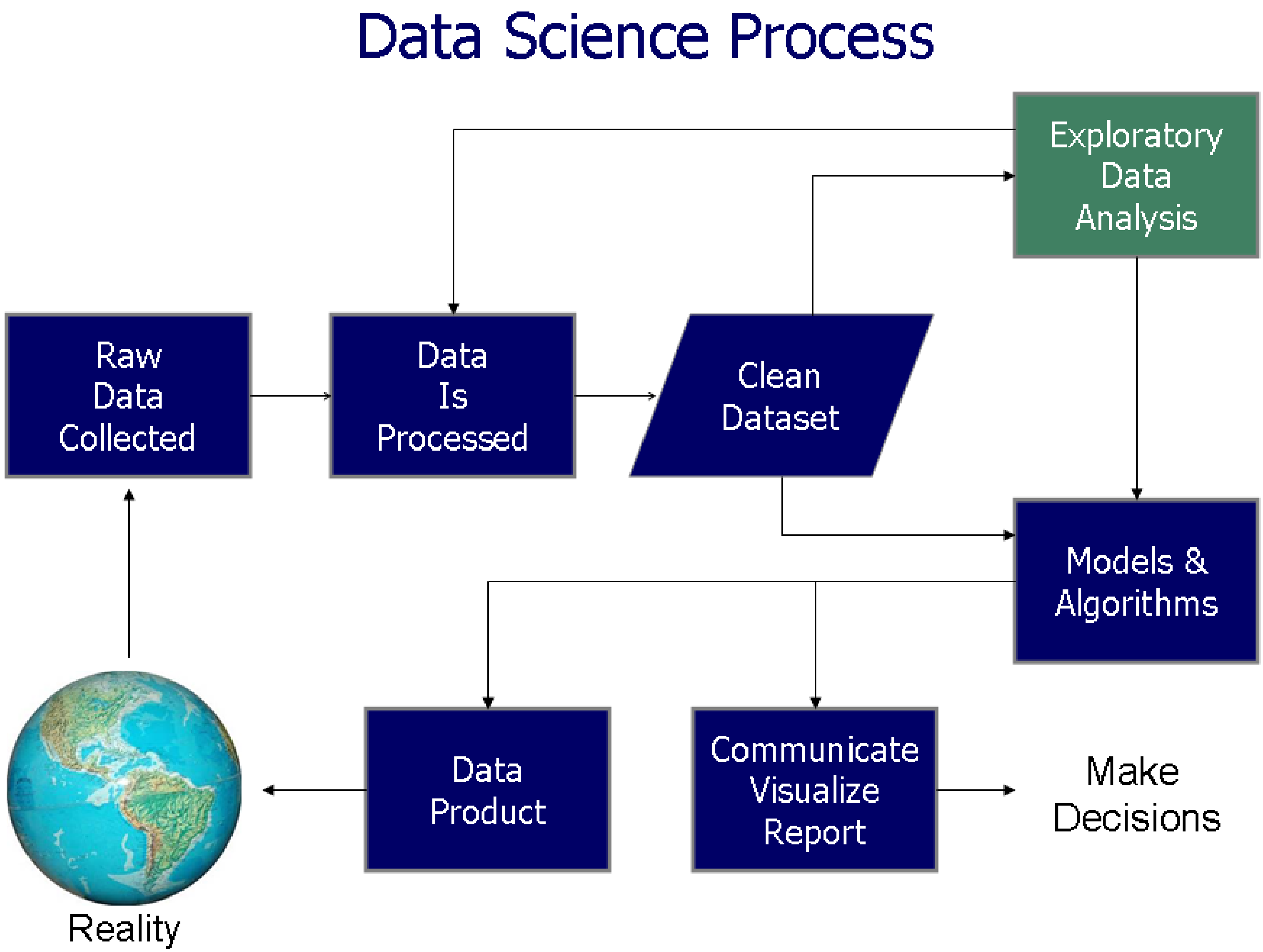
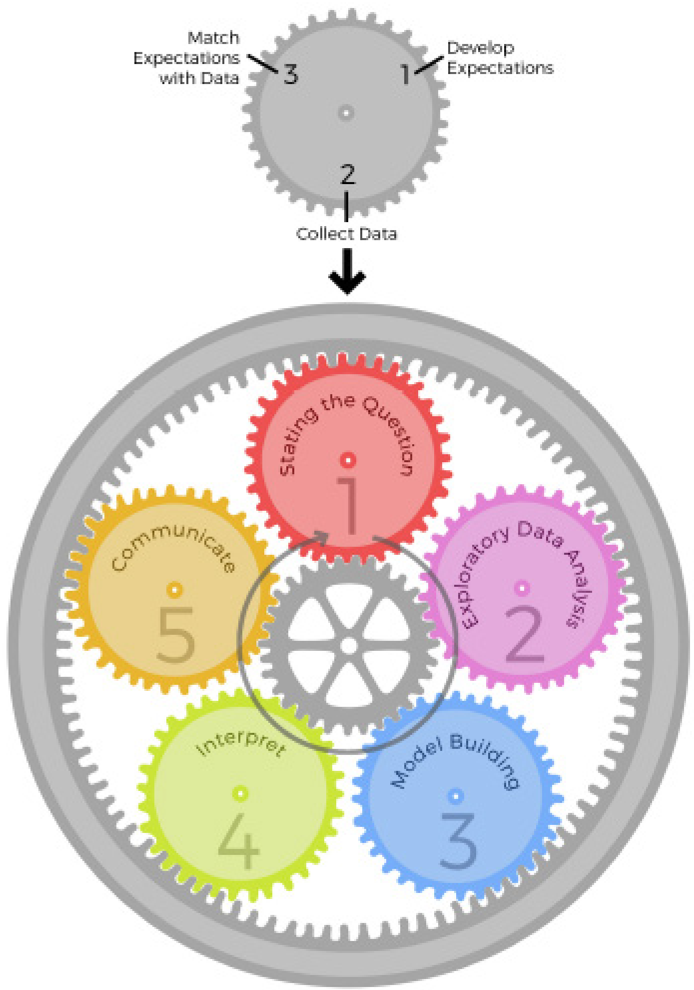
Week02

Data Analysis Process

Data Science Process



Epicycles of Data Analysis



Peng, R. D., & Matsui, E. (2016). The Art of Data Science. Lulu.com.

Epicycles of Data Analysis

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Peng, R. D., & Matsui, E. (2016). The Art of Data Science. Lulu.com.

Stating the Questions

- What business problem do you think you're trying to solve?

How to reduce churn to maintain profits?

Identify high-value customers based on recent purchase data

Build a model by using available customer data to predict the likelihood of churn for each customer

Rank customer based on churn propensity and customer value

Data Collection

- Data collection is the process of gathering and measuring information

- Sources of data
  - Internal data: information generated from within the business, covering areas such as operations, maintenance, personnel, and finance
  - External data: data comes from the governments and the market, including customers and competitors

- Importance questions for data collection
  - Which data to collect
  - How to collect the data
  - Who will collect the data
  - When to collect the data

Data Collection

Past

- Surveys
- Interviews
- Observations

Present

- Sensors
- Log
- Web

Data Preprocessing

- Data preprocessing
  - Data mining technique that involves transforming raw data into an understandable format

Data Cleaning
- Filling in missing values
- Smoothing the noisy data
- Resolving the inconsistencies in data

Data Integration
- Merging data with different representations
- Resolving the conflicts after merging

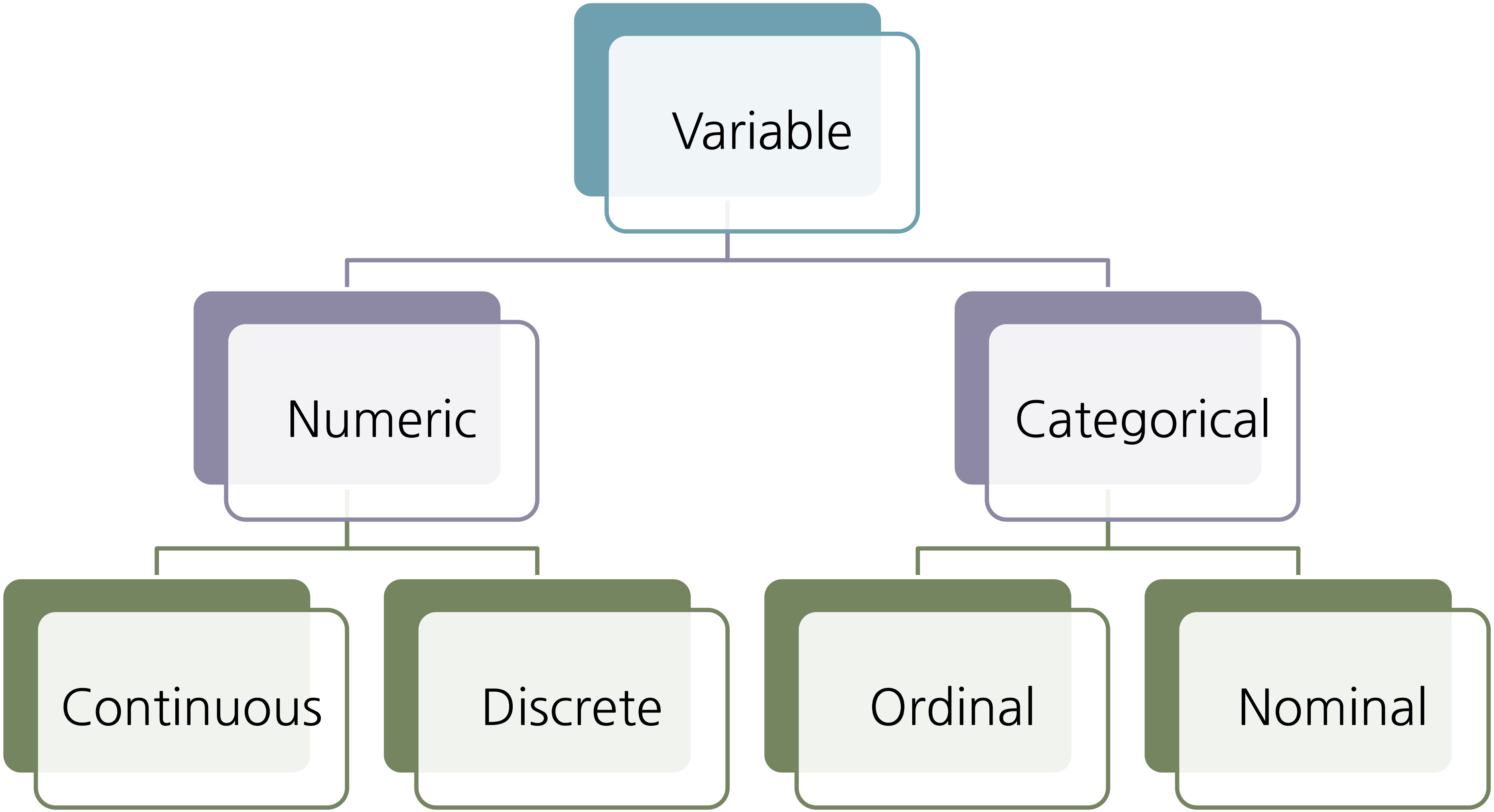
Data Transformation
- Normalization
- Aggregation

Data Reduction
- Feature selection
- Feature extraction

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods
  - Suggest hypotheses about the causes of observed phenomena
  - Assess assumptions on which statistical inference will be based
  - Support the selection of appropriate statistical tools and techniques
  - Provide a basis for further data collection through surveys or experiments

Types of Variables



Types of Variables

- Numeric (Quantitative)
  - A broad category that includes any variable that can be counted, or has a numerical
- Continuous
  - A variable with infinite number of values
  - Example
    - Many numeric variables: temperature, weight, height, pressure and etc.
- Discrete
  - A variable that can only take on a certain number of values or have a countable number of values between any two values
  - Example
    - The number of cars in a parking lot
    - the number of flaws or defects

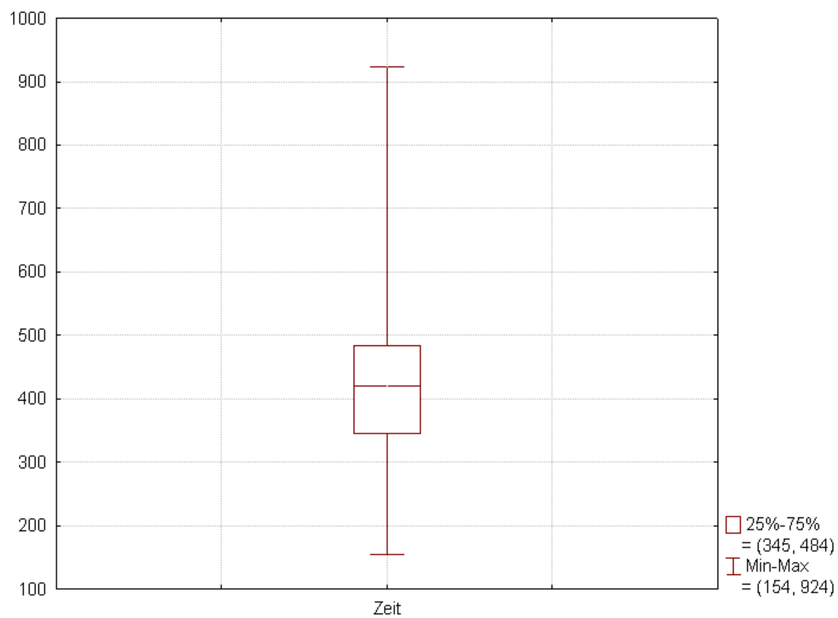
Types of Variables

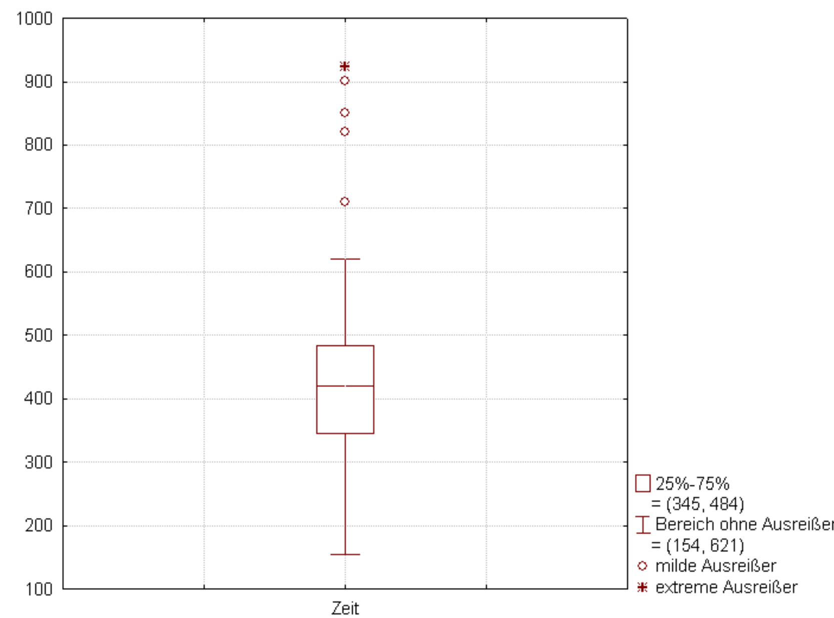
- Categorical
  - A variable that contains a finite number of categories or distinct groups
- Nominal
  - A Variable that has two or more categories, but there is no intrinsic ordering to the categories.
  - Example
    - (Male, Female), (Class 1, Class 2, Class 3), (Red, Yellow, Green)
- Ordinal
  - Similar to a nominal variable, but the difference between the two is that there is a clear ordering of the variables.
  - Example
    - Score: A+,A,A-,B+,B,B-,C+,C,C-,D,F
    - Size: S, M, L, XL, XXL

Exploratory Data Analysis: Numeric Data

- Data distribution
  - For numeric data, calculate summary statistics and obtain a boxplot or a histogram

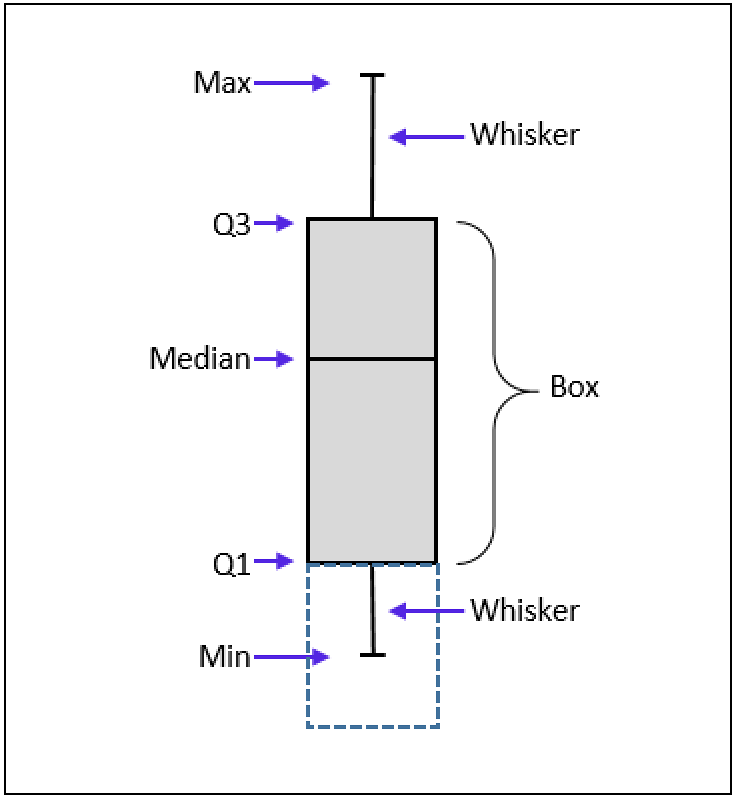
Boxplot



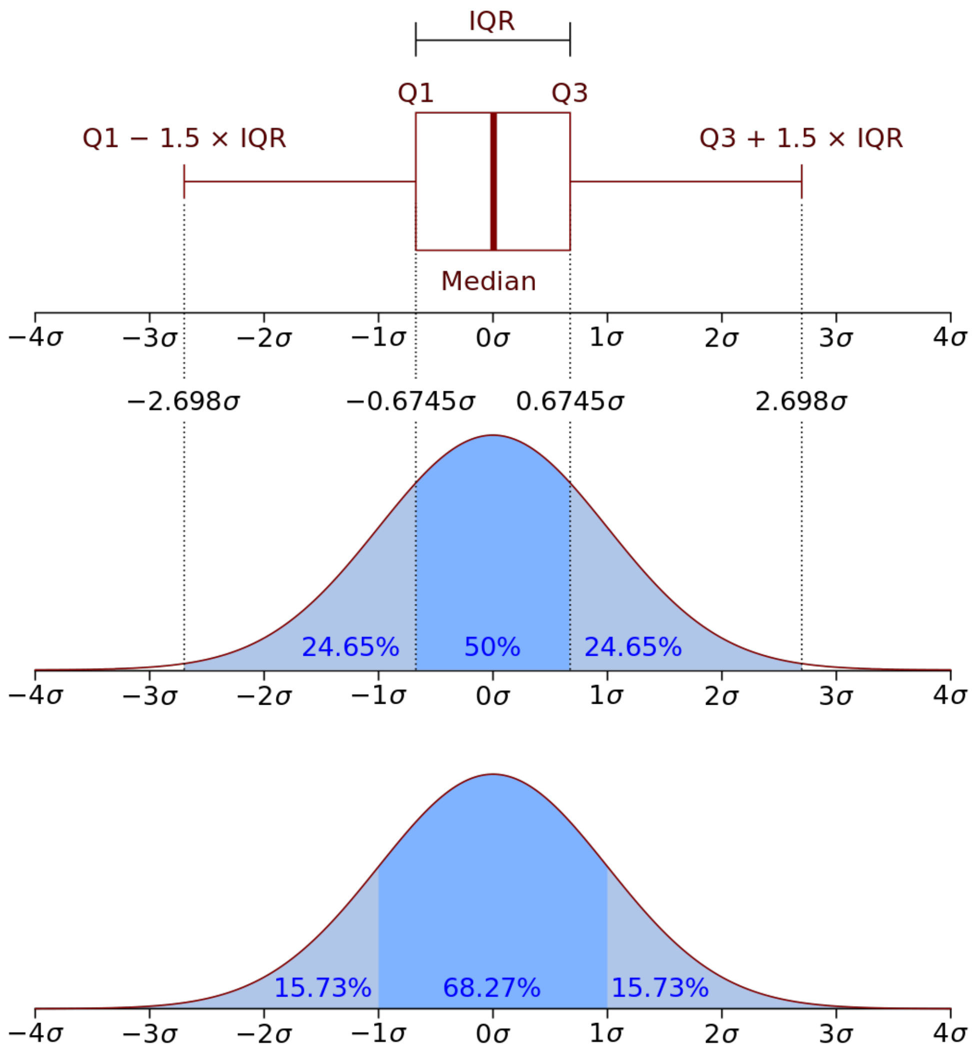


- Boxplot is a method for graphically depicting groups of numerical data through their quartiles

Boxplot

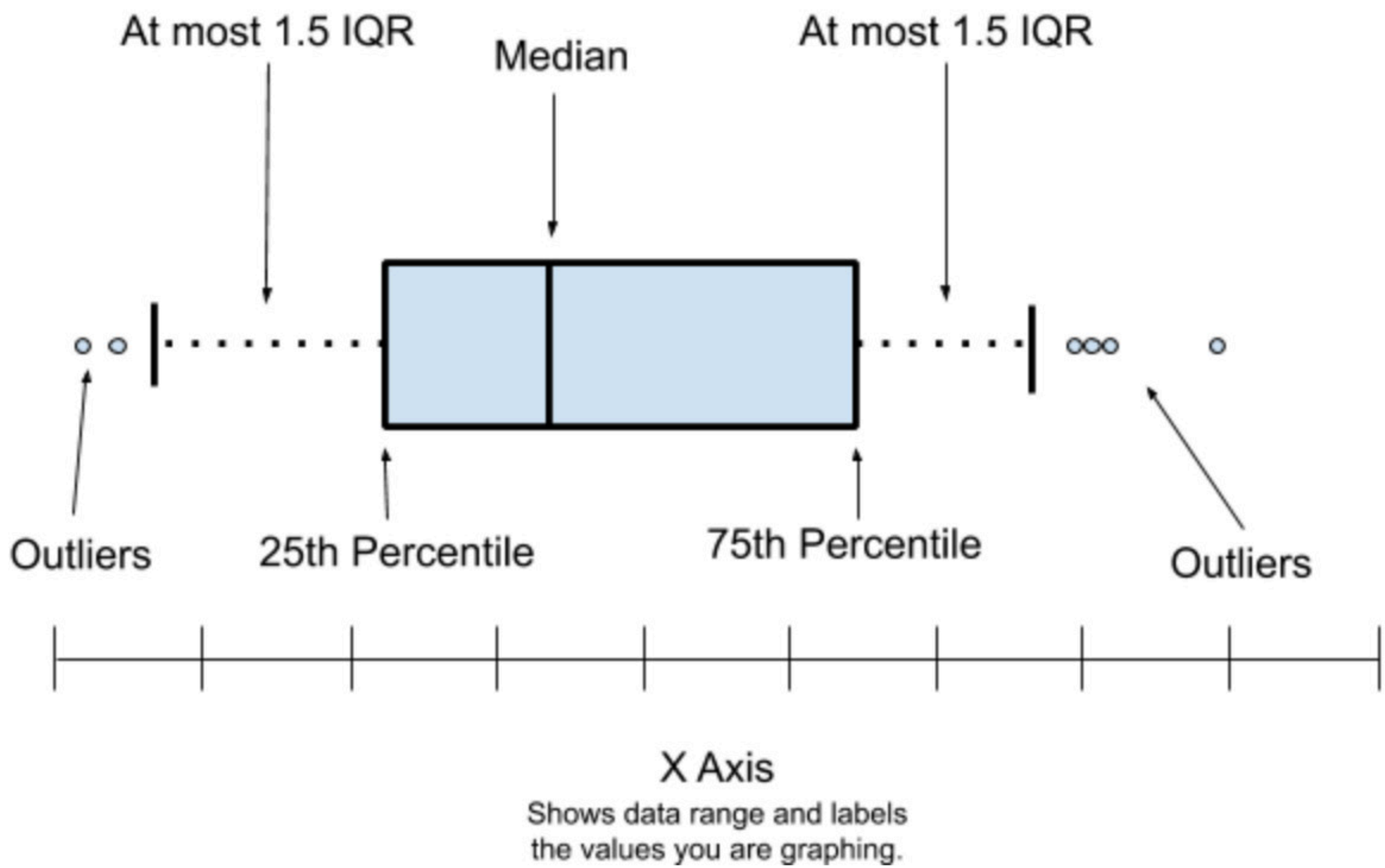


Max
Whisker
Q3
Median
Box
Q1
Whisker
Min



IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73%
68.27%
15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

Boxplot


At most 1.5 IQR
Median
At most 1.5 IQR
Outliers
25th Percentile
75th Percentile
Outliers
X Axis
Shows data range and labels
the values you are graphing.

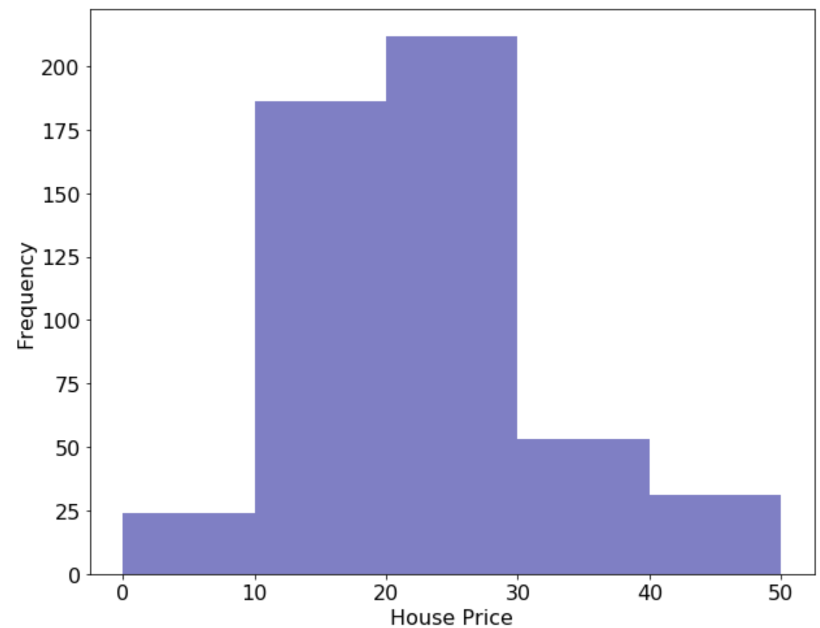
Exploratory Data Analysis: Numeric Data

- Data distribution
  - Histogram is an approximate representation of the distribution of numerical data

24.0,21.6,34.7,33.4,36.2,28.7,22.9,
27.1,16.5,18.9,15.0,18.9,21.7,20.4,
18.2,19.9,23.1,17.5,20.2,18.2,13.6,
19.6,15.2,14.5,15.6,13.9,16.6,14.8
… …

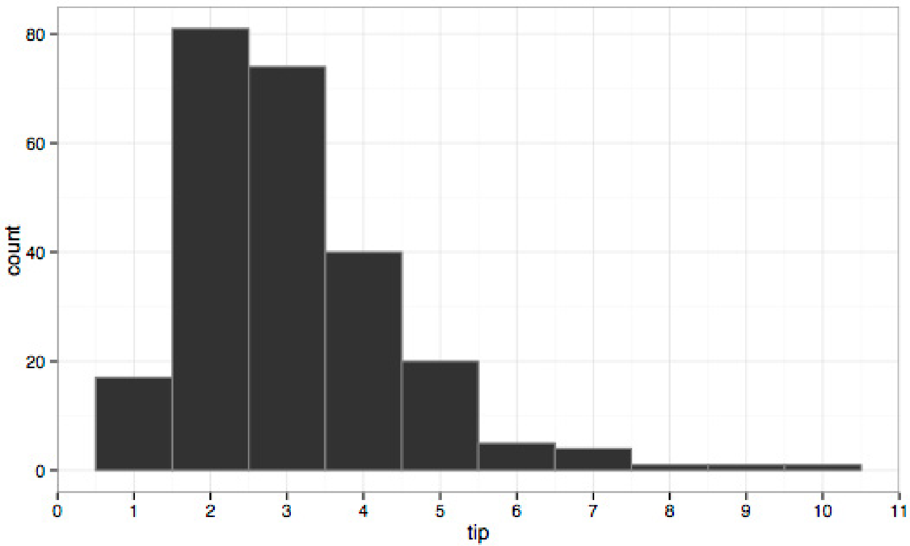
Bin	Count
0 to 10	24
10 to 20	186
20 to 30	212
30 to 40	53
40 to 50	31

Histogram

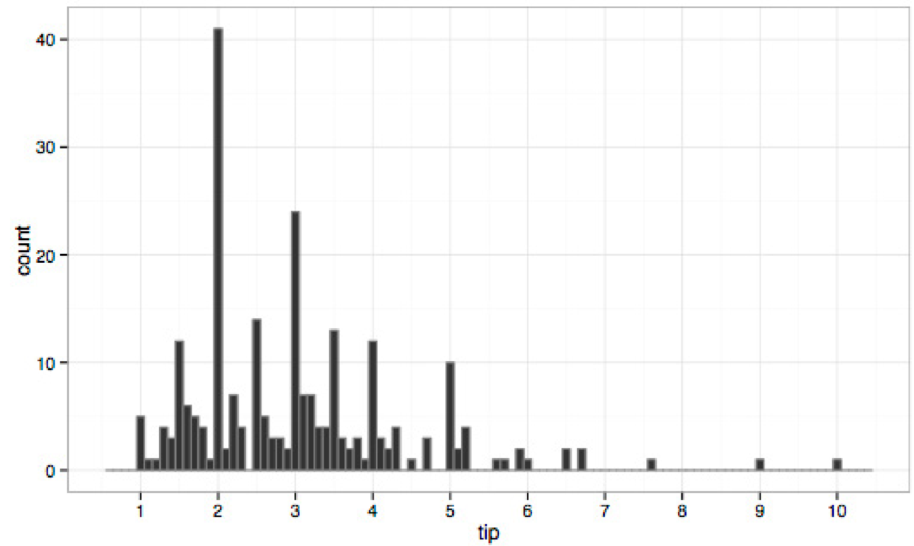


Exploratory Data Analysis: Numeric Data

- Same data, different bin widhts



1$



10C

Kernel Density Estimation

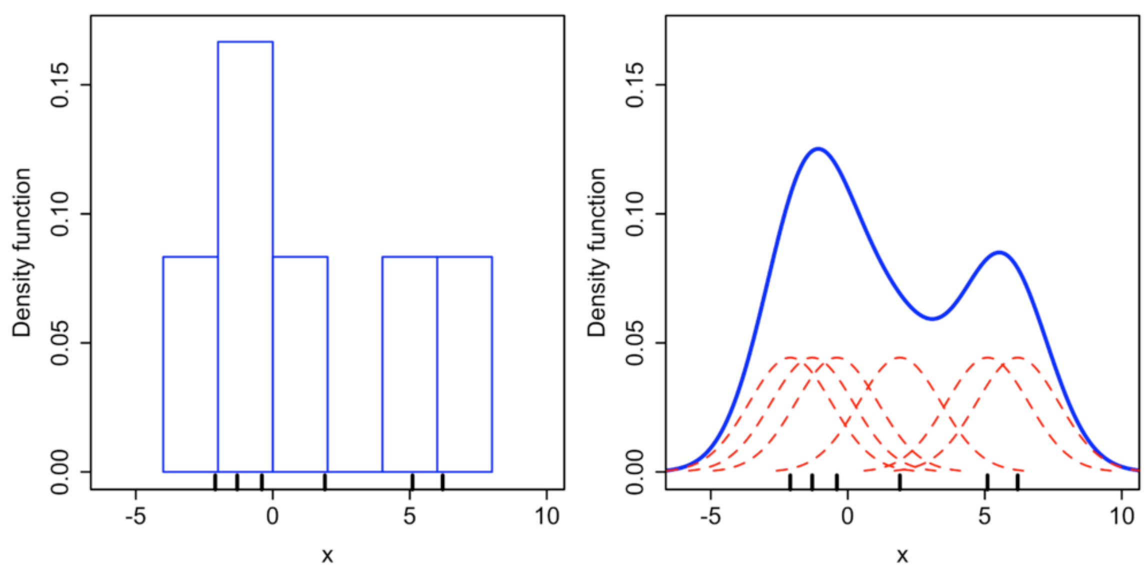
- KDE is a nonparametric density estimation method

What is a nonparametric method?

- Parametric method
  - Assume that data are drawn from a specific form of function up to unknown parameters
  - Linear regression, logistic regression, naïve Bayes classifier
- Nonparametric method
  - Do not rely on assumptions that the data drawn from a specific form of function up to unknown parameters
  - Unlike parametric methods, there is no single global model
  - Learn to find patterns from training set and interpolate
  - Heavier computational cost than parametric ones
  - k-nearest neighbors regression and classification, decision tree

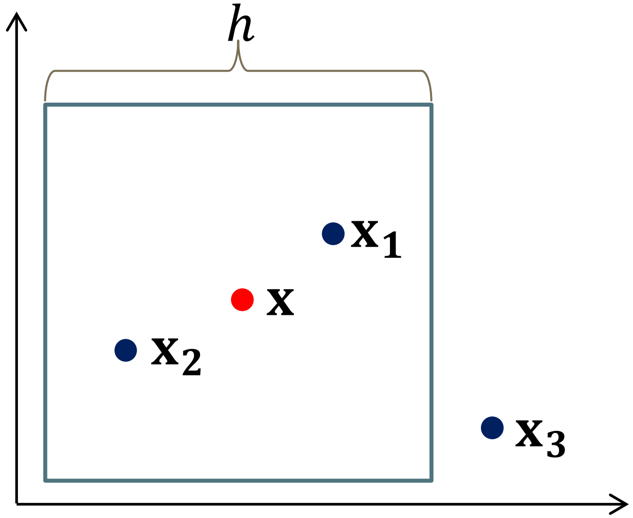
Density Estimation

- Density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function
  - Parametric method assumes a certain probability distribution function in advance and parameters are estimated based on observed data
    - For example, Gaussian distribution, Chi-square distribution and etc.
  - Nonparametric method does not set any probability distribution functions for observed data
    - Data samples determine shape of probability distribution functions



Parzen-Window Density Estimation

- The basis of Parzen-window density estimation is to count how many samples fall within a specified region (window)
  - If the number of data samples in a specified region is large, probability density is also large in the region



  - Window function(also called as kernel function)

$$\phi(\mathbf{u}) = \begin{cases} 1, & if \ |u_j| \le \frac{1}{2}; j = 1, \dots, p \\ 0, & otherwise \end{cases}$$

    - This function is assigning a value to a sample point if it lies within ½ of the edges of the hypercube

Parzen-Window Density Estimation

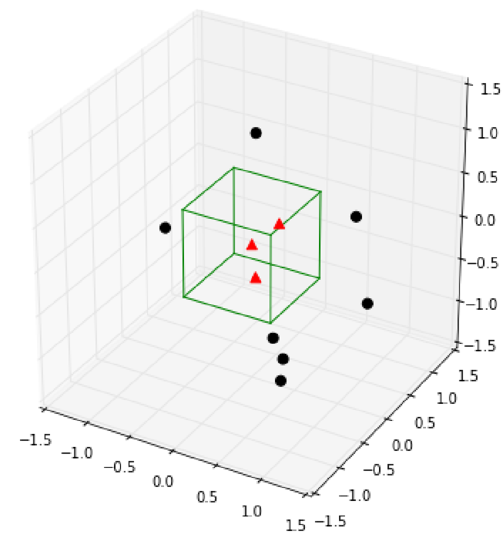
- To estimate the density at point $\mathbf{x}$, simply center the region at $\mathbf{x}$ and count the number of samples in the region

$$p(\mathbf{x}) \approx \frac{k_n/n}{V}$$

  - V is volume of the region

- The number of data samples within the hypercube

$$k_n = \sum_{i=1}^{n} \phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$



- Based on the window function, Parzen-window estimation can be formulated as follows

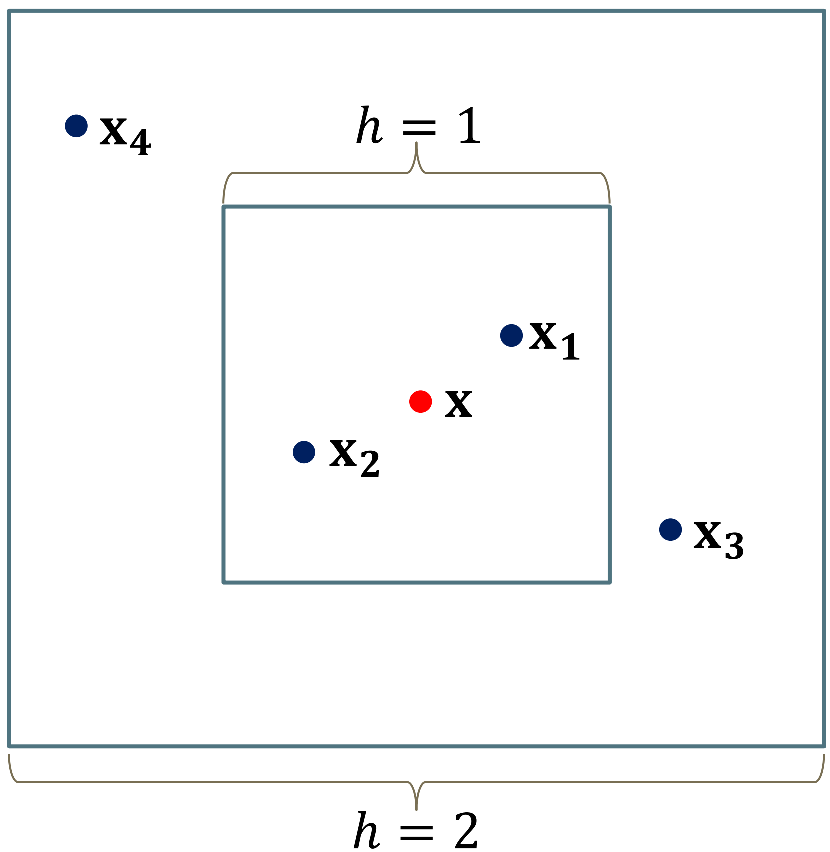
$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^p} \phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

  - h^p is the volume of the region(hypercube)

Parzen-Window Density Estimation

- 2-dimensional example

$$\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \begin{cases} 1, & if \ \left|\frac{x_j - x_{ij}}{h}\right| \leq \frac{1}{2}; j = 1, \dots, p \\ 0, & otherwise \end{cases}$$



$\mathbf{x} = (0,0)$

$\mathbf{x_1} = (0.25, 0.25)$

$\mathbf{x_3} = (0.7, -0.4)$

Parzen-Window Density Estimation

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

- Check $p_n(\mathbf{x})$ is in fact a density function
  - $p_n(\mathbf{x}) \geq \mathbf{0} \quad \forall\, \mathbf{x}$
  - $\int p_n(\mathbf{x})d\mathbf{x} = 1$

$$\int p_n(\mathbf{x})d\mathbf{x} = \int \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)d\mathbf{x} = \frac{1}{h^p n}\sum_{i=1}^{n}\overbrace{\int \phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)d\mathbf{x}}^{\text{volume of hypercube}}$$

$$= \frac{1}{n}\frac{1}{h^p}\sum_{i=1}^{n}h^p = 1$$

Parzen-Window Density Estimation

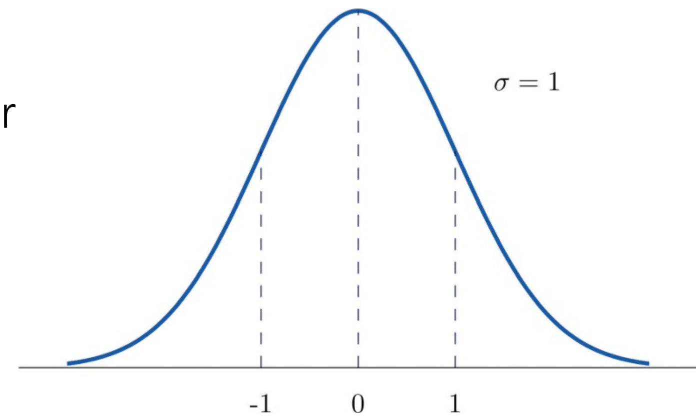
$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

- Use a general window function ϕ
  - Any ϕ that makes $p_n(\mathbf{x})$ legitimate density can be used
    - $p_n(\mathbf{x}) \geq \mathbf{0} \quad \forall\, \mathbf{x}$
    - $\int p_n(\mathbf{x})d\mathbf{x} = 1$

  - The most popular example of window function is $N(\mathbf{0}, \boldsymbol{I})$

$$\phi(\mathbf{u}) = \frac{1}{\sqrt{2\pi}}e^{-\mathbf{u}^T\mathbf{u}/2}$$

    - Point $\mathbf{x}$ closer to data sample point $\mathbf{x}_i$, receives higher weight
    - In this case, obtained density function is smooth



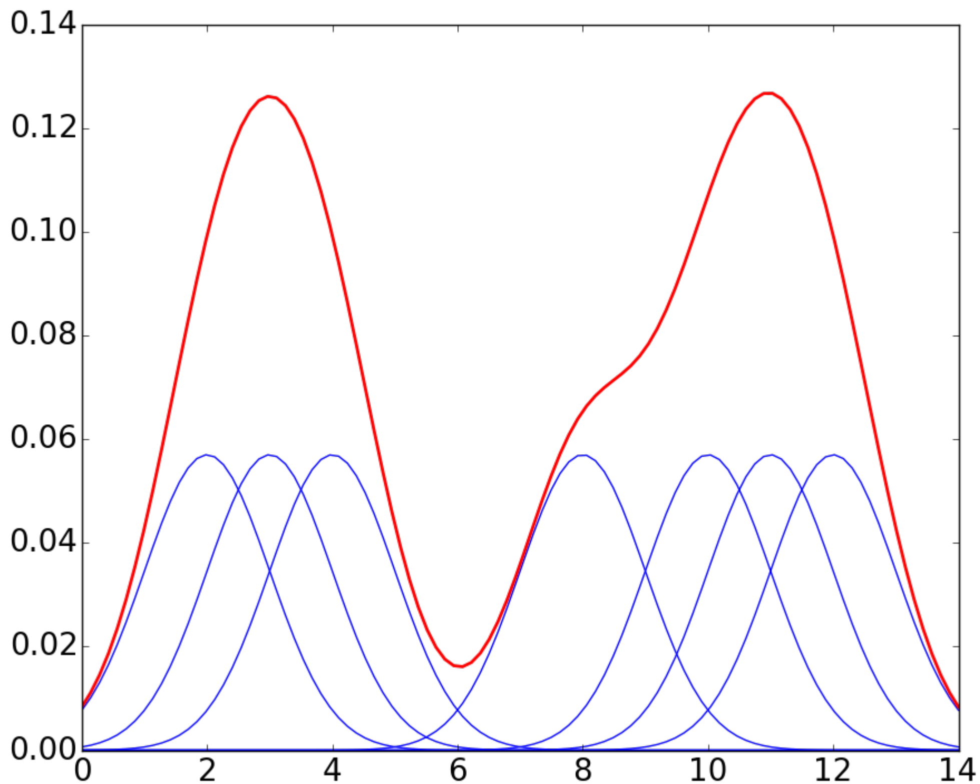
1D case

Example: Parzen-Window Density Estimation

- 1-D data set consisting 7 samples

$$D = \{2,3,4,8,10,11,12\}$$

  - Use standard normal distribution as window function and set window size $h = 1$

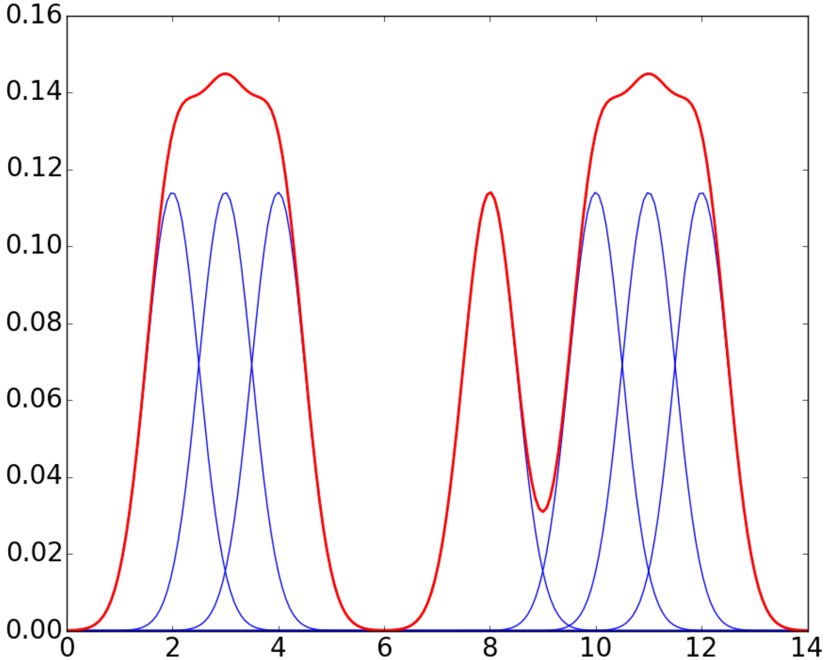


Example: Parzen-Window Density Estimation

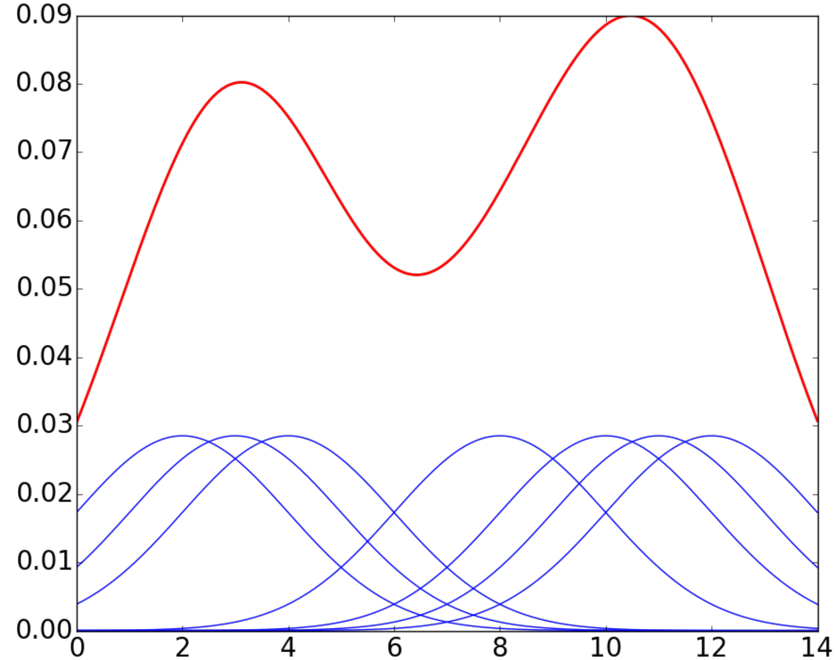
- 1-D data set consisting 7 samples

$$D = \{2,3,4,8,10,11,12\}$$

  - Same data set, but use different window size $h = 0.5, 2$



$h = 0.5$



$h = 2$

Parzen-Window Density Estimation

- Other window functions for 1D
  - Epanechnikov

$$\phi(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{if } |u| \le 1 \\ 0 & \end{cases}$$

  - Exponential

$$\phi(u) = \frac{1}{2}\exp(-u)$$

  - Triangular

$$\phi(u) = \begin{cases} (1-|u|) & \text{if } |u| \le 1 \\ 0 & \end{cases}$$

  - Cosine

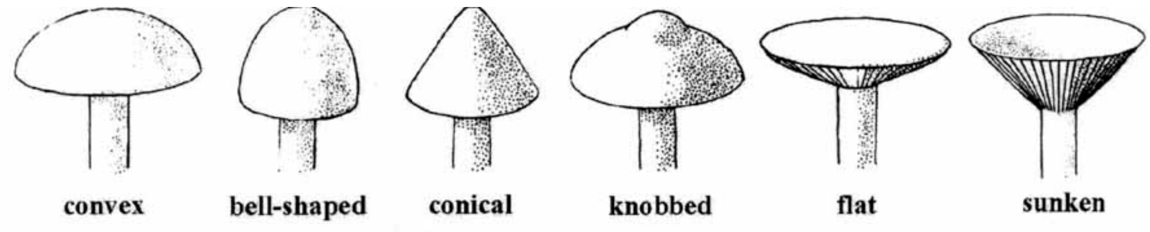
$$\phi(u) = \begin{cases} \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) & \text{if } |u| \le 1 \\ 0 & \end{cases}$$

Exploratory Data Analysis: Categorical Data

- Data distribution
  - For categorical data, calculate frequencies of attributes or categories

Do you remember mushroom data?

cap shape



Category	Count
Convex	3656
Bell-shaped	452
Conical	4
Knobbed	828
Flat	3152
Sunken	32

Programming Exercise

Exploratory Data Analysis: Example

- House Sales Prices in King County
  - Data source: https://www.kaggle.com/harlfoxem/housesalesprediction

Exploratory Data Analysis: Example

- Variables: 19 house features and the ID and price of houses
  - id: a notation for a house
  - date: Date house was sold
  - **price: the price of houses**
  - bedrooms: the number of bedrooms
  - bathrooms: the number of bathrooms, where .5 accounts for a room with a toilet but no shower
  - sqft_living: Square footage of the apartments interior living space
  - sqft_lot: Square footage of the land space
  - floors: total floors (levels) in house
  - waterfront: house which has a view to a waterfront
  - view: An index from 0 to 4 of how good the view of the property was
  - condition: An index from 1 to 5 on the condition of the apartment
  - grade: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design

Exploratory Data Analysis: Example

- Variables: 19 house features and the ID and price of houses
  - sqft_above: square footage of house apart from basement
  - sqft_basement: square footage of the basement
  - yr_built: built year
  - yr_renovated: year when house was renovated
  - zipcode: zip
  - lat: latitude coordinate
  - long: longitude coordinate
  - sqft_living15: The square footage of interior housing living space for the nearest 15 neighbors
  - sqft_lot15: The square footage of the land lots of the nearest 15 neighbors

Summary Statistics: Numeric

- Summary statistics
  - Calculate mean, variance (standard deviation), min, max and so on

```
house['bedrooms'].mean()
house['bedrooms'].var()
house['bedrooms'].std()
house['bedrooms'].min()
house['bedrooms'].max()
house['bedrooms'].median()
```

	mean	std	median	min	max
bedrooms	3.37	0.93	3.00	0.00	33.00

```
house['bedrooms'].describe()
```

Exploration: Plot

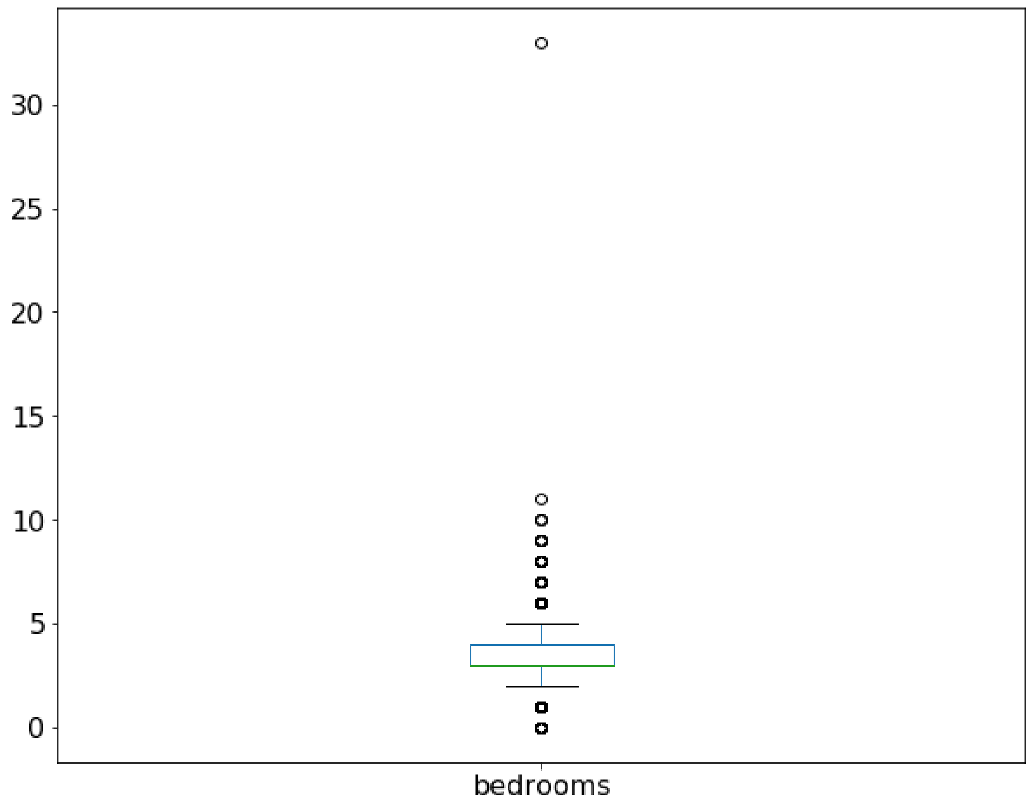
- Pandas provides the basic features for plots
  - plot
  - bar or barh
  - hist
  - box
  - kde
  - scatter

- Documentation
  - https://pandas.pydata.org/pandas-docs/stable/visualization.html

Exploration: Plot

- Box plot

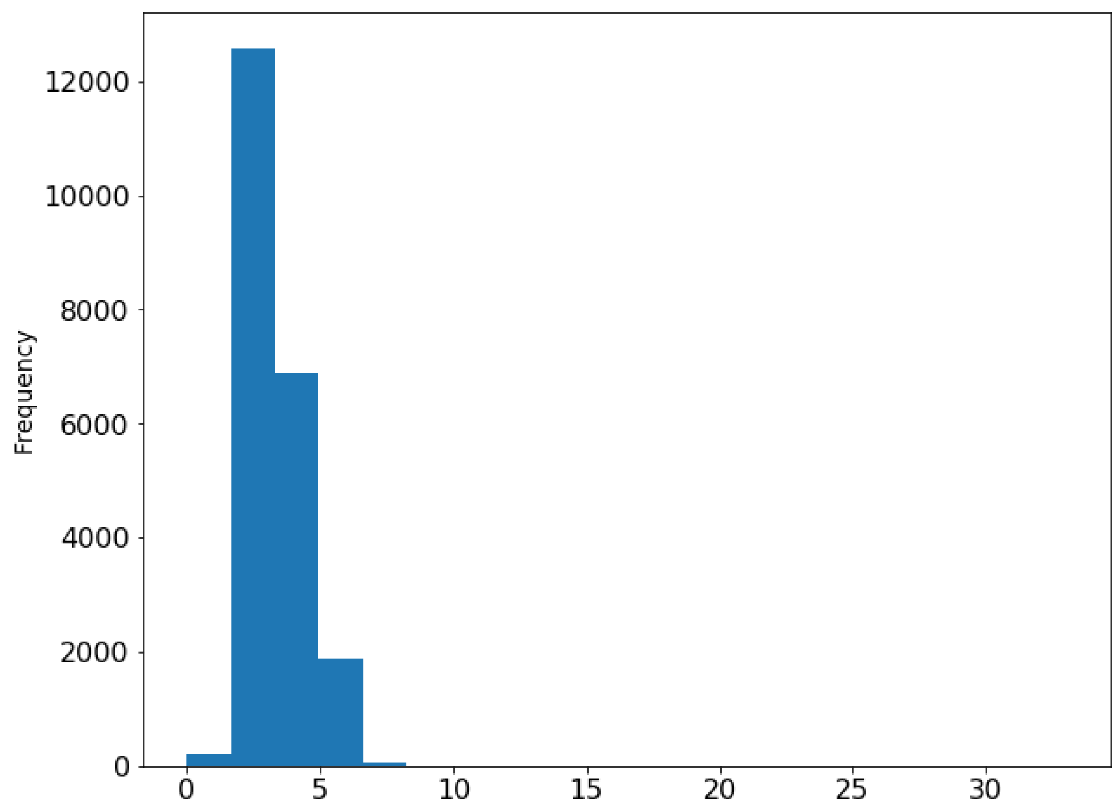
```
house['bedrooms'].plot.box()
```



Exploration: Plot

- Histogram

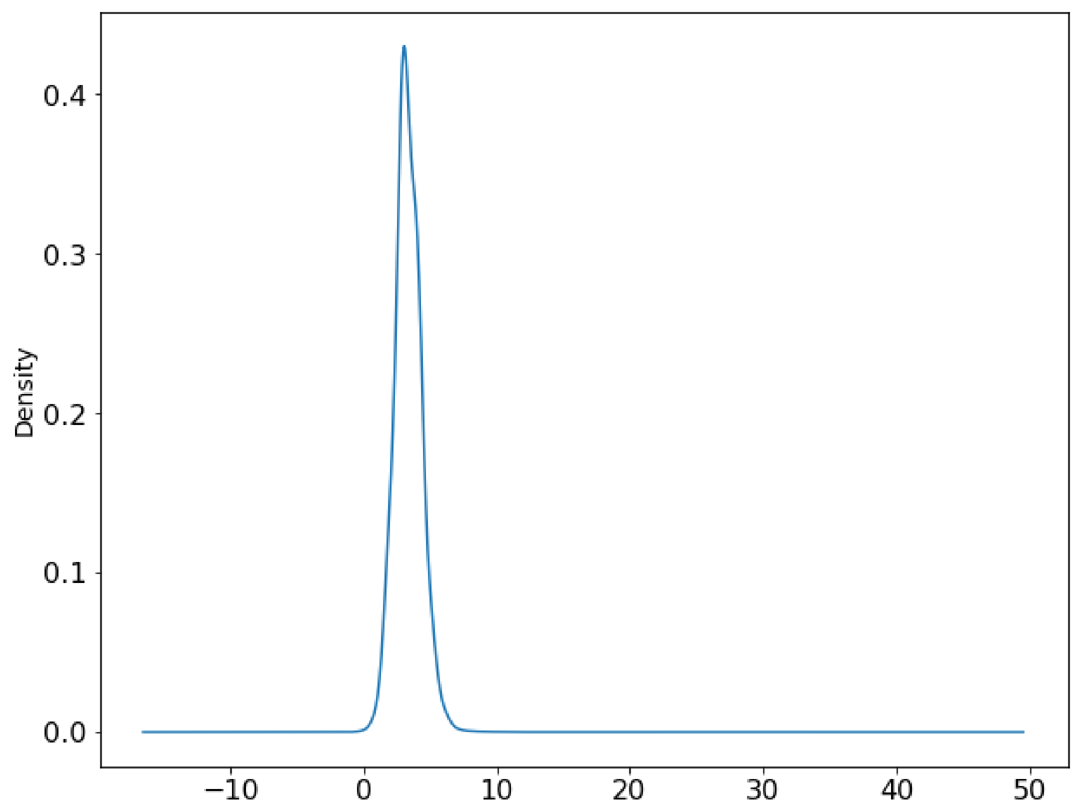
```
house['bedrooms'].plot.hist(bins=20)
```



Exploration: Plot

- Kernel density estimation plot

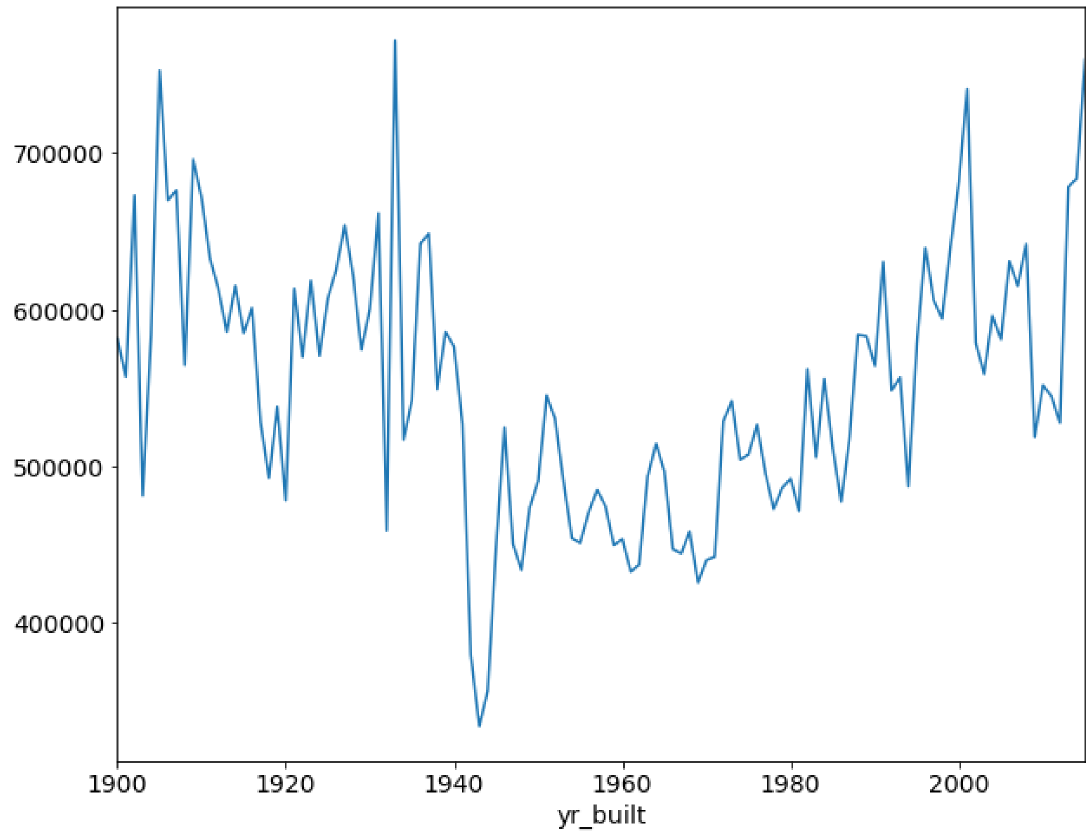
```
house['bedrooms'].plot.kde(bw_method=0.5)
```



Exploration: Plot

- Line plot

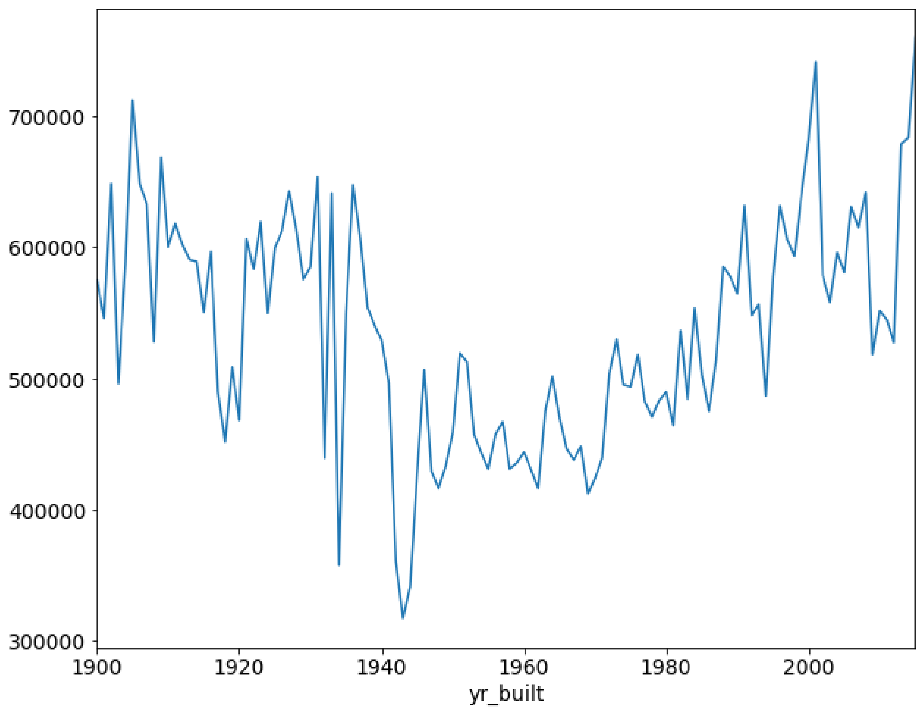
```
house.groupby('yr_built')['price'].mean().plot()
```



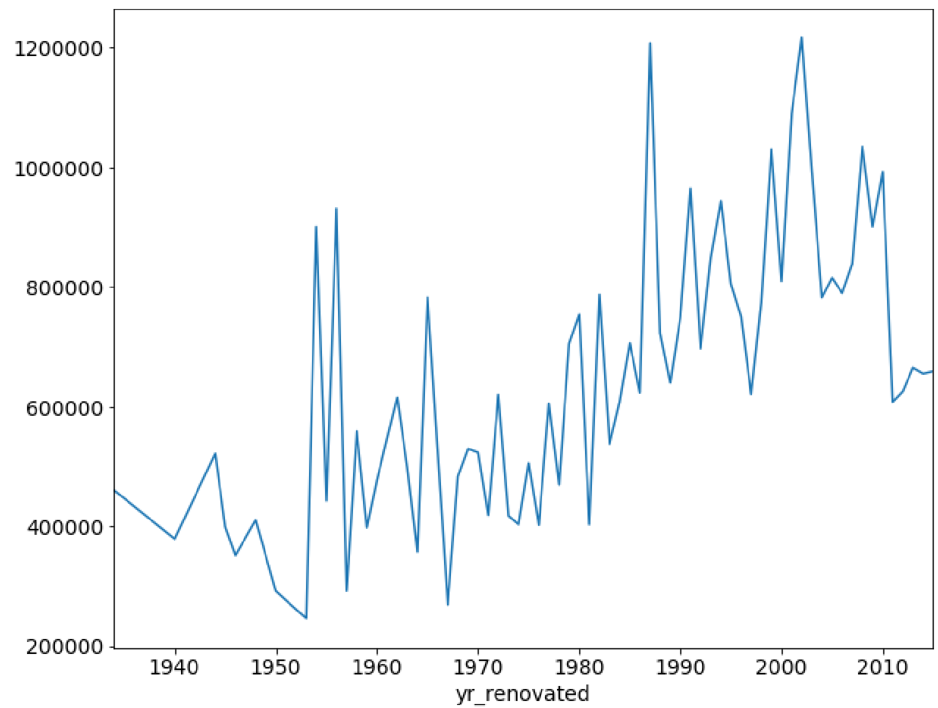
Exploration: Plot

Renovated?

No



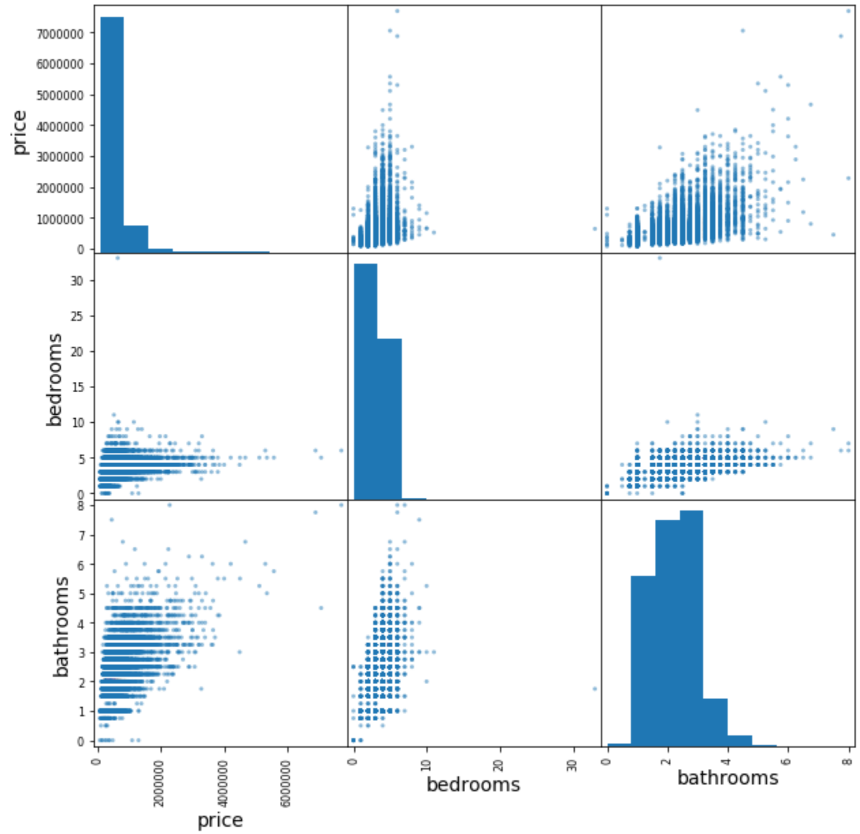
Yes



Exploration: Plot

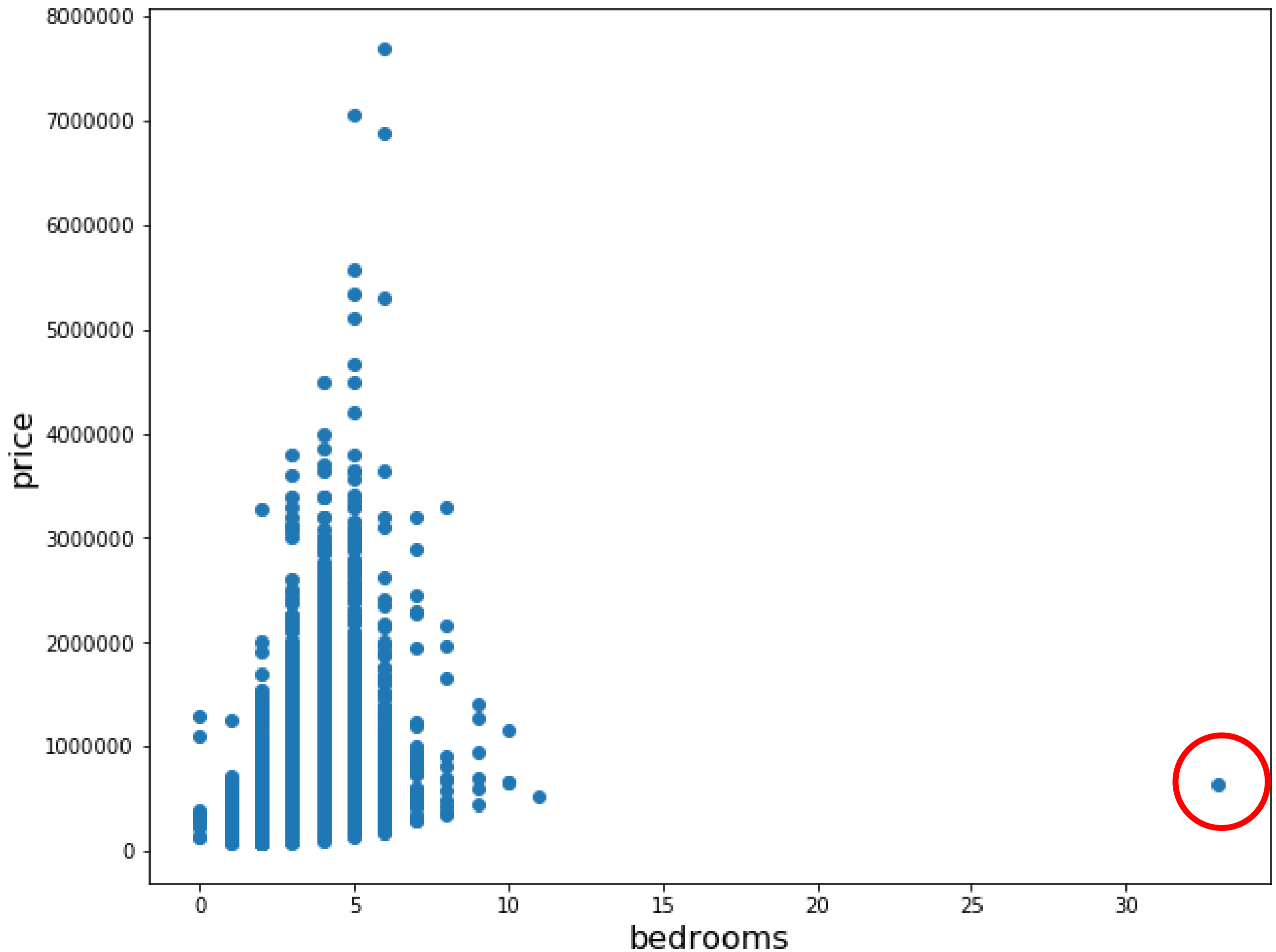
- Scatter matrix
  - Pandas version ≥ 0.21

```
from pandas.plotting import scatter_matrix
scatter_matrix(house[['price','bedrooms','bathrooms']],figsize=(10,10))
```



Exploration: Plot

- Scatter plot: Find outliers



Exploration: Correlation

- For linear regression, the existence of multicollinearity on data should be checked
  - Obtain correlation matrix

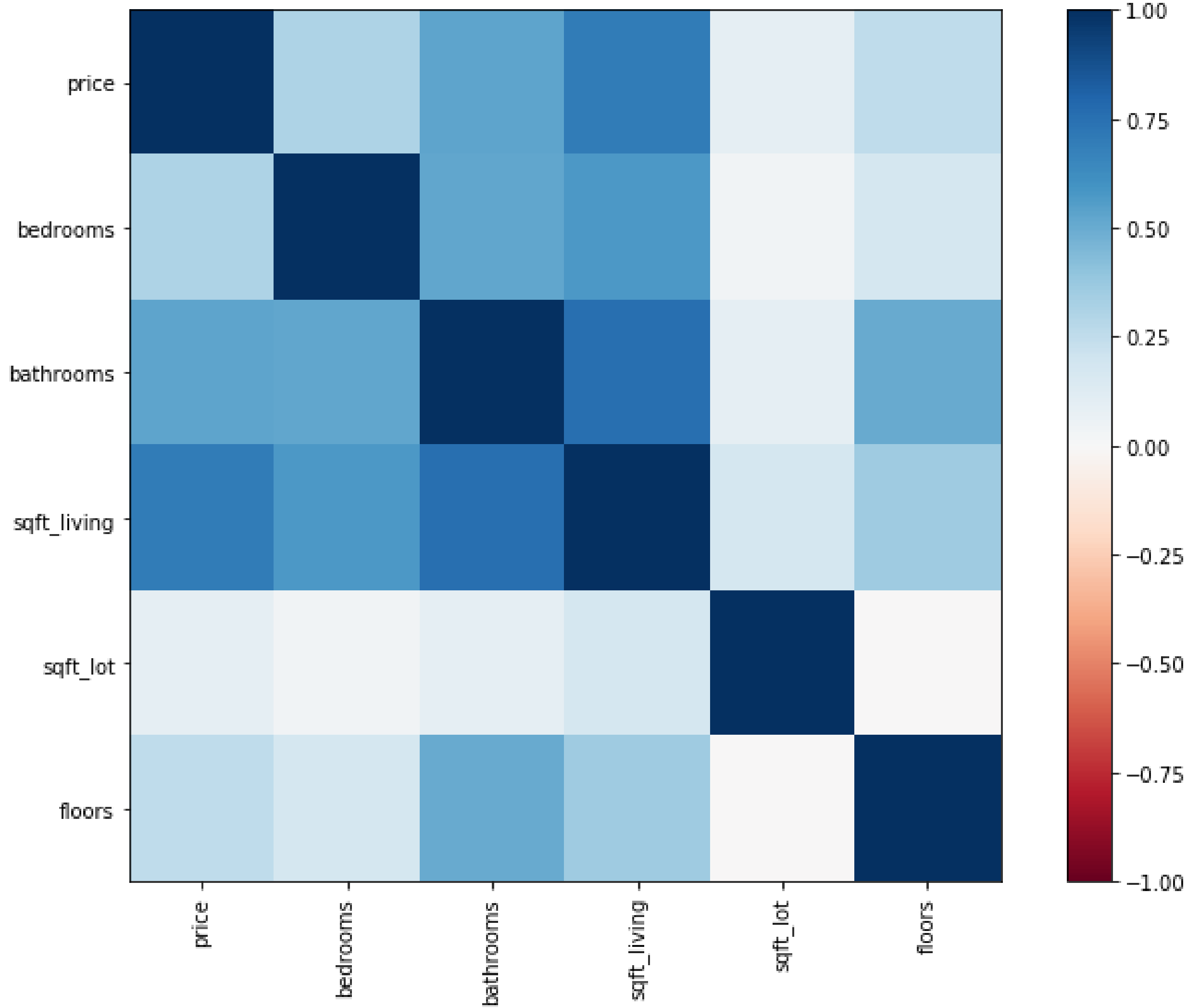
```
corr=house[['price','bedrooms','bathrooms','sqft_living','sqft_lot','fl
oors']].corr()
```

  - Correlation plot

```
fig=plt.figure(figsize=(12,8))
cax=plt.imshow(corr, vmin=-1, vmax=1, cmap=plt.cm.RdBu)
ax=plt.gca()
ax.set_xticks(range(len(corr)))
ax.set_yticks(range(len(corr)))
ax.set_xticklabels(corr,fontsize=10,rotation='vertical')
ax.set_yticklabels(corr,fontsize=10)
plt.colorbar(cax)
```

Exploration: Correlation

- Correlation plot



Summary Statistics: Categorical

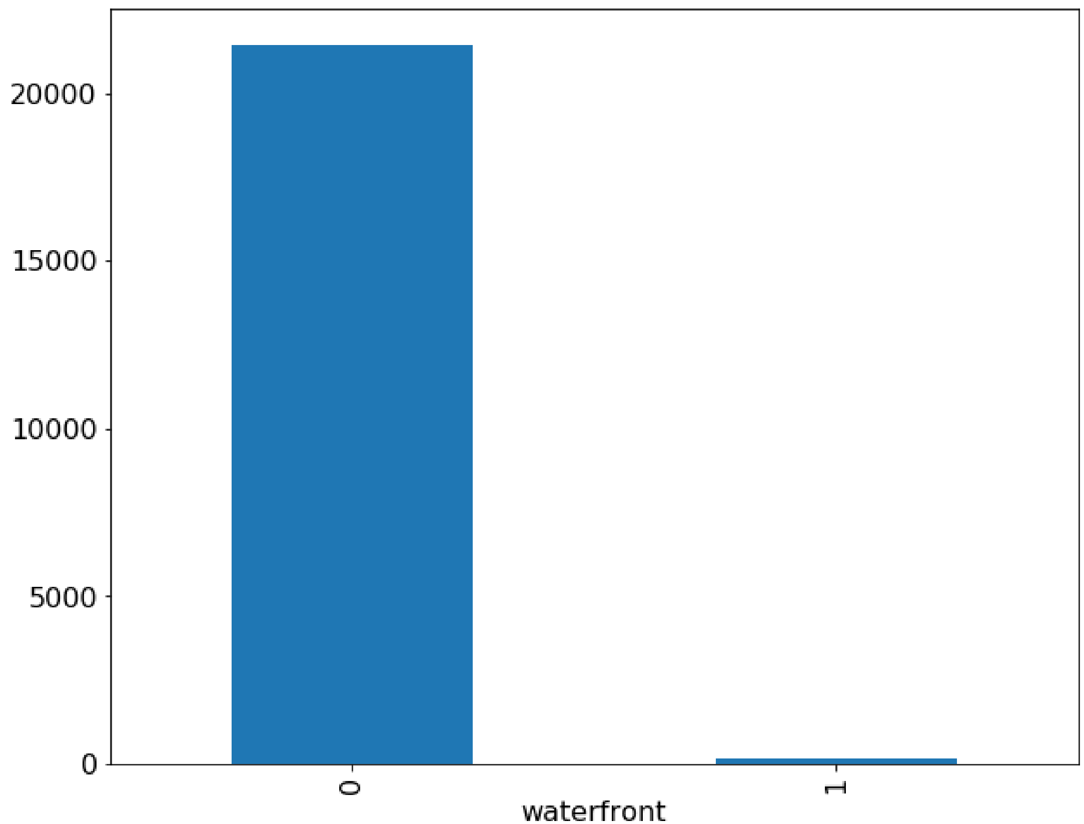
- Frequency

```
house['grade'].value_counts()
```

Exploration: Plot

- Bar

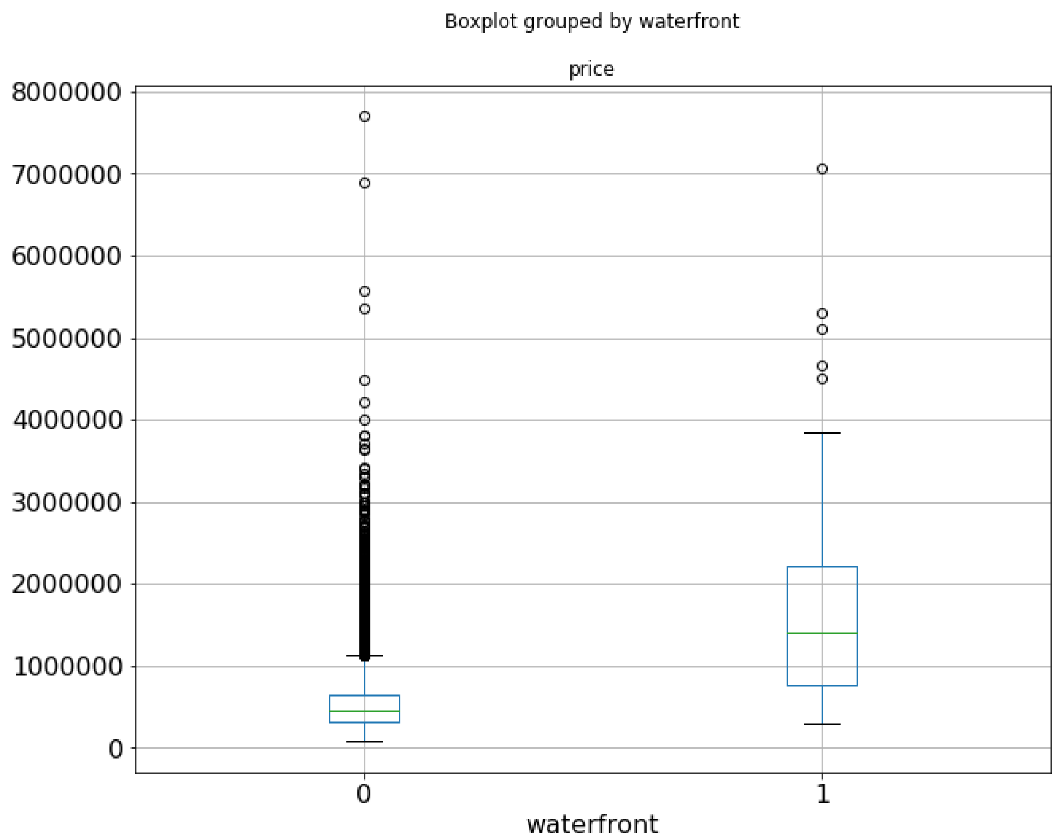
```
house['waterfront'].value_counts().plot.bar()
```



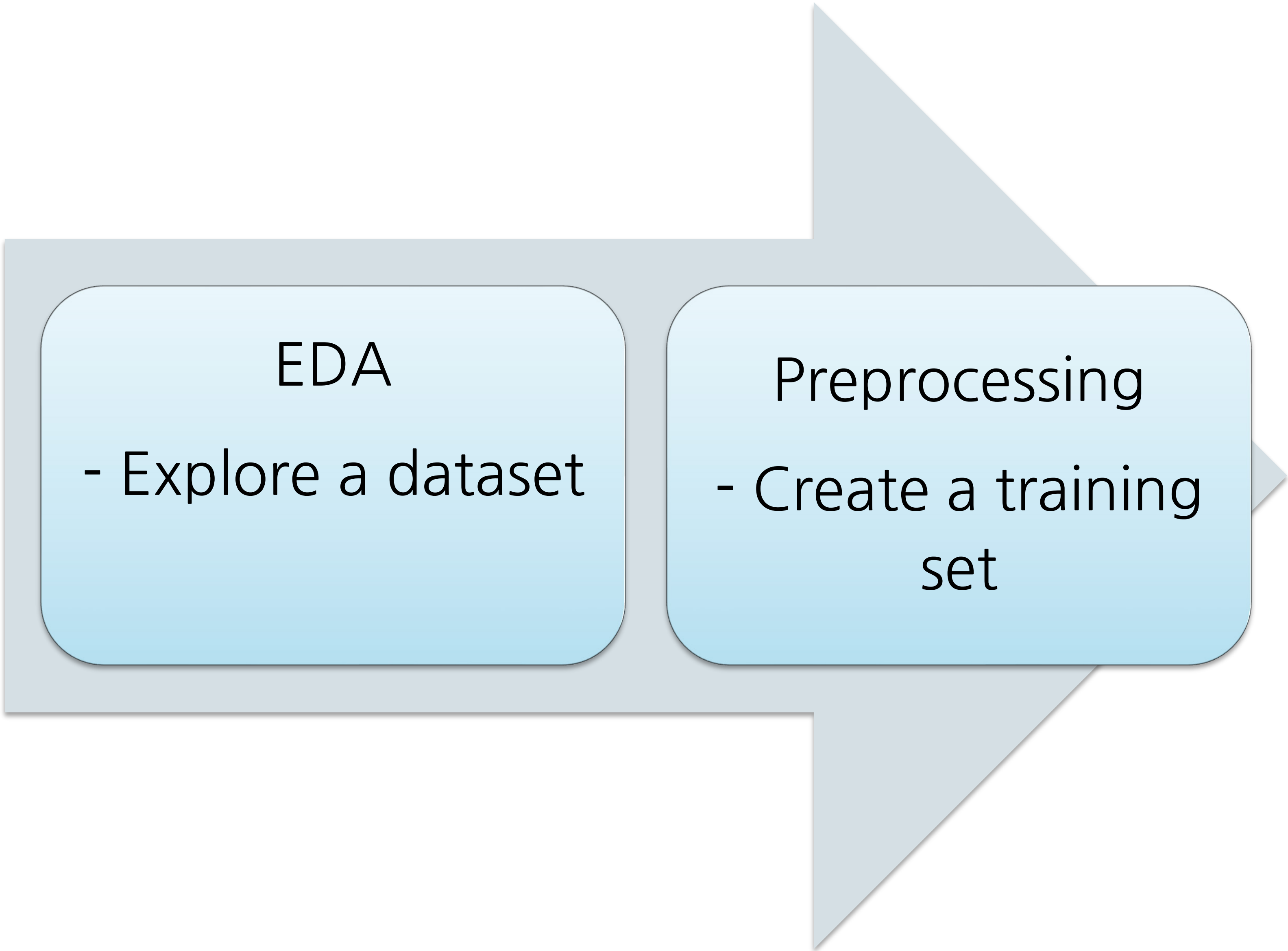
Exploration: Plot

- Boxplot

```
house.boxplot(column=['price'], by='waterfront',ax=ax)
```



What is Next Step of EDA?



Create Train Dataset

- Check whether there are irrelevant samples
  - Based on the results of EDA, are there any samples that seem to be removed for training?

- Select appropriate independent variables to predict the target
  - This process may depend on regression algorithms to be applied
    - Ex) Linear regression is used → Select variables which seem to be linearly related with the target and check multicollinearity
  - Variable transformation can be applied to generate better independent variables

Multicollinearity

- For linear regression, the existence of multicollinearity on data should be checked
  - Calculate variation inflation factor (VIF)
    - Variance inflation factor(VIF) quantifies the severity of multicollinearity in a least square method
    - Calculate VIF
      - Step 1) Apply least square method to regression problem that i-th input variable is regressed by the remained input variables

$$x_i = \alpha_1 x_1 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_p x_p + \alpha_0 + \epsilon$$

      - Step 2) Calculate R^2 for above regression problem and set the value as R_i^2
      - Step 3) Calculate VIF from R_i^2

$$VIF = \frac{1}{1 - R_i^2}$$

Learn a Model

- Linear regression

```
from sklearn.linear_model import LinearRegression
reg=LinearRegression()
```

  - Apply linear regression
  - Check the estimates of coefficients
  - $F-test, t-test$
  - Calculate R^2
  - Examine residuals
    - Q-Q plot
    - Normality test
    - Homoscedasticity test

Learn a Model

- Linear regression
  - For linear regression, there is a good alternative for sklearn

```
import statsmodels.api as sm
X=sm.add_constant(X)
model=sm.OLS(y, X)
result=model.fit()

result.summary()
```

Assignment

Assignment 02

- Build linear regression model
  - Among numeric variables, select input variables
    - Describe reasons for variable selection
  - You can apply variable transformation
    - Describe reasons of variable transformation
  - You can discard some rows satisfying specific conditions
    - Specify conditions

- Summarize the process and result using Power Point
  - Some students have to create video clip to explain their results

- Submit both slide and python code