FEATURE ENGINEERING

FEATURE SELECTION/SCALING - FORMAL

Most machine learning algorithms require your data to respect some formal requirements. Common examples are:

- **No null entries (we have seen how to handle these in the data cleaning class)**
- **Only numerical values**
- **Only categorical values**
- **No negative values**
- **Comparable features**

We can often transform our features to make them amenable to certain types of algorithms

FEATURE SELECTION - NUMERICAL VS CATEGORICAL

- The techniques used to handle numerical and categorical variables in preparing a dataset for machine learning, are quite different

- Numerical Features are usually used by the algorithms to quantify "distances" between different data points

- On the other hand, we cannot define such distances for categorical variables

If a feature is the colour of an object, how "different" red from blue?
how about red from green?

FEATURE SELECTION - ONE HOT ENCODING

- **One hot encoding: replace a categorical variable with one of more new features that have the values of 0 or 1.**

- **What are the consequences of this technique in terms of the cardinality of the variable?**

	age	workclass	education	gender	hours-per-week	occupation	income
0	39	State-gov	Bachelors	Male	40	Adm-clerical	<=50K
1	50	Self-emp-not-inc	Bachelors	Male	13	Exec-managerial	<=50K
2	38	Private	HS-grad	Male	40	Handlers-cleaners	<=50K
3	53	Private	11th	Male	40	Handlers-cleaners	<=50K
4	28	Private	Bachelors	Female	40	Prof-specialty	<=50K
5	37	Private	Masters	Female	40	Exec-managerial	<=50K
6	49	Private	9th	Female	16	Other-service	<=50K
7	52	Self-emp-not-inc	HS-grad	Male	45	Exec-managerial	>50K
8	31	Private	Masters	Female	50	Prof-specialty	>50K
9	42	Private	Bachelors	Male	40	Exec-managerial	>50K
10	37	Private	Some-college	Male	80	Exec-managerial	>50K

FEATURE SELECTION - ONE HOT ENCODING

- **One hot encoding: replace a categorical variable with one of more new features that have the values of 0 or 1.**

- **What are the consequences of this technique in terms of the cardinality of the variable?**

	age	workclass	education	gender	hours-per-week	occupation	income
0	39	State-gov	Bachelors	Male	40	Adm-clerical	<=50K
1	50	Self-emp-not-inc	Bachelors	Male	13	Exec-managerial	<=50K
2	38	Private	HS-grad	Male	40	Handlers-cleaners	<=50K
3	53	Private	11th	Male	40	Handlers-cleaners	<=50K
4	28	Private	Bachelors	Female	40	Prof-specialty	<=50K
5	37	Private	Masters	Female	40	Exec-managerial	<=50K
6	49	Private	9th	Female	16	Other-service	<=50K
7	52	Self-emp-not-inc	HS-grad	Male	45	Exec-managerial	>50K
8	31	Private	Masters	Female	50	Prof-specialty	>50K
9	42	Private	Bachelors	Male	40	Exec-managerial	>50K
10	37	Private	Some-college	Male	80	Exec-managerial	>50K

FEATURE SELECTION - ONE HOT ENCODING

- **One hot encoding for the working class column**

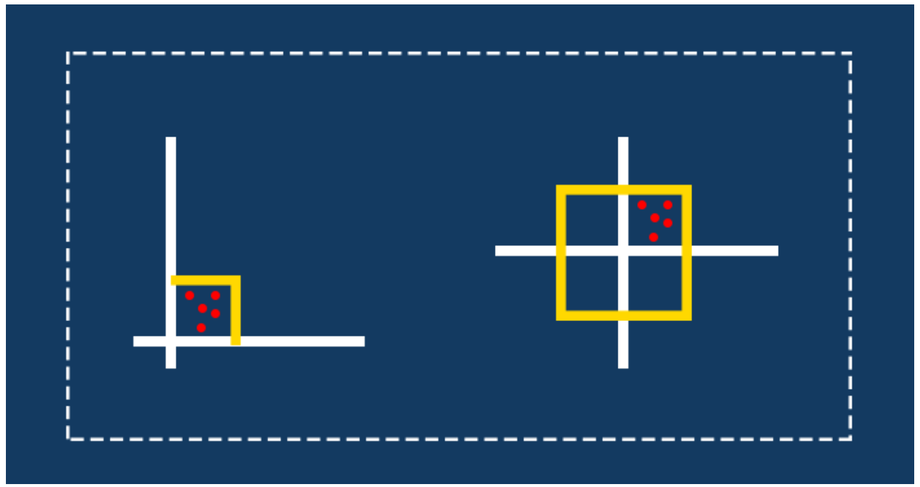
workclass	Government Employee	Private Employee	Self Employed	Self Employed Incorporated
Government Employee	1	0	0	0
Private Employee	0	1	0	0
Self Employed	0	0	1	0
Self Employed Incorporated	0	0	0	1

- **Watch out that numbers can encode categorical variables**

FEATURE SELECTION - BINNING

- Create groups of lower granularity from elements of a variable
- Also known as descritization when applied to numerical variables
- In Python can be done with cut and qcut

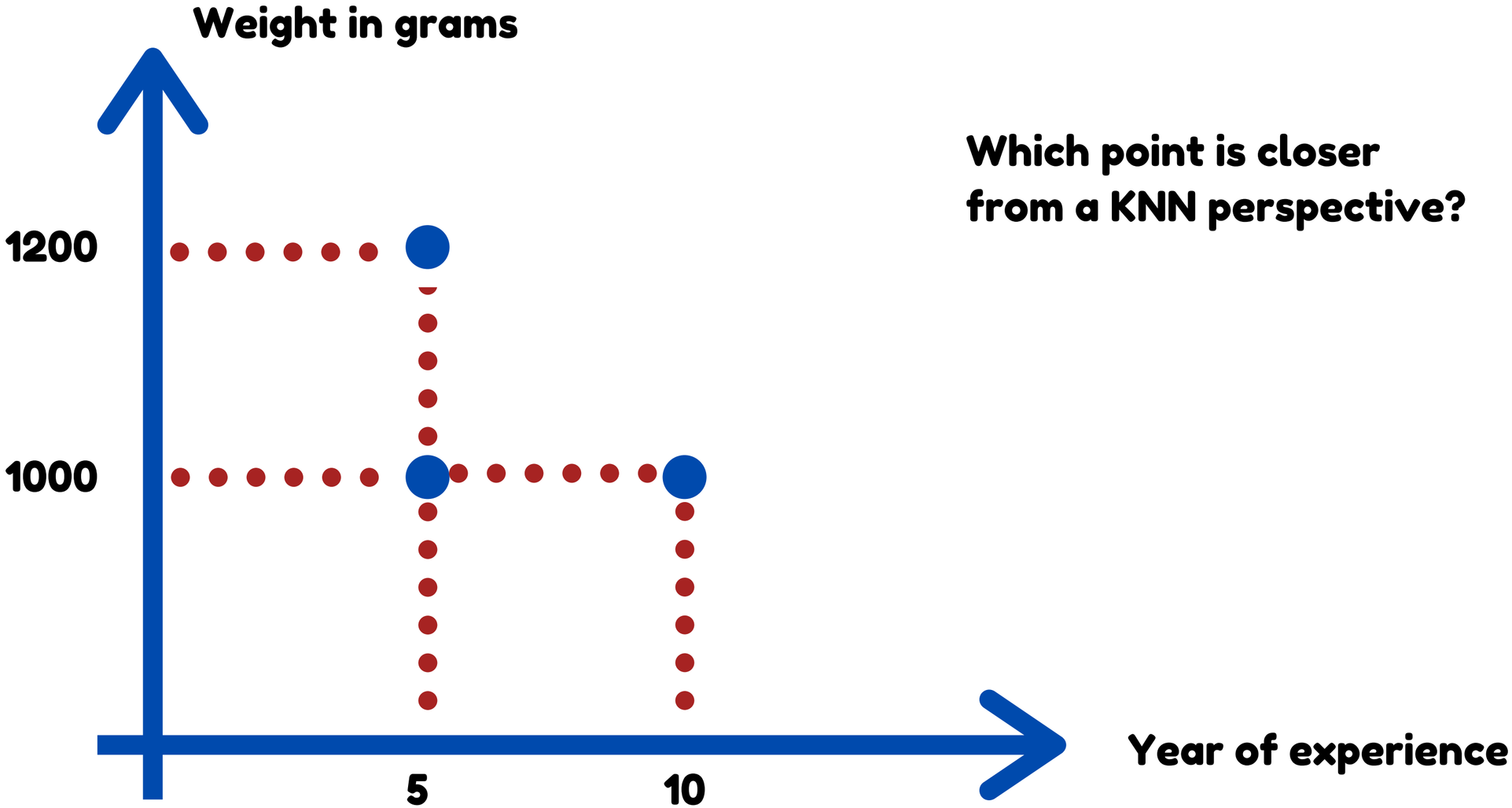
FEATURE SCALING



- The process of processing numerical features and converting them into a uniform scale is known as feature scaling

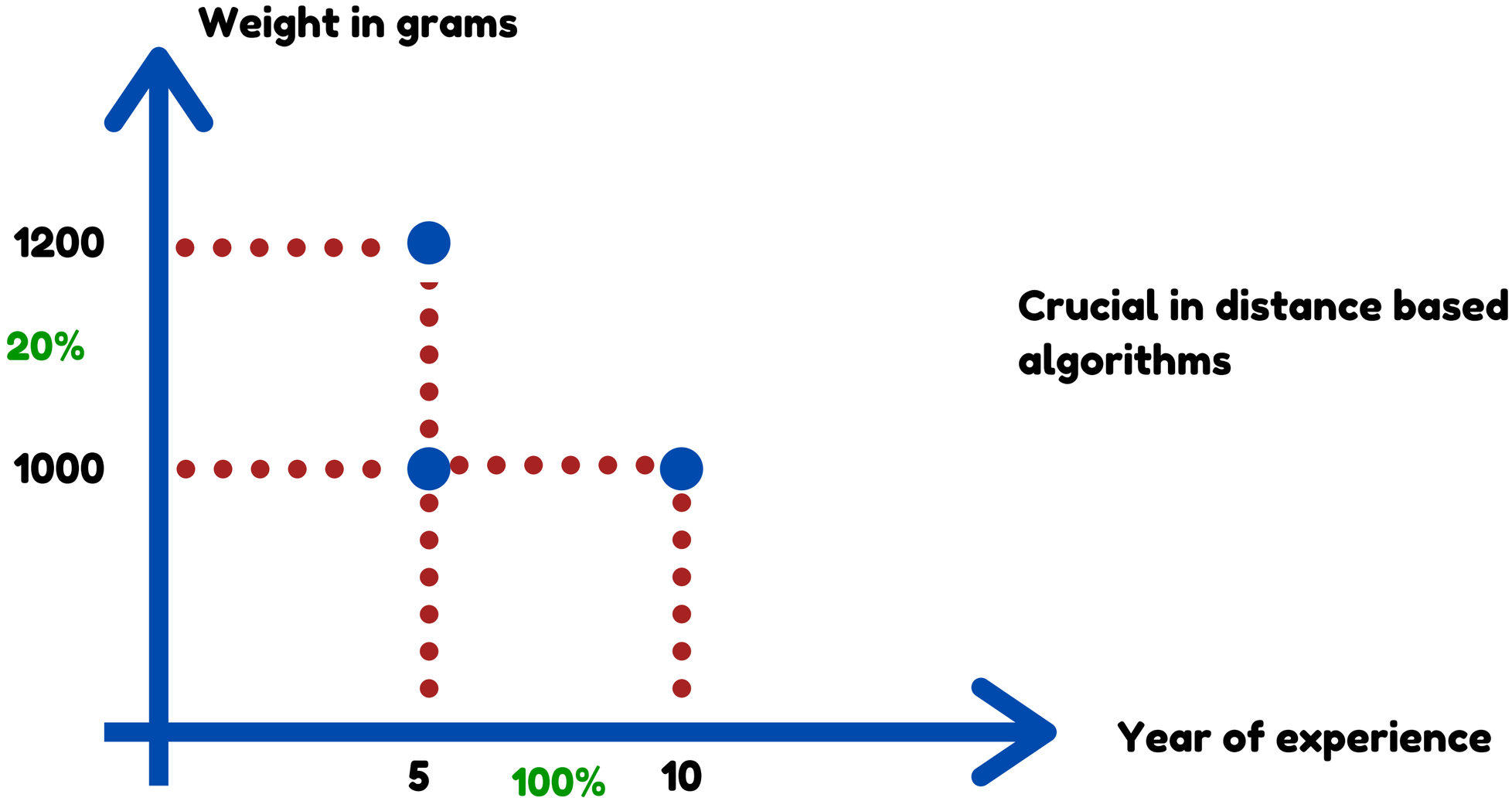
- whilst some algorithms are virtually invariant to this technique, for other it is determinant

FEATURE SCALING - WHY IS IT IMPORTANT?



Which point is closer from a KNN perspective?

FEATURE SCALING - WHY IS IT IMPORTANT?



Crucial in distance based algorithms

Z-SCORE : STANDARDIZATION

THE Z-SCORE IS A WAY TO STANDARDIZE/NORMALIZE ALL YOUR DATA IN A WAY THAT TELLS YOU HOW MANY STANDARD DEVIATIONS EACH POINT IS FROM THE MEAN

IF THE Z-SCORE IS LESS THAN 1: THAT DATA POINT IS WITHIN 1 STANDARD DEVIATION OF MEAN

ALLOWS YOU TO BRING DIFFERENT DISTRIBUTIONS TO THE SAME SCALE

$$Z = \frac{x - \mu}{\sigma}$$

EXAMPLE OF USAGE: TO STANDARDIZE GRADES ACROSS DIFFERENT SCHOOLS

VERY IMPORTANT IN MACHINE LEARNING

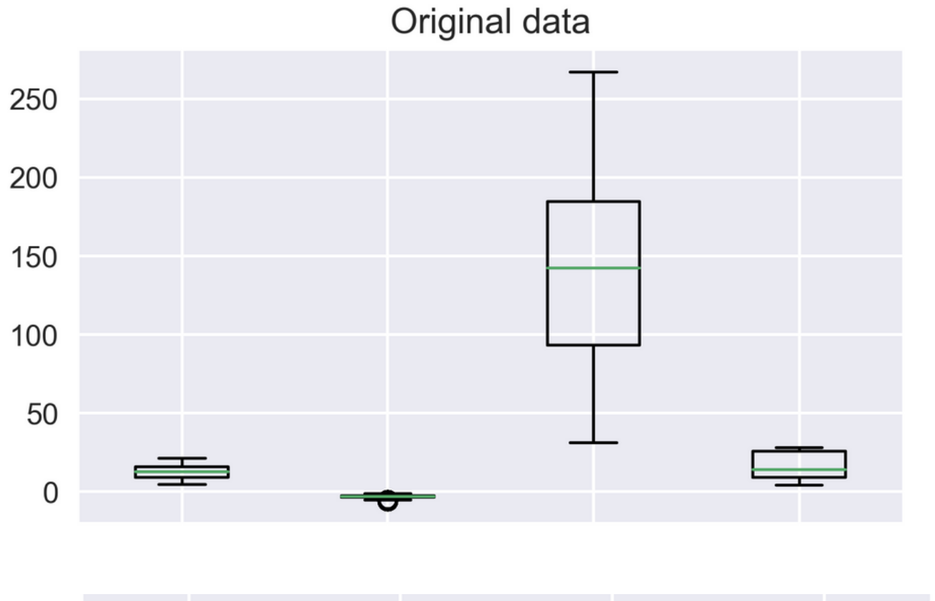
NORMALIZATION

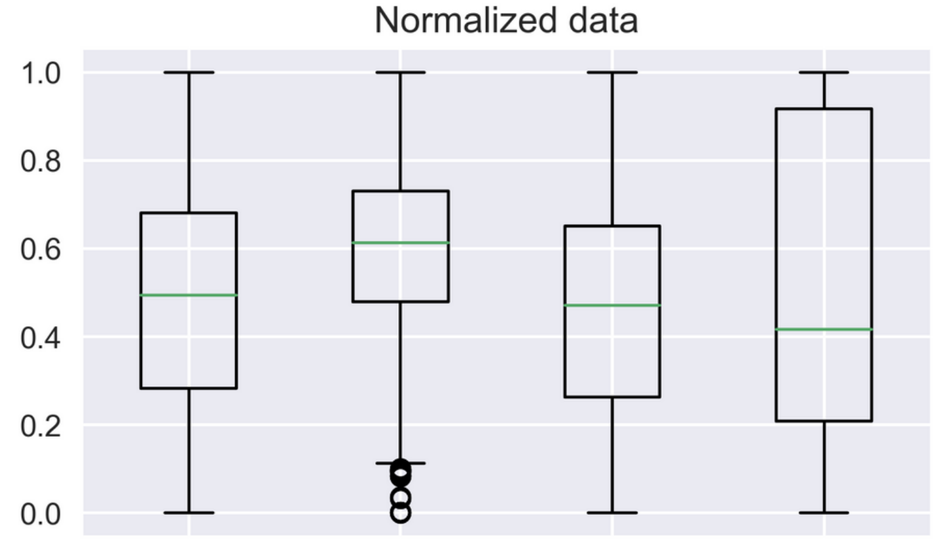
RE-SCALE THE DISTRIBUTION TO BE BETWEEN ZERO AND ONE!

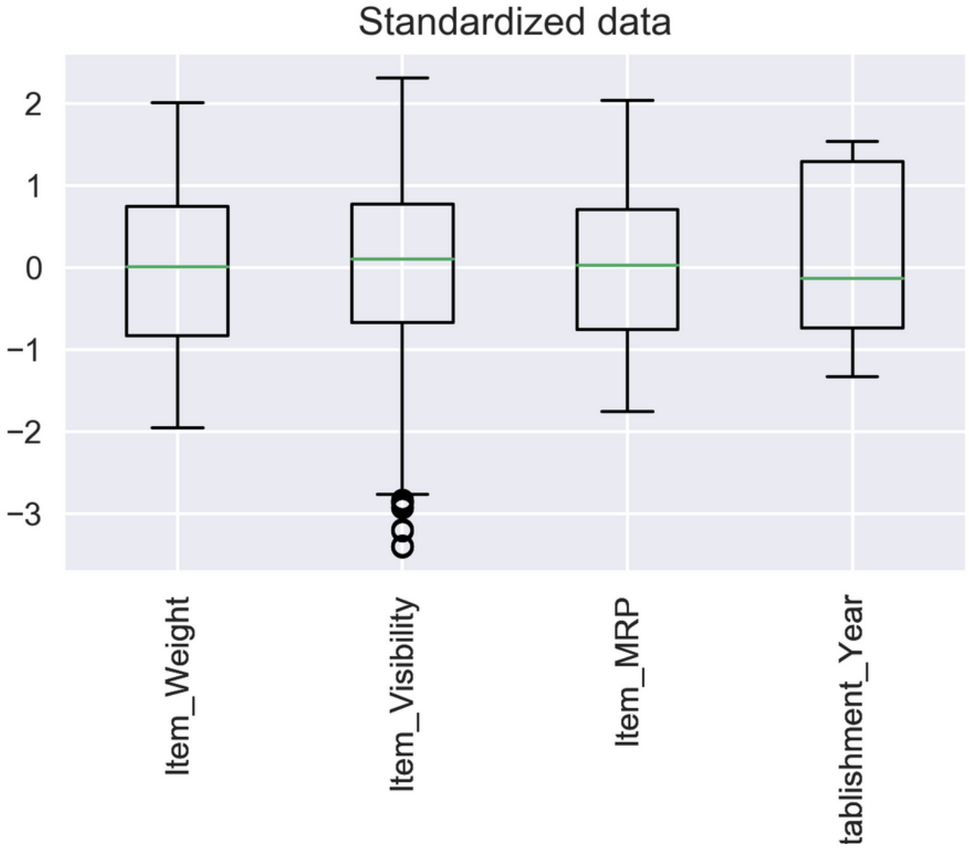
EACH VALUE OF THE DATASET X WILL BE CONVERTED TO A NEW "NORMALIZED" VALUE Z

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

SHAPE OF THE DISTRIBUTION OF DATA IS UNCHANGED


Original data
250
200
150
100
50
0

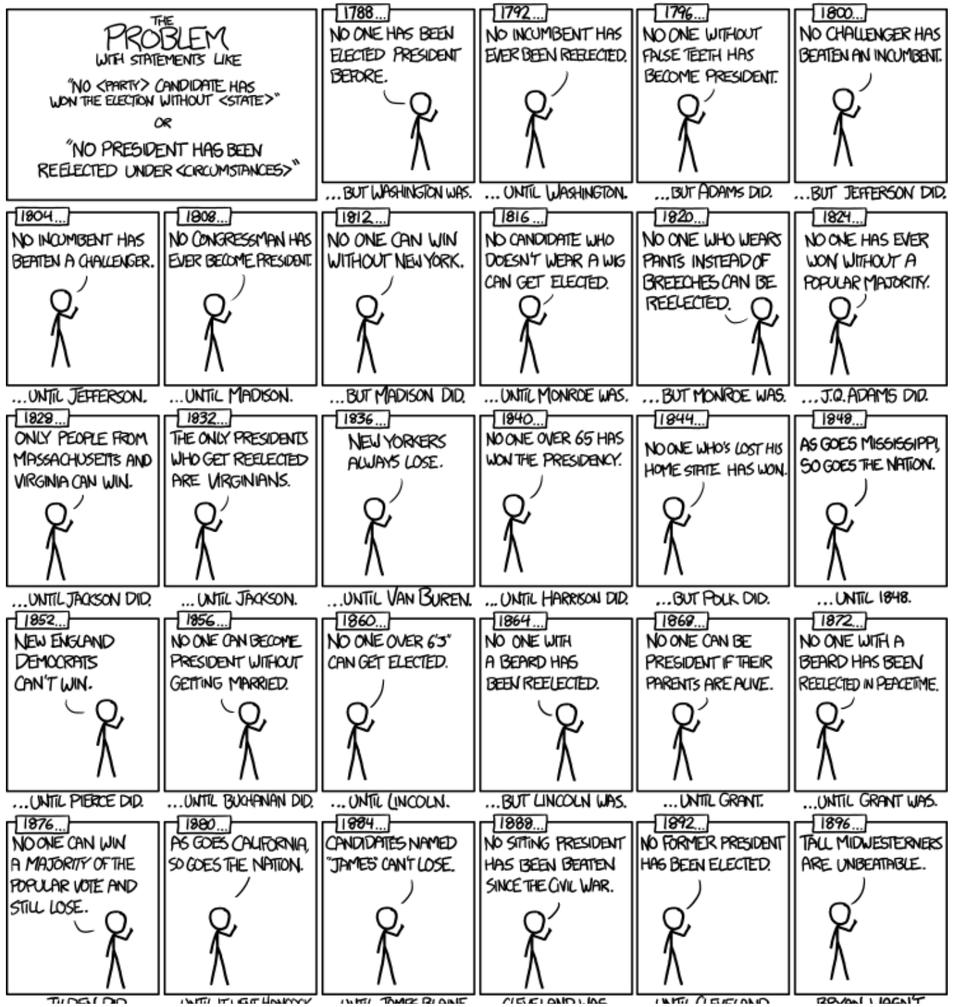

Normalized data
1.0
0.8
0.6
0.4
0.2
0.0

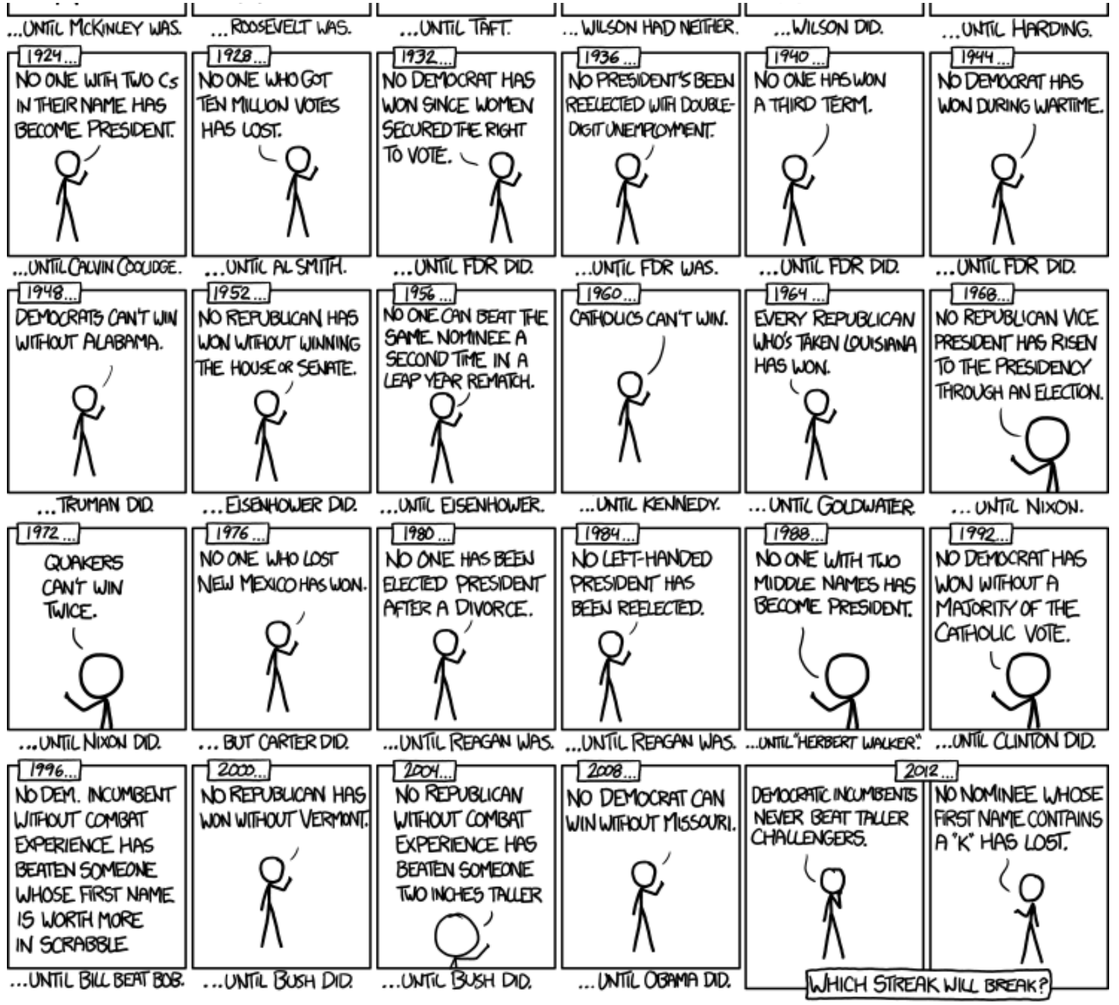

Standardized data
2
1
0
−1
−2
−3
Item_Weight
Item_Visibility
Item_MRP
tablishment_Year

FEATURE SELECTION - SEMANTIC

TIWYEYP

FEATURE SELECTION - AVOID OVERFITTING!



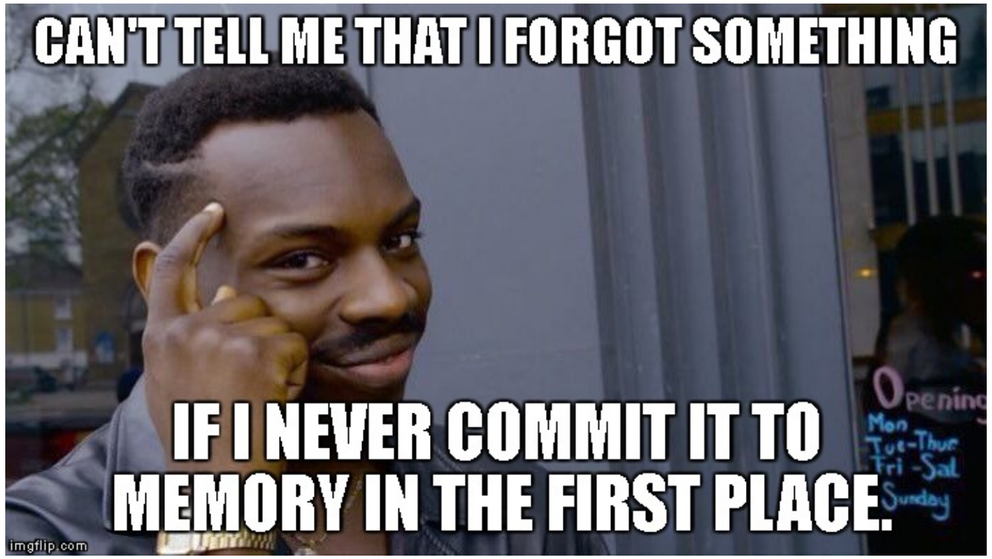


FEATURE SELECTION - USE INDEPENDENT VARS

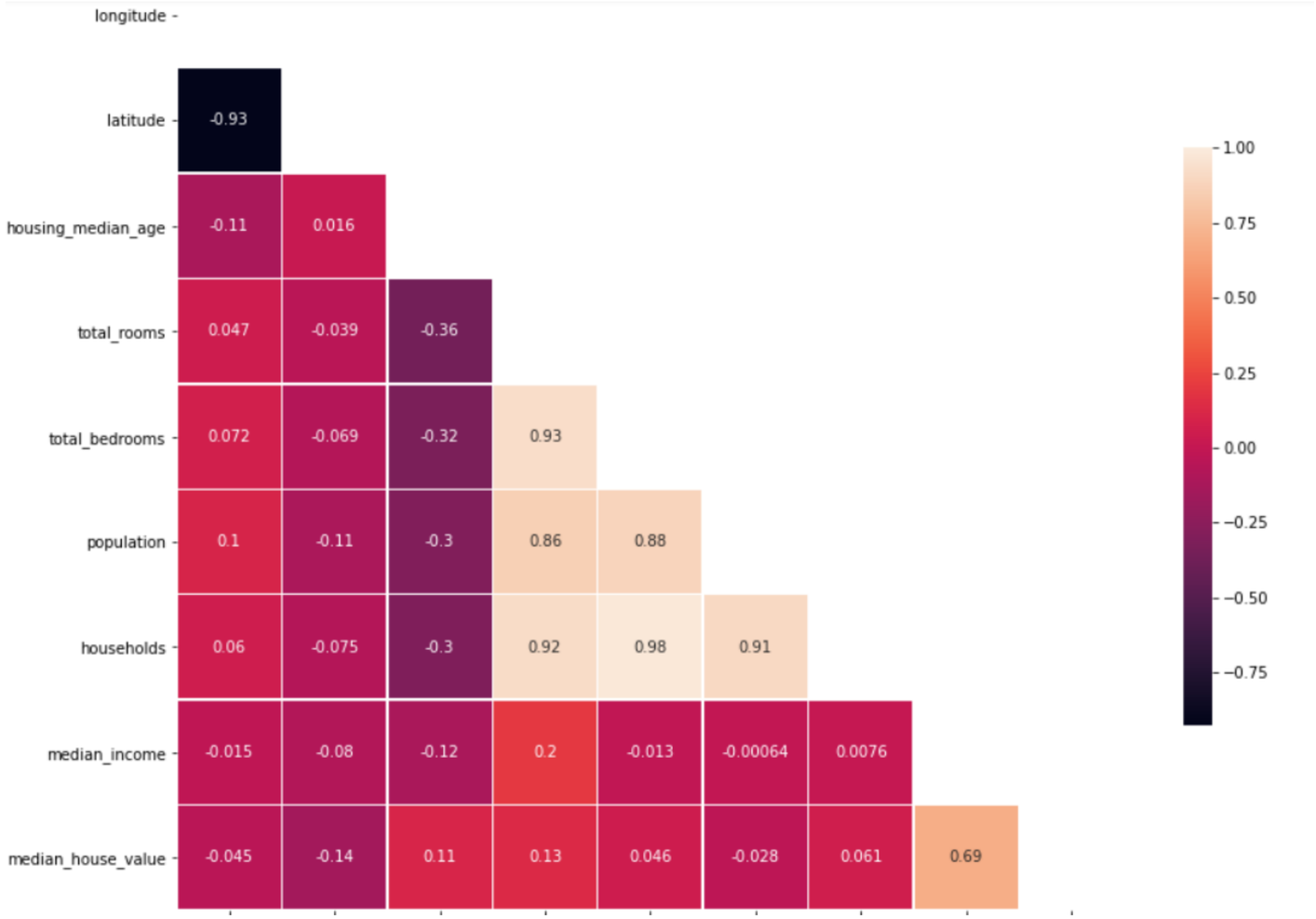
- **Remember Module 2 techniques:**

- Chi-Squared goodness of fit tests of contingency

- ANOVA tests/ 2-sample t-tests to see if you can perform further aggregations



CORRELATION THRESHOLDS



DEEP RELATION WITH THE CONCEPT OF MUTUAL INFORMATION

THE BEST NUMERICAL FEATURES ARE THOSE WITH A HIGH EXPLANATORY POWER OF THE TARGET VARIABLE, BUT LOW MUTUAL INFORMATION AMONGST THEMSELVES

CORRELATION THRESHOLDS ALLOW US CREATE THE ABOVE CRITERIA AND THUS FILTER OUT REDUNDANT VARIABLES

FEATURE ENGINEERING - INCREASE SIGNAL

OFTEN YOU WILL KNOW THAT A PARTICULAR VARIABLE IS RELEVANT IN SOME SEMANTIC WAY:

- TRAFFIC ON COMMUTING ROADS WILL DECREASE ON WEEKENDS
- THE OPTIMAL CONCENTRATION OF A CERTAIN CHEMICAL IS 10 G/L
- SALES INCREASE NEAR THE HOLIDAYS
- HUMIDITY AND TEMPERATURE INCREASE CHANCE OF BREAKDOWN
- ...

YOU CAN (AND SHOULD) TRANSFORM YOUR VARIABLES IN WAYS THAT EVIDENCE THESE "SIGNALS" RATHER THAN EXCPECTING YOUR ALGORITHM TO "PICK THEM UP". FOR THIS, YOU NEED TO UNDERSTAND THE UNDERLYING MECHANICS OF THE PHENOMENA YOU ARE ANALYSING.

TISWEYP