Exploratory Data Analysis

DS 203 Programming for Data Science

Amit Sethi, EE, IITB

Learning outcomes

- Define exploratory data analysis
- Perform basic EDA of single variables
- Perform basic EDA of pairs of variables
- Select EDA appropriate to the type of variable

EDA is about taking stock of data

- Understand the main characteristics of your data to plan for downstream analyses

- Spot any issues with the data early

- Think about the type of analysis techniques, approaches, and the experts needed

What do we analyze in EDA

- The entire data at a glance
- Each variable in isolation
- Pairs of variables

Types of questions about the entire data

- Number of samples
- Number of variables per sample
- Samples with missing variables
- Corrupted samples

Hypothetical dataset

Make	Model	Year	kmpl	Top-speed	0-60 kmph	Drivability
Hyundai	i-20	2017	18	120	13s	"3"
Hyundai	i-20	2018	17	130	11s	"4"
Hyundai	i-20	2019	19	130	13	"3"
Hyundai	i-10	2017	20	120	12s	"4"
Hyundai	i-10	2018	19	130	10	"5"
Hyundai	i-10	2019	20	120	12	"4"
…	…	…	…	…	…	…
…	…	…	…	…	…	…
Datsun		2019	20	110	15	"2"
w•ÿ	Baleno	2019	20	120	17	"3"
	Nano	2018	30	80	55	"2"

Types of questions about each variable

- Type and coding
  - Nominal (may be coded as numerical)
  - Ordinal (may be coded as numerical)
  - True numerical (integer, quantized, float)
- Distribution
  - Descriptive statistics
  - Histograms
- Utility and ethics
  - Variability
  - Availability
  - Should it be used?

Type and coding of variables can be different

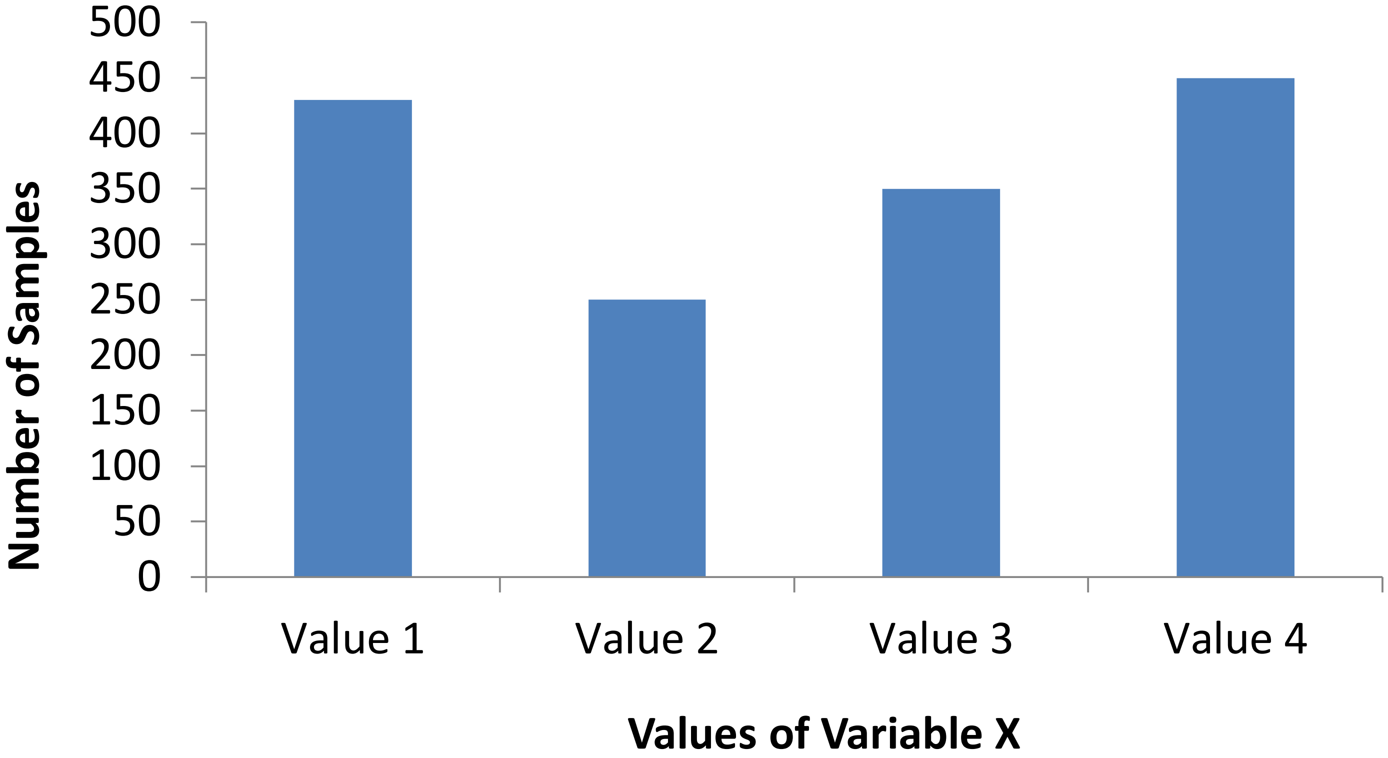
- Integers can be used to code:
  - Nominal / Categorical (species, postal codes)
  - Binary categorical (face or not-face)
  - Ordinal (very good, good, normal, bad, very bad)
  - Numerical (age in years)
  - Temporal (date)
- Text can be used to code:
  - Nominal / Categorical (species, postal codes)
  - Numerical saved as text
  - Temporal saved as text (“Sept 5, 2020”)

Description of discrete variables

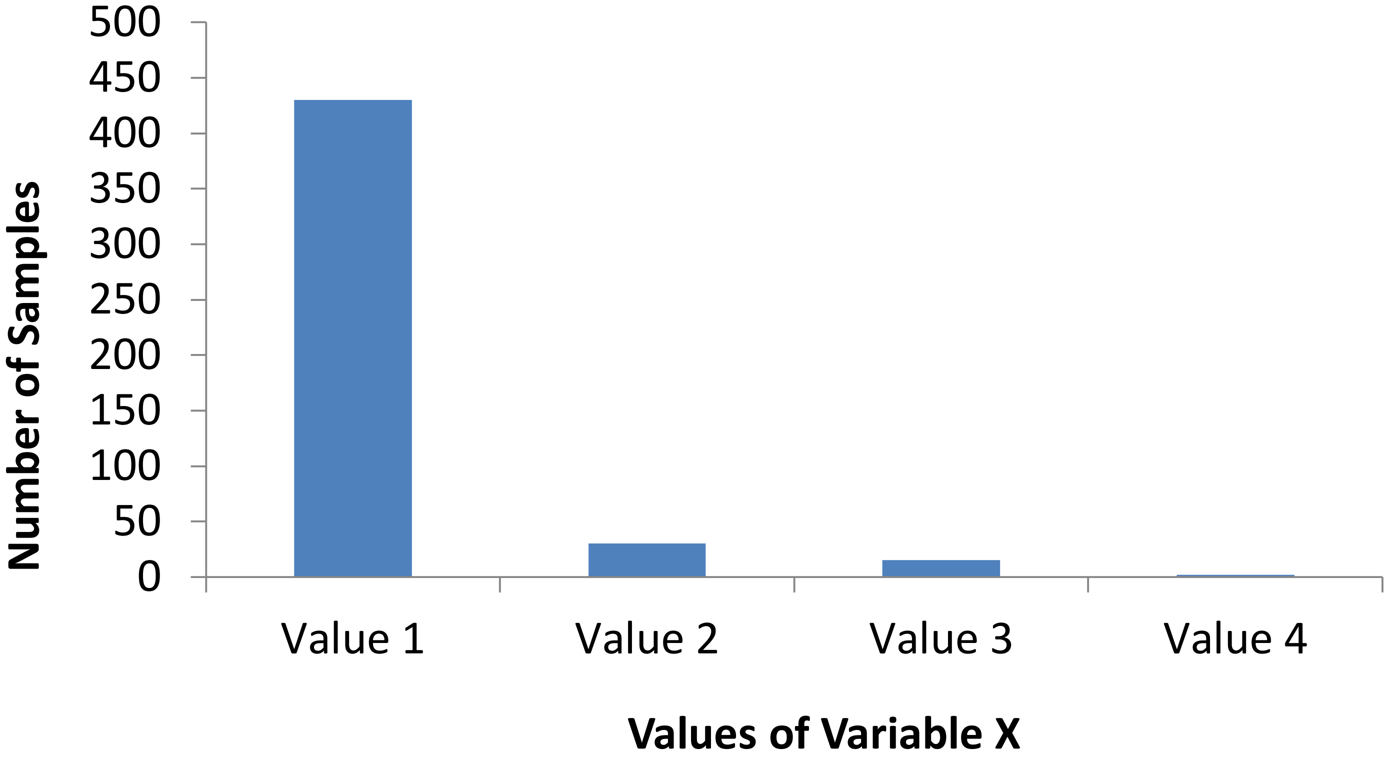
- List of unique values

- Order of values for ordinal

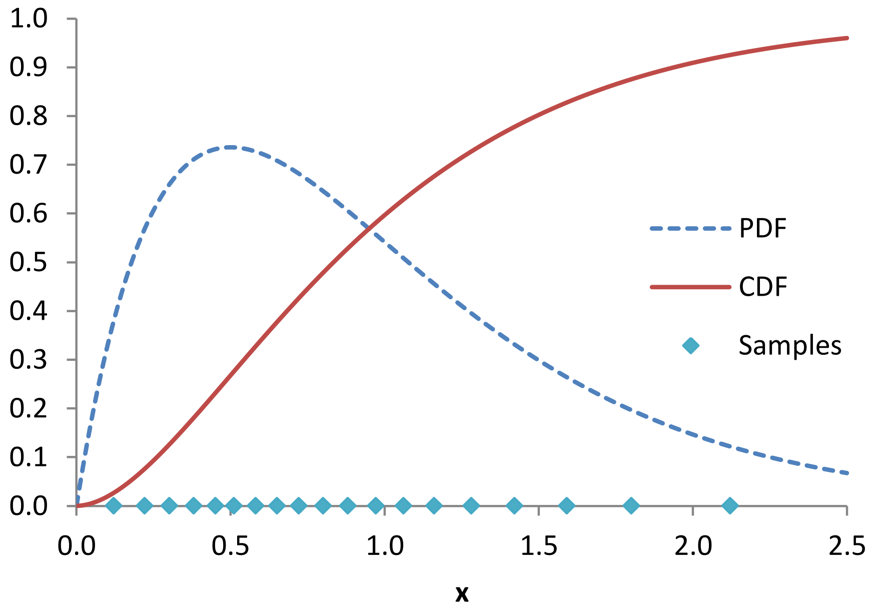
Histogram of a discrete variable is like a probability mass function



Histogram can indicate problems



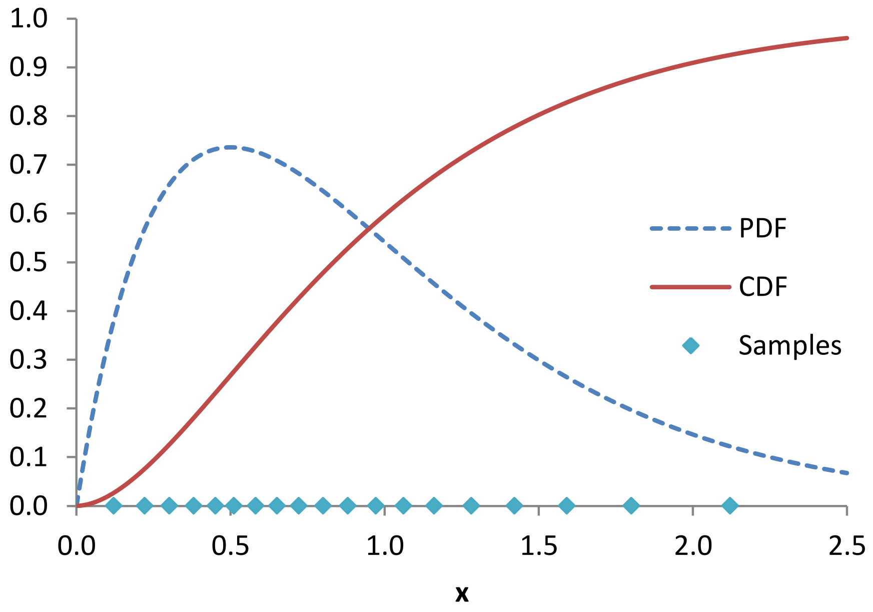
A continuous variable is described by its probability density function



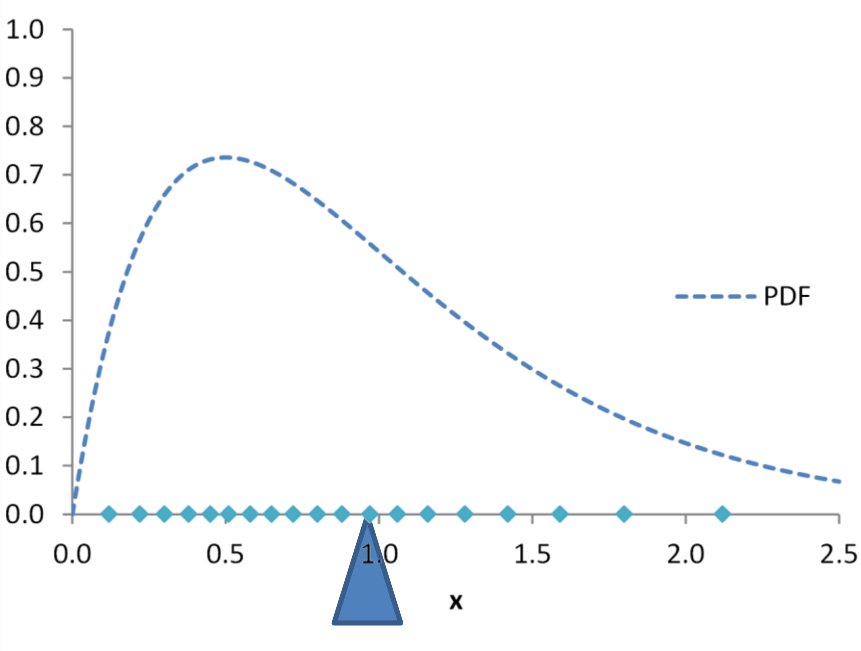
CDF is the integral of the PDF

It is monotonic, and rises from 0 to 1

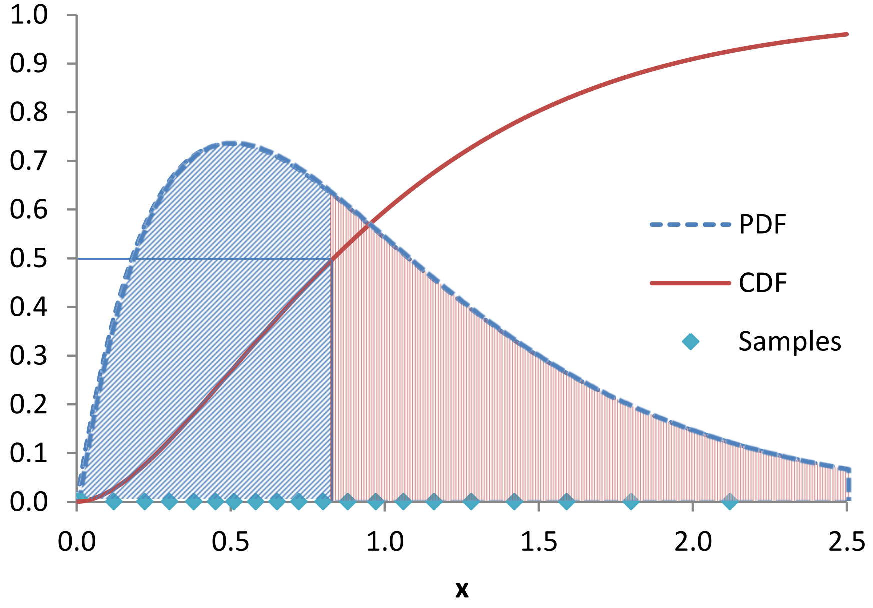
A continuous variable is sampled



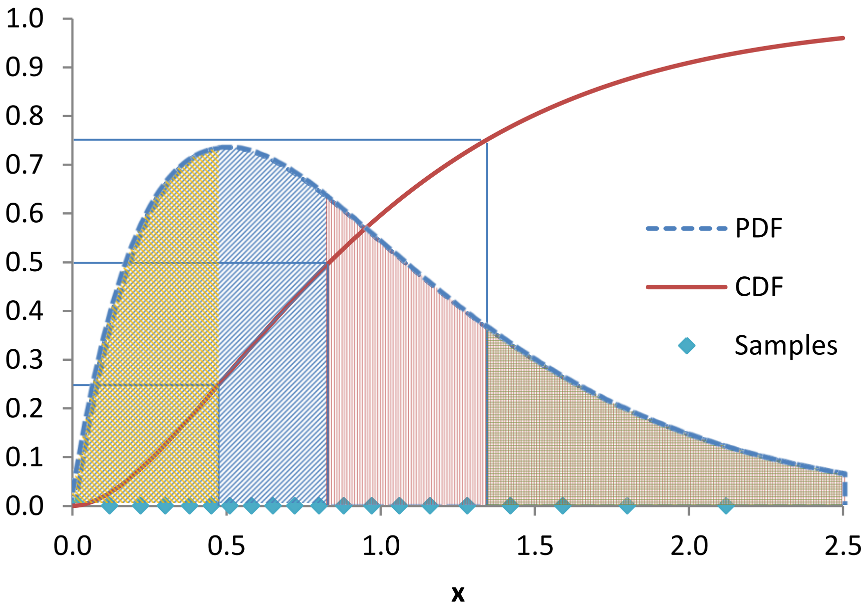
Mean is center of gravity of the PDF; Sample mean is not population mean



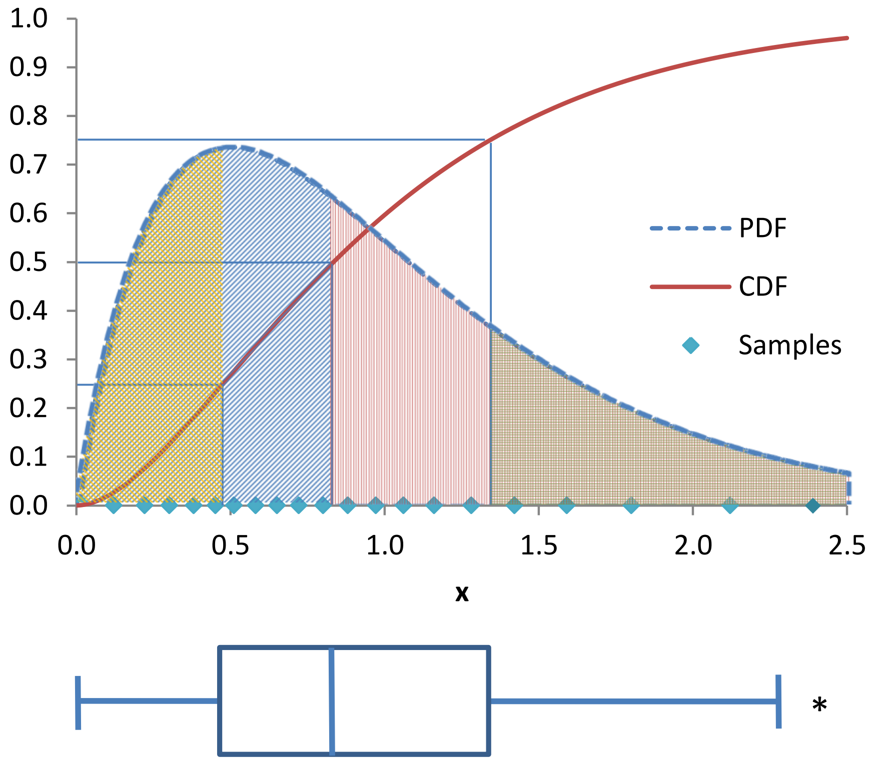
Median divides the PDF into two equal areas



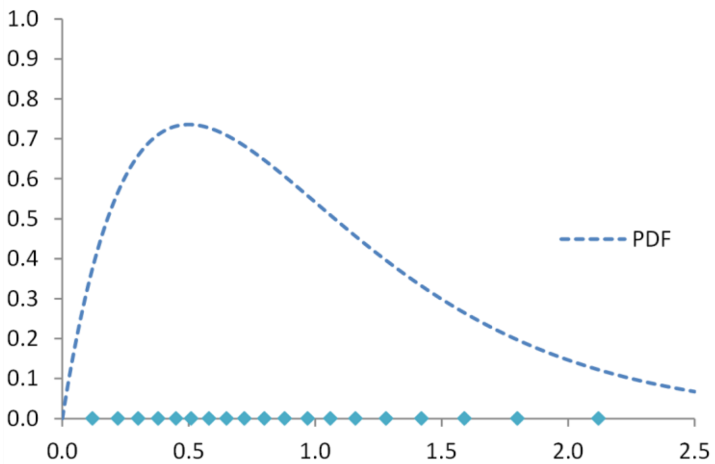
Quartiles divides the PDF into four equal areas

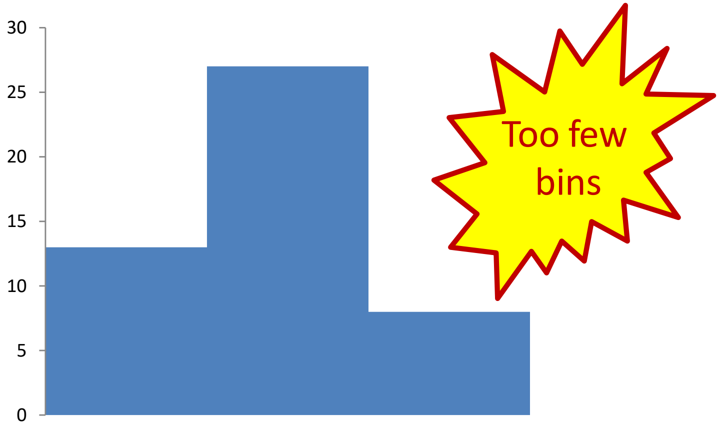


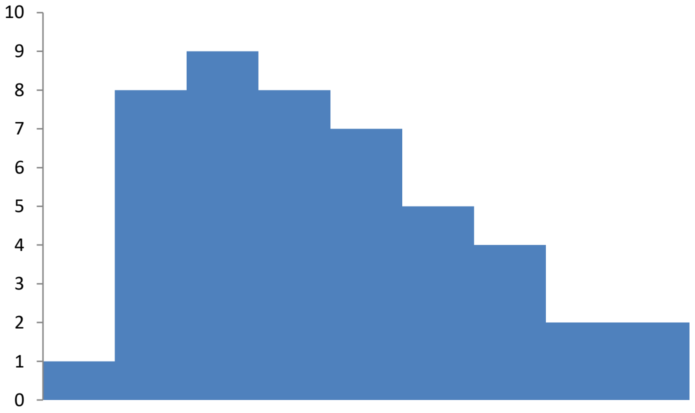
Box and whiskers plot summarizes the PDF

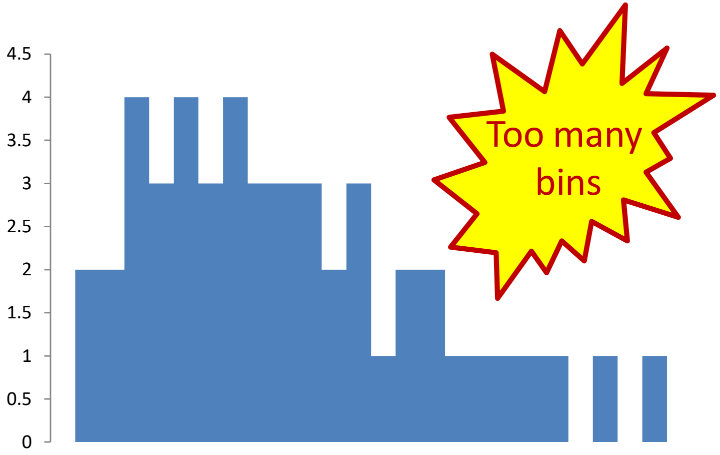


Histogram divides the range into discrete bins for counting samples









Types of questions about pairs of variables

- Relation between variables
  - Are some variables correlated?
  - Are there other strong relations between variables?
  - Are some variables redundant?

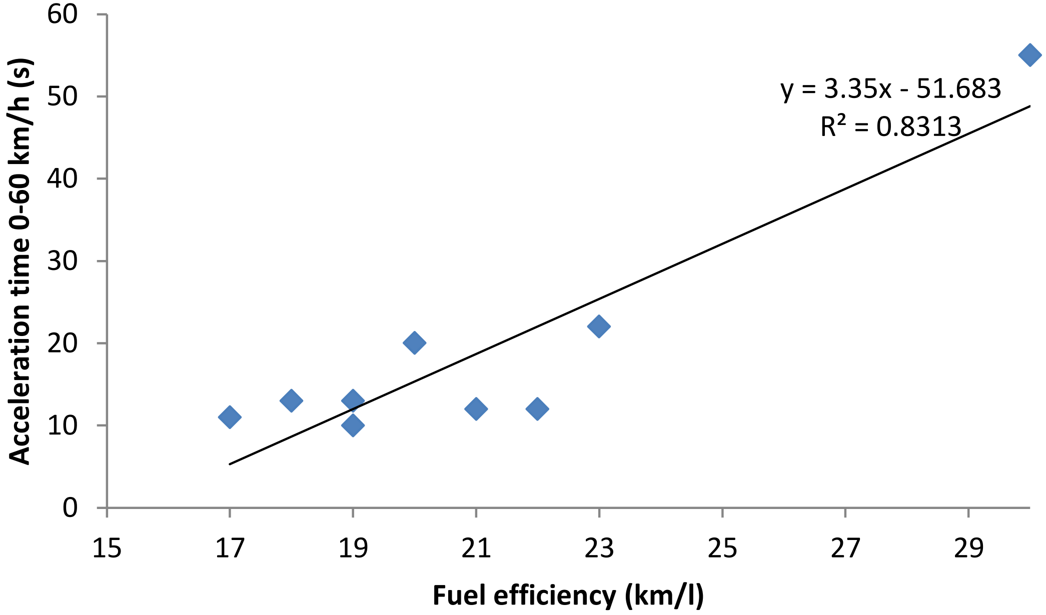
Cross-tab is viewed between discrete variables

Example: “Economic satisfaction” survey respondents by income and gender

Income↓; Gender →	Male	Female
None	0	0
Low	300	0
Medium	10,000	3,000
High	5,000	2,000

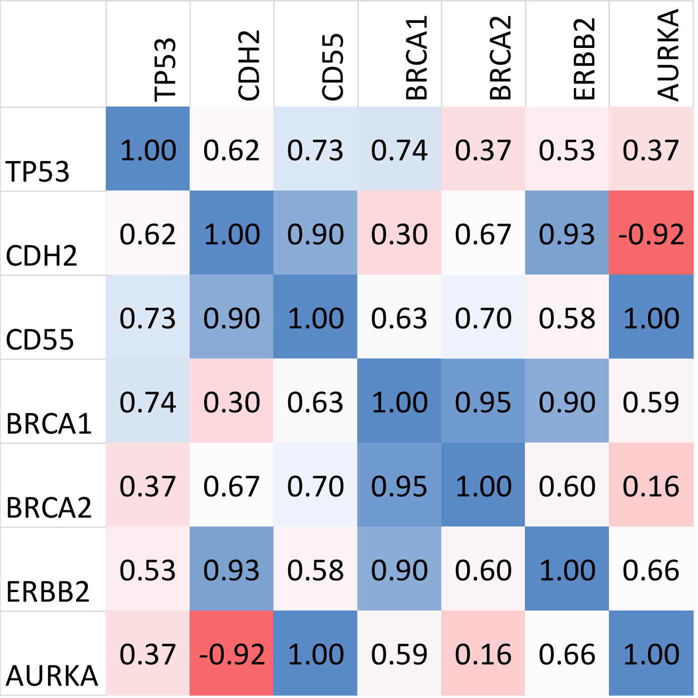
- No women with “low” income?
- Very few men with “low” income?
- Is there a sampling bias (e.g. email survey)?

Correlation and scatter plots are between pairs of continuous variables



- There is an outlier; otherwise, the relation is not strong, which indicates hidden factors

Correlation matrix can be computed for all continuous variables together



	TP53	CDH2	CD55	BRCA1	BRCA2	ERBB2	AURKA
TP53	1.00	0.62	0.73	0.74	0.37	0.53	0.37
CDH2	0.62	1.00	0.90	0.30	0.67	0.93	-0.92
CD55	0.73	0.90	1.00	0.63	0.70	0.58	1.00
BRCA1	0.74	0.30	0.63	1.00	0.95	0.90	0.59
BRCA2	0.37	0.67	0.70	0.95	1.00	0.60	0.16
ERBB2	0.53	0.93	0.58	0.90	0.60	1.00	0.66
AURKA	0.37	-0.92	1.00	0.59	0.16	0.66	1.00

- Highly (positively or negatively) correlated variables can create problems later