NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 13: Lecture 2
Regression

Center for
Data Science

NYU

How can we connect the points in a scatter-plot to generate a line-chart matching the pattern in the data?

DS-UA 111
Data Science for Everyone

Week 13: Lecture 2

Regression

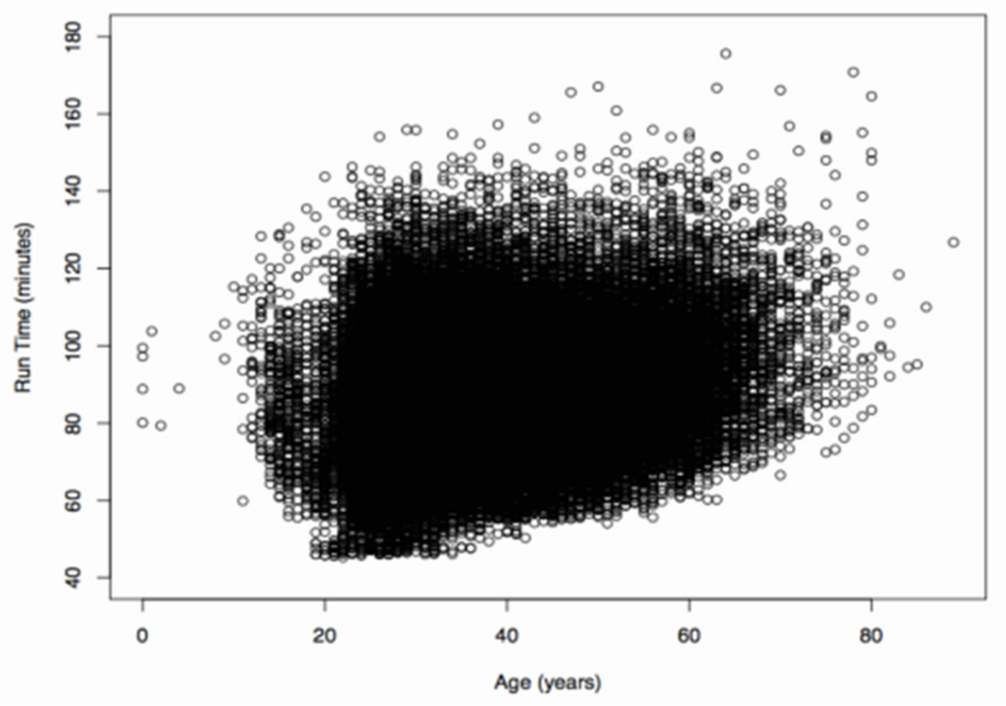
Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

- Please check Week 13 agenda on NYU Classes
  - Homework 3/4
  - Lab 8
  - Project Milestone
- Refer to the Calendar linked to NYU Classes

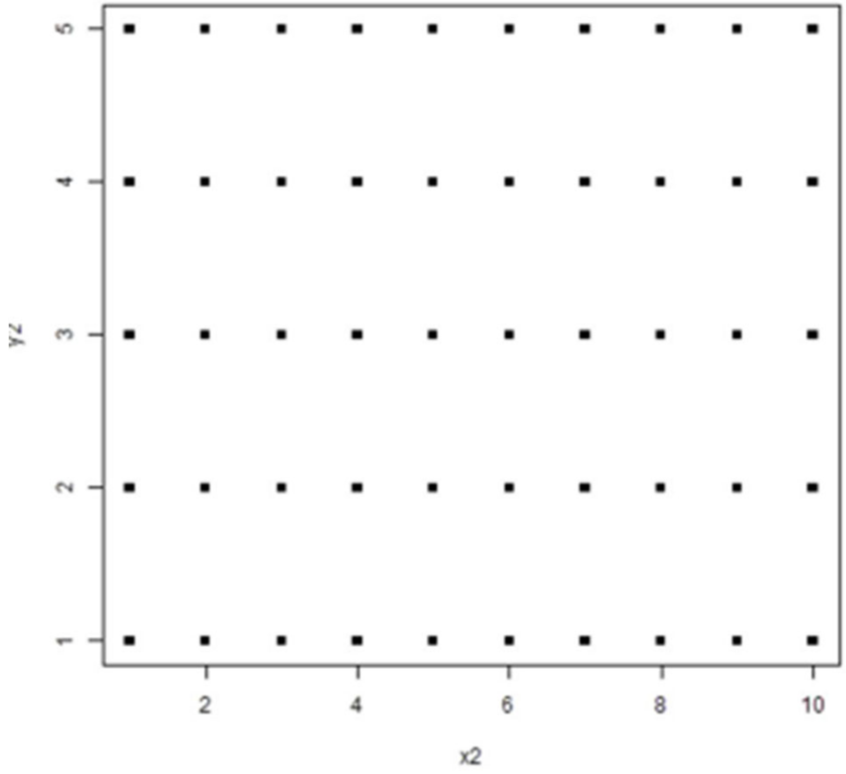
Review

- Scatter-plots allow use to visualize two quantitative variables
- Be careful of over-plotting
  - With duplicate values we will miss data in the chart
  - With nearly duplicate values we will have a cluttered chart



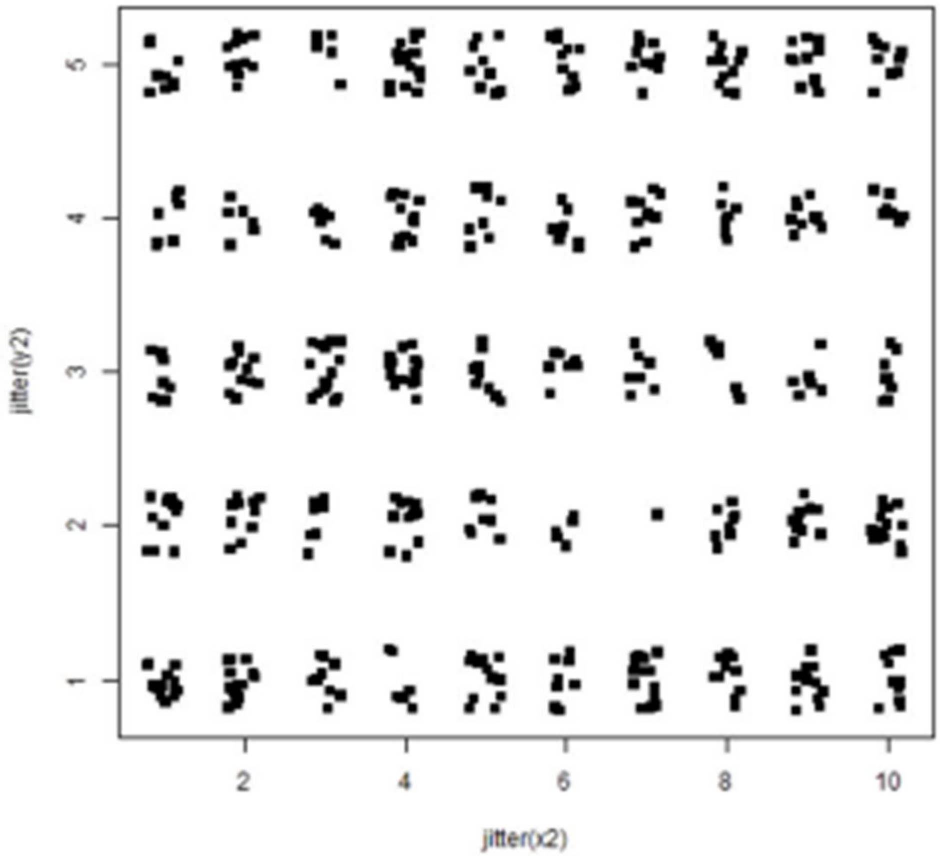
Review

- Scatter-plots allow use to visualize two quantitative variables
- Be careful of over-plotting
  - With duplicate values we will miss data in the chart
  - With nearly duplicate values we will have a cluttered chart



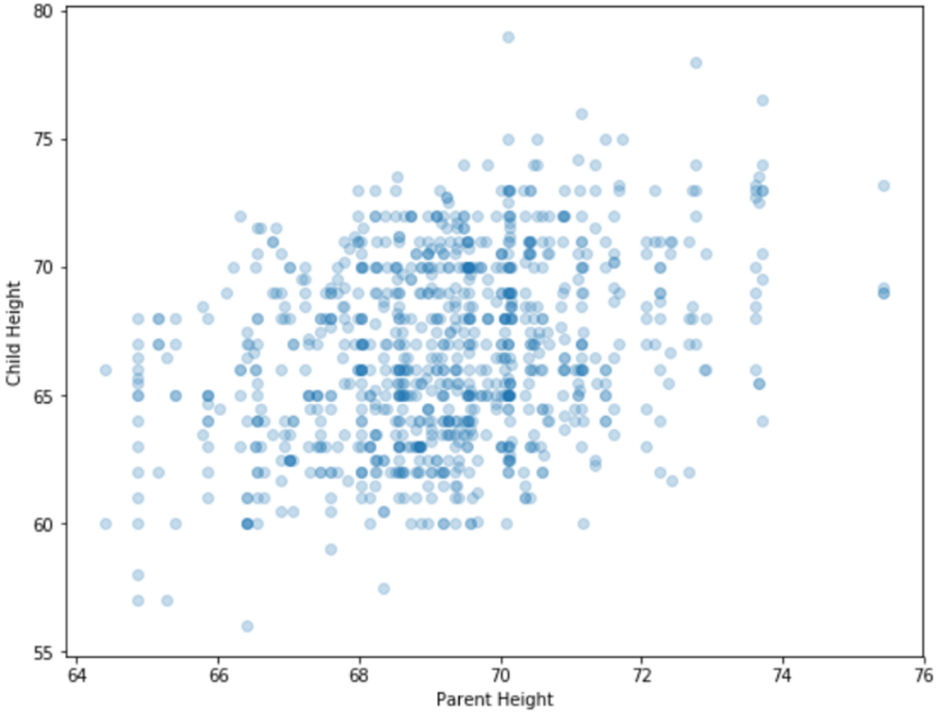
Review

- We can try to fix over-plotting in different ways
  - Jittering the points
  - Adjusting the saturation of the colors
  - Splitting the data between different charts
  - Grouping together the records. Calculating statistics like mean and median to summarize the data



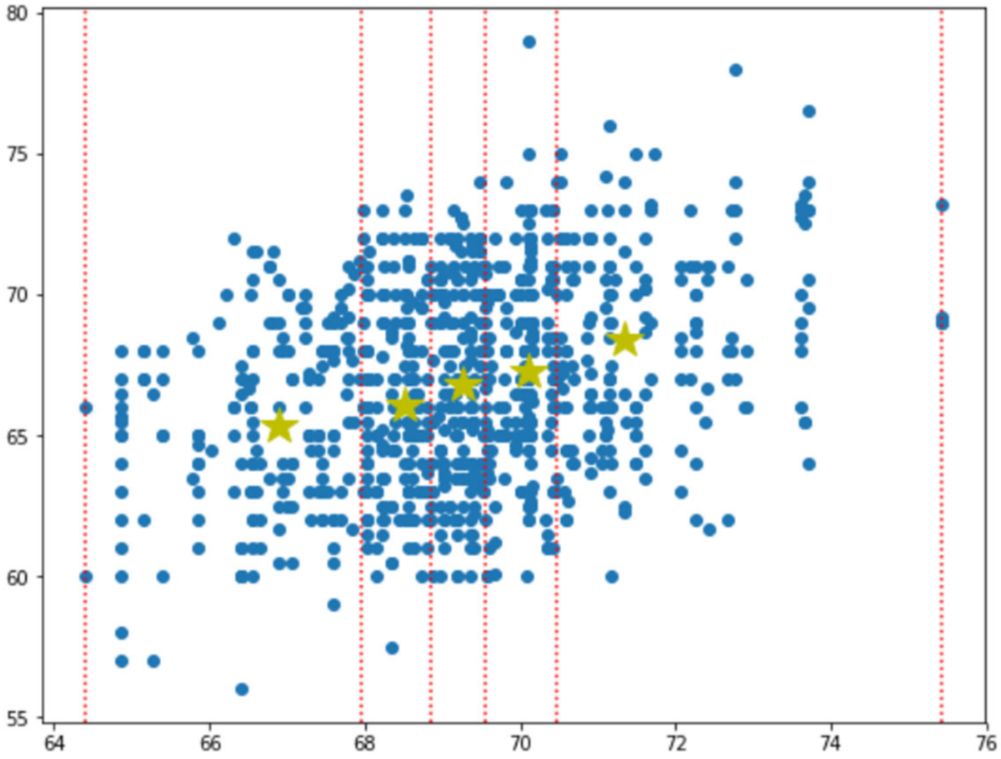
Review

- We can try to fix over-plotting in different ways
  - Jittering the points
  - Adjusting the saturation of the colors
  - Splitting the data between different charts
  - Grouping together the records. Calculating statistics like mean and median to summarize the data



Review

- We can try to fix over-plotting in different ways
  - Jittering the points
  - Adjusting the saturation of the colors
  - Splitting the data between different charts
  - Grouping together the records. Calculating statistics like mean and median to summarize the data

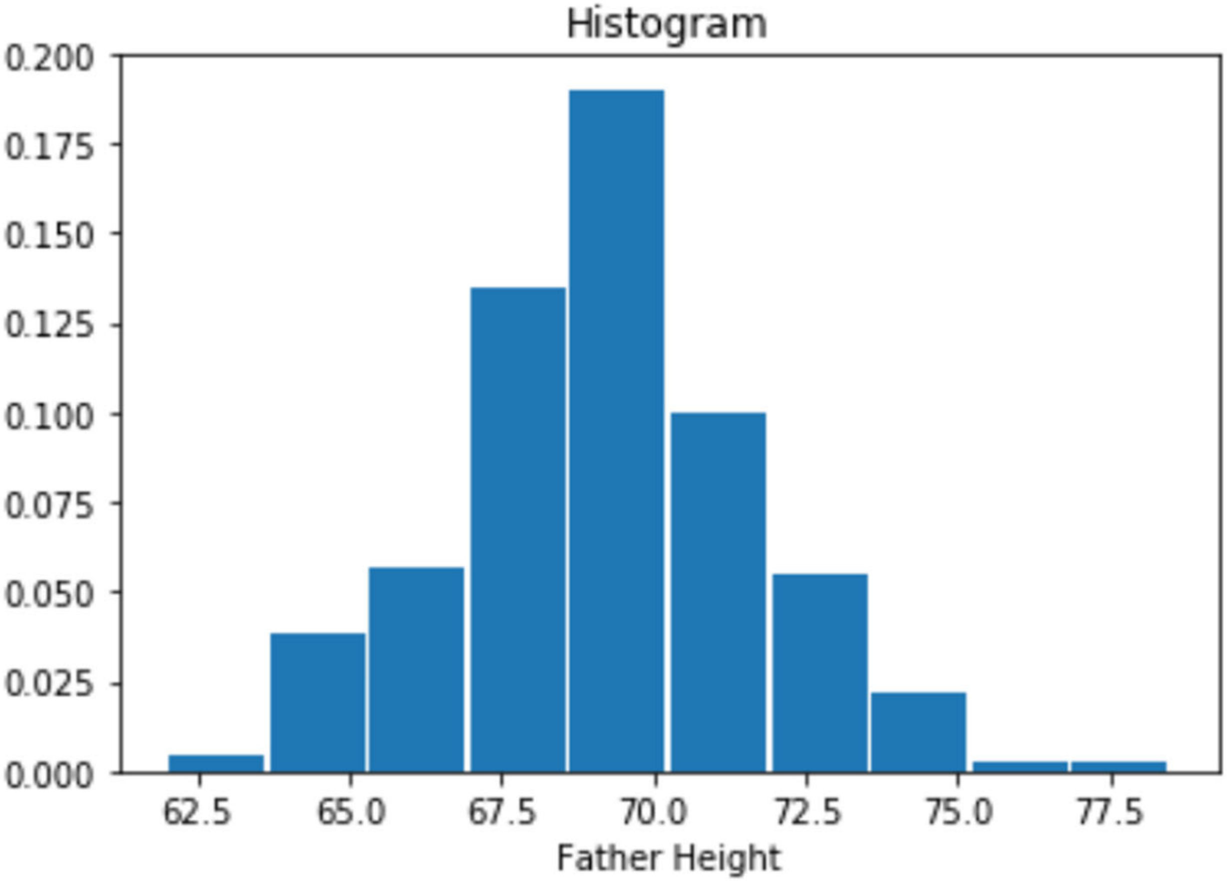


Exercise

- Which of the following plots can be used to depict a single qualitative variable?
  1. histograms
  2. bar chat
  3. box plots
  4. scatter plots
  5. line chart
- What are the functions for these charts in matplotlib?

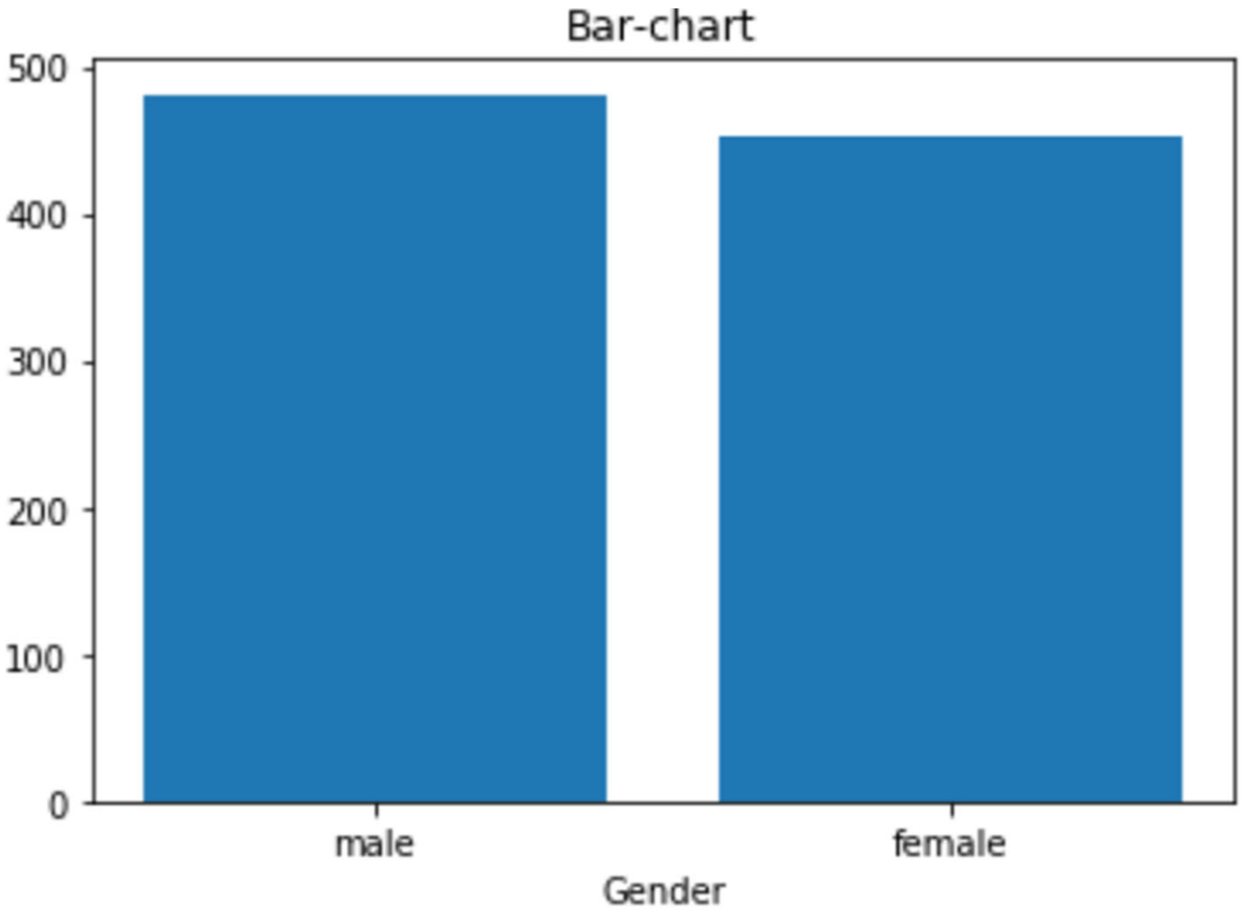
Exercise

plt.hist



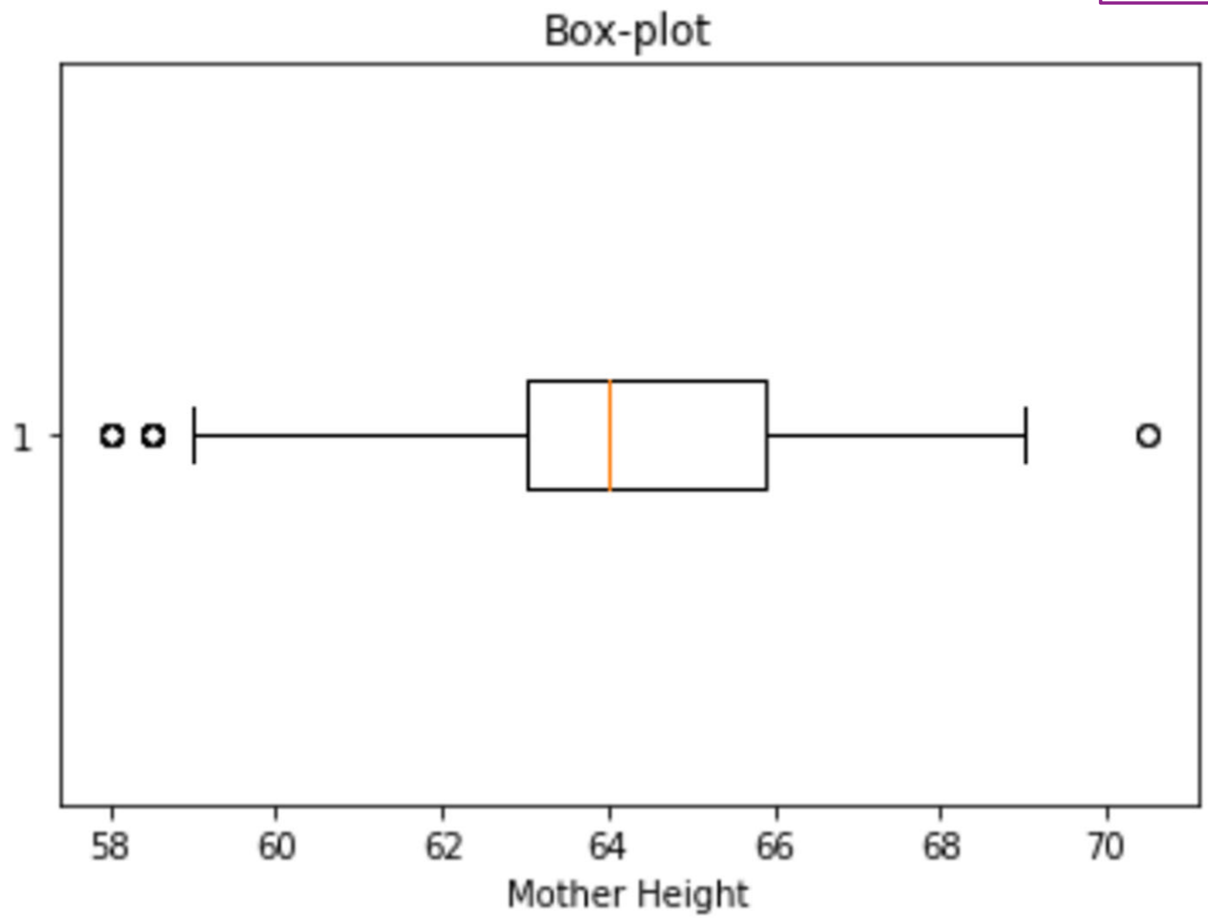
Exercise

plt.bar



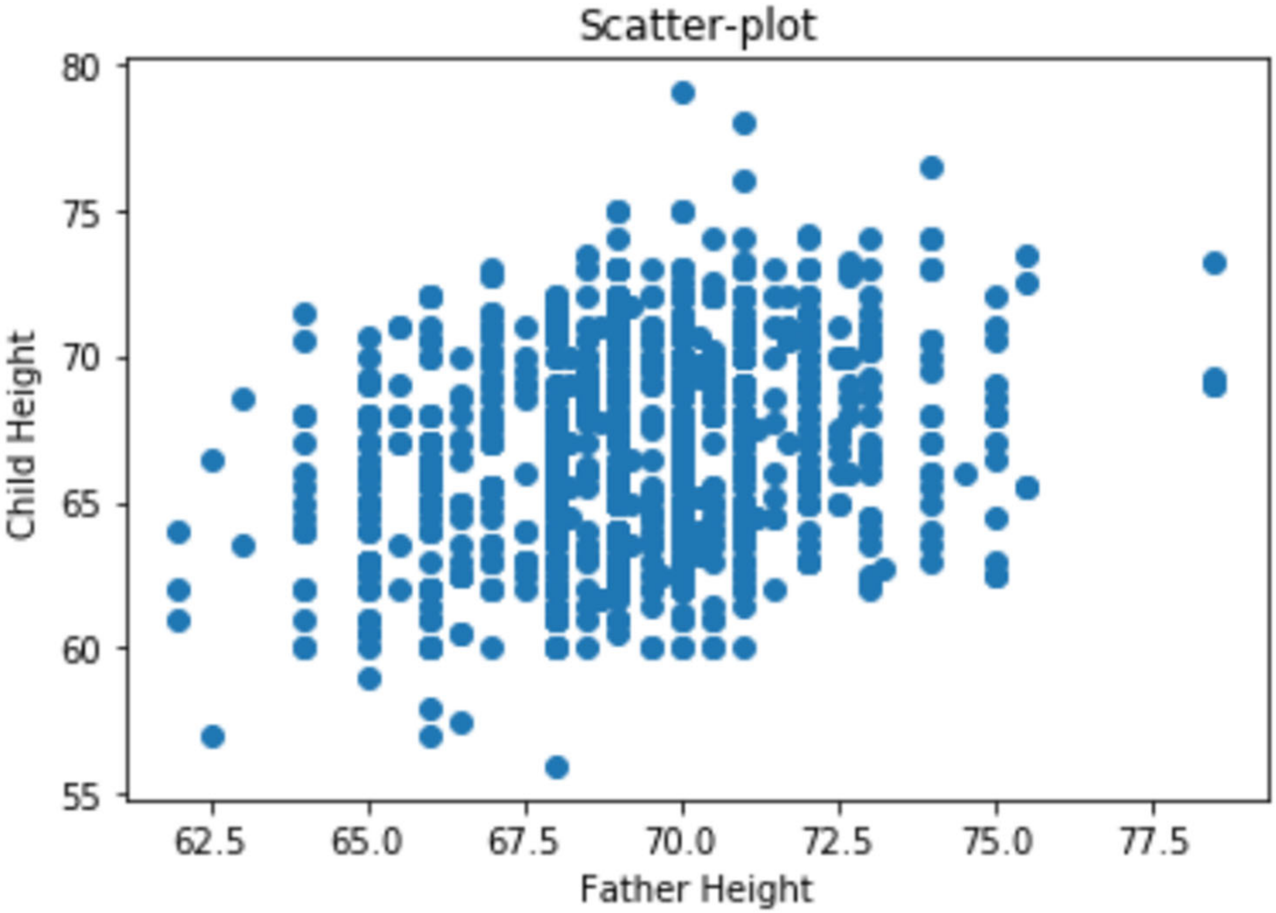
Exercise

plt.boxplot

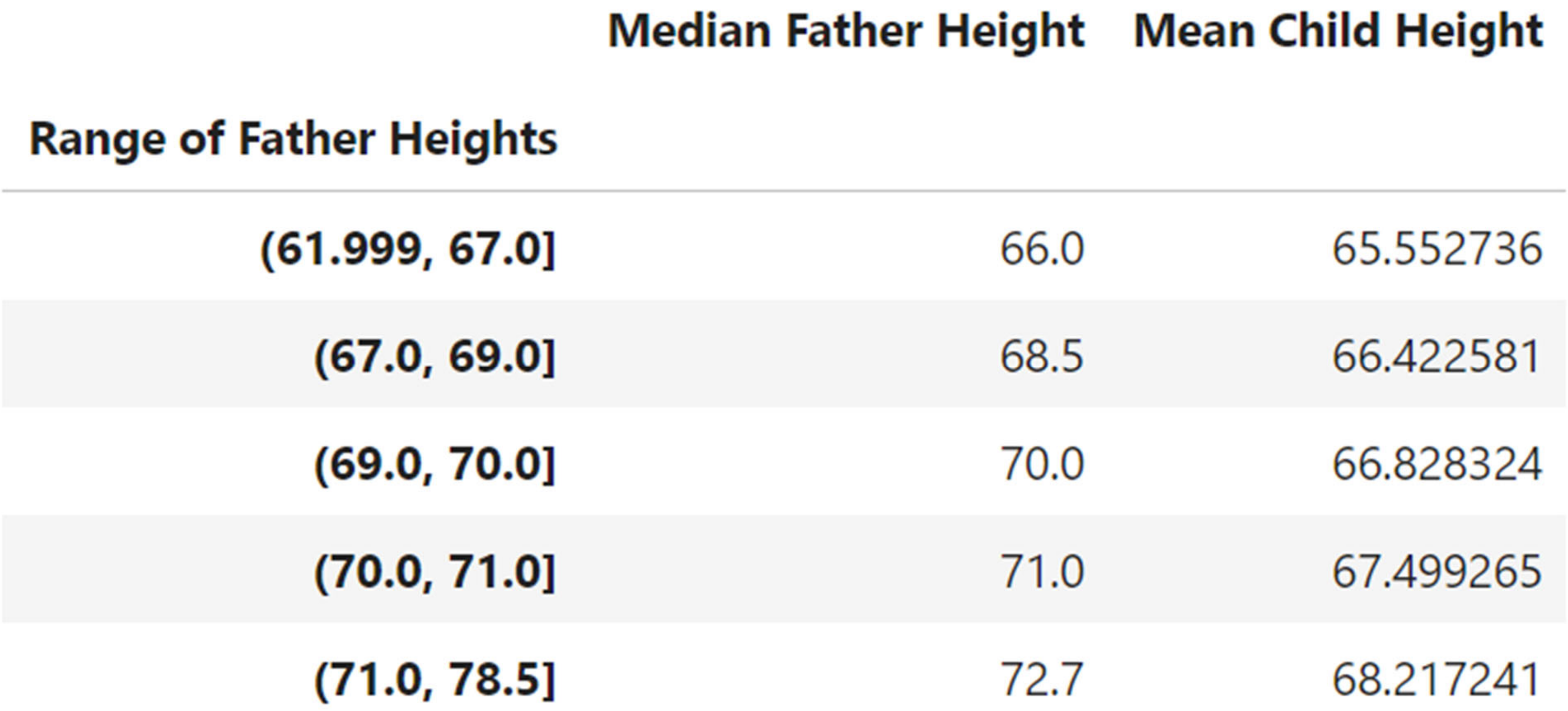


Exercise

plt.scatter



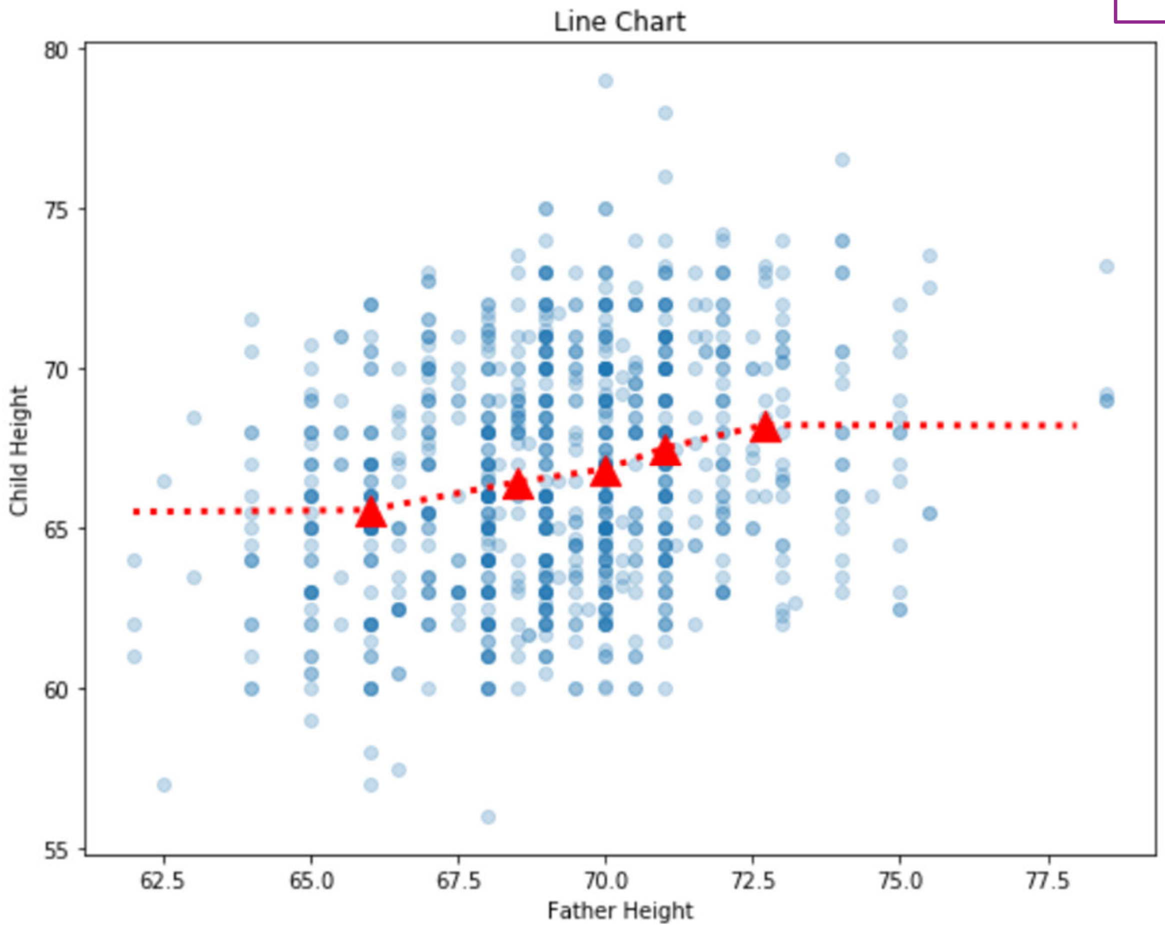
Exercise



	Median Father Height	Mean Child Height
Range of Father Heights		
(61.999, 67.0]	66.0	65.552736
(67.0, 69.0]	68.5	66.422581
(69.0, 70.0]	70.0	66.828324
(70.0, 71.0]	71.0	67.499265
(71.0, 78.5]	72.7	68.217241

Exercise

plt.plot



Agenda

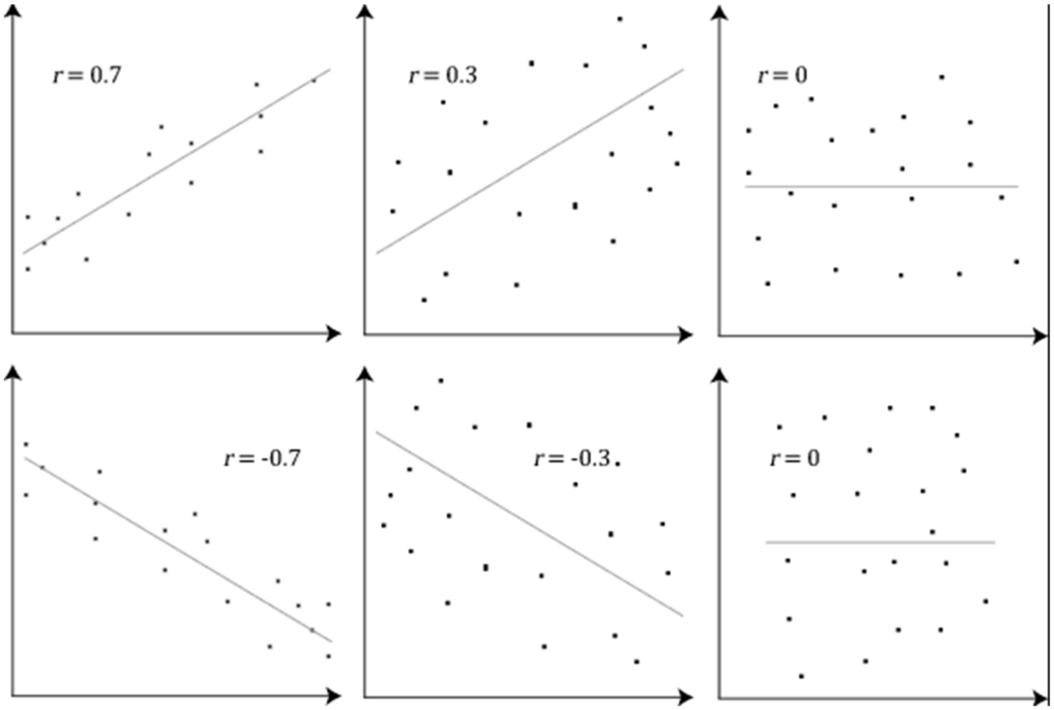
- Understanding Associations between Variables
  - Correlation
- Describing the Pattern in the Association between Variables
  - Regression

References

- Prediction
  - Chapter 15.2-15.4

Correlation

- Correlation measures the linear association between variables
- Linear means shaped like a line



Correlation

- Correlation comes from the transformation of the data to standard units
  - Average of…
  - Product of…
    - x in standard units and
    - y in standard units

Correlation

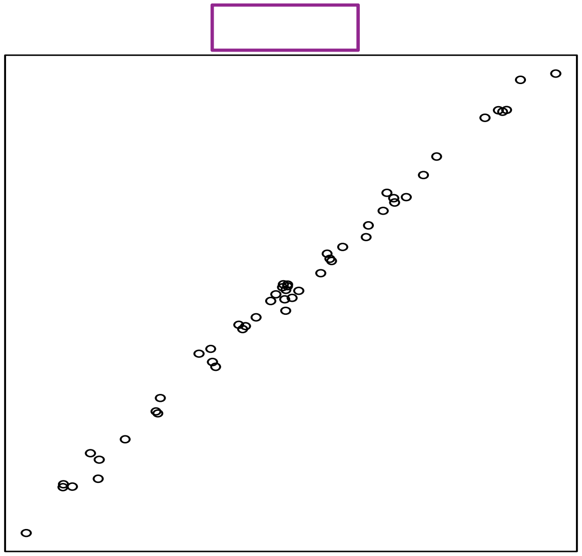
- Correlation comes from the transformation of the data to standard units
  - Average of…
  - Product of…
    - x in standard units and
    - y in standard units
- The x variable lies on the horizontal axis. A better name would be independent variable or explanatory variable
- The y variable lies on the vertical axis. A better name would be dependent variable or response variable

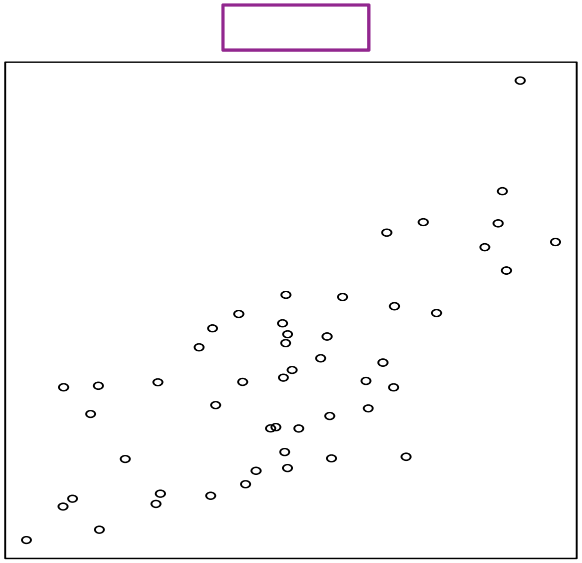
Correlation

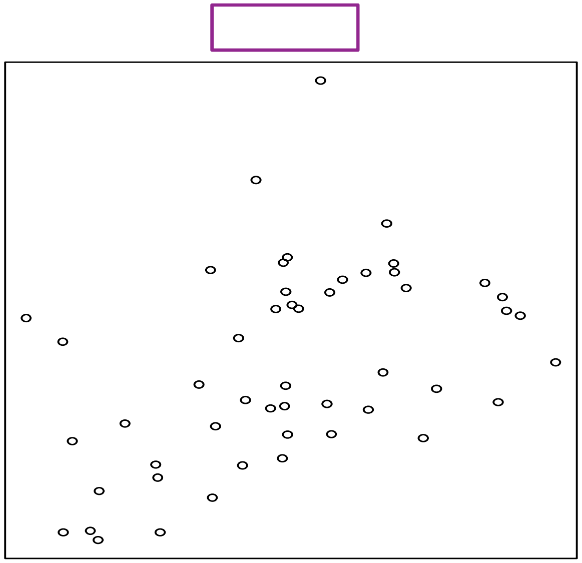
- The values of *r* range from $-1 \le r \le 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
  - $r = 0$: No linear association; uncorrelated

Exercise

- Match the correlations to possible values 0.95, 0.75, 0.50, 0.30, 0.10

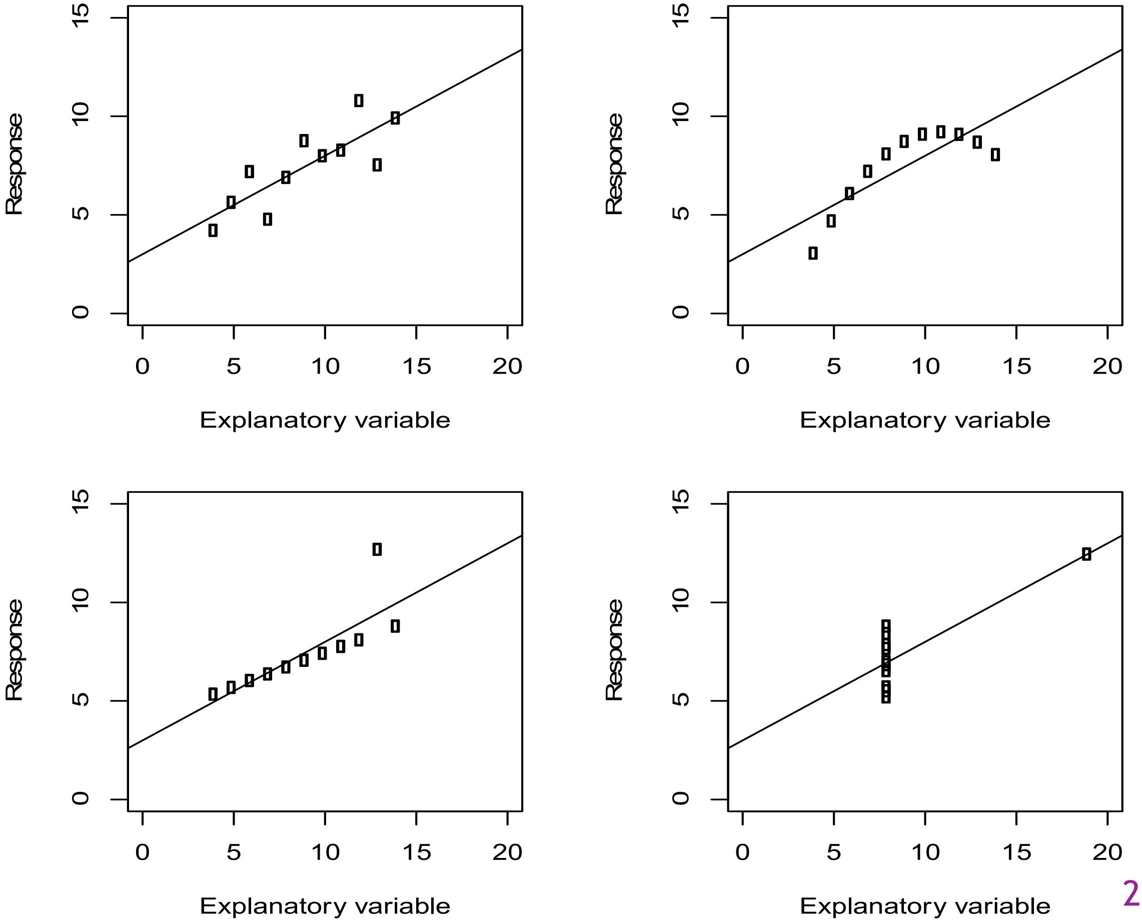






Limitations of Correlation

- We have four datasets with different patterns. However many statistics are equal Mean
    - 9 (explanatory)
    - 7.5 (response)
  - Standard deviation
    - 3.31 (explanatory)
    - 2.02 (response)



Regression

- Suppose we want to draw a line-chart through the scatter-plot to fit the pattern.
- The equation for a line is

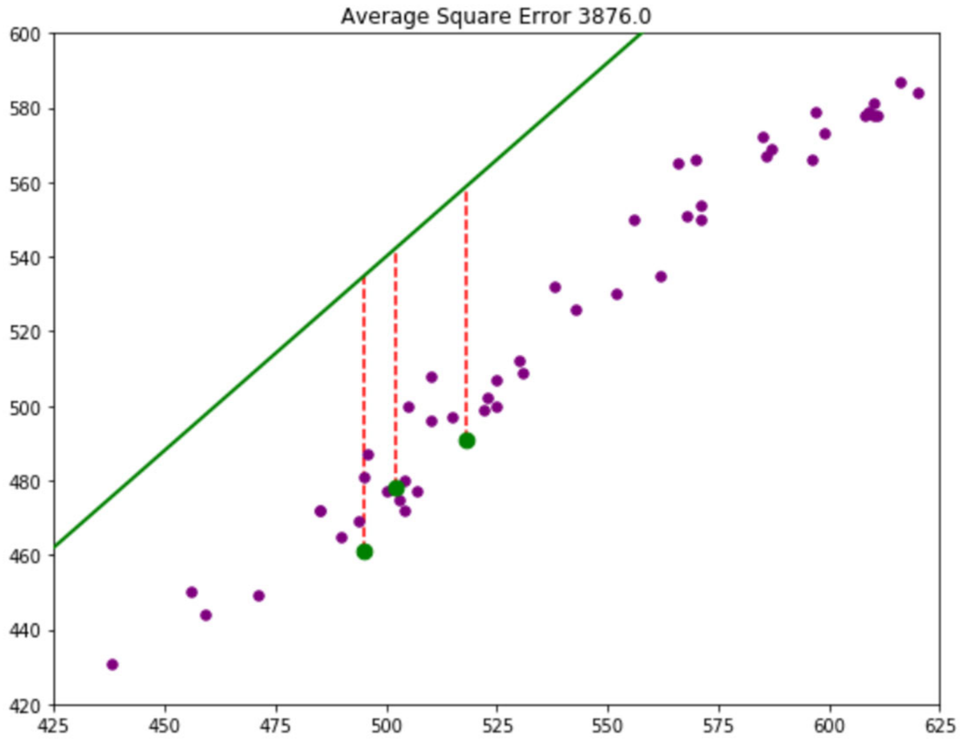
Output = Intercept + Slope * Input

- The correlation helps us determine the slope of the line in standard units

$$\underbrace{\frac{\text{estimate of } y \;-\; \text{average of } y}{\text{SD of } y}}_{\text{estimate of } y \text{ in standard units}} = r \times \underbrace{\frac{\text{the given } x \;-\; \text{average of } x}{\text{SD of } x}}_{x \text{ in standard units}}$$

Residuals

- We call the difference between observation and prediction a residual.
- In least squares regression we fit a line to the scatter-plot by minimizing the mean square error
- The mean square error is the average of the squared residuals.



Summary

- Understanding Associations between Variables
  - Correlation
- Describing the Pattern in the Association between Variables
  - Regression

Goals

- Understand limitations of correlation
- Use correlation to fit a line to a scatter-plot