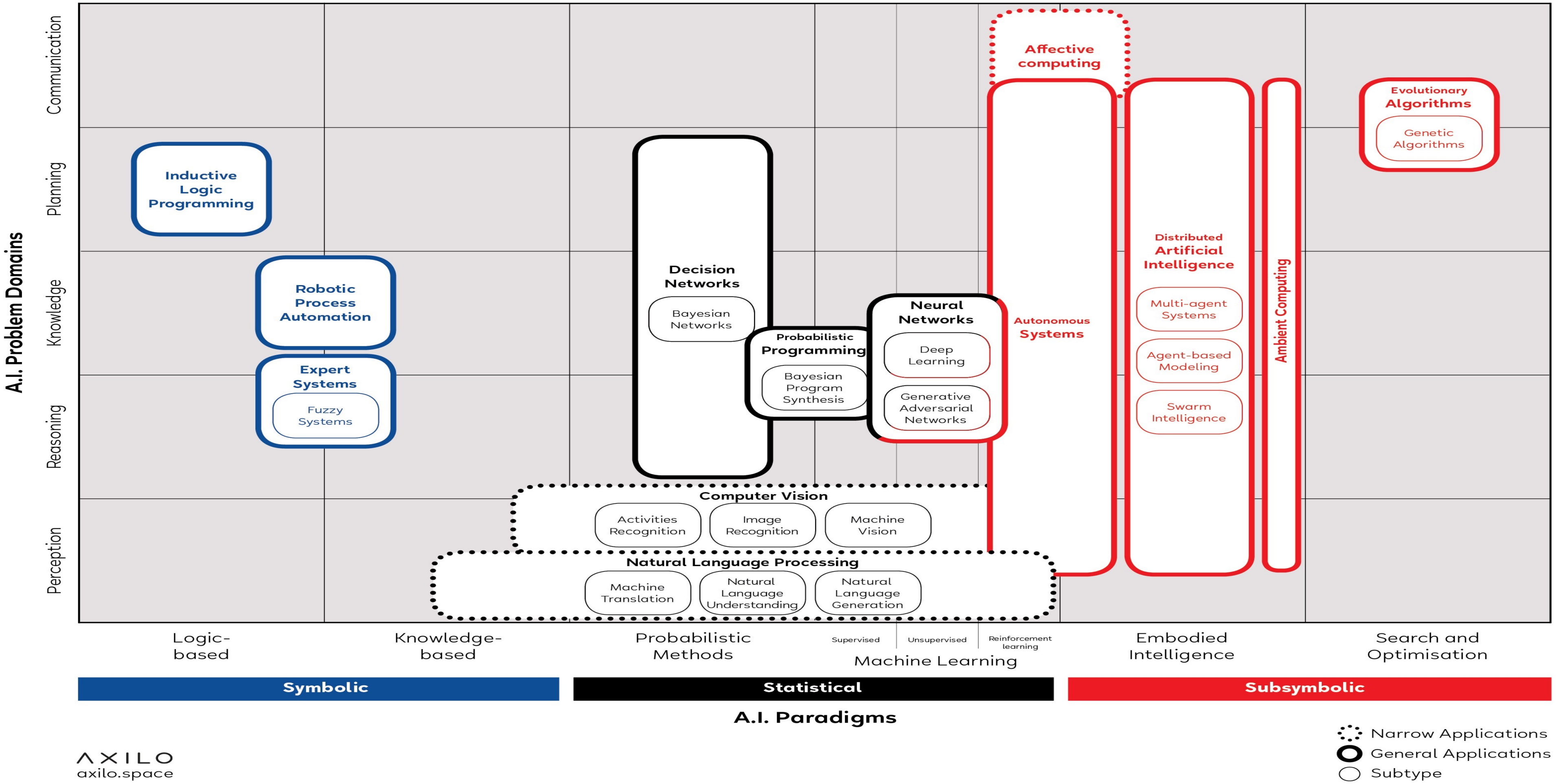
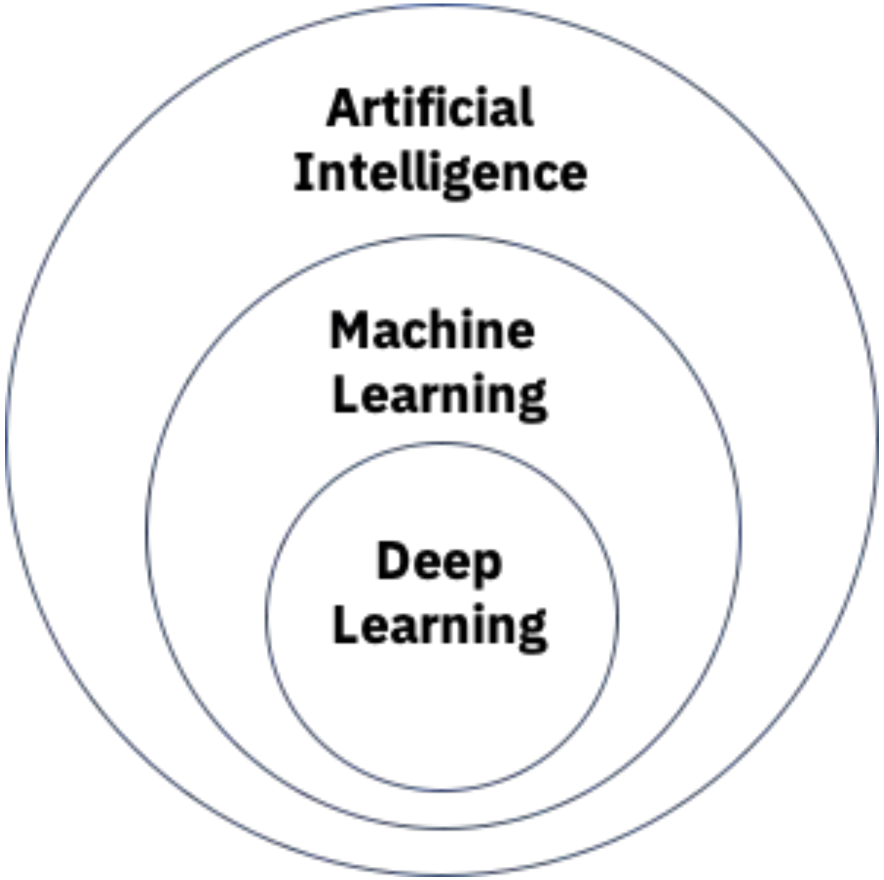
AI and Data

Vered Aharonson

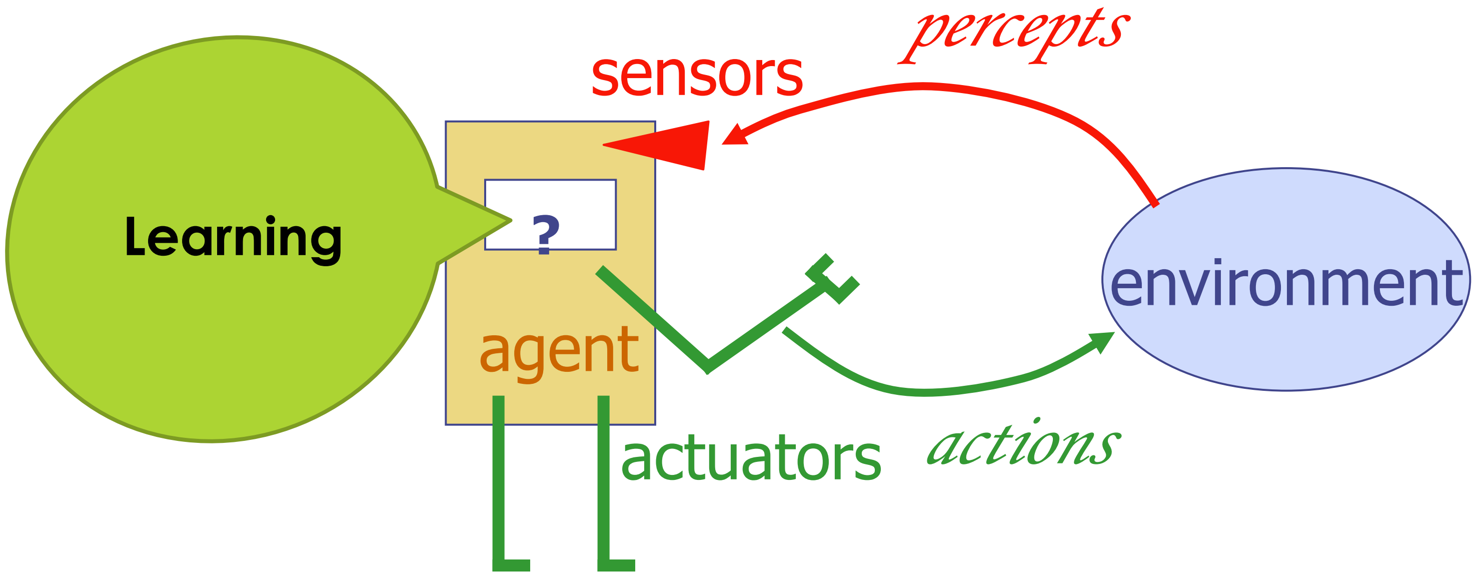
AI Paradigms



The Artificial Intelligence (AI) World



Intelligent Agents and Learning



Based on: Russel and Norvig, "Artificial Intelligence: a Modern Approach"

Data: The Root of Learning

- Machines (and Humans) Learn from Data
- AI and Human sensors capture the data from the Physical and Cyber spaces.
- Still, this Raw data should be prepared before it is used for learning.
- This preparation can be one of the most difficult steps in any machine learning project:
- Each dataset is different and highly specific to the project. Nevertheless, there are enough commonalities across predictive modeling projects that we can define a loose sequence of steps and subtasks that you are likely to perform.
- This process provides a context in which we can consider the data preparation required for the project, informed both by the definition of the project performed before data preparation and the evaluation of machine learning algorithms performed after.

Learning Objectives

- Understand data preparation as a step in a broader predictive modeling machine learning project.
  - which are common steps performed on each project.
  - What are the unknown underlying structure of the problem to learning algorithms.
  - What data preparation methods to apply, or explore.

Goal-based search for methods

- Each ML project is unique, but all need to deal with the same questions:
  - what the best results are or might be ?
  - what algorithms to use to achieve them ?
- Methods generally
  - Establish a baseline in performance as a point of reference
  - Compare several models to discover what algorithm works best for your specific dataset.

Different names, same steps..

- Data preparation techniques are dealt with in the following domains
  - “*applied machine learning process*“
  - “*data science process*“,
  - “*knowledge discovery in databases*” (KDD) - older name..
- Descriptions and names vary, but steps are similar

The ML steps

- **Common steps**

1: Define the Problem.

2: Prepare the Data.

3: Build, Optimize and Evaluate Models.

4: Finalize a Model.

We'll focus now on 1 and 2…

1. Define the problem

- What do we need to achieve and what data do we need to achieve it?
- Experts, Web and Crowd sourcing
- Collect/download data
- Investigate the data

2. Prepare the data

- Why do we need preparation?

- Many names here too.. : “*data cleaning*“, “*data pre-processing*” “*data wrangling*“, “*feature engineering*“ …
- Prepare the data before using it, such that it’ll be fit to create and evaluate a machine learning model

Data preparation steps

- **Data Cleaning**: Identifying and correcting mistakes or errors in the data.
- **Data Transforms**: Changing the scale or distribution of variables.
- **Feature Selection**: Identifying those input variables that are most relevant to the task.
- **Feature Engineering**: Deriving new variables from available data.
- **Dimensionality Reduction**: Creating compact representation/projections of the data.

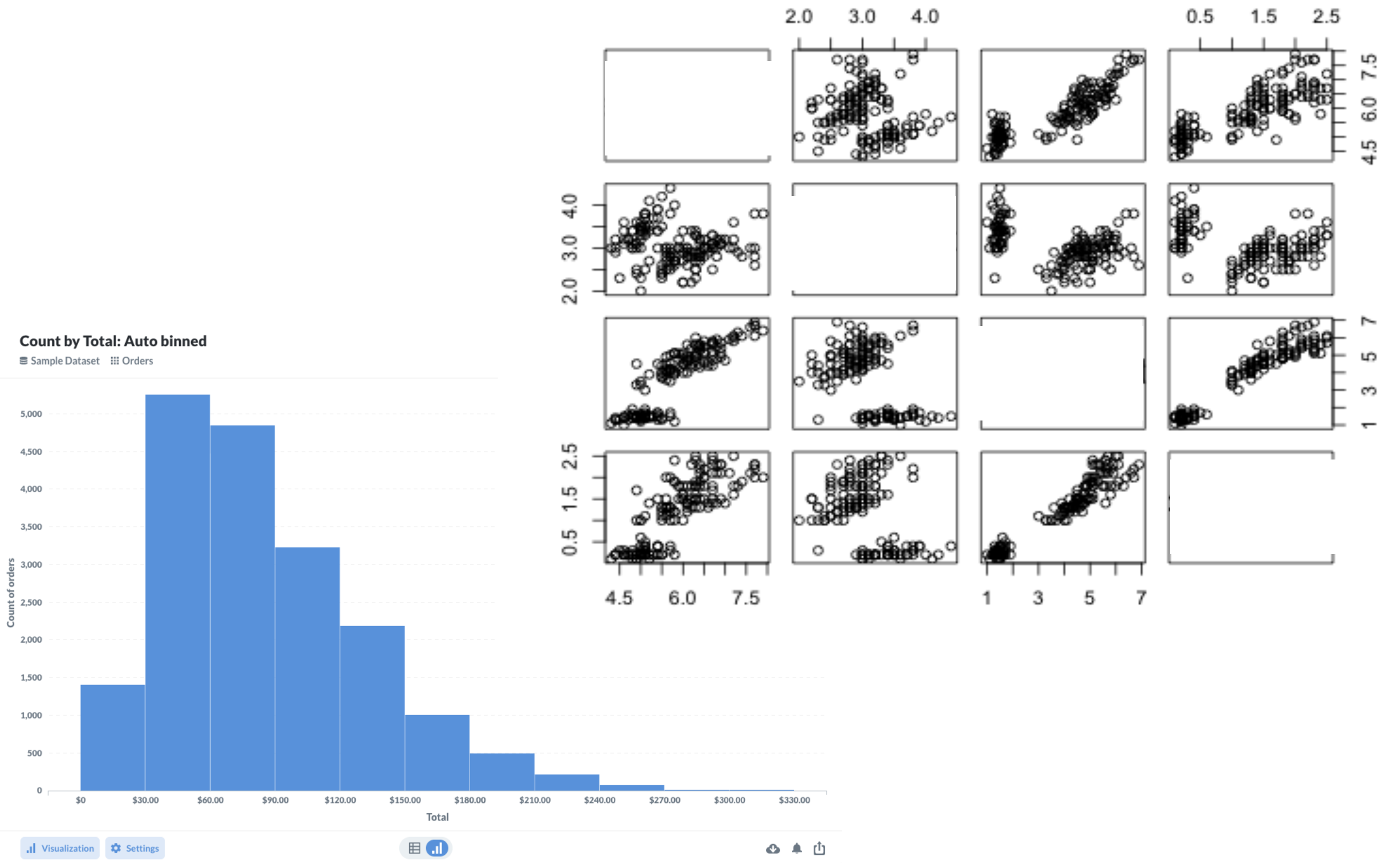
- Guiding rule: how to best expose the characteristics and patterns in the data

Tasks involved

- Review the data that has been collected.
- Summarize the collected data using statistical methods.
- Visualize the collected data using plots and charts.
- Discuss the project with subject matter experts.
- Decide which data to use
- Select variables that will be used as inputs and outputs for the model.

Tools

- Statistical methods
- Descriptive statistics
- Hypothesis tests
- Pairwise plots
- Histograms
- ….



Count by Total: Auto binned

Sample Dataset Orders

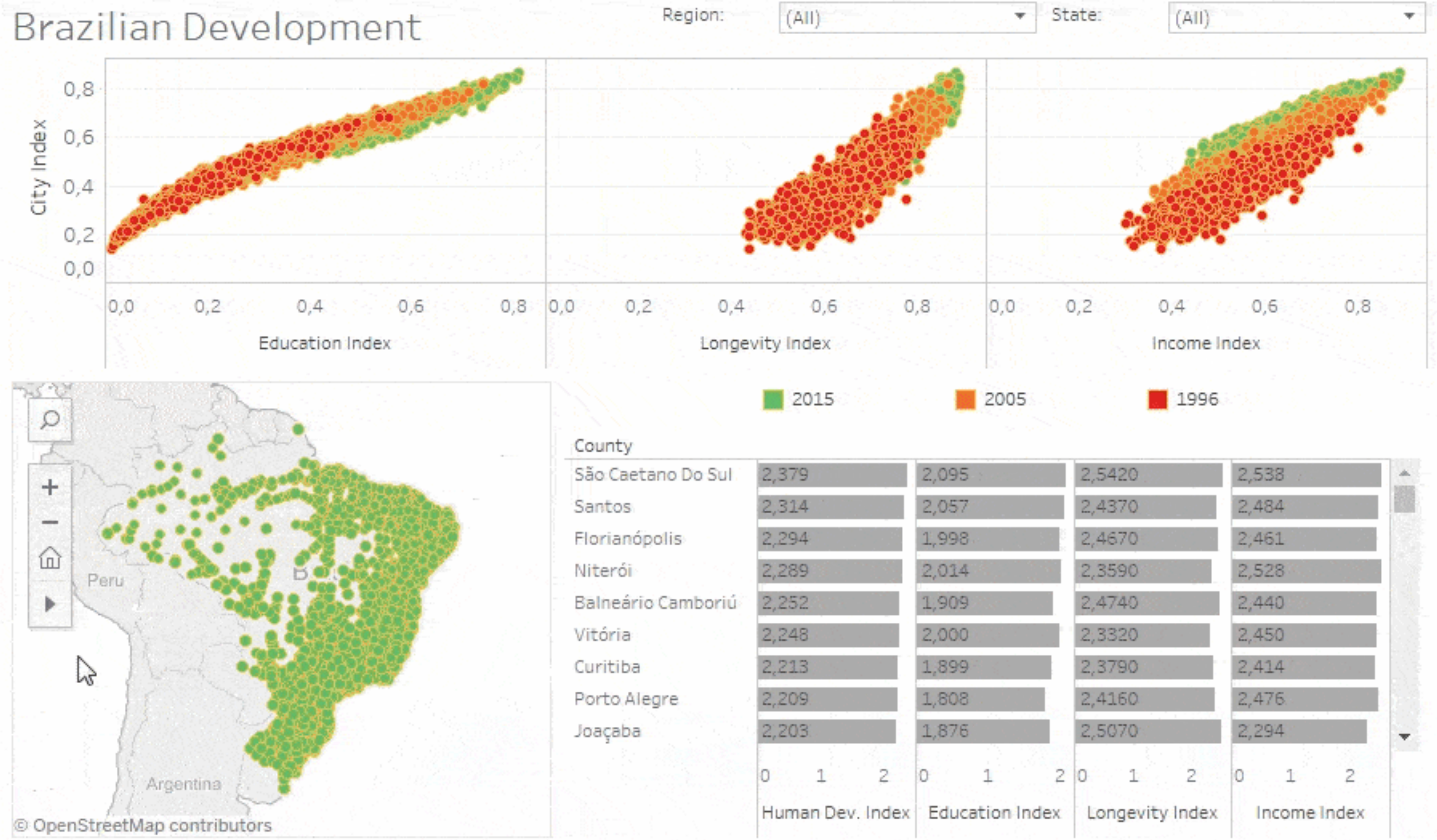
Definitions

- **Data Types**: Machine learning algorithms require data to be numbers.
- **Data Requirements:** Some machine learning algorithms impose requirements on the data.
- **Data Errors**: Statistical noise and errors in the data may need to be corrected.
- **Data Complexity**: Complex nonlinear relationships may be teased out of the data.

Data Integrity checks and corrections

- Missing data:
  - Empty cells , NaNs..
  - Fill them up: Data imputation
  - A popular approach : calculate a statistical value for each column (i.e.the mean) and replace all missing values for that column with the statistic.
- Unnecessary Data : Duplicates/ irrelevant
- Irregular Data - Outliers or anomalies
- Inconsistent Data -contradicting values Capitalization, formats..

Data Visualization

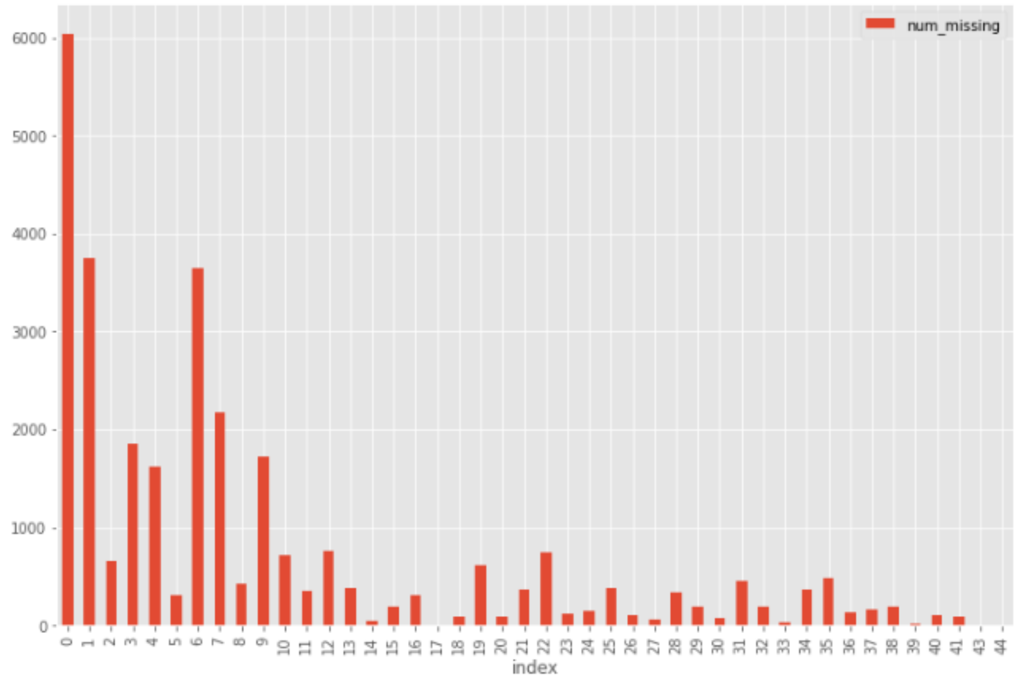


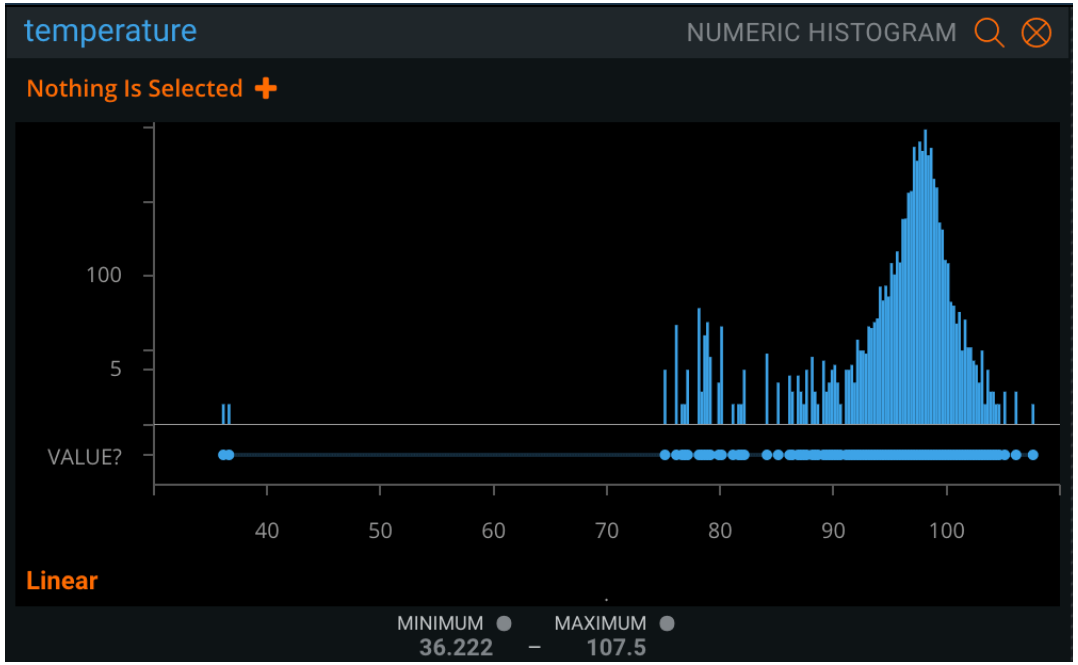
County	Human Dev. Index	Education Index	Longevity Index	Income Index
São Caetano Do Sul	2,379	2,095	2,5420	2,538
Santos	2,314	2,057	2,4370	2,484
Florianópolis	2,294	1,998	2,4670	2,461
Niterói	2,289	2,014	2,3590	2,528
Balneário Camboriú	2,252	1,909	2,4740	2,440
Vitória	2,248	2,000	2,3320	2,450
Curitiba	2,213	1,899	2,3790	2,414
Porto Alegre	2,209	1,808	2,4160	2,476
Joaçaba	2,203	1,876	2,5070	2,294

Sensitivity: Internal

Visualization (cont.)

```
<matplotlib.axes._subplots.AxesSubplot at 0x16ec982e8>
```



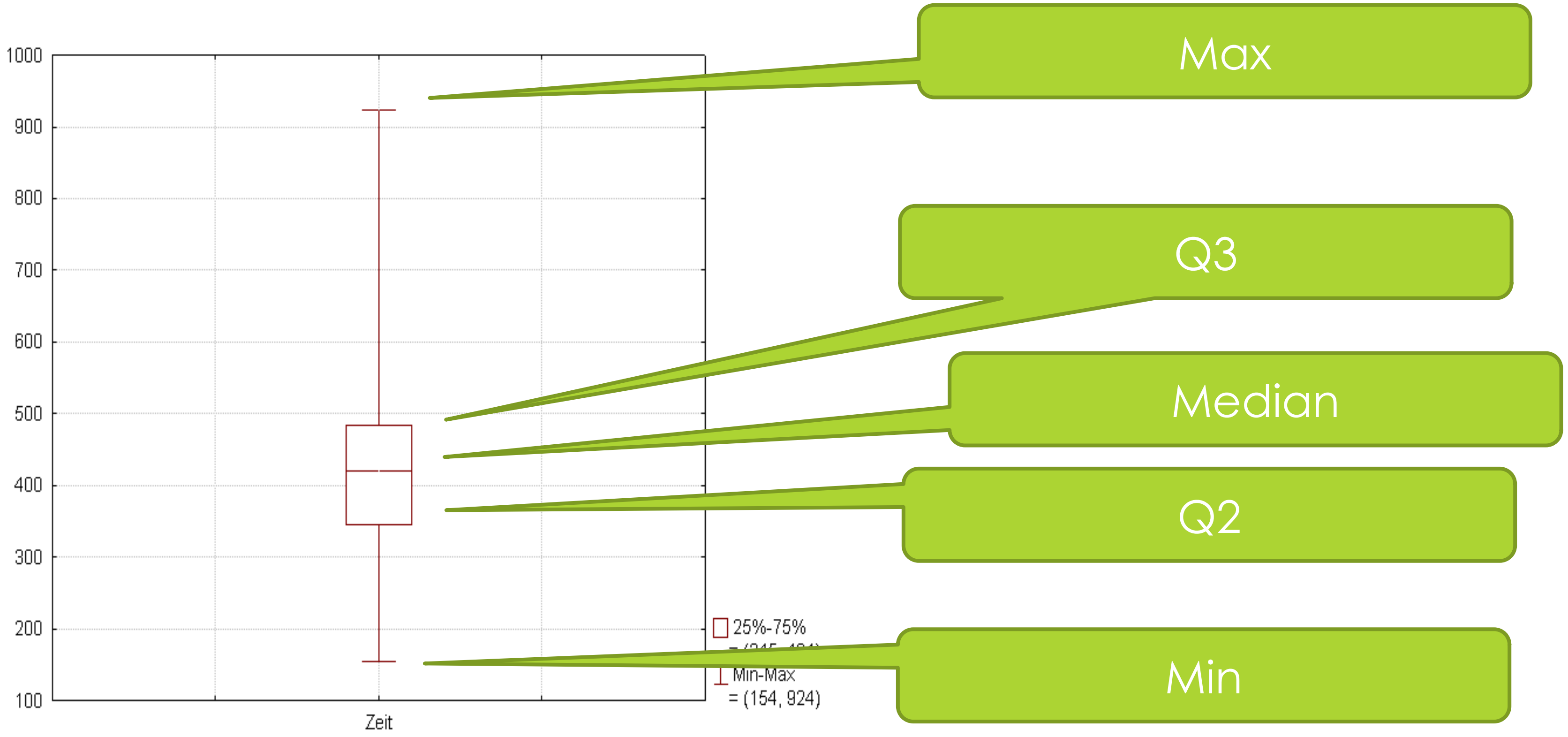


Nothing Is Selected ✚

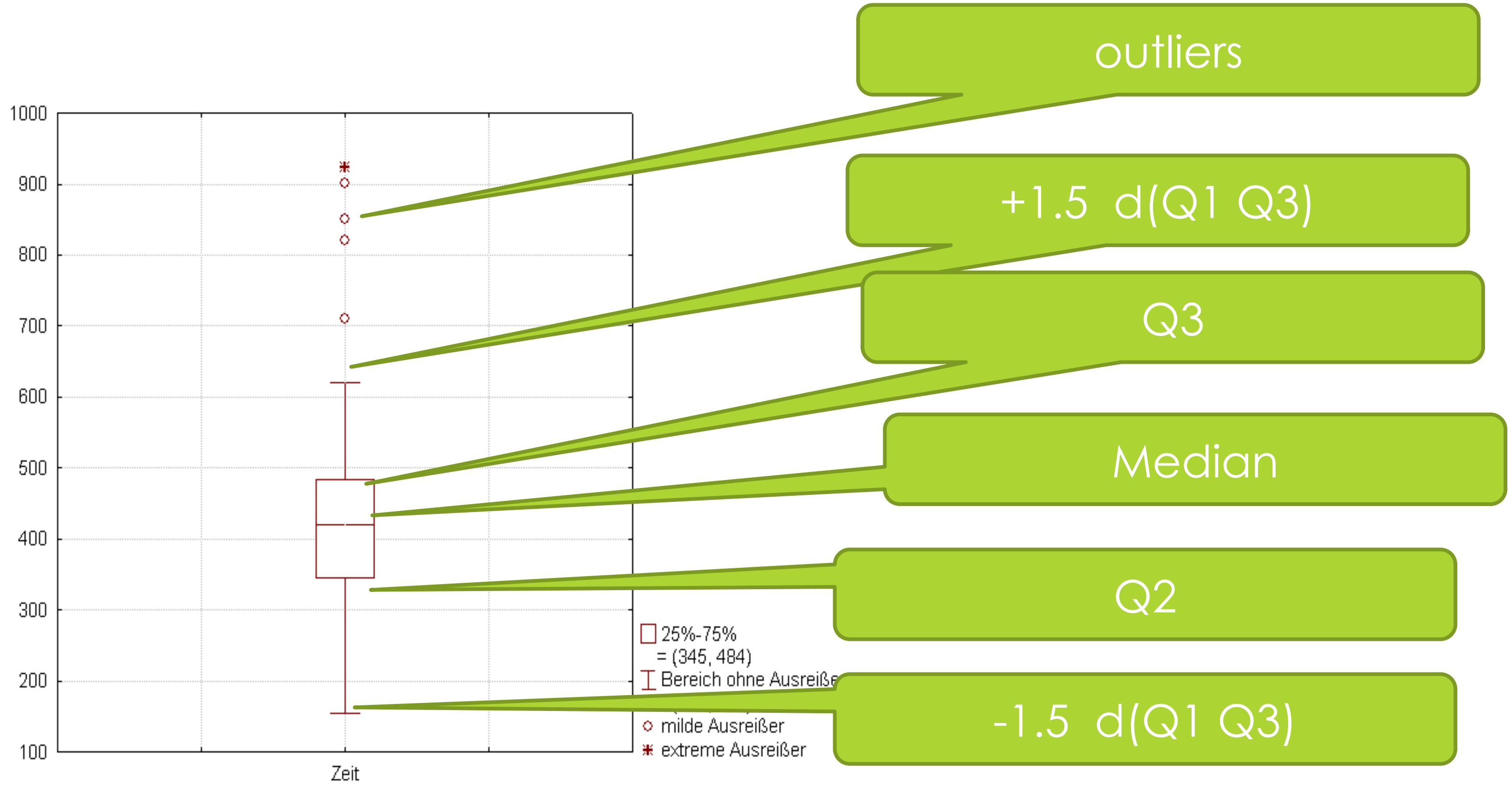
Linear

MINIMUM ● 36.222 – MAXIMUM ● 107.5

Box Plot (V1)


Max
Q3
Median
Q2
Min
1000
900
800
700
600
500
400
300
200
100
Zeit
25%-75%
= (345, 484)
Min-Max
= (154, 924)

Box Plot (V2)



outliers

+1.5 d(Q1 Q3)

Q3

Median

Q2

-1.5 d(Q1 Q3)

Sensitivity: Internal

Data Scaling and Normalization

- Many machine learning algorithms perform better when numerical input variables are scaled to a standard range
- A popular techniques: Normalization.
- Scales each input variable separately to the range 0-1
  - the range for floating-point values with the maximal precision.
  - Requires to know or to accurately estimate the minimum and maximum observable values for each variable.
  - Estimation can be done from the available data.

Transformations

- Machine learning models require numeric input and output variables.
- If the data contains categorical variables, they need to be transformed (encoded).
    - Ordinal Data: The categories have an inherent order
    - Nominal Data: The categories do not have an inherent order
- A popular technique for transforming categorical variables into numbers is the one-hot encoding.

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Transformation (cont.)

- Some machine learning algorithms prefer or require categorical or ordinal input variables.
- A simple transform: Discretisation/binning
- Values for the variable are grouped together into discrete bins and each bin is assigned a unique integer such that the ordinal relationship between the bins is preserved.
- Popular techniques for grouping the values into k discrete bins :
- **Uniform**: Each bin has the same width in the span of possible values for the variable.
- **Quantile**: Each bin has the same number of values, split based on percentiles.
- **Clustered**: Clusters are identified and examples are assigned to each group.

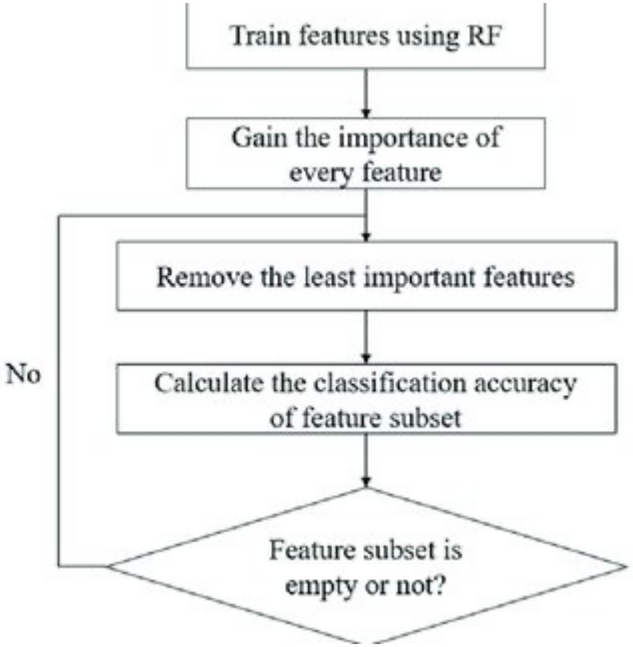
Transformation (cont.)

- Creating new features from existing ones:
- Single feature transformations
  - DOB -> age
  - Zip code –> Socio-economic status
- Multiple features transformations
  - Length * width -> plot size
  - Zip code, Education, age -> Socio-economic status

Feature Selection

- Select the most important features in a dataset.
- The process of reducing the number of input variables when developing a model.
- Both reduces the computational cost of modeling and, (sometimes) improves the model performance
- A popular feature selection algorithm is the Recursive Feature Elimination (RFE) .
- Easy to configure and use and effective at selecting the features (columns) in a training dataset that are more or most relevant in predicting the target variable.
- The scikit-learn Python machine learning library provides an implementation of RFE for ML. RFE is a transform. To use it, first, the class is configured with the chosen algorithm specified via the “estimator” argument and the number of features to select via the “*n_features_to_select*” argument.

RFE



Train features using RF

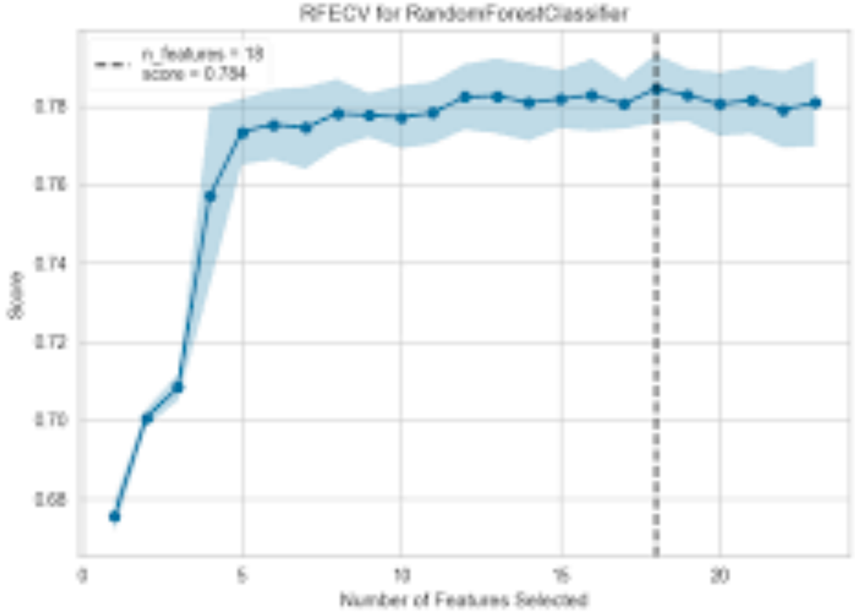
Gain the importance of every feature

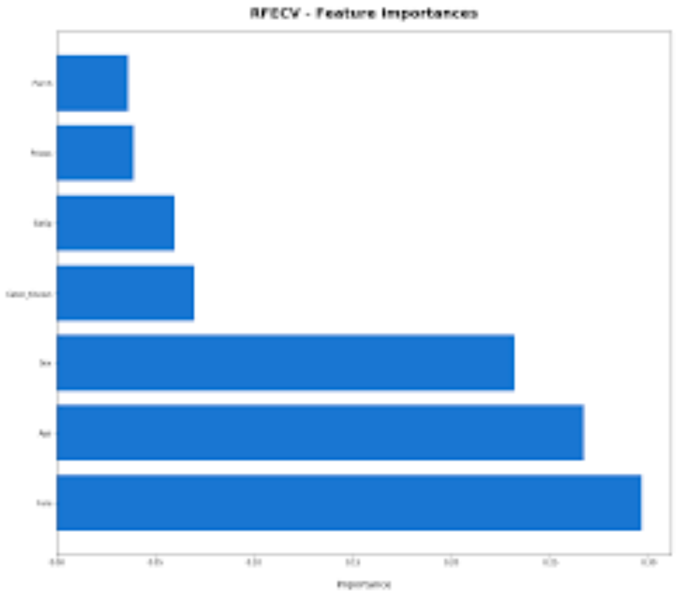
Remove the least important features

Calculate the classification accuracy of feature subset

Feature subset is empty or not?

No



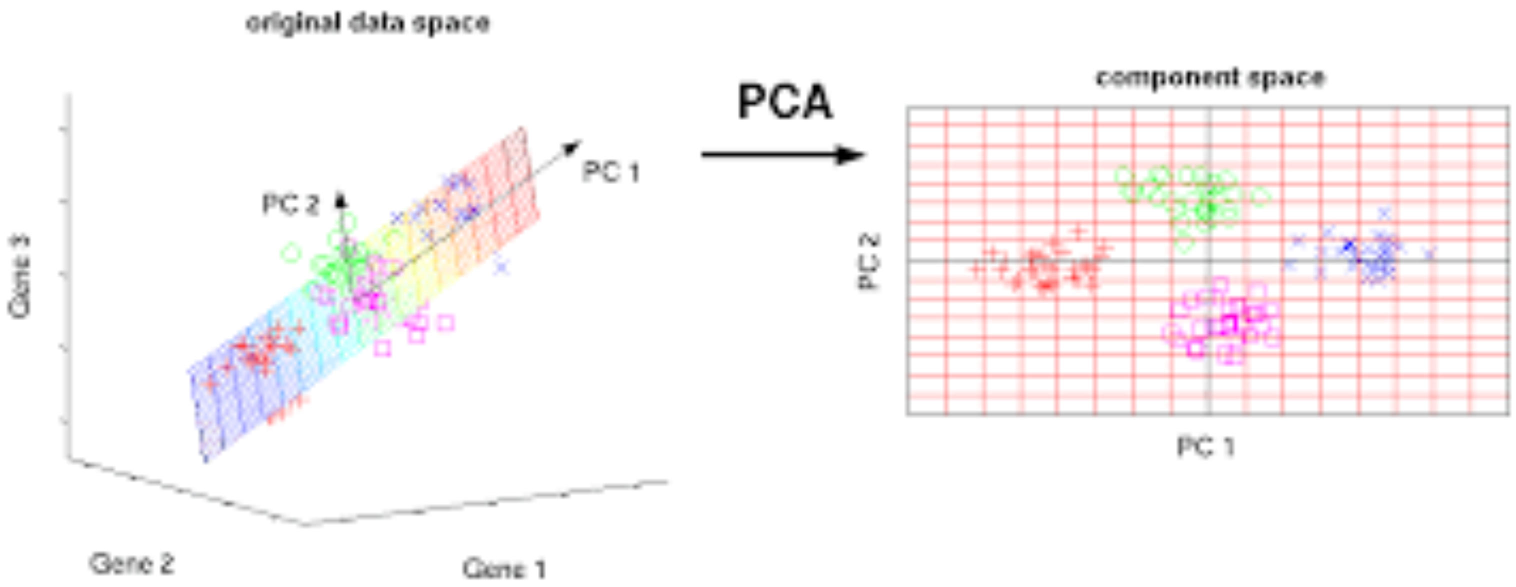


Sensitivity: Internal

Feature reduction

- Dimensionality.= The number of input variables or features for a dataset
- More input features often make a predictive modeling task more challenging to model, \
  - “The curse of dimensionality”
- Dimensionality reduction = techniques that reduce the number of input variables in a dataset.
- the most popular technique for dimensionality reduction in machine learning is Principal Component Analysis, or PCA for short.
- This is a technique that comes from the field of linear algebra - Linear transformation of the feature-space.

PCA


original data space
PC 2
PC 1
Gene 3
Gene 2
Gene 1
PCA
component space
PC 2
PC 1

Example

- Task: Predict a person’s salary
- Given:
  - The person’s age,
  - education (# years),
  - Year of data collection,
  - average salary at that year,

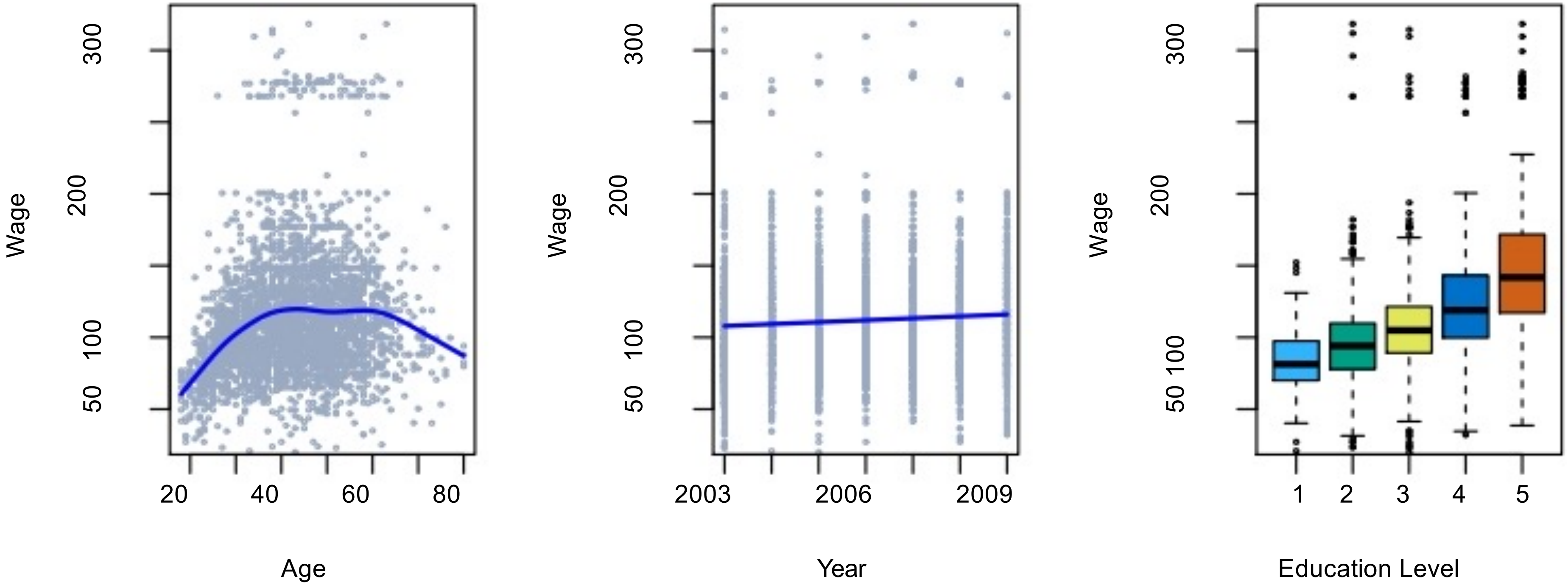
- Definitions:
- Explanatory variables = age, education (# years), Year of data collection, average salary at that year,
- Target Variable = the salary

Prediction Model

- First observe the data: Is there a connection/relation between the explanatory variables and the target variable?

Visualizations..

- Salary vs. age
- Average Salary vs. year
- Salary vs. education


Wage
Age
Wage
Year
Wage
Education Level