Train In Data

Random Sample Imputation

Random sample imputation: definition

- Random sampling consist in taking a random observation from the pool of available observations of the variable, and using that randomly extracted value to fill the NA.

- Suitable for both numerical and categorical variables

Random sample imputation: example

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	
	Kia
60	Ford
120	BMW
200	Kia

Random Sample

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	**Ford**
100	Kia
60	Ford
120	BMW
90	**Kia**
200	Kia

Random sample imputation: Assumptions

- Data is missing at random
- The idea is to replace the population of missing values with a population of values with the same distribution of the original variable.

Random sample imputation: Advantages

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated in production (during model deployment)
- Preserves the variance of the variable

Random sample imputation: Limitations

- Randomness
- The relationship of imputed variables with other variables may be affected if there are a lot of NA
- Memory heavy for deployment, as we need to store the original training set to extract values from and replace the NA in coming observations.

Random sample imputation: Randomness

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	
	Kia
60	Ford
120	BMW
200	Kia

Random Sample 1

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	**Ford**
100	Kia
60	Ford
120	BMW
90	**Kia**
200	Kia

Prediction 1

Prediction
1000
1200
500
4000
2000
1000
900
1600
3000
1100
500

Random sample imputation: Randomness

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	
	Kia
60	Ford
120	BMW
200	Kia

Random Sample 2

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	**Kia**
90	Kia
60	Ford
120	BMW
120	**BMW**
200	Kia

Prediction 2

Prediction
1000
1200
500
4000
2000
900
110
1600
3000
3000
500

Random sample imputation: Randomness

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	
	Kia
60	Ford
120	BMW
200	Kia

Random Sample 3

Price	Make
100	Ford
90	Ford
50	Fiat
40	BMW
20	Ford
100	BMW
200	Kia
60	Ford
120	BMW
120	Ford
200	Kia

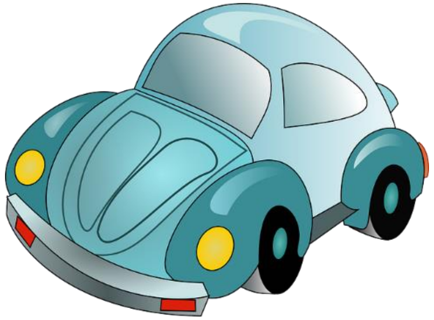
Prediction 3

Prediction
1000
1200
500
4000
2000
3500
500
1600
3000
800
500

Random sample imputation: Randomness

- Every time we score the same observation, we may obtain a different prediction
- Unwanted side-effect
- **Set the seed using other variables in the dataset**

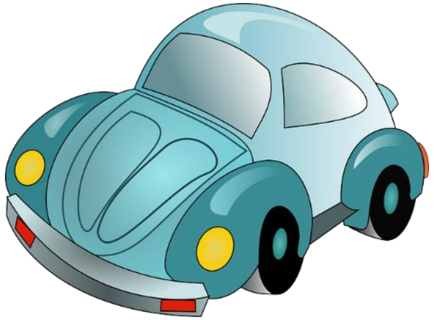
Single observation – round 1



Today

Make	Model	Color	Engine	Fuel
Ford		Blue	100	

Single observation – round 1



Today

Make	Model	Color	Engine	Fuel
Ford		Blue	100	

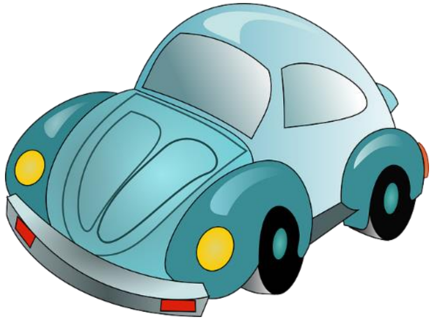
→

Make	Model	Color	Engine	Fuel
Ford	**Fiesta**	Blue	100	**Gas**

→

15 minutes

Single observation – round 2



Tomorrow

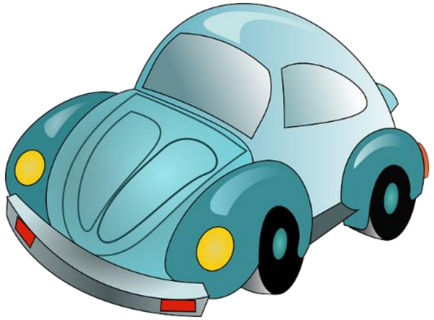
Make	Model	Color	Engine	Fuel
Ford		Blue	100	

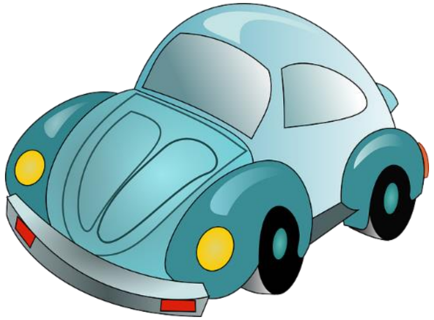
→

Make	Model	Color	Engine	Fuel
Ford	Ka	Blue	100	efuel

→ 5 minutes

identical observations ≠ predictions

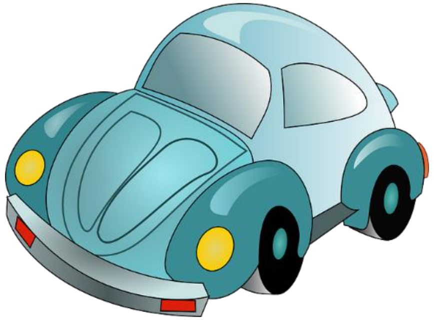


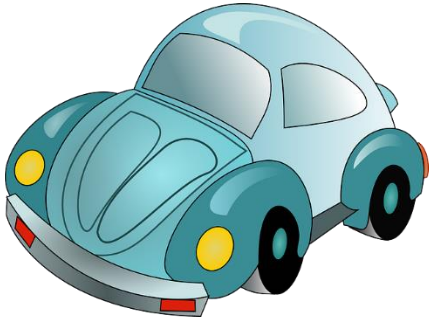


Make	Model	Color	Engine	Fuel
Ford		Blue	100	

Make	Model	Color	Engine	Fuel
Ford		Blue	100	

identical observations ≠ predictions





Make	Model	Color	Engine	Fuel
Ford		Blue	100	

⬇

Make	Model	Color	Engine	Fuel
Ford	Fiesta	Blue	100	Gas

15 minutes

Make	Model	Color	Engine	Fuel
Ford		Blue	100	

⬇

Make	Model	Color	Engine	Fuel
Ford	Ka	Blue	100	efuel

5 minutes

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
  - Random Sample imputation with pandas
  - Effect of the imputation on:
    - Variable distribution
    - Outliers

Random sample Imputation

- The population of values used to replace NA should be the train set.
- To avoid over-fitting

THANK YOU

www.trainindata.com