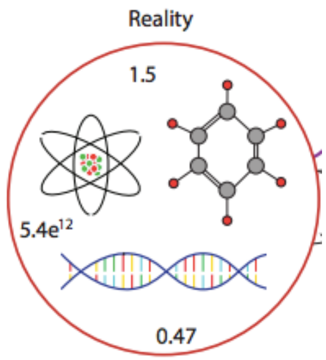
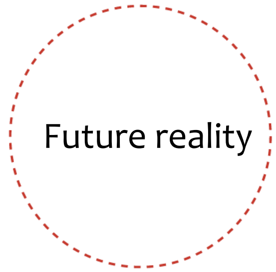
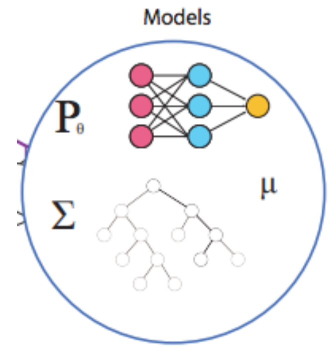






Statistics 154, Spring 2019

Modern Statistical Prediction and Machine Learning

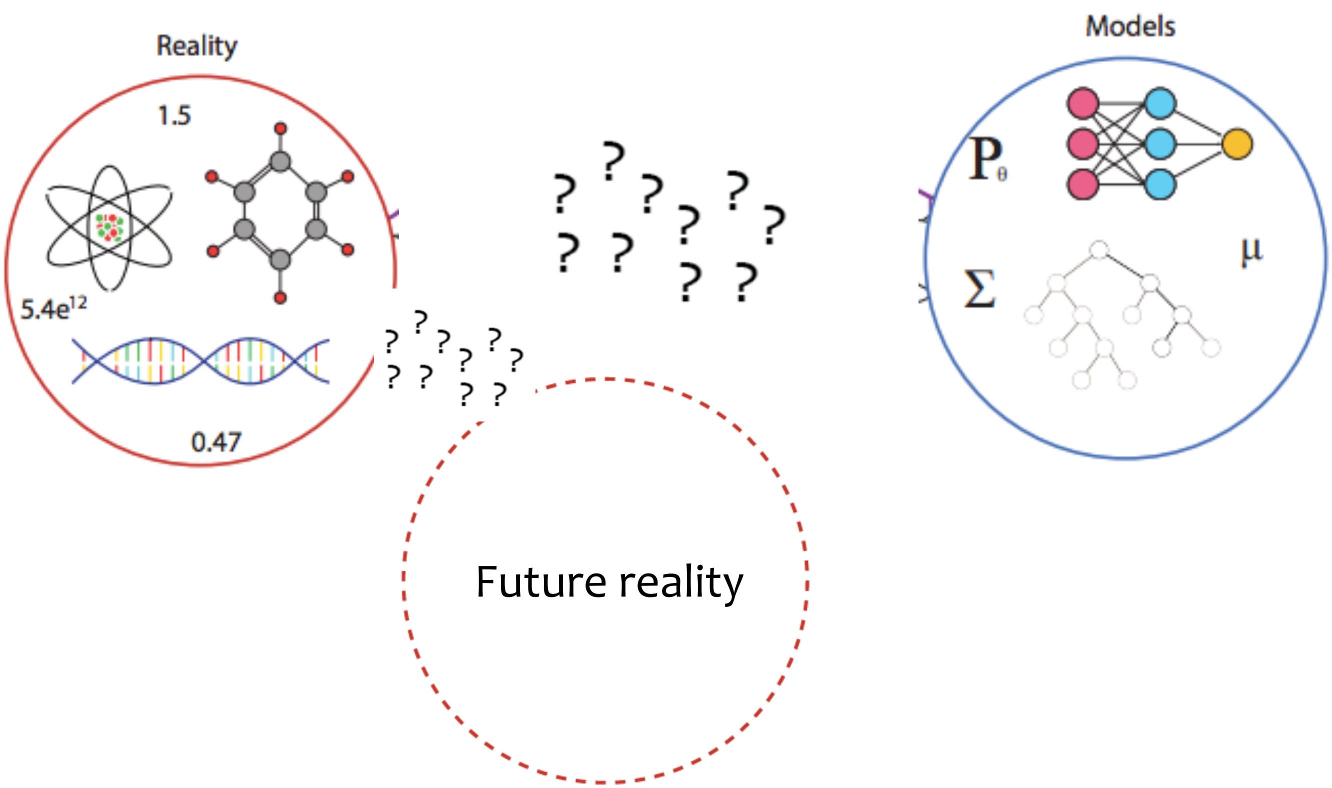
Lecture 4: Cross-validation -- pros and cons

Instructor: Bin Yu
(binyu@berkeley.edu); office hours: Tu: 9:30-10:30 am; Wed:**1:30-2:30 pm (change)**
office: 409 Evans

GSIs: Yuansi Chen (Mon: 10-12; 12-2); Raaz Dwivedi (Mon: 2-4; 4-6)
yuansi.chen@berkeley.edu; raaz.rsk@berkeley.edu
(office hours to be announced)

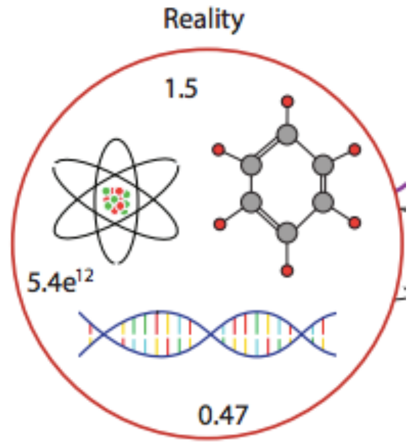
Recap: **Why are we here?**

To solve prediction problems in real world by connecting the two solid circles below in a justifiable way to say things about the third cycle



Recap: **What will you learn?**

Future reality

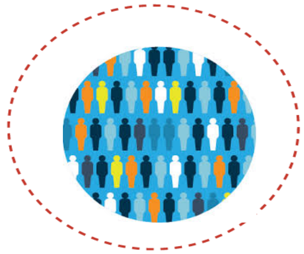


- Problem formulation
- Data collection, causality, data cleaning
- Cross-validation
- EDA (exploratory data analysis, visualization)
- Unsupervised learning (e.g. PCA, clustering)
- Supervised learning (LS, regularized LS, Kernel regression, logistic regression, Support Vector Machines (SVMs), Nearest Neighbor (NN), Decision trees, Random Forests, Deep Learning)
- Data results, validation, conclusions

I. PQR-S: conceptual framing to start a prediction problem

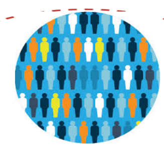
or a check-list

PQR-S helps you think straight!



- **P**opulation

- **Q**uestion

- **Representative** data collection (data neutral, fairness) is  similar to  ?

- **-** (to be filled in throughout the course)

- **S**crutinize or validate data results

Recap: association is not causation

- Association is not causation
- Confounding factors often at play
- Randomization is the gold standard (randomized experiments)
- Many observational studies (not randomized experiments)

Data quality

- Data checking
- Data cleaning
- Data reproducibility

Live data vs. dead data

- Our project uses live data, which is defined as data that can be enriched by more data and human knowledge and with human domain experts interests

- Dead data: from the web with unknown data collection process and/or purpose, no interests from domain experts on its analysis

The above definitions were inspired by G. Gelman's distinction between "live problem" and mere "real data". He said in an email

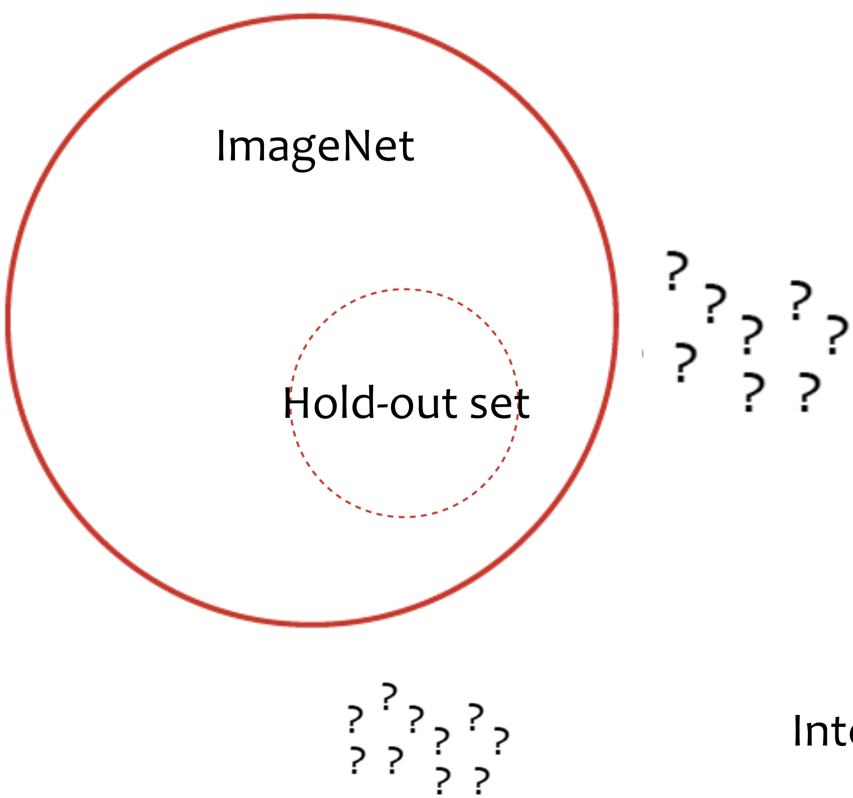
"It's not enough for the data to be `real'; the data also should connect to some live question of interest."

Read his blog at

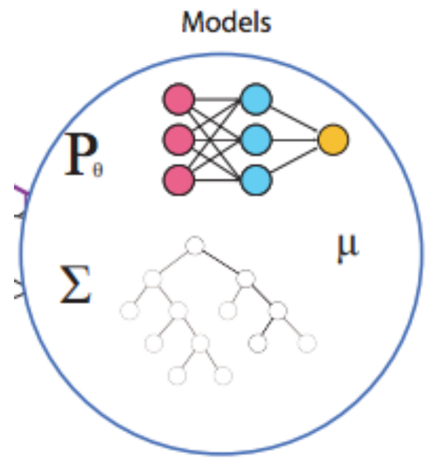
https://statmodeling.stat.columbia.edu/2009/07/23/that_modeling_f/

Internal validity is a minimal requirement ??

Training data set

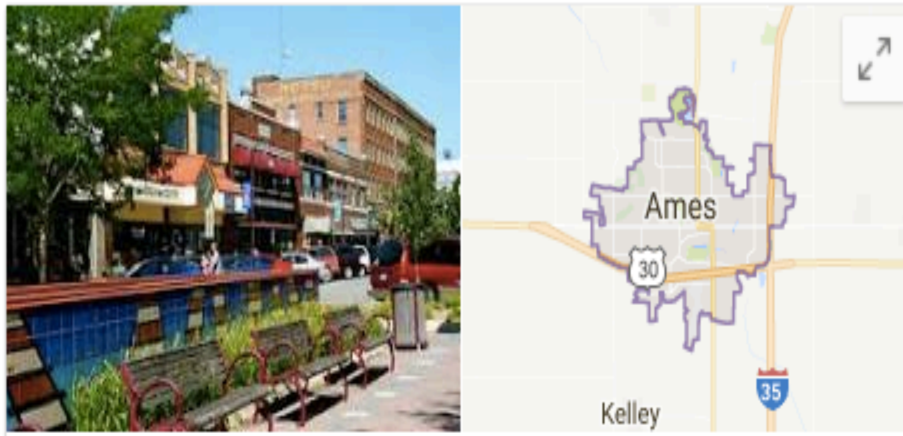


Stats/ML algorithms



Internal validity: algorithm predicts well on hold-out data from the red circle

Q: how much does a house in Ames Iowa sell?



Ames

City in Iowa

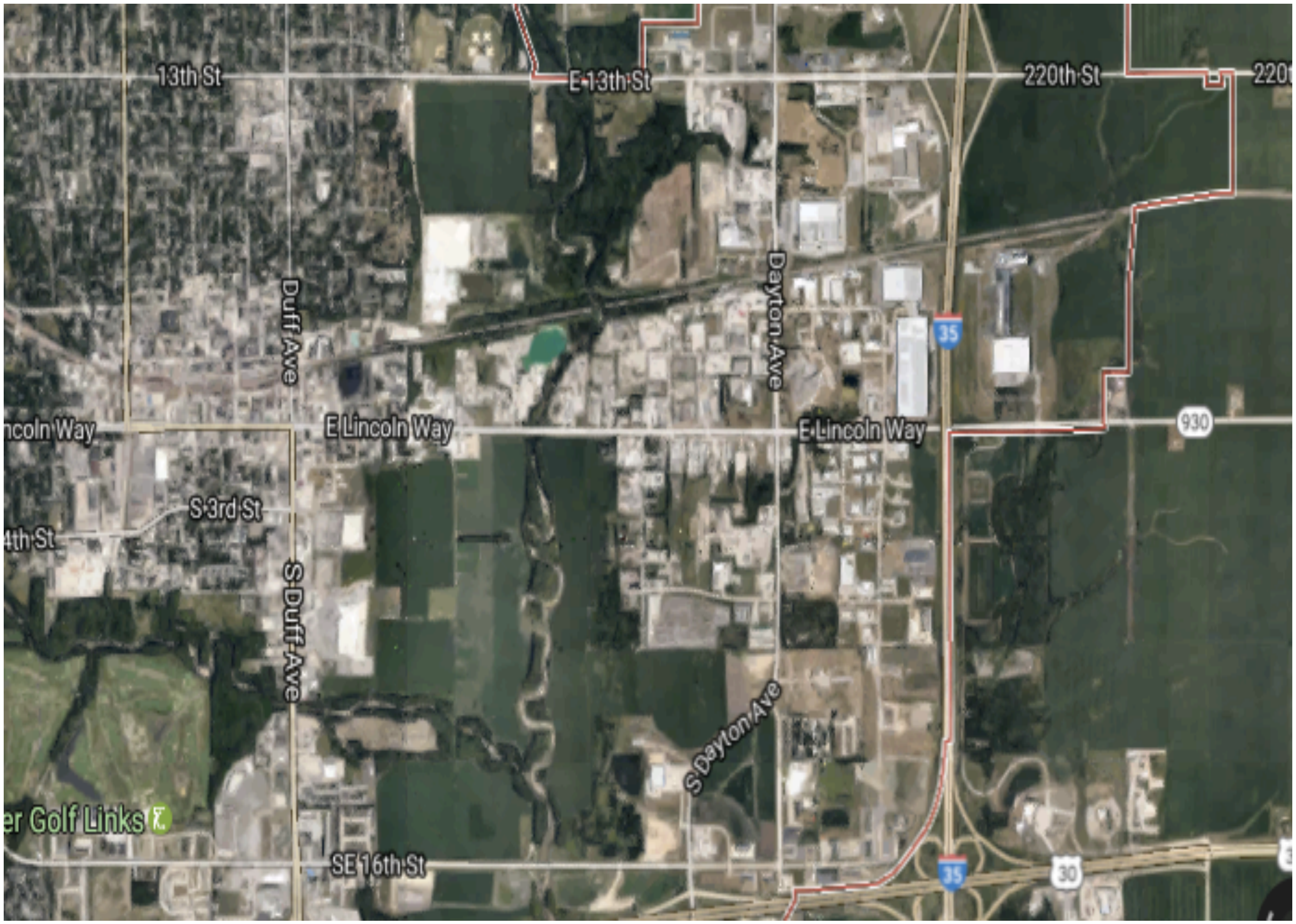
Ames is a city located in the central part of the U.S. state of Iowa in Story County. Lying approximately 30 miles north of Des Moines, it had a 2010 population of 58,965. Wikipedia

Weather: 24°F (-4°C), Wind N at 9 mph (14 km/h), 48% Humidity

Population: 61,792 (2013)

Local time: Saturday 10:54 PM

All houses in Ames, Iowa



Some math notations

Given n data units (or observations) indexed by i:

x_i predictor vector (or feature, or covariate, or attribute)

y_i response (variable) (continuous or discrete)

$$\{(x_i, y_i)\}_{i=1}^{n}$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \in R^p$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n \times p} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in R^n$$

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$$

Prediction function and mean sq. error (MSE)

Ex: $\hat{y}$ = mean, or median of y_i, i=1,…, n

$$prediction \quad function$$

$$\hat{y} = \hat{f}(x), x \in R^p$$

$$MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2$$

Test MSE
or prediction
error at (x_0, y_0)

$$(\hat{f}(x_0) - y_0)^2$$

why hat on f?

Why additive cross data units or observations?

When is it not a good Idea?

Why sq. error?

When is it not a good idea?

Prediction error is also called loss function

Comments on data: “dead” data?

Data is provided by Dean De Cock who said

“The data came to me directly from the Assessor’s Office in the form of a data dump from their records system. The initial Excel file contained 113 variables describing 3970 property sales that had occurred in Ames, Iowa between 2006 and 2010. The variables were a mix of nominal, ordinal, continuous, and discrete variables used in calculation of assessed values and included physical property measurements in addition to computation variables used in the city’s assessment process. For my purposes, a “layman’s” data set that could be easily understood by users at all levels was desirable; so I began my project by removing any variables that required special knowledge or previous calculations for their use. Most of these deleted variables were related to weighting and adjustment factors used in the city’s current modeling system.”

Dean De Cock said:

"After removal of these extraneous variables, 80 variables remained that were directly related to property sales. Although too vast to describe here individually (see the documentation file http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt), I will say that the 80 variables focus on the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property (e.g. When was it built? How big is the lot? How many square feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?).

Dean De Cock said:

"In general the 20 continuous variables relate to various area dimensions for each observation. In addition to the typical lot size and total dwelling square footage found on most common home listings, other more specific variables are quantified in the data set. Area measurements on the basement, main living area, and even porches are broken down into individual categories based on quality and type. The large number of continuous variables in this data set should give students many opportunities to differentiate themselves as they consider various methods of using and combining the variables…."

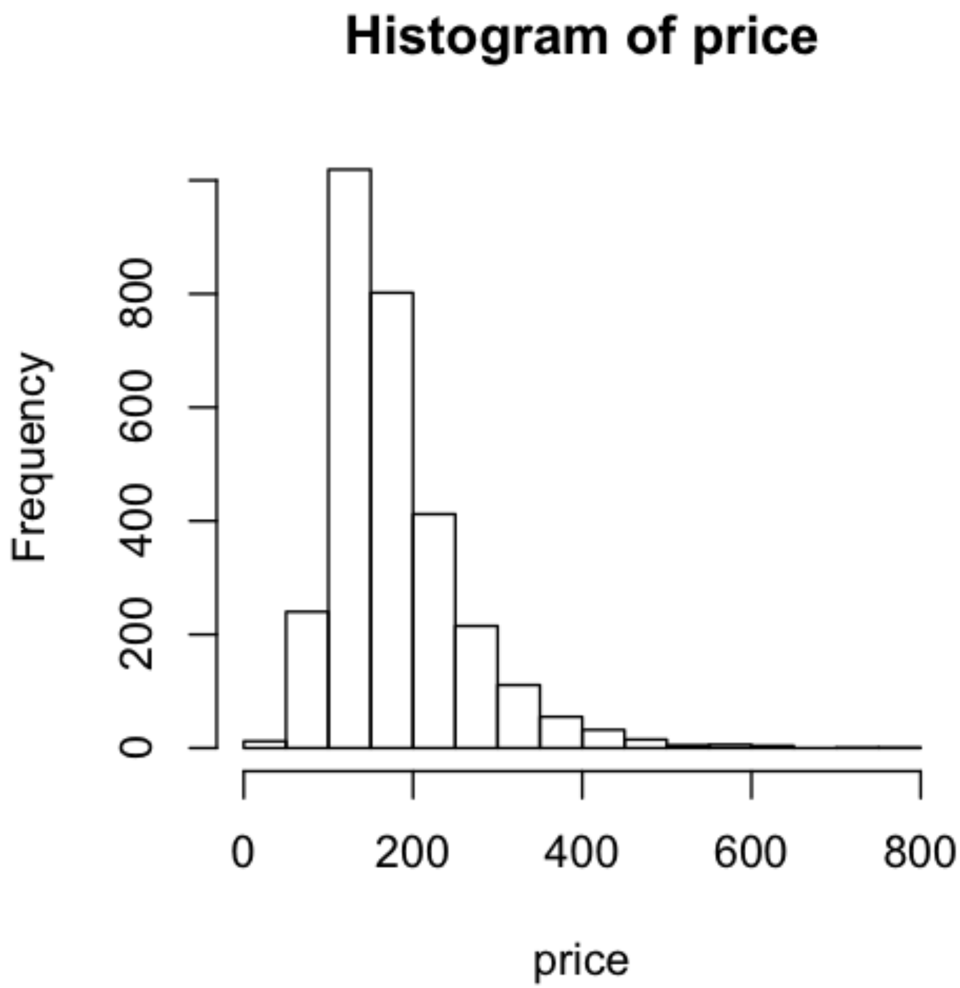
Sale prices of houses from 2006-2010

```
> price
  [1] 215.000 105.000 172.000 244.000 189.900 195.500 213.500 191.500 236.500 189.000 175.900 185.000 180.400 171.500 212.000 538.000 164.000 394.432 141.000 210.000
 [21] 190.000 170.000 216.000 149.000 149.900 142.000 126.000 115.000 184.000  96.000 105.500  88.000 127.500 149.900 120.000 146.000 376.162 306.000 395.192 290.941
 [41] 220.000 275.000 259.000 214.000 611.657 224.000 500.000 320.000 319.900 205.000 175.500 199.500 160.000 192.000 184.500 216.500 185.088 180.000 222.500 333.168
 [61] 355.000 260.400 325.000 290.000 221.000 410.000 221.500 204.500 215.200 262.500 254.900 271.500 233.000 181.000 205.000 143.000 189.000  99.500 125.000 194.500
 [81] 152.000 171.000  67.500 112.000 148.000 138.500 122.000 133.000 127.000 169.000 190.000 362.500 285.000 260.000 190.000 155.000 151.000 149.500 152.000 222.000
[101] 177.500 177.000 155.000 147.110 267.916 254.000 155.000 206.000 130.500 230.000 218.500 243.500 205.000 212.500 196.500 197.500 171.000 142.250 143.000 128.950
[121] 159.000 178.900 136.300 180.500 137.500  84.900 142.125 197.600 172.500 116.500  76.500 128.000 153.000 132.000 178.000 154.300 180.000 190.000 135.000 214.000
[141] 136.000 165.500 145.000 148.000 142.000 167.500 108.538 159.500 108.000 135.000 122.500 119.000 109.000 105.000 107.500 144.900 129.000  97.500 144.000 162.000
[161] 242.000 132.000 154.000 166.000 134.800 160.000 148.000 192.000 155.000  80.400  96.500 109.500 115.000 143.000 107.400  80.000 119.000 130.000 119.000 129.000
[181] 100.000  12.789 105.900 150.000 139.000 240.000  76.500 149.700 125.500 122.500 140.750 128.500 209.500  87.000 134.000 128.000 132.000 139.900 123.900 138.400
[201] 109.500 140.000 149.500 159.900 122.000 110.000  55.000 140.000 244.400 173.000 107.500 100.000  95.000  93.369 114.900  94.000 136.000 136.500 131.500 121.500
[221] 125.000 154.000 137.900 158.000 137.250 160.250 163.000 158.900 328.000 270.000  85.000 128.000 260.000 230.000 124.000  83.000 144.500 129.000 127.000 128.000
[241] 186.000 308.030 114.000  84.900 178.000 270.000 218.000 236.000 147.000 245.350 206.000 198.900 187.000 320.000 138.500 155.000 159.000 191.000 200.500 150.000
[261] 161.750 128.200 127.000 318.000 272.000 237.000 240.000 224.900 143.750 143.000 232.000 213.000 185.500  84.900 155.891 100.000 144.000  64.000 125.200 107.000
[281]  90.000 140.000 113.000  80.000 144.500 104.000 128.000  58.500 127.000 126.000 160.000 100.000 169.000 257.500 215.000 266.500 335.000 203.135 185.000 162.500
[301] 289.000 125.500  82.000 110.000  68.400 102.776  55.993  50.138 246.000 254.900 190.000 201.000 169.900 170.000 160.000 220.000 179.781 174.000 269.500 214.900
[321] 202.900 378.500 169.000 173.500 139.000 166.500  83.500 119.500  85.000  76.000  75.500  88.250  85.500 130.000 157.900 149.900 159.000 136.000 161.000 285.000
[341] 231.000 124.500 157.000 345.000 189.500 270.000 189.000 377.500 168.500 375.000 278.000 240.000 239.500 177.500 185.000 191.000 178.000 185.000 181.316 166.000
[361] 178.000 174.000 173.000 225.000 180.500 187.500 501.837 372.500 260.000 185.000 260.000 181.000  82.500 215.000 154.000 200.000 249.000 187.500 184.000 278.000
[381] 157.000 152.000 197.500 240.900 263.435 220.000 235.000 213.000 167.900 158.000 165.000 158.000 136.000 148.500 156.000 128.000 143.000  76.500 120.500 124.500
[401]  97.000 130.000 111.000 125.000 112.000  97.000 118.000 119.500 143.750 146.000 148.500 123.000 147.000 137.900 147.000 148.500 138.000 128.500 100.000 148.800
[421] 337.500 462.000 485.000 555.000 325.000 256.300 253.293 398.800 335.000 404.000 402.861 451.950 610.000 582.933 360.000 296.000 409.900 255.500 335.000 274.900
[441] 300.000 324.000 350.000 280.000 284.000 269.500 233.170 386.250 445.000 290.000 255.900 213.000 196.000 184.500 212.500 230.000 552.000 382.500 320.000 248.500
[461] 286.500 254.000 173.000 173.000 184.000 167.800 174.000 174.000 174.000 175.900 192.500 181.000 180.000 160.200 188.500 200.000 170.000 189.500 184.100 195.500
[481] 192.000 178.000 207.500 236.000 257.500 244.000 167.000 179.000 190.000 156.000 245.000 181.000 214.000 168.000 337.000 403.000 327.000 340.000 336.000 265.000
[501] 315.000 260.000 260.000 263.550 402.000 248.000 244.600 275.000 257.500 287.090 275.500 245.000 253.000 468.000 252.678 210.000 208.300 225.000 229.456 229.800
[521] 250.000 370.878 238.500 310.000 270.000 252.000 241.000 264.500 291.000 263.000 185.000 234.500 209.000 159.000 152.000 143.500 193.000 203.000 184.900 159.000
[541] 142.000 153.000 224.243 220.000 257.000 189.000 171.500 120.000 145.000 184.000 162.000 160.000  82.000  76.000 110.000 135.000 141.000 122.000 124.100 129.000
[561] 131.400  62.383 123.000 275.000 235.000 280.750 164.500 173.733 222.000 195.000 172.500 180.000 156.000 172.500 318.750 211.500 241.600 180.500 150.000 154.000
[581] 185.000 185.750 200.000 206.000 162.000 256.900 197.900 163.000 113.000 230.000 167.900 213.250 227.000 130.000 143.000 117.500 168.500 172.500 161.500 141.500
[601] 118.000 127.500 140.000 177.625 110.000 167.000 153.000 145.100 154.000 177.500 158.000 124.500 174.500 122.000  82.500 110.000 149.500 175.000 167.000 128.900
[621] 140.000 147.000 124.000 187.500 159.000 256.000 205.000 193.500 110.000 104.900 150.000 156.500 176.000 149.500 139.000 155.000 120.000 153.000 144.000 176.000
[641] 153.000 135.000 131.000 123.000 126.000 115.000 164.900 113.000 145.500 102.900  95.000 152.500 129.900 132.000  99.900 135.000 149.000 114.000 109.500 125.000
[661] 142.900 156.500  59.000 105.000 106.000  78.500 190.000 154.000 163.000 200.000 143.500 135.000 153.000 157.500 113.500 133.000  92.900 128.500  90.000 138.000
[681] 128.000 139.000 118.900 138.000 132.500 133.500 135.000 144.750 145.000 127.000 109.500 115.000 110.000 128.900 103.500  66.500 130.000 129.000 150.000 107.500
[701]  94.550 124.500 135.000 103.000  93.000 129.500  93.000  80.000  45.000  37.900  91.300  99.500 113.000  87.500 110.000 106.000 265.979 160.000 119.000 168.000
[721]  58.500 143.000  85.000 124.900 119.000 146.500  34.900  44.000 223.500 149.000 205.000 137.000 121.000 128.000 134.900 117.000 132.500  93.000 119.000 100.000
[741] 141.500 133.000  60.000 105.000 115.000 150.000 126.500 214.500 167.500 155.000 155.000 179.900 104.000  62.500 149.000 103.000 123.000  97.500 135.000  70.000
[761] 116.000  88.750 179.000 179.000 159.900  61.000 103.600  63.000 175.000 139.000 172.500 113.500 130.000 149.900 134.900 137.000 139.000 165.000 148.000 165.000
[781]  63.900 161.500 143.000 135.000  82.500 190.000 139.600 122.000 127.500 121.500  60.000 154.400 113.000 125.000  84.000 139.500 131.000 105.000 108.000 162.000
[801] 156.500 316.600 271.900 213.000 239.900 239.500 131.000 118.964 153.337 147.983 118.858 118.858 142.953 148.325 113.722 269.500 269.500 269.500 323.262 297.000
```

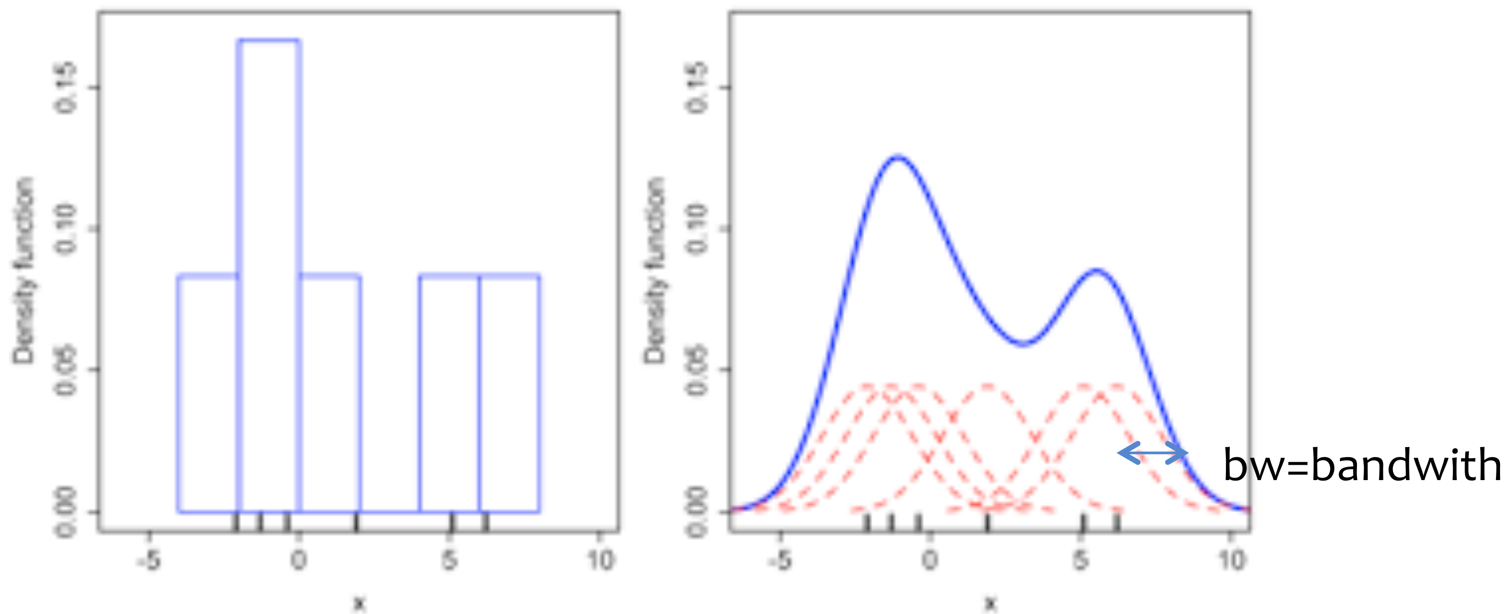
Questions about the data

- What is the data unit?
- Data has been checked and “cleaned” by Dean De Cock
- Could two prices correspond to the same house?

Plot the entire data



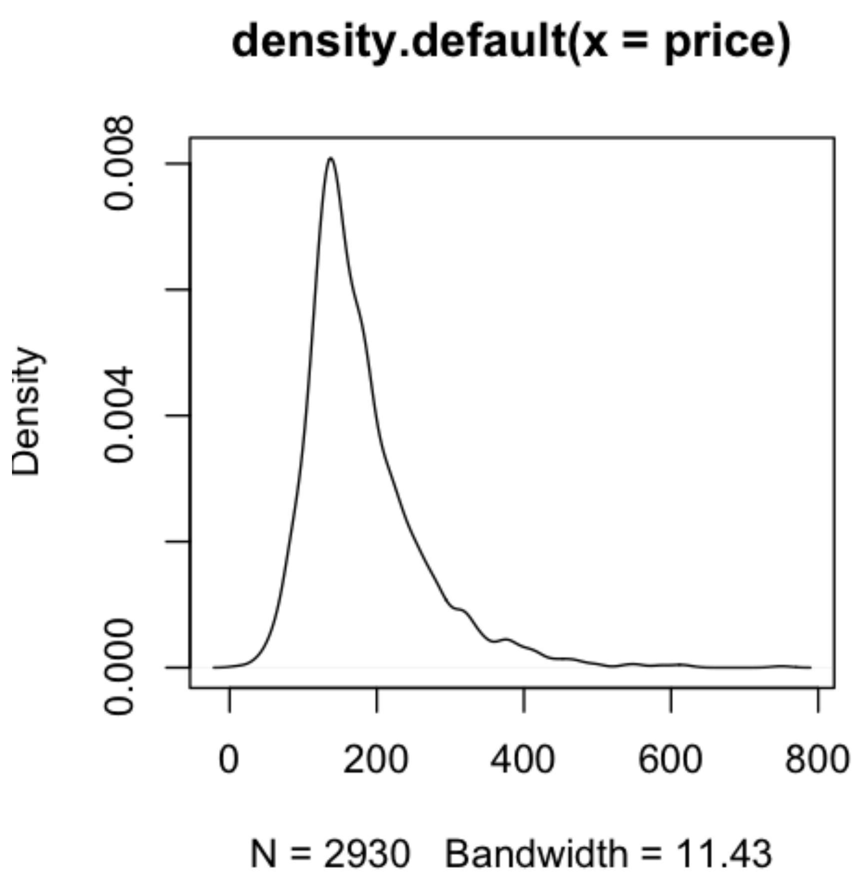
Kernel density curve

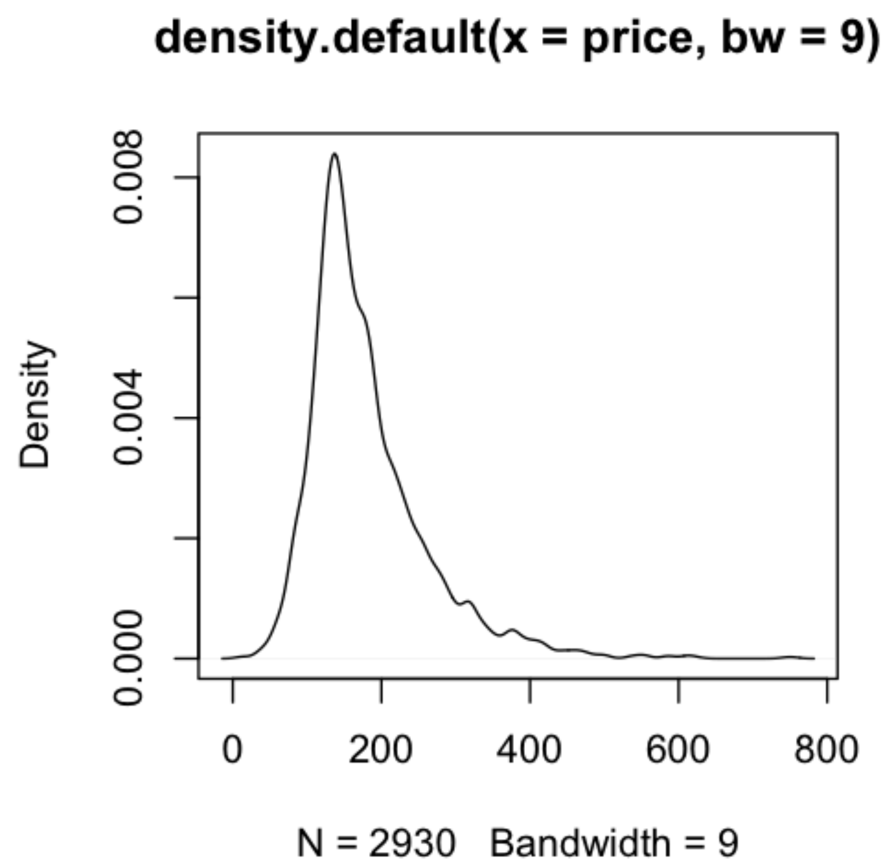


The larger the bw, the smoother the density curve, the less info of data Is retained.

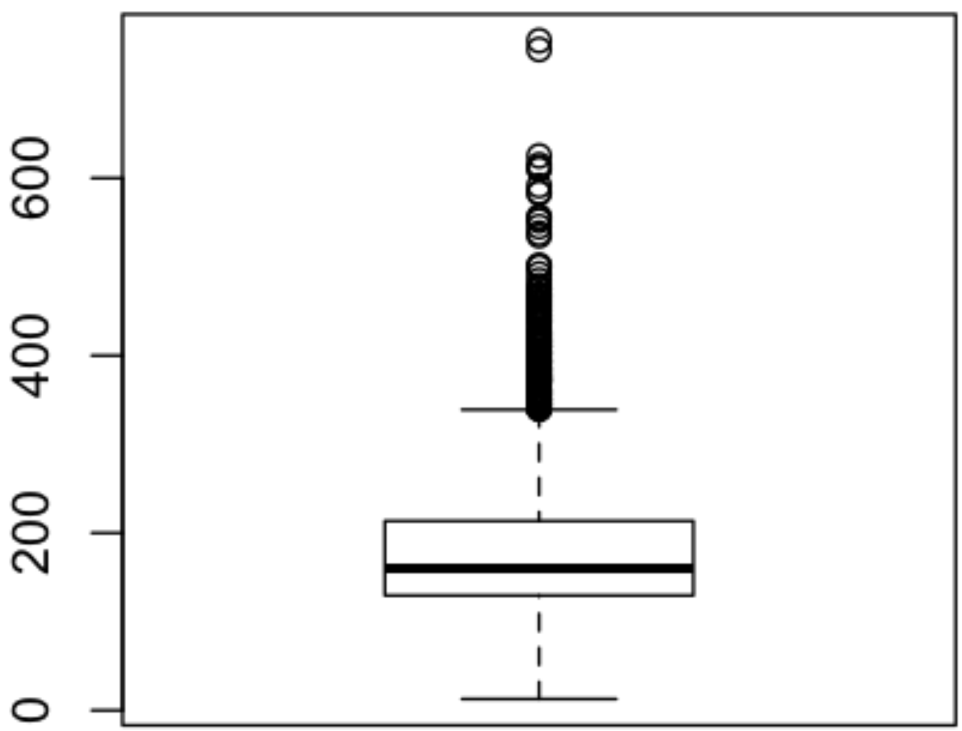
Wikipedia

Plot the entire data





Boxplot
box: median, 75th and 25th (diff=Interquartile range)
outside bars: 1.5 interquartile range
points: outliers



A new comer, Jane, to Ames in 2019

- Suppose Joe has access to the 2006-2010 sale price data
- Realtor Joe wants to give Jane a sense on the housing price while talking to her – hard to convey graph in a conversation
- Which number do you suggest Joe give Jane?

A new comer, Jane, to Ames in 2019

- Mean vs. median as "data center"

Black board derivations to show mean minimizes squared error (or L2) and median minimizes absolute value error (or L1)

Mean is sensitive to outliers, or expensive house prices.

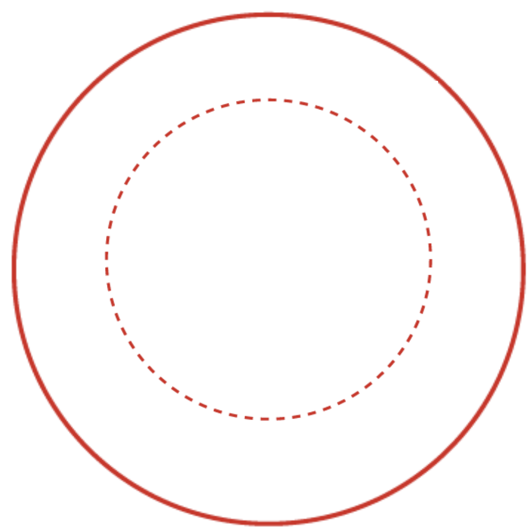
Median is robust to outliers. That is, high sale prices change the mean a lot, but not the median (unless more than half of the prices are changed).

Is it always a good idea to use median as "data center" or

a one-number summary?

A simulated prediction problem

- Joe has data as a random sample of 100 from the entire sale price data set, and he wants to choose between mean or median as the ball-park number to give to Jane as a prediction for the next random draw (as a proxy to the next house getting on the market) from the entire data set, with squared error as a prediction performance metric



Translated in math terms

- Joe has data as iid random samples $X_1, \ldots, X_n$ (n=100), he wants to predict Y, indep and identically distributed as $X_1, \ldots, X_n$
- He wants to use data to predict Y with expected squared error as a performance metric $E(\hat{Y} - Y)^2$, which is minimized by
  the expected value of Y = EY= μ
- What is we use expected absolute error?

Translated in math terms

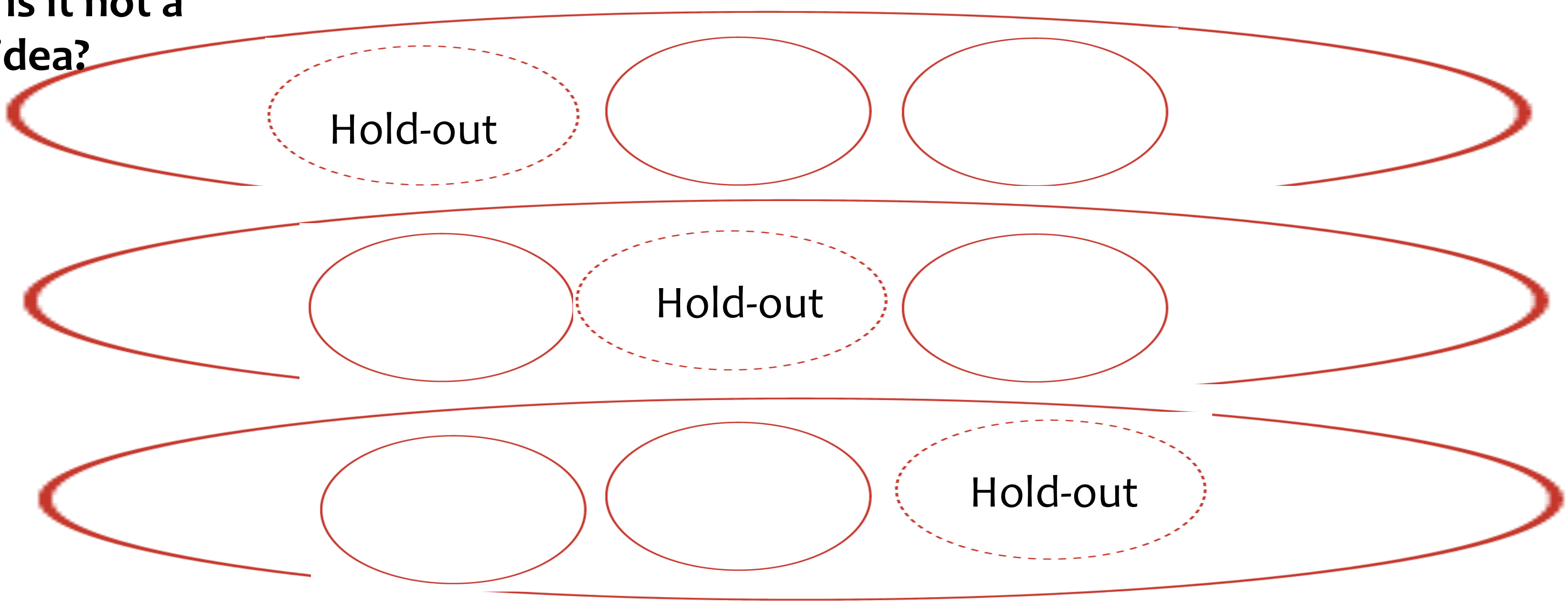
- Q1: given an predictor $\hat{Y}$,
  how to estimate prediction error $E(\hat{Y} - Y)^2$?

- Q2: can we use the estimated prediction error in Q1 to choose
  between mean and median?

Cross-validation (CV): hold-out sets are re-used and in model fitting, to estimate prediction error within one data set

Given a prediction problem with an "exchangeable" data set, CV creates k "pseudo-replicated" prediction problems or it creates K hold-out sets. K=3 below.

When is it not a good idea?



CV prediction error is the average over K-fold
(not always a good estimate of the pred. error)

Blackboard work on math notations for CV

- For K-fold CV, divide the data into K blocks of size m $Z_1, \ldots, Z_K$ indexed by i=1,…, K, where m=n/K is the number of data units in The ith block

$$Z_i = (X_i, Y_i) \in R^{m\times(p+1)}$$

- Denote the (K-1) blocks without the ith block by Z_{-i}

- One can develop a predictor based on Z_{-i} : $\hat{f}(Z_{-i})$ to predict the ith block Z_i , for example, the mean or median of the Y's in Z_{-i}

Blackboard work on math notations for CV

- We have a loss function $\ell(\cdot,\cdot)$ to measure the prediction error, then we get prediction error (PE) on the ith block.
- For $u, w \in R^m$, define $\ell(u,w) := \sum_{j=1}^{m} \ell(u_j, w_j)$
- Let $PE_i = \frac{1}{m}\ell(\hat{f}(Z_{-i}), Y_i)$
- The CV (estimated) prediction error is $CV_{\hat{f}}(\ell) = \frac{1}{K}\sum_{i=1}^{K} PE_i(\ell)$
- When the loss is squared error, we get $CV_{\hat{f}}(MSE) = \frac{1}{K}\sum_{i=1}^{K} MSE_i$

CV with K=10 (often used)

- The expected value of the CV prediction error is the prediction error if we only have 90% of the data or 90 data point when n=100

- The prediction error of sample mean with n samples is

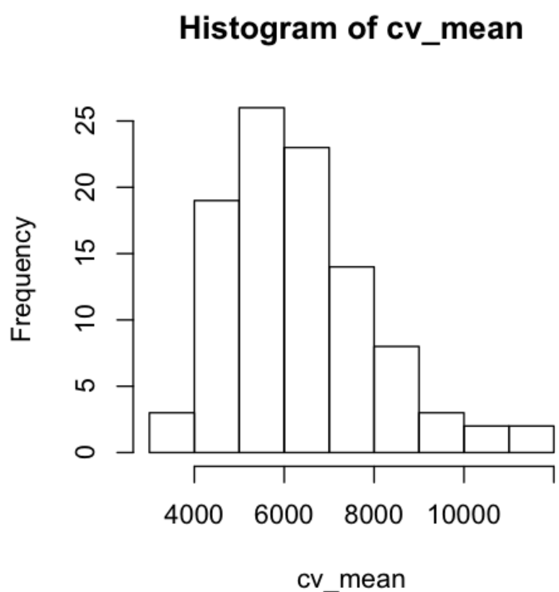
  var(Y)(1+1/n) (**black board** derivation)

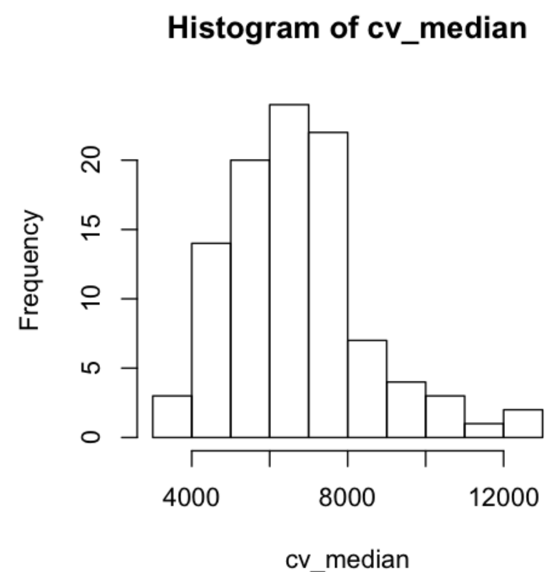
  compared with

  var(Y)(1+1/(0.9n)) for expected CV prediction error

CV K=10, correct prediction error is around 6,000, but CV error could be as small as 4,000 or as large as 10,000 or 12,000

- When we have different samples of data, the CV prediction error has quite a big variability, worse for median

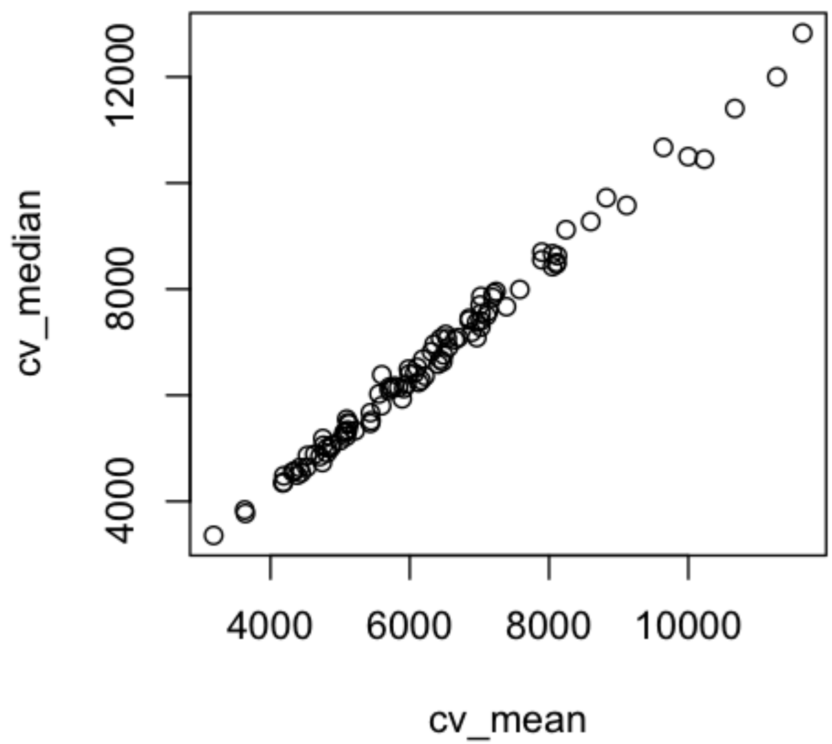




Where cv_mean is the vector that contains the CV prediction errors of mean for 100 runs of 100 samples from the price population; similarly for cv_median.

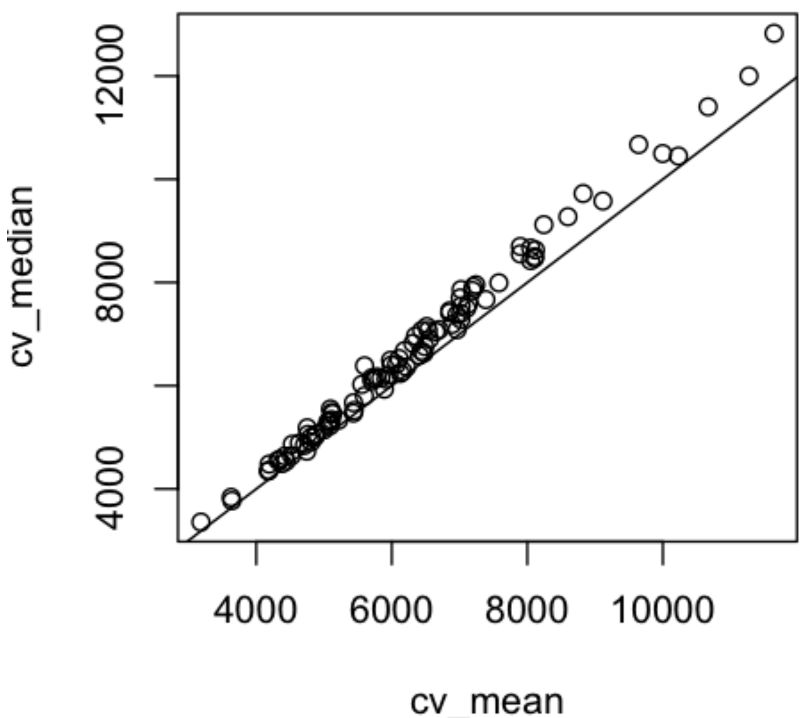
CV K=10: CV recommends mean if cv_mean is less than cv_median

- But for each run (or each 100 samples), it seems that CV can't help us decide on mean vs. median since the points fall on a line, is it the case?



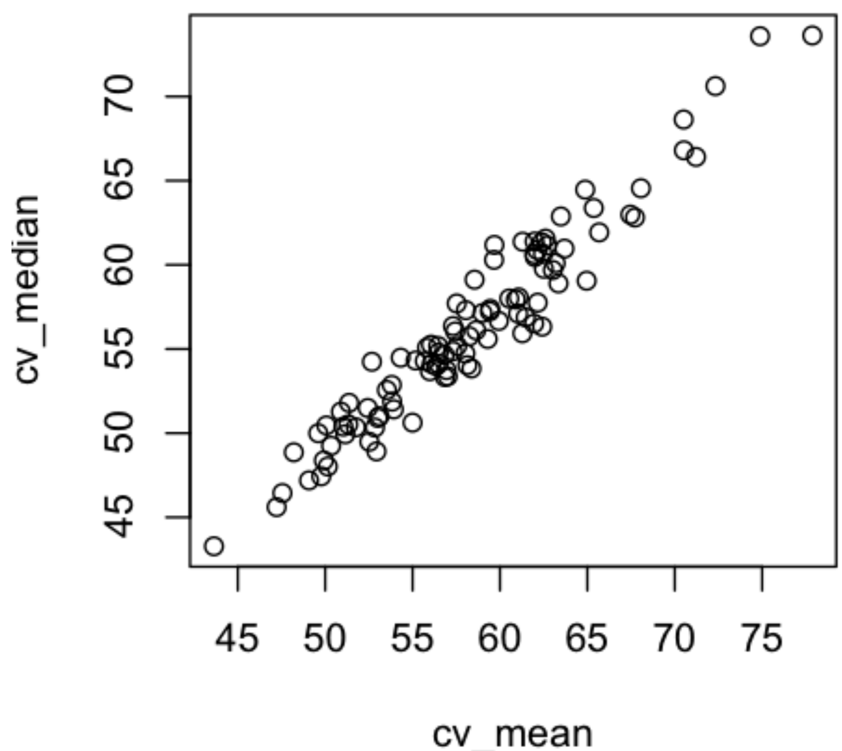
CV K=10: adding a diagonal line via abline(0,1)

- Actually for each run (or each 100 samples), CV CAN help us decide on mean vs. median almost all the time since the points are almost all above the line, which means that cv_median is larger than cv_mean – implying that mean is chosen by CV.



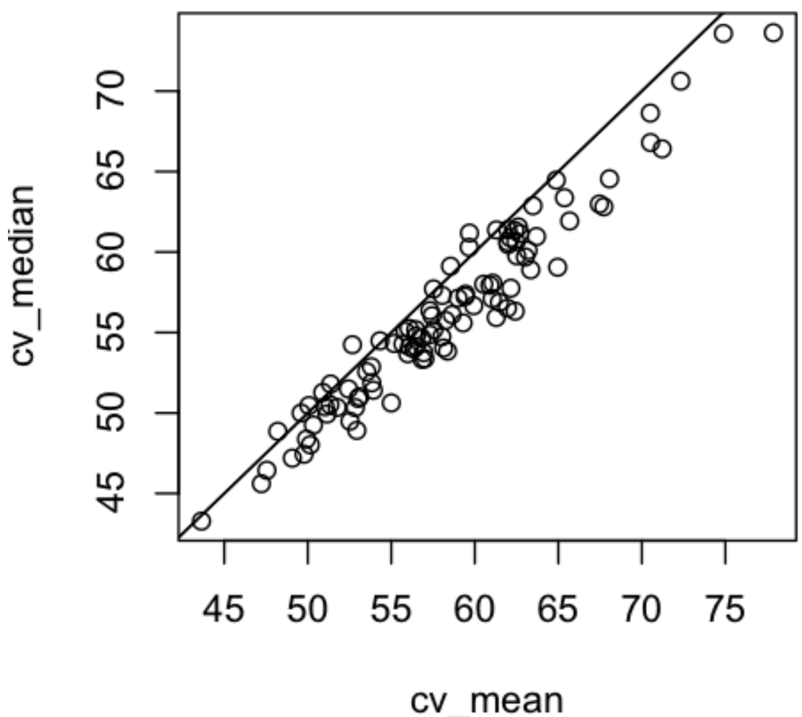
What if we use absolute value error?

Then CV recommends median, even though is is not clear in the plot below without the diagonal line.



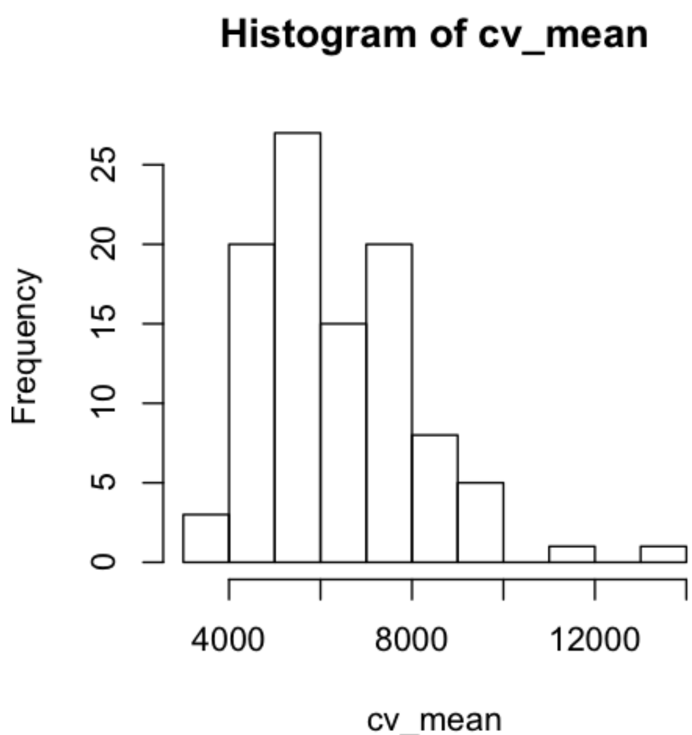
What if we use absolute value error?

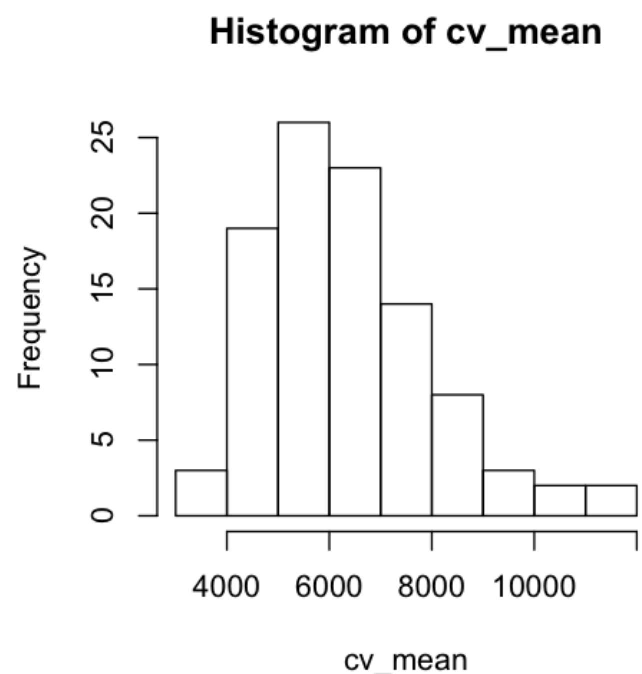
Adding the line, it is clear that CV recommends median almost all the time since the points are almost all below the line – recall that each point corresponds to one run or one set of 100 samples



CV with K=n: leave-one-out CV -mean

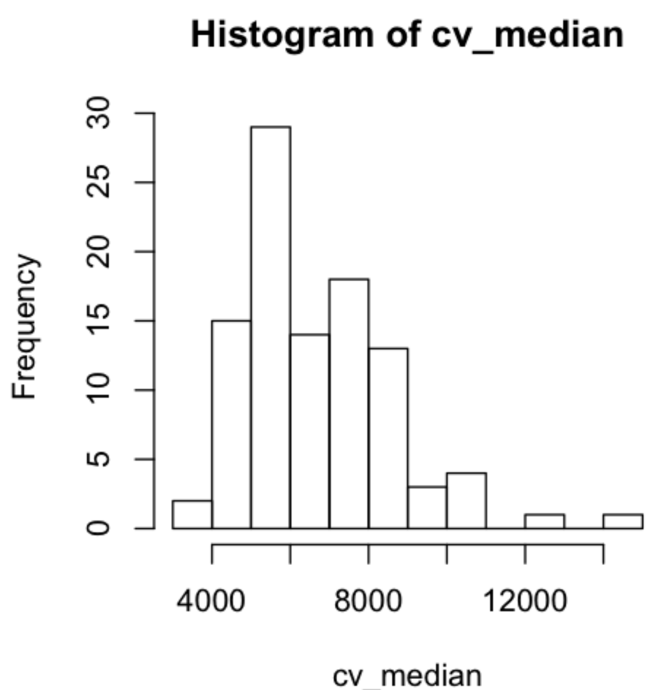
K=n has CV prediction error
closest to the n-sample prediction error
but with a larger variance, when compared to K=10

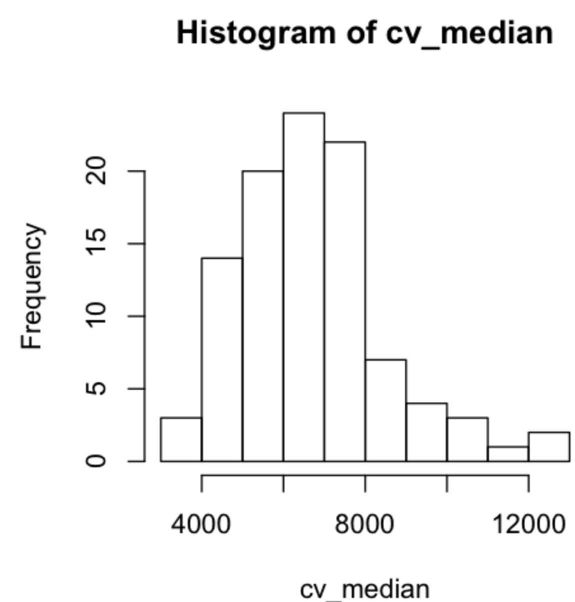




CV with K=n: leave-one-out CV -- median

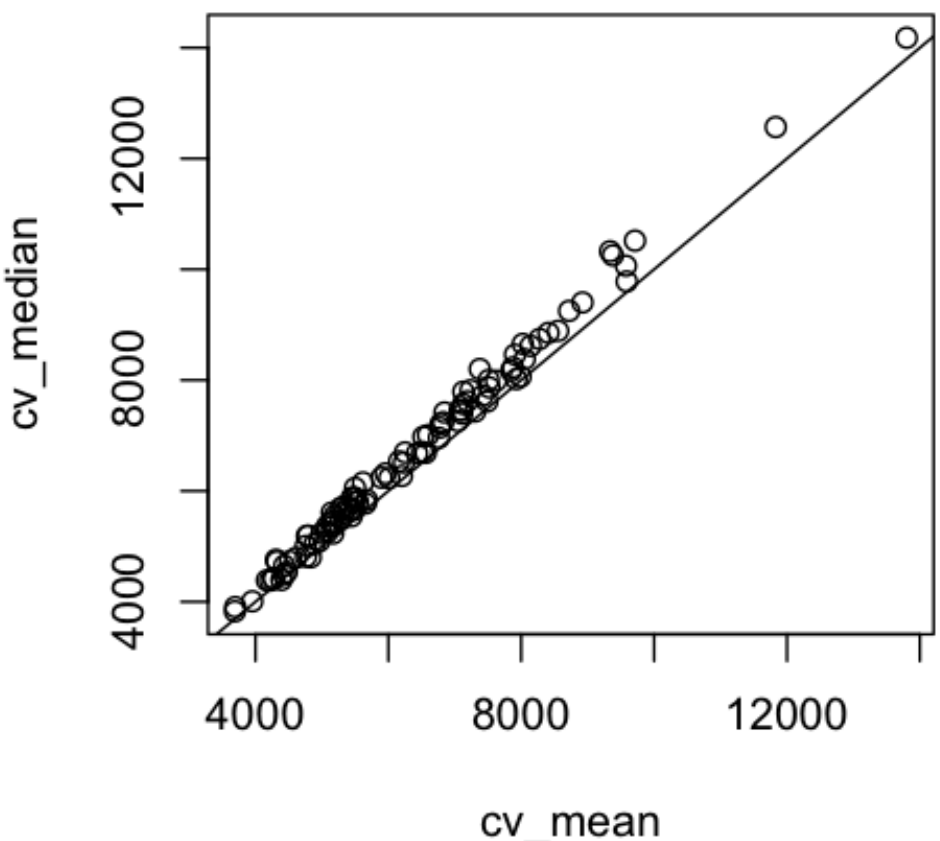
K=n has CV prediction error
closest to the n-sample prediction error
but with a larger variance (not as pronounced), when compared to K=10





But CV still chooses mean almost all the time

- K=n, leave-one-out CV



A good practice is 3-partition of data

- For exchangeable data and approx. not dependent
- First thing, set aside a test set (20-30%): that is set aside not used in model fitting and comparisons -- only for estimating the prediction error at the very end
- Divide training data into two parts
  validation set (set aside to get prediction error and used many times)
  fitting set (run CV on)

Reading assignments

- Review of eigen-value decomposition (SVD too)
- Reading of James et al book chapter on cross validation