

DATA SCIENCE

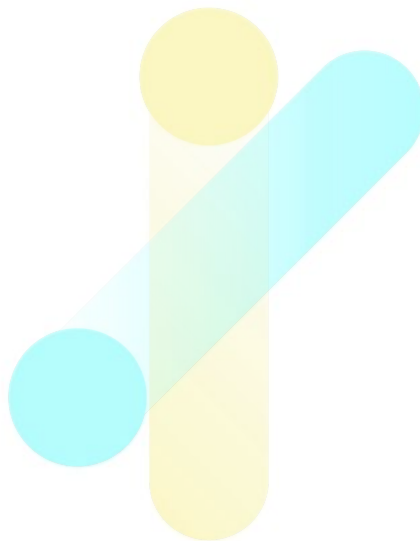
TRAINING PROGRAM

Lab 4: Hypothesis Testing and Visualization

Taneli Vähäkangas, Kodit.io

Dataset 1

- Real estate transactions
<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>
- Register to Kaggle and download
- Features:
 - SalePrice: sale price in dollars (target variable to predict)
 - LotArea: lot size in square feet
 - Utilities: type of utilities available
 - BldgType: type of dwelling
 - YearBuilt: original construction date
 - ...



Lab 4: Hypothesis Testing

- Which of these hypotheses make sense? Can they be tested?
 - Bigger apartments are more expensive
 - Is this the default position? What would be the alternate position?
 - Could we use dataset 1 to test the hypothesis?
 - How strong evidence do we want for our hypothesis?
 - There is a linear correlation between size and price
 - Why would this be a default position?
 - Is this a testable hypothesis?
- Select some of the categorical variables and formulate a hypothesis
 - Is your hypothesis testable?
 - Do you find support for rejecting the null hypothesis?
 - What are the populations?

Lab 4: Visualization

- How would you show (not tell) the effect of different attributes
 - On the size of an apartment?
 - On the price of an apartment?
 - Regarding each other?
- How would you visualize development of prices over time?
 - Or maybe the changes in pool sizes over time? Or number of above grade bathrooms ...
- Housing economy theory says that apartment size affects its price per area
 - Can you visualize if this relation exists?
- Would these be easier to implement using Plot.ly or Highcharts services?
 - Instead of matplotlib or seaborn or bokeh libraries