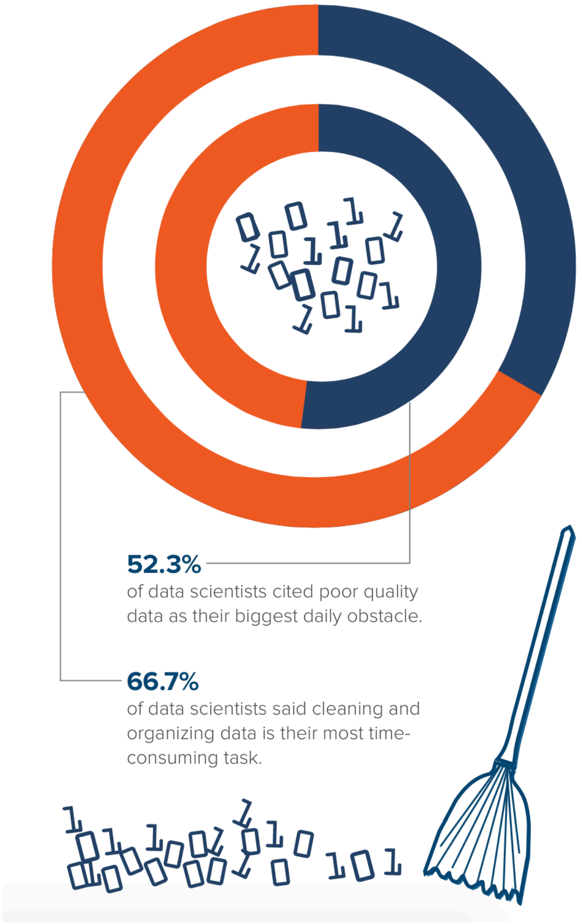
Data Wrangling


52.3%
of data scientists cited poor quality data as their biggest daily obstacle.
66.7%
of data scientists said cleaning and organizing data is their most time-consuming task.

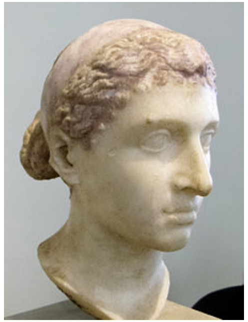
Imputation

Imputation

Name	Place of Residence	Native Language
Tutankhamun	Egypt	Egyptian
Ramses I	Egypt	Egyptian
Imhotep	Egypt	Egyptian
Cleopatra	Egypt	**?**
Plato	Greece	Greek
Socrates	Greece	Greek
Aristophanes	Greece	Greek
Euclid	**?**	Greek

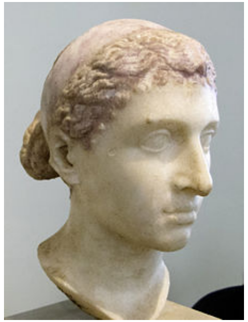
Imputation

Name	Place of Residence	Native Language
Tutankhamun	Egypt	Egyptian
Ramses I	Egypt	Egyptian
Imhotep	Egypt	Egyptian
Cleopatra	Egypt	**Greek**
Plato	Greece	Greek
Socrates	Greece	Greek
Aristophanes	Greece	Greek
Euclid	**?**	Greek



Imputation

Name	Place of Residence	Native Language
Tutankhamun	Egypt	Egyptian
Ramses I	Egypt	Egyptian
Imhotep	Egypt	Egyptian
Cleopatra	Egypt	**Greek**
Plato	Greece	Greek
Socrates	Greece	Greek
Aristophanes	Greece	Greek
Euclid	**Egypt**	Greek



Imputation

Imputation may be a preprocessing step

Imputation may be ‘the whole point’

- matrix completion
- Netflix prize

Imputation

matrix completion

1000 x 100 matrix of 1-5 star preferences

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]
[1,]	3	4	NA	3	2	3	3	3	5	4	4	2	1	5	4	4	4	NA	5	2
[2,]	5	4	NA	NA	4	4	4	3	5	5	4	4	4	5	4	4	4	4	5	NA
[3,]	5	NA	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
[4,]	5	5	3	5	5	4	4	5	5	NA	3	NA	NA	NA	3	5	5	NA	5	4
[5,]	3	4	3	4	3	NA	4	4	4	4	5	3	3	4	3	4	4	4	3	4
[6,]	4	4	3	5	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	4
[7,]	4	NA	5	2	4	NA	4	1	3	3	5	5	3	4	2	5	5	4	4	NA
[8,]	5	3	5	5	5	5	5	4	4	5	4	4	4	4	4	4	5	5	4	5
[9,]	2	4	5	4	4	3	4	4	4	3	5	5	3	4	4	4	5	4	4	4
[10,]	4	4	3	3	3	3	3	3	5	4	4	NA	3	3	3	3	4	3	5	3
[11,]	4	4	3	4	5	4	5	4	5	5	4	5	4	5	NA	4	NA	3	4	4
[12,]	4	5	2	3	4	4	4	4	5	5	5	4	4	4	3	5	5	4	5	4
[13,]	3	4	3	4	4	4	3	4	5	5	5	3	3	4	2	5	5	3	5	3
[14,]	5	5	5	NA	5	NA	5	NA	5	5	NA	5	5	5	5	5	NA	5	5	5
[15,]	NA	4	NA	4	5	4	3	4	4	5	5	3	3	4	3	4	5	3	4	4
[16,]	4	5	3	5	5	3	3	4	5	4	2	4	5	5	5	5	5	5	NA	2
[17,]	4	3	4	4	3	4	4	3	5	3	4	4	3	4	4	4	4	3	4	4
[18,]	5	5	4	NA	4	5	4	NA	5	4	5	5	5	4	5	5	5	2	5	4
[19,]	4	3	4	5	4	5	4	4	4	4	4	4	3	4	3	5	4	4	4	4

Imputation

Missing Completely at Random (**MCAR**)?

Imputation

matrix completion

1000 x 100 matrix of 1-5 star preferences

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]
[1,]	3	4	NA	3	2	3	3	3	5	4	4	2	1	5	4	4	4	NA	5	2
[2,]	5	4	NA	NA	4	4	4	3	5	5	4	4	4	5	4	4	4	4	5	NA
[3,]	5	NA	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
[4,]	5	5	3	5	5	4	4	5	5	NA	3	NA	NA	NA	3	5	5	NA	5	4
[5,]	3	4	3	4	3	NA	4	4	4	4	5	3	3	4	3	4	4	4	3	4
[6,]	4	4	3	5	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	4
[7,]	4	NA	5	2	4	NA	4	1	3	3	5	5	3	4	2	5	5	4	4	NA
[8,]	5	3	5	5	5	5	5	4	4	5	4	4	4	4	4	4	5	5	4	5
[9,]	2	4	5	4	4	3	4	4	4	3	5	5	3	4	4	4	5	4	4	4
[10,]	4	4	3	3	3	3	3	3	5	4	4	NA	3	3	3	3	4	3	5	3
[11,]	4	4	3	4	5	4	5	4	5	5	4	5	4	5	NA	4	NA	3	4	4
[12,]	4	5	2	3	4	4	4	4	5	5	5	4	4	4	3	5	5	4	5	4
[13,]	3	4	3	4	4	4	3	4	5	5	5	3	3	4	2	5	5	3	5	3
[14,]	5	5	5	NA	5	NA	5	NA	5	5	NA	5	5	5	5	5	NA	5	5	5
[15,]	NA	4	NA	4	5	4	3	4	4	5	5	3	3	4	3	4	5	3	4	4
[16,]	4	5	3	5	5	3	3	4	5	4	2	4	5	5	5	5	5	5	NA	2
[17,]	4	3	4	4	3	4	4	3	5	3	4	4	3	4	4	4	4	3	4	4
[18,]	5	5	4	NA	4	5	4	NA	5	4	5	5	5	4	5	5	5	2	5	4
[19,]	4	3	4	5	4	5	4	4	4	4	4	4	3	4	3	5	4	4	4	4

Imputation

Missing Completely at Random (**MCAR**)?

Ukrainian Presidential Election 2014

Choice	**Votes**
Petro Poroshenko	9,857,308
Other Candidates	8,162,196
? (Eligible votes that were not cast)	12,079,742

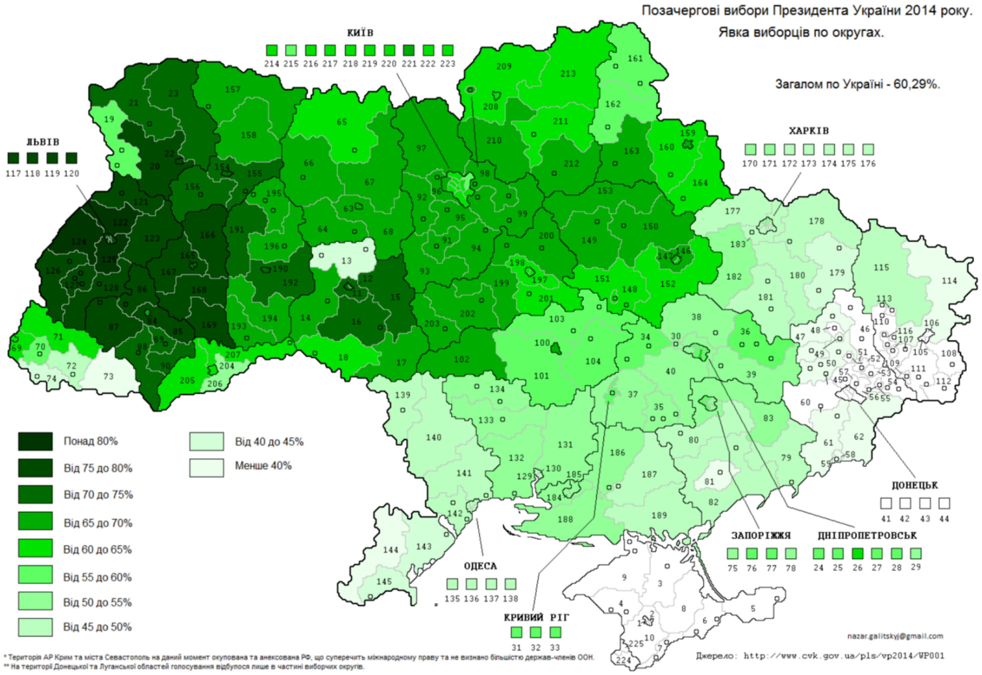
from wikipedia article, “Ukrainian presidential election, 2014”

Imputation

Missing Completely at Random (**MCAR**)?

2014 Ukrainian Presidential Election Turn-out

from wikipedia article, “Ukrainian presidential election, 2014”

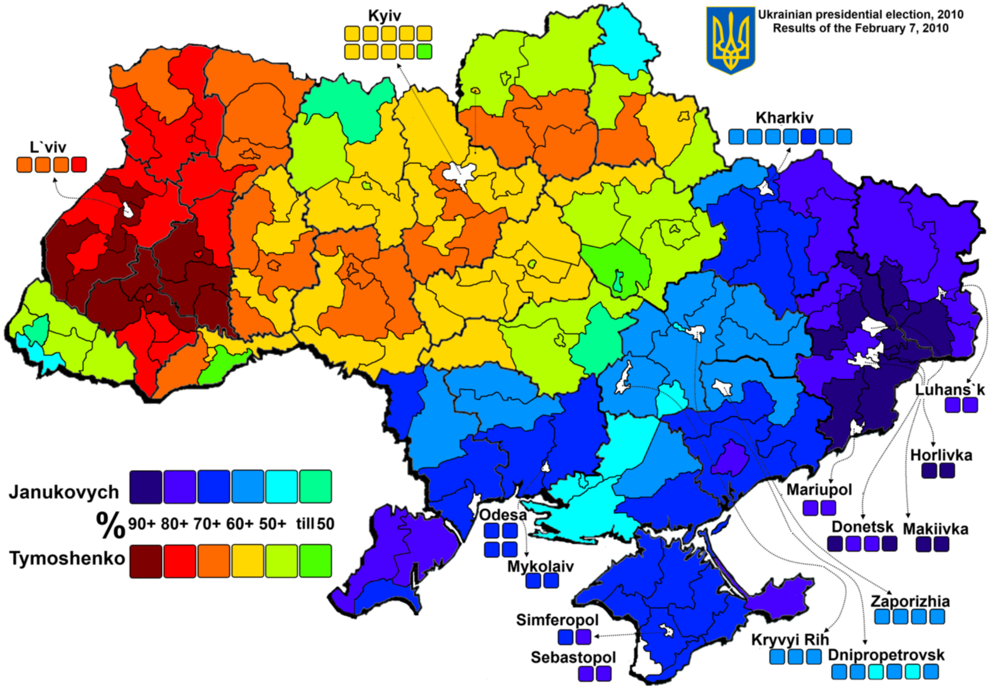


Imputation

Missing Completely at Random (**MCAR**)?

2010
Ukrainian
Presidential
Election
Results

from wikipedia article, "Ukrainian presidential election, 2014"



Imputation

Missing Completely at Random (**MCAR**)?

A fundamental assumption in most imputation.

To test this assumption, for categorical data, code missing data as "missing" to see whether it is really MCAR, or whether a supervised learning method can find a pattern to the missingness.

Imputation

- delete all observations that are missing data
- fill in missing values using the mean or median value for that feature
- find procedures that can utilize the correlations and structure in the data to predict the missing values

Imputation

1. Knn Imputation

2. SVD Imputation

3. Regression Imputation (CART works well)

4. Directly use machine learning methods that can accommodate missing data (surrogate splits, CART, MARS) without needing any imputation

Imputation

K nearest neighbors algorithm

"impute" package in R

1. fill-in missing values using the mean or median for that variable

2. compute the distance between the observation missing a value and all other observations to find the k closest observations

3. ignore the variable that is missing the value when computing the distance

4. average the values for that variable of the k nearest neighbors

Imputation

SVD Imputation

"softImpute" package in R

1. Initialize the missing data with variable means.

2. Use a rank-k SVD of the data matrix X to impute the missing locations. Repeat.

   Or, equivalently

   Find the k largest eigenvectors of the sample covariance matrix X^TX, which are the first k principal components of the data matrix X. Project all observations onto this principal component subspace. Use these projected values to update the missing data values. Repeat.

Imputation

1000 x 100 matrix of preferences

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]
[1,]	3	4	NA	3	2	3	3	3	5	4	4	2	1	5	4	4	4	NA	5	2
[2,]	5	4	NA	NA	4	4	4	3	5	5	4	4	4	5	4	4	4	4	5	NA
[3,]	5	NA	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
[4,]	5	5	3	5	5	4	4	5	5	NA	3	NA	NA	NA	3	5	5	NA	5	4
[5,]	3	4	3	4	3	NA	4	4	4	4	5	3	3	4	3	4	4	4	3	4
[6,]	4	4	3	5	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	4
[7,]	4	NA	5	2	4	NA	4	1	3	3	5	5	3	4	2	5	5	4	4	NA
[8,]	5	3	5	5	5	5	5	4	4	5	4	4	4	4	4	4	5	5	4	5
[9,]	2	4	5	4	4	3	4	4	4	3	5	5	3	4	4	4	5	4	4	4
[10,]	4	4	3	3	3	3	3	3	5	4	4	NA	3	3	3	3	4	3	5	3
[11,]	4	4	3	4	5	4	5	4	5	5	4	5	4	5	NA	4	NA	3	4	4
[12,]	4	5	2	3	4	4	4	4	5	5	5	4	4	4	3	5	5	4	5	4
[13,]	3	4	3	4	4	4	3	4	5	5	5	3	3	4	2	5	5	3	5	3
[14,]	5	5	5	NA	5	NA	5	NA	5	5	NA	5	5	5	5	5	NA	5	5	5
[15,]	NA	4	NA	4	5	4	3	4	4	5	5	3	3	4	3	4	5	3	4	4
[16,]	4	5	3	5	5	3	3	4	5	4	2	4	5	5	5	5	5	5	NA	2
[17,]	4	3	4	4	3	4	4	3	5	3	4	4	3	4	4	4	4	3	4	4
[18,]	5	5	4	NA	4	5	4	NA	5	4	5	5	5	4	5	5	5	2	5	4
[19,]	4	3	4	5	4	5	4	4	4	4	4	4	3	4	3	5	4	4	4	4
[20,]	5	5	NA	5	5	3	NA	5	5	5	5	5	4	5	5	5	4	5	5	4

Imputation

1000 x 100 matrix of preferences

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]
[1,]	NA	NA	NA	NA	NA	3	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA
[2,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[3,]	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	NA	NA
[4,]	NA	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	NA	NA	NA	NA
[5,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	NA	NA	NA	NA
[6,]	NA	NA	NA	NA	NA	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	NA
[7,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[8,]	NA	NA	NA	5	NA	NA	NA	NA	NA	5	NA	NA	4	NA	NA	NA	NA	NA	NA	NA
[9,]	NA	4	NA	NA	NA	NA	NA	4	NA	NA	NA	NA	NA	NA	NA	4	NA	NA	NA	NA
[10,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	NA	NA
[11,]	NA	NA	NA	4	NA	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[12,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	NA	NA	5	4	NA	NA
[13,]	NA	NA	3	NA	NA	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[14,]	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[15,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	NA	NA	NA	NA	NA	NA	NA	NA	4	NA
[16,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	NA	NA	NA	NA	NA	NA
[17,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[18,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	NA	NA	NA	NA	NA	NA
[19,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4
[20,]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Imputation

Knn

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	4.000000	4.000000	3.000000	3.714286	4.142857	3.000000	4.000000	4.142857	4.571429	4.000000
[2,]	4.142857	4.142857	2.571429	3.428571	4.142857	3.285714	4.142857	4.285714	4.571429	4.000000
[3,]	5.000000	4.571429	2.714286	4.142857	4.714286	3.285714	3.857143	4.142857	4.714286	4.142857
[4,]	3.714286	5.000000	2.714286	3.428571	4.142857	2.428571	3.857143	4.142857	4.285714	4.000000
[5,]	4.142857	4.285714	2.571429	3.571429	4.571429	3.571429	4.571429	3.714286	4.714286	4.571429
[6,]	3.428571	3.857143	2.857143	4.142857	4.571429	4.000000	3.571429	4.285714	4.714286	4.428571
[7,]	4.428571	3.857143	3.571429	3.857143	4.428571	2.857143	4.142857	4.142857	4.285714	4.428571
[8,]	3.714286	3.571429	3.285714	5.000000	3.857143	4.000000	3.285714	4.428571	4.571429	5.000000
[9,]	3.714286	4.000000	2.428571	3.857143	4.000000	4.428571	3.428571	4.000000	3.714286	4.000000
[10,]	4.000000	4.142857	3.000000	3.428571	4.142857	4.000000	3.571429	4.285714	4.142857	4.428571

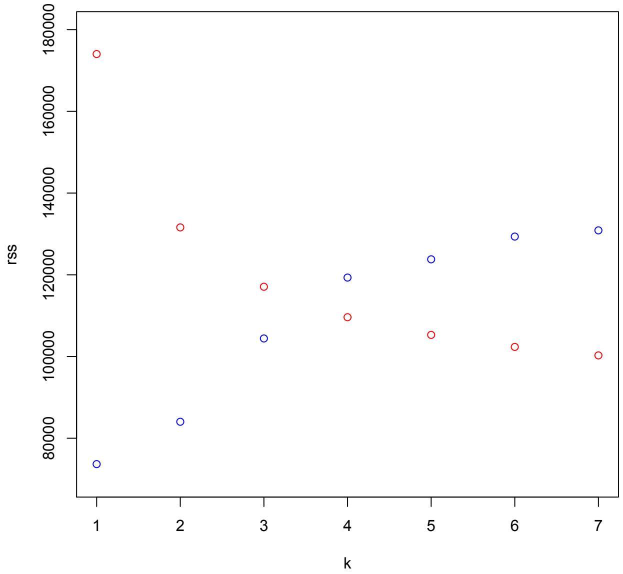
Imputation

Choosing a test set,

How do I determine test error for tuning/choosing my imputation technique?

Imputation

Since the rankings are on a discrete 1-5 scale, and assuming each ranking is equally probable on each value {1,2,3,4,5}, then the expected sum of squared errors for the 80492 missing values should be 321,968.



Red – Knn
Blue – SVD

Imputation

1. Knn Imputation
2. SVD Imputation
3. Regression Imputation (CART works well)
4. Directly use machine learning methods that can accommodate missing data (surrogate splits, CART, MARS) without needing any imputation

Questions?