unique    3
top       a
freq      2
dtype: object
```

Python pandas describe

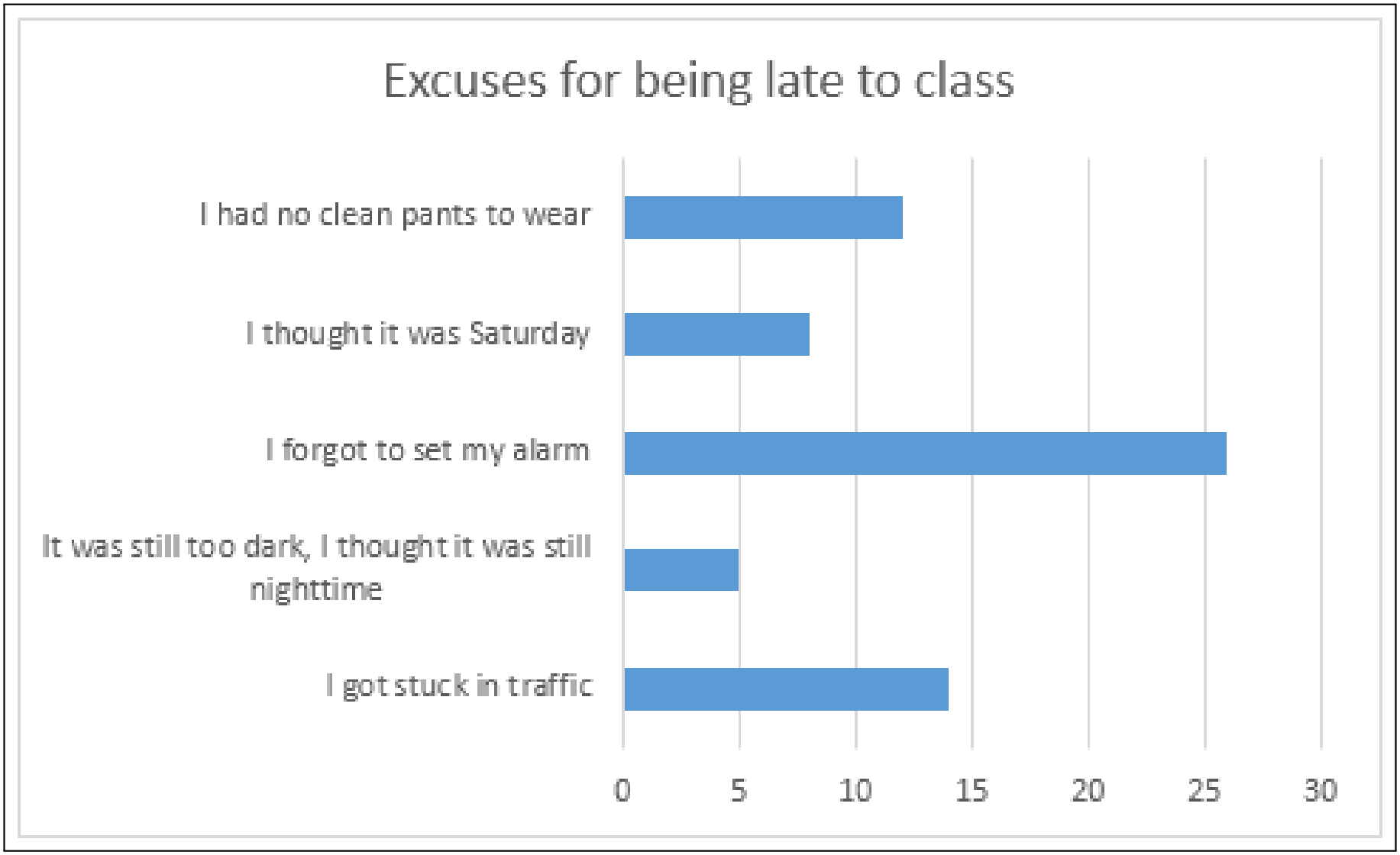
```
>>> df = pd.DataFrame({'categorical': pd.Categorical(['d','e','f']),
...                    'numeric': [1, 2, 3],
...                    'object': ['a', 'b', 'c']
...                   })
```

Describing all columns of a `DataFrame` regardless of data type.

```
>>> df.describe(include='all')
        categorical  numeric object
count             3      3.0      3
unique            3      NaN      3
top               f      NaN      c
freq              1      NaN      1
mean            NaN      2.0    NaN
std             NaN      1.0    NaN
min             NaN      1.0    NaN
25%             NaN      1.5    NaN
50%             NaN      2.0    NaN
75%             NaN      2.5    NaN
max             NaN      3.0    NaN
```

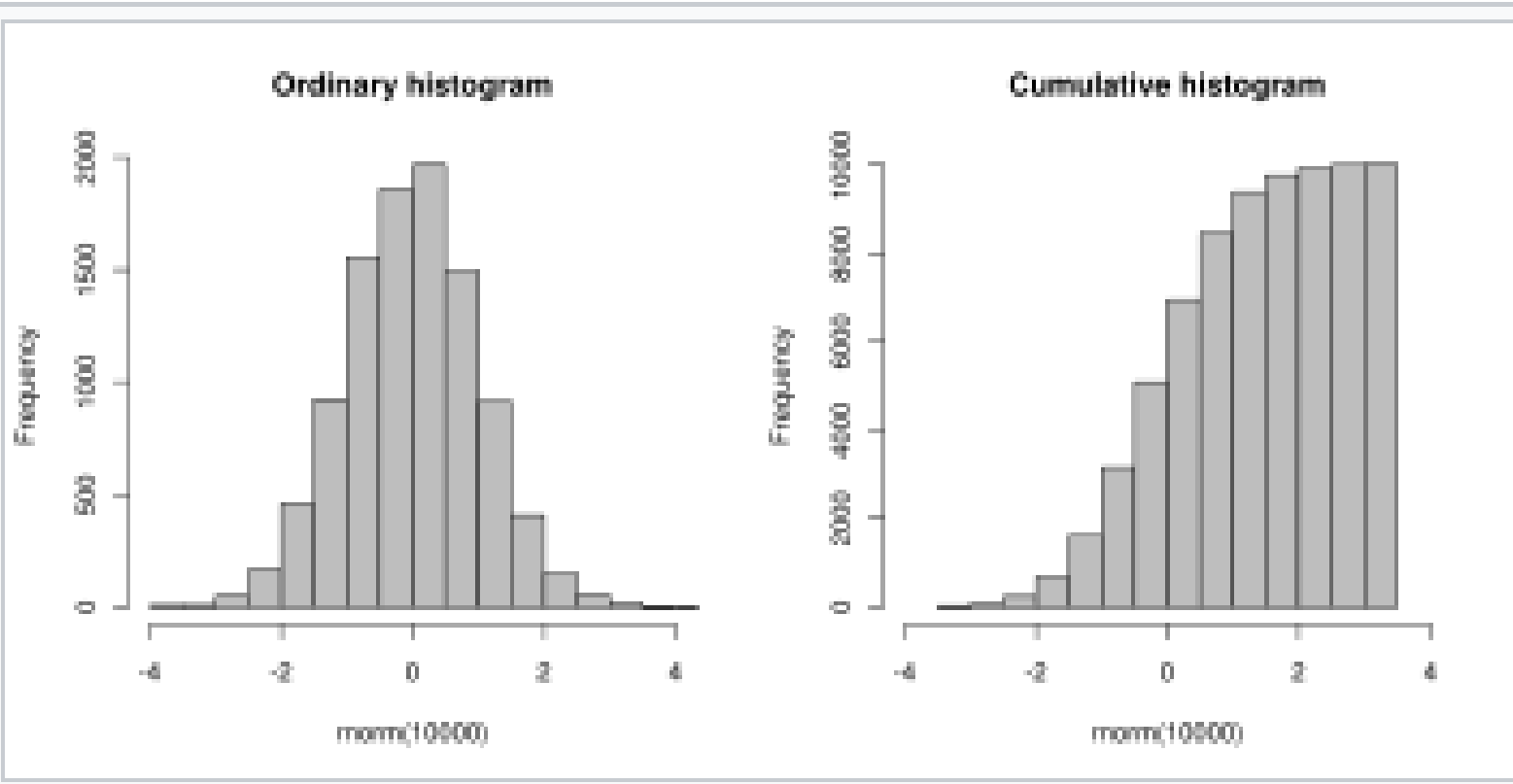
Bar Chart

Bar Chart



Histogram

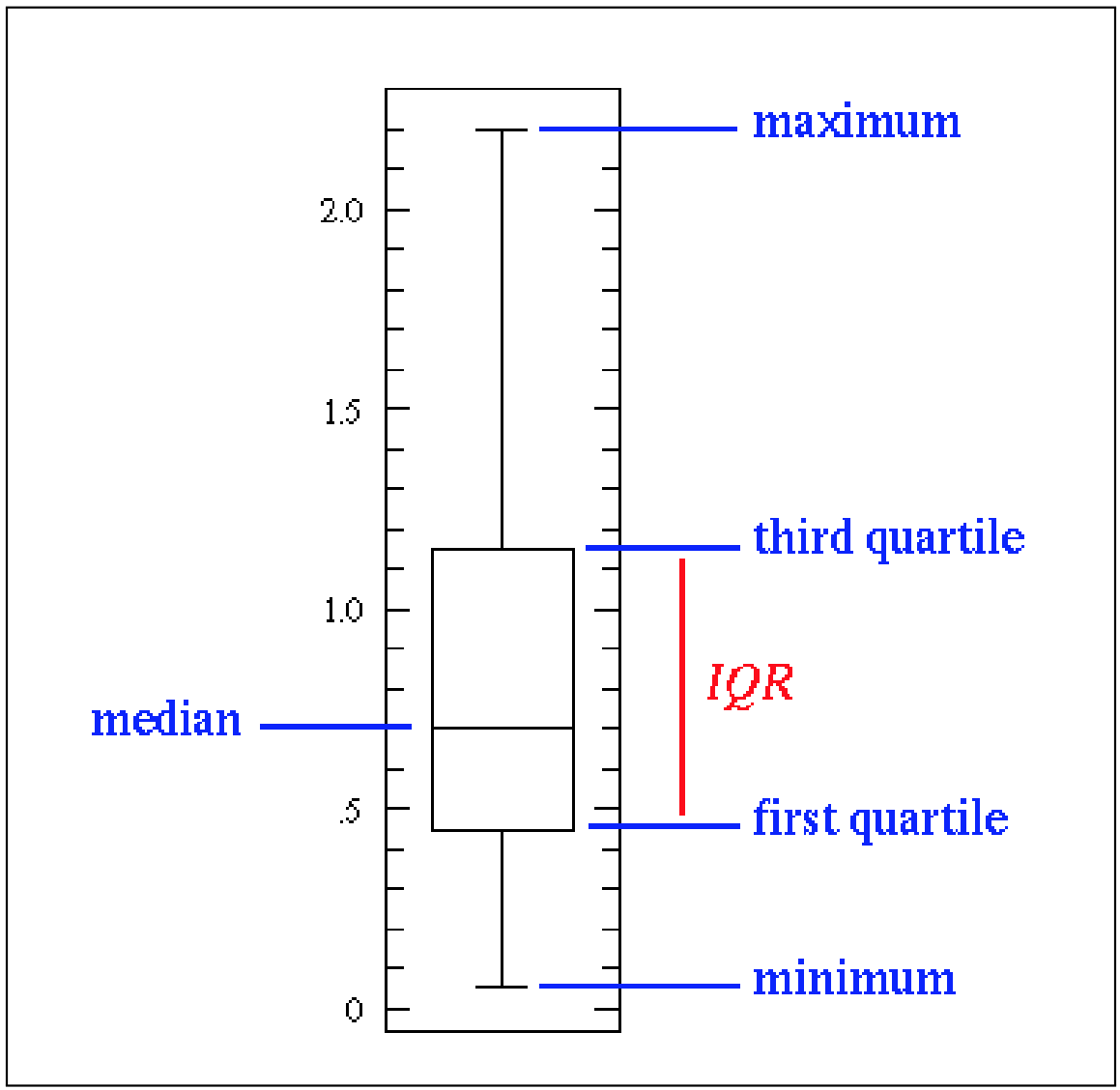
Histogram



An ordinary and a cumulative histogram of the same data. The data shown is a random sample of 10,000 points from a normal distribution with a mean of 0 and a standard deviation of 1.

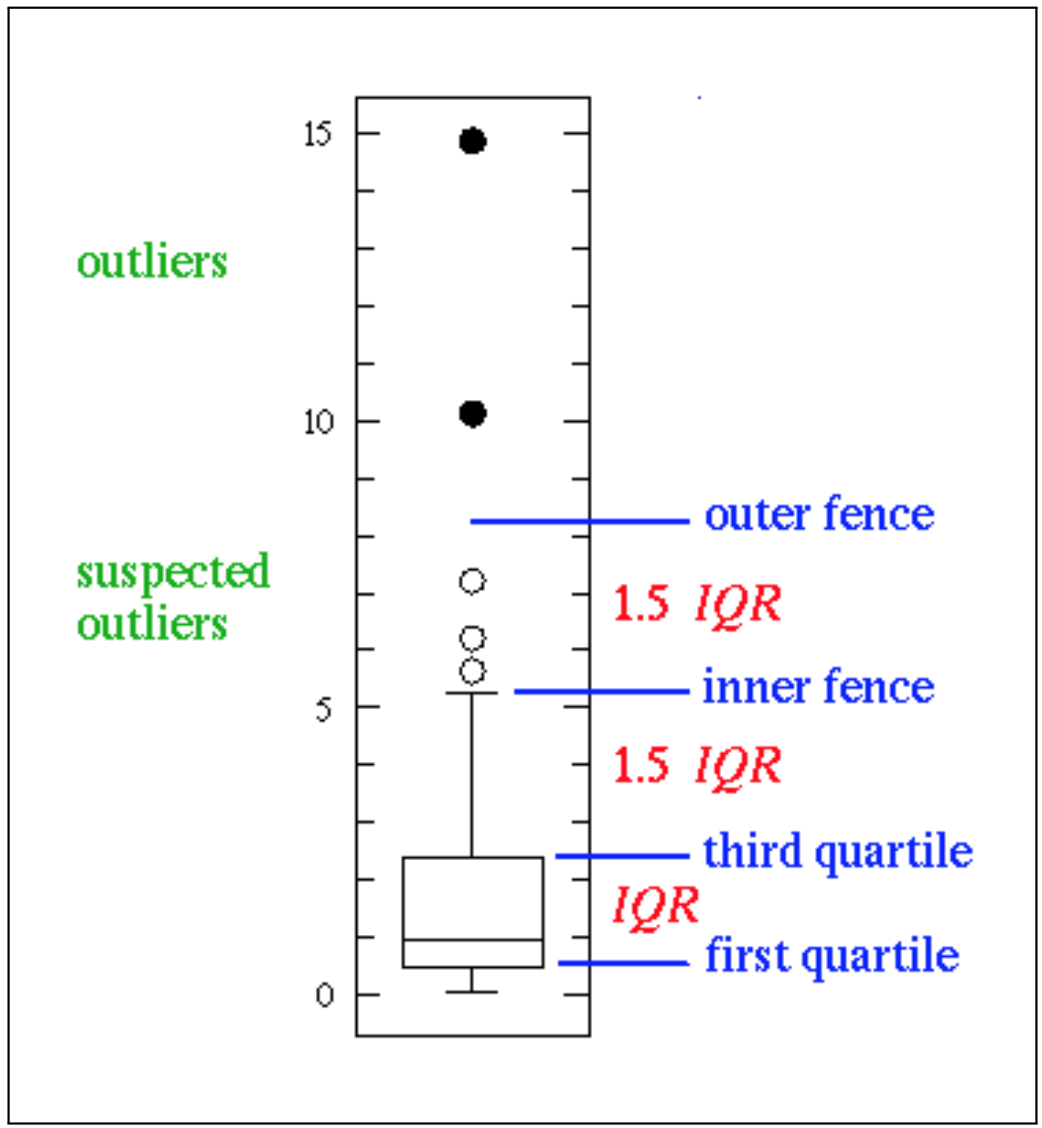
Box Plot

Box Plot



Box Plot

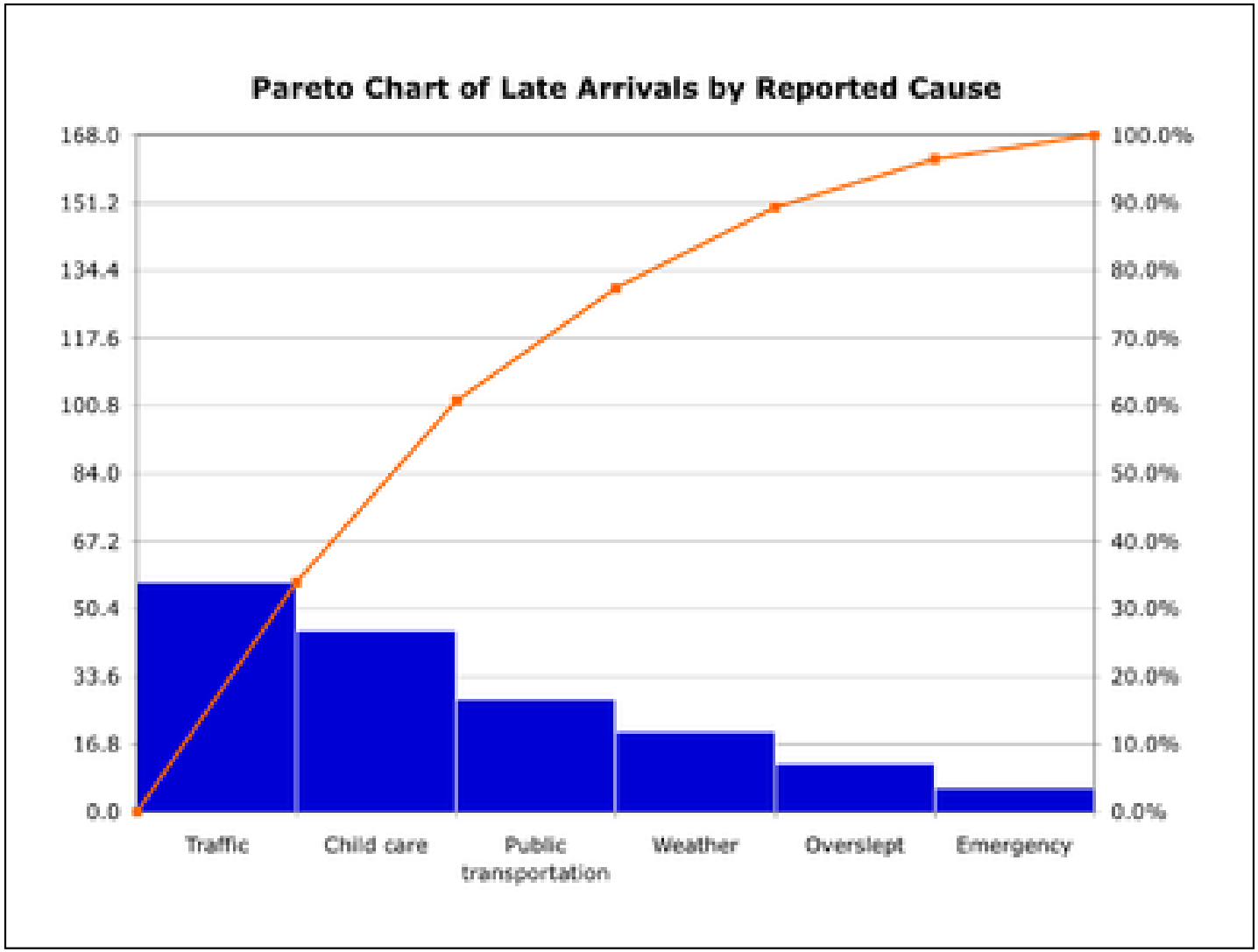
Box Plot



Pareto Diagram

Pareto Diagram

Order by decreasing frequency

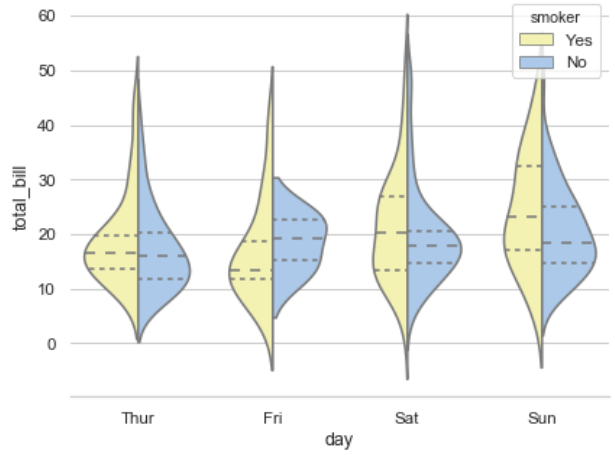


Violin Plot

Violin Plot

shows frequency too

Grouped violinplots with split violins



Python source code: [download source: grouped_violinplots.py]

```
import seaborn as sns
sns.set(style="whitegrid", palette="pastel", color_codes=True)

# Load the example tips dataset
tips = sns.load_dataset("tips")

# Draw a nested violinplot and split the violins for easier comparison
sns.violinplot(x="day", y="total_bill", hue="smoker",
               split=True, inner="quart",
               palette={"Yes": "y", "No": "b"},
               data=tips)
sns.despine(left=True)
```

Normalization and Z-scores

Normalization of Numbers

means getting them on the same scale

so they can be compared *apples* to *apples*

eg use frequency rather than count

eg use Z-scores of a normal distribution
to allow for different mean and variance

Comparing Two Attributes

Adapted from Frank E. Harrell Jr. on graphics:

http://biostat.mc.vanderbiltedu/twiki/pub/Main/StatGraphCourse/graphscourse.pdf

Two categorical variables

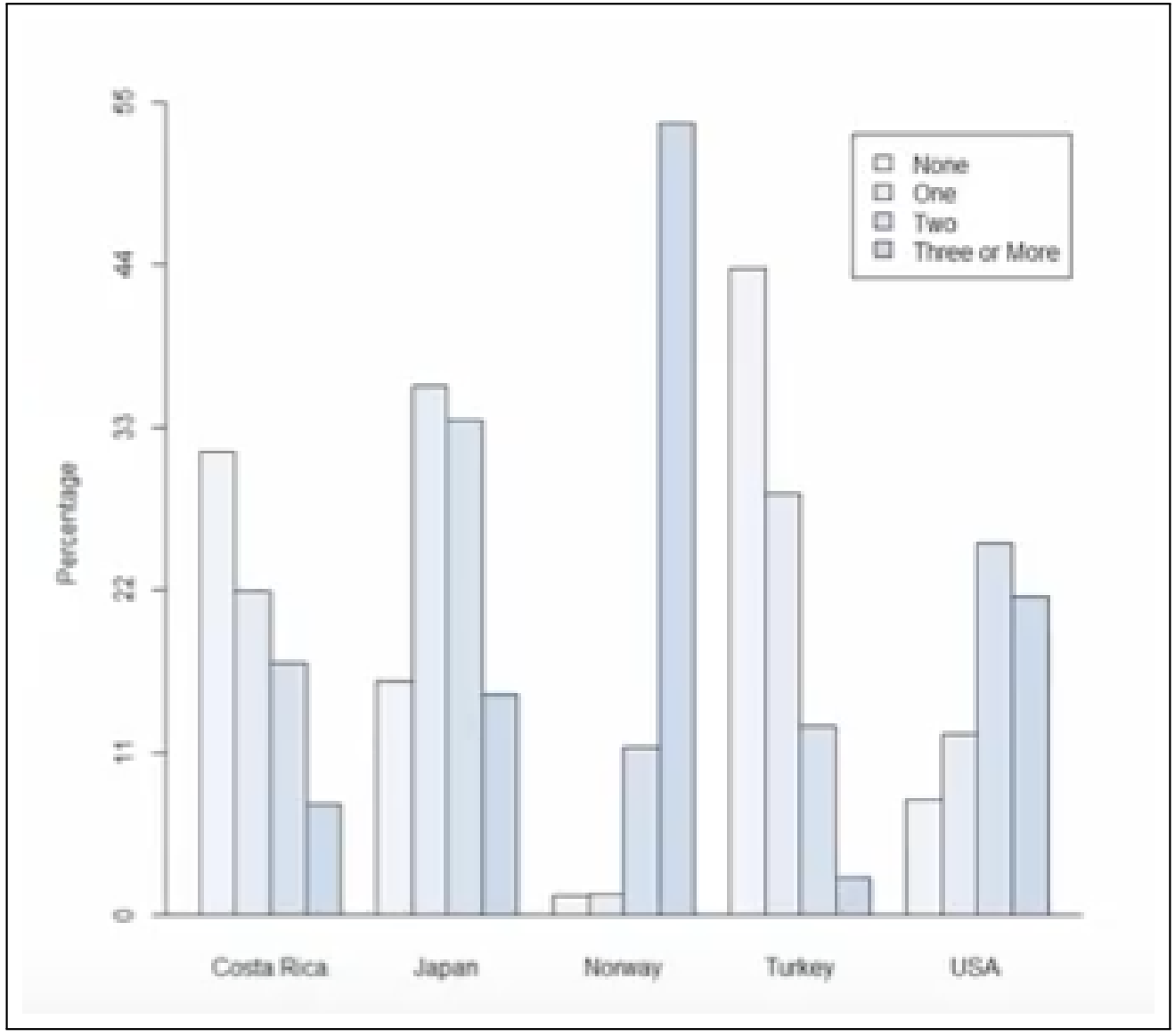
- Use frequency table
  - One categorical variable and other continuous variable
- Box plots of continuous variable values for each category of categorical variable
- Side-by-side dot plots (means + measure of uncertainty, SE or confidence interval)
  - Do not link means across categories!

Two continuous variables

- Scatter plot of raw data if sample size is not too large
- Prediction with confidence bands

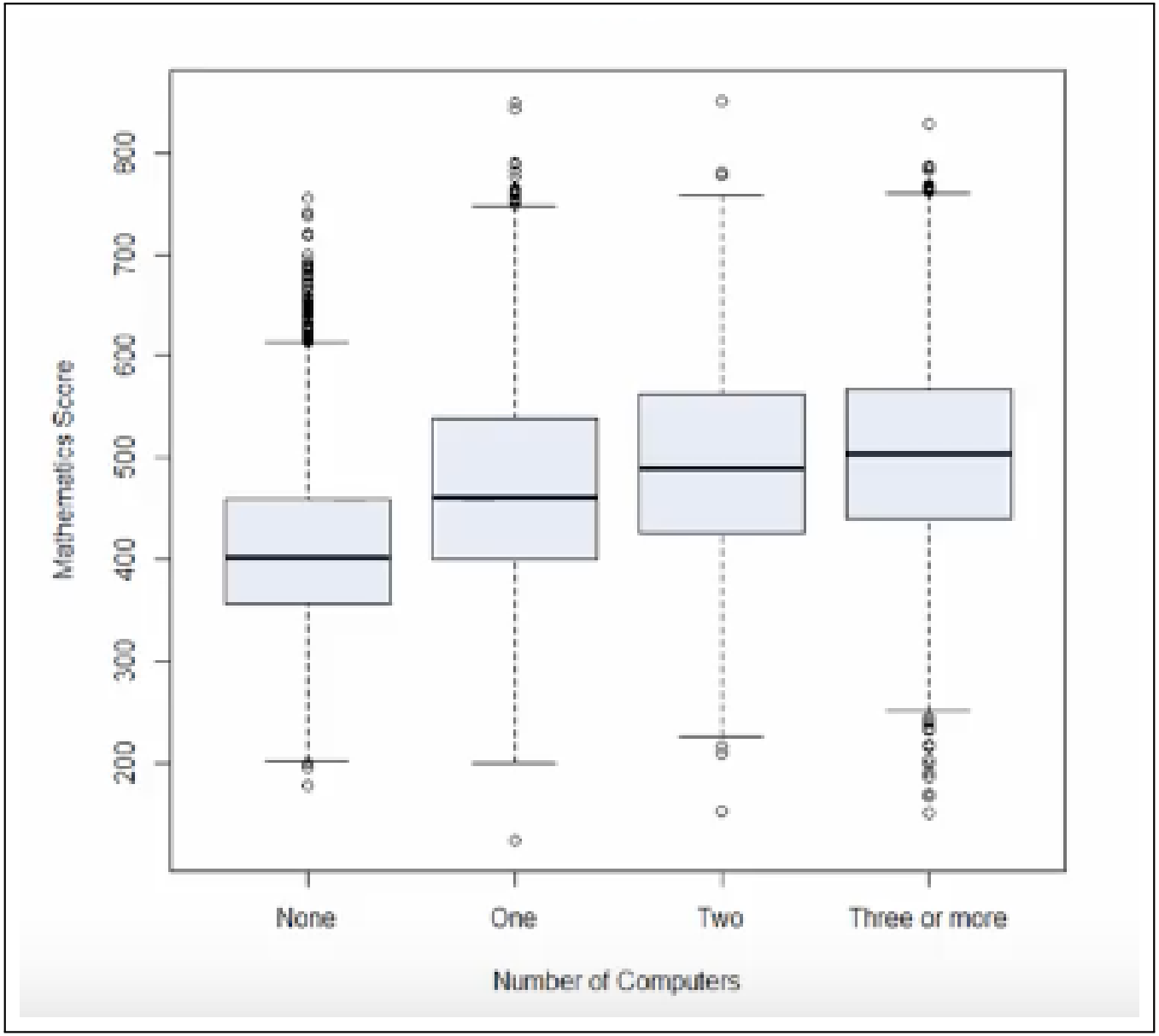
Comparing Two Attributes

Compare categorical and categorical



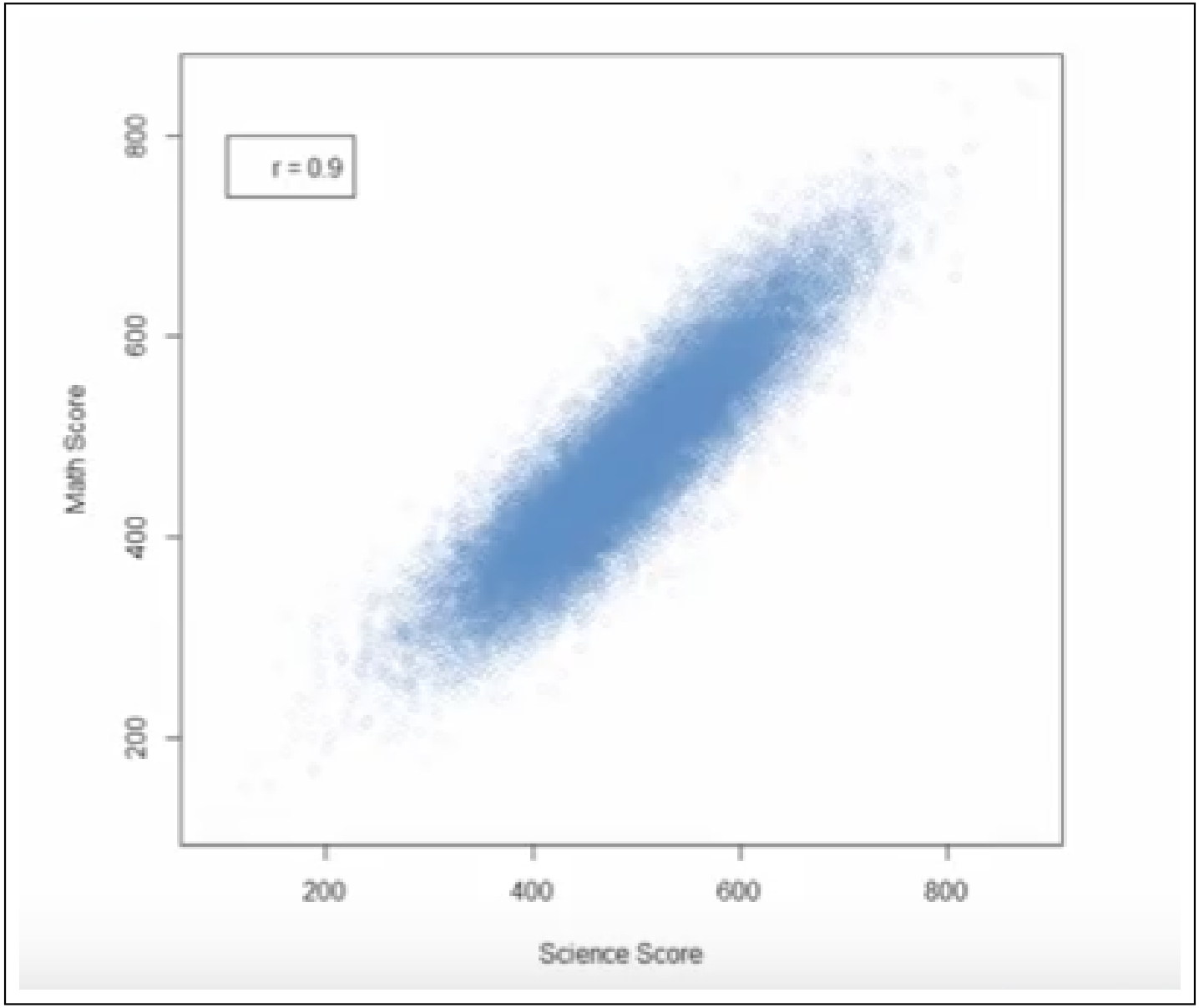
Comparing Two Attributes

Compare categorical and continuous



Comparing Two Attributes

Compare continuous and continuous



Correlation is not Causality

These are different concepts
and
correlation does not imply causality