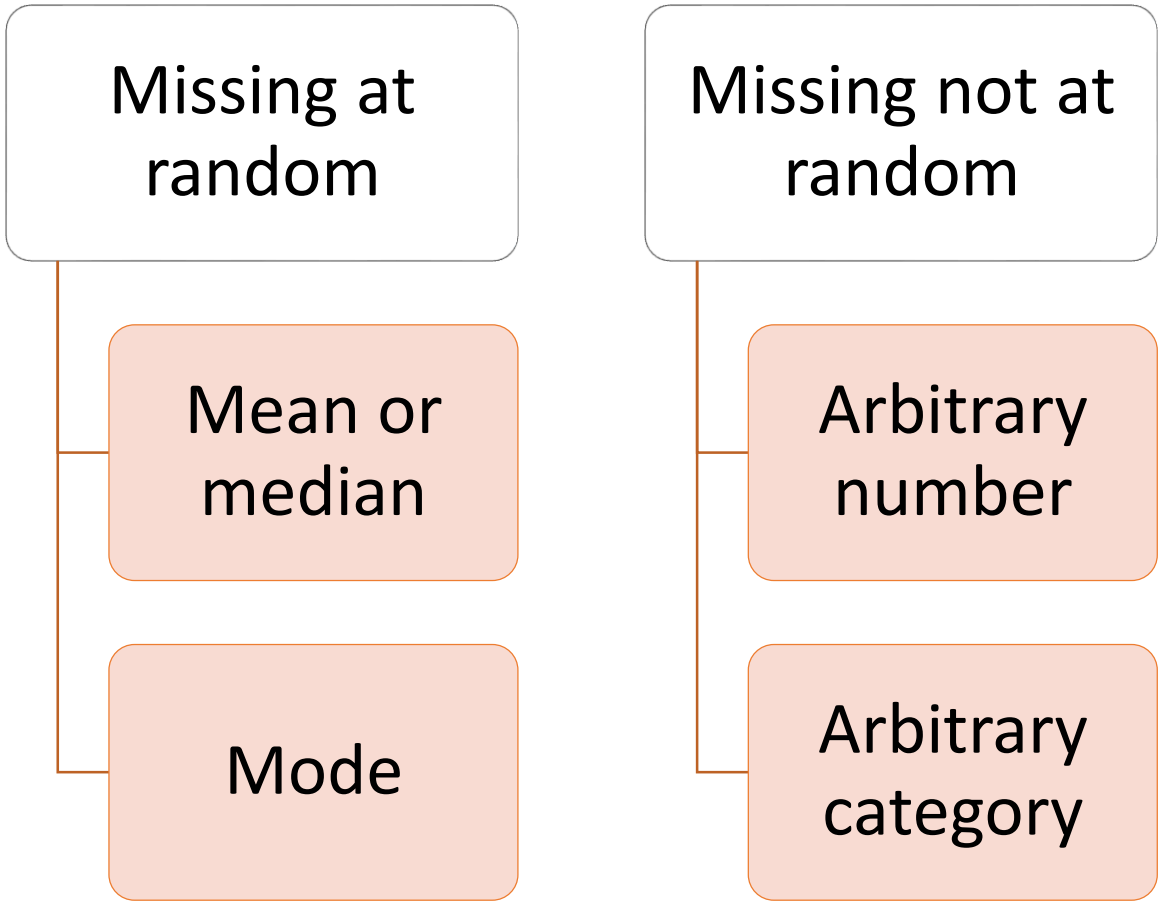
Train In Data

Imputation: considerations

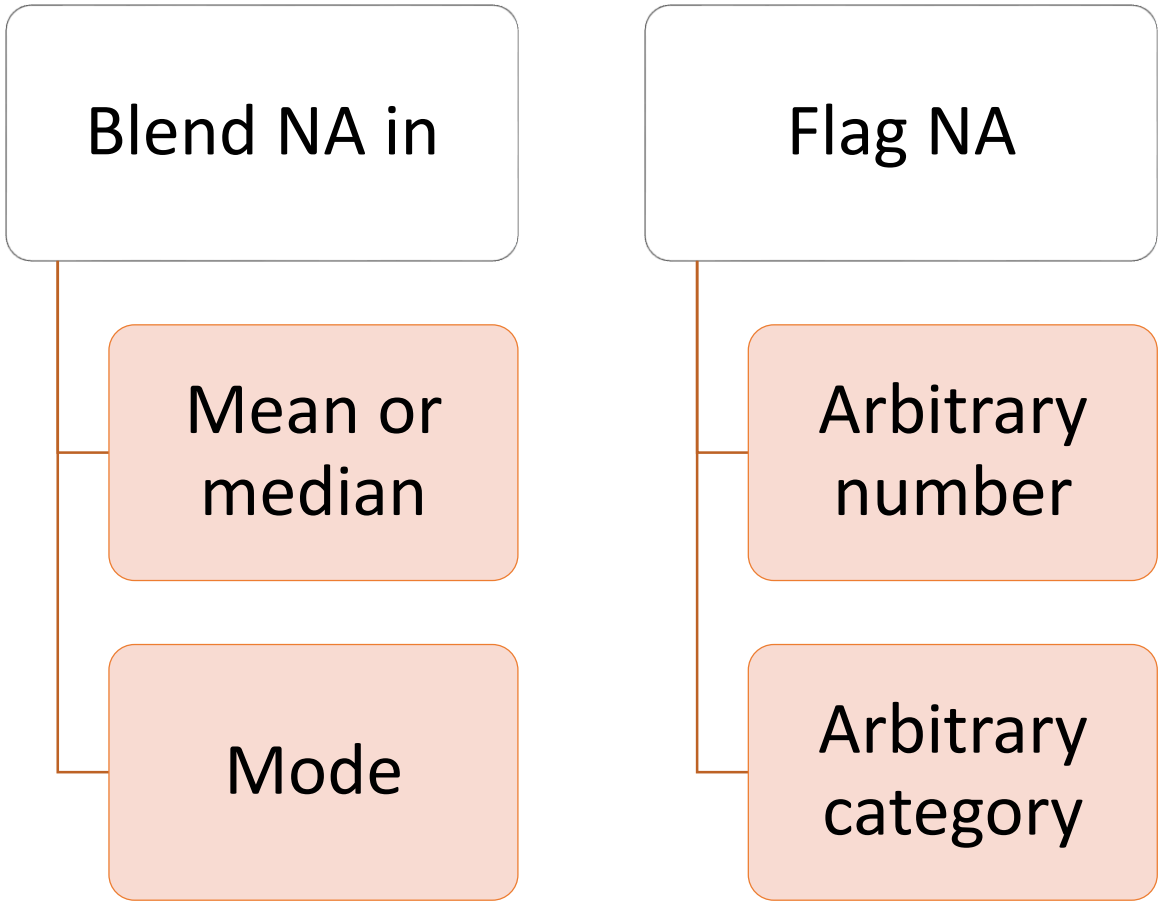
Basic imputation: Advantages

- Easy to implement
- Fast way to obtain complete datasets

Basic imputation: Assumptions



Basic imputation: Assumptions



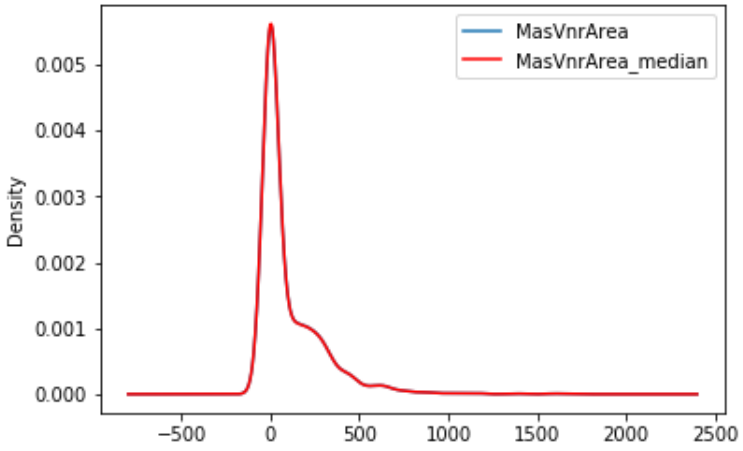
Statistic methods + missing indicator

Basic imputation - effects

- Distort the variable distribution (including variance)
- Distort relationship to other variables (i.e., covariance)
- **More missing values ➔ greater distortions**
- Particularly important if we are interested in joint distributions

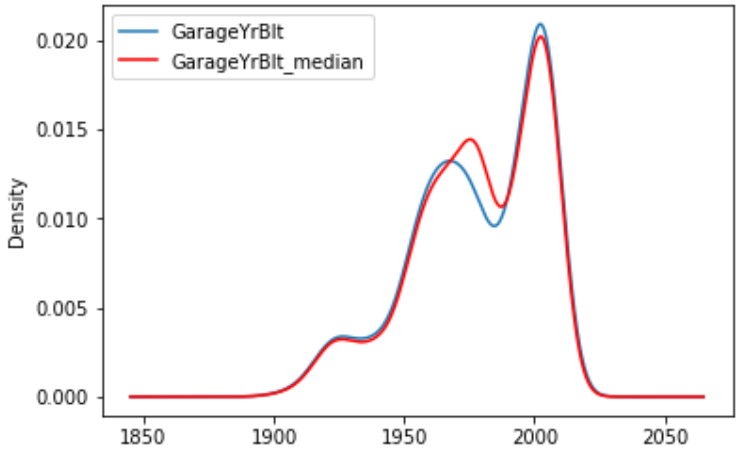
Mean / Median Imputation effects

MasVnrArea 0.5% missing obs



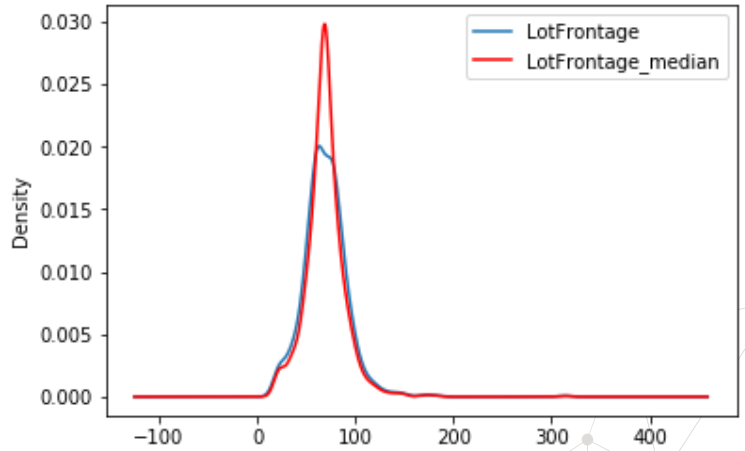
Variance: 32983
Variance after imputation: 32874

GarageYrBlt 5.5% missing obs



Variance: 624
Variance after imputation: 591

LotFrontage 17% missing obs

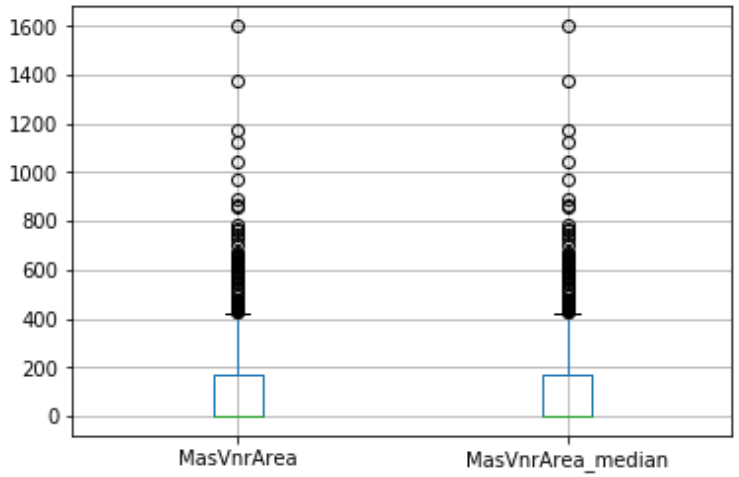


Variance: 532
Variance after imputation: 434

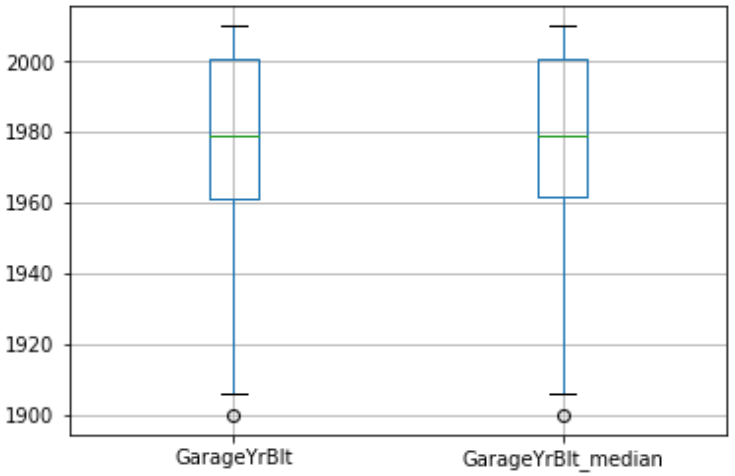
- On variance and distribution shape

Mean / Median Imputation effects

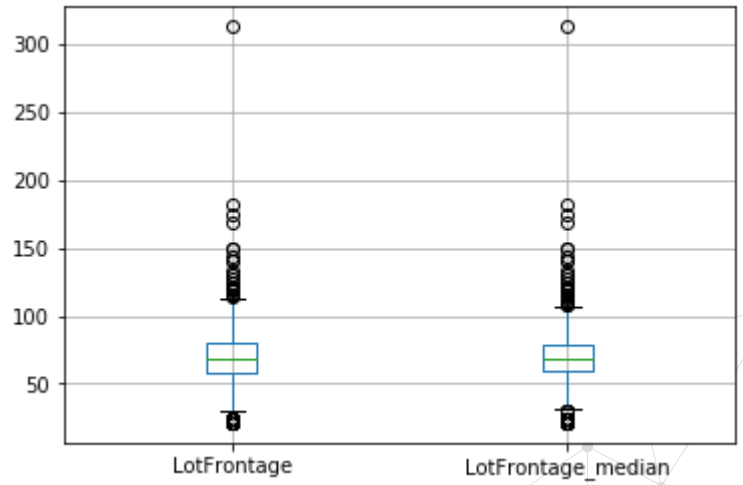
MasVnrArea 0.5% missing obs



GarageYrBlt 5.5% missing obs



LotFrontage 17% missing obs



More “outliers” after imputation

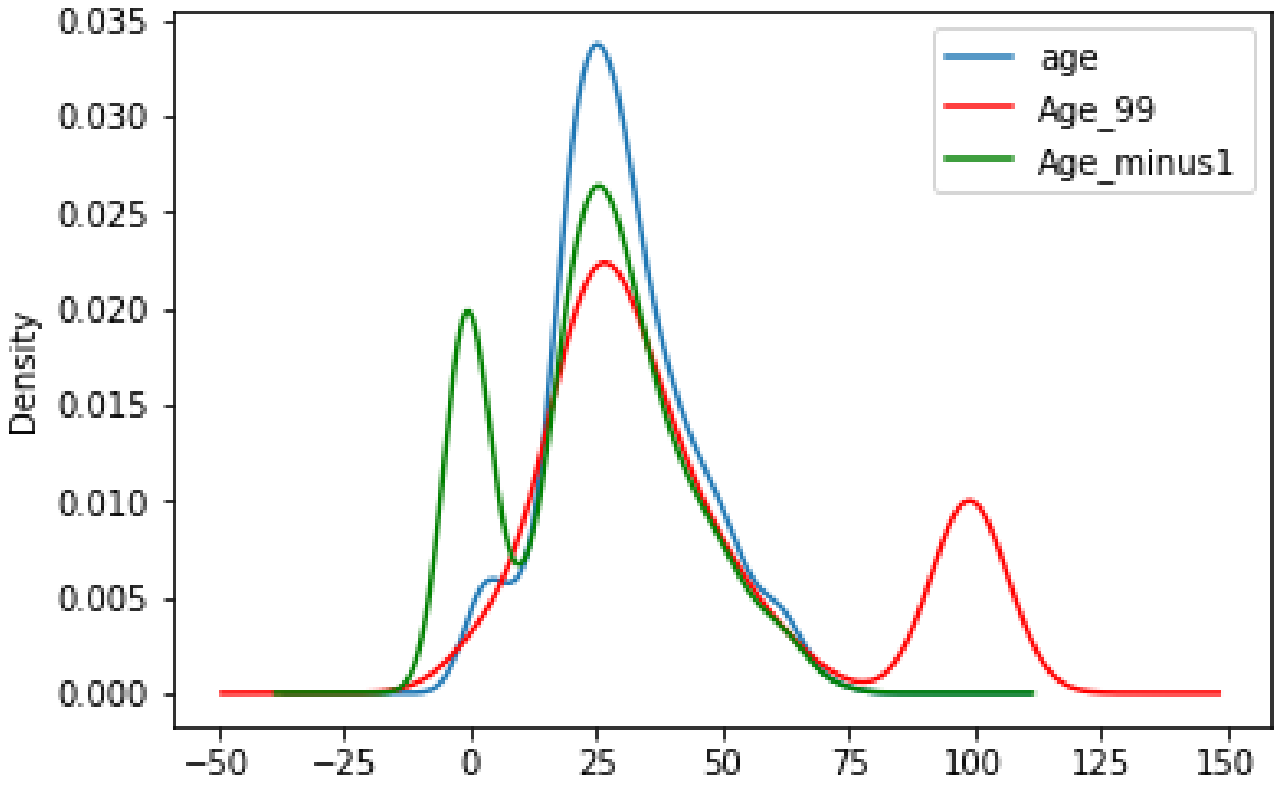
- On distribution, statistical parameters and outliers.

Mean / Median Imputation effects

	LotFrontage	OverallQual	MasVnrArea	BsmtUnfSF	TotalBsmtSF	1stFlrSF	GrLivArea	GarageYrBlt	WoodDeckSF	SalePrice
LotFrontage	532.587202	6.587119	6.805603e+02	9.496573e+02	2.908856e+03	3.379794e+03	3.919952e+03	30.611717	1.347414e+02	6.689645e+05
OverallQual	6.587119	1.843859	1.014970e+02	1.746147e+02	2.886241e+02	2.242973e+02	4.091242e+02	17.902809	3.168557e+01	8.320132e+04
MasVnrArea	680.560330	101.496976	3.298354e+04	7.540788e+03	2.478877e+04	2.086595e+04	3.520785e+04	1203.583792	3.208924e+03	6.836439e+06
BsmtUnfSF	949.657293	174.614725	7.540788e+03	1.875241e+05	7.513307e+04	4.987449e+04	5.203392e+04	1823.065167	-1.833201e+03	6.833028e+06
TotalBsmtSF	2908.855504	288.624075	2.478877e+04	7.513307e+04	1.682931e+05	1.212079e+05	8.615192e+04	3173.042442	1.227966e+04	2.003928e+07
1stFlrSF	3379.793504	224.297266	2.086595e+04	4.987449e+04	1.212079e+05	1.398656e+05	1.044401e+05	2009.195552	1.109406e+04	1.783631e+07
GrLivArea	3919.951834	409.124216	3.520785e+04	5.203392e+04	8.615192e+04	1.044401e+05	2.681277e+05	2738.982988	1.558395e+04	2.934477e+07
GarageYrBlt	30.611717	17.902809	1.203584e+03	1.823065e+03	3.173042e+03	2.009196e+03	2.738983e+03	624.305948	6.658911e+02	9.309355e+05
WoodDeckSF	134.741376	31.685571	3.208924e+03	-1.833201e+03	1.227966e+04	1.109406e+04	1.558395e+04	665.891118	1.648582e+04	3.029981e+06
SalePrice	668964.454191	83201.317781	6.836439e+06	6.833028e+06	2.003928e+07	1.783631e+07	2.934477e+07	930935.489321	3.029981e+06	6.105731e+09
LotFrontage_median	532.587202	5.384774	5.539213e+02	7.880954e+02	2.370929e+03	2.750747e+03	3.189686e+03	24.755173	1.060091e+02	5.448388e+05
MasVnrArea_median	674.423263	100.533003	3.298354e+04	7.472110e+03	2.465436e+04	2.080136e+04	3.496714e+04	1182.673336	3.212101e+03	6.790442e+06
GarageYrBlt_median	28.095264	16.875386	1.134381e+03	1.724142e+03	2.989473e+03	1.890272e+03	2.576346e+03	624.305948	6.276246e+02	8.774854e+05

- On covariance.

Arbitrary value imputation - effects

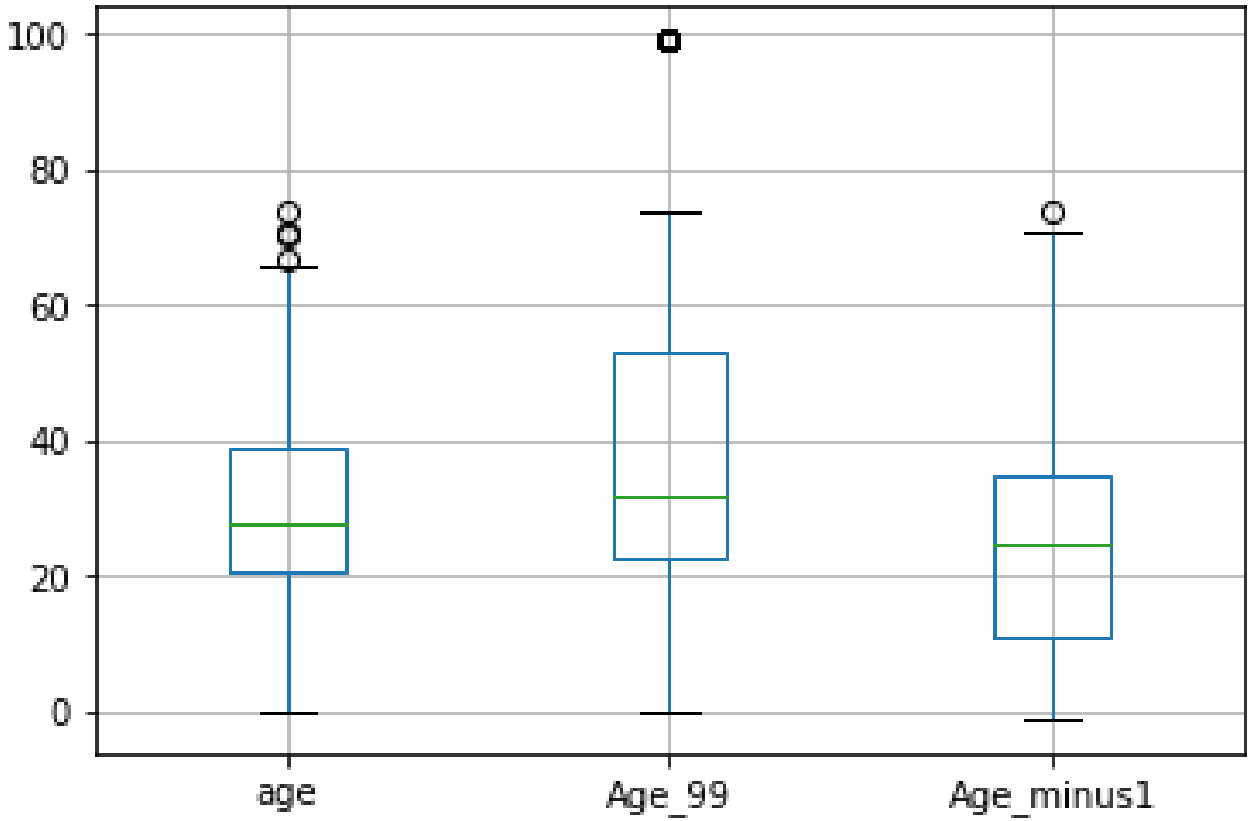


- ~20% of data is missing in Age

Original variable variance: 194
Variance after 99 imputation: 888
Variance after -1 imputation: 307

- On variance and distribution shape

Arbitrary value imputation and outliers



Masks outliers

- On distribution, statistical parameters and outliers.

Arbitrary value imputation: effects

	fare
fare	2248.326729
age	136.176223
Age_99	-38.722001
Age_minus1	177.733891

- On covariance.

Mode imputation: effects

- Distortion the relation of the most frequent label with other variables within the dataset
- May lead to an over-representation of the most frequent label if there is a big number of NA
- **More missing values ➔ greater distortions**

Effect on covariates

- Houses with NA have lower prices
- Houses with B have highest prices
- B is most frequent category
- After replacing NA with B, the average price decreases

Accompanying Jupyter Notebook

- Jupyter Notebooks in **introduction folder**

- Effect of the imputation on:
  - Variable distribution - variance
  - Covariance
  - Outliers

Train In Data

THANK YOU

www.trainindata.com