Data Pre-processing for Machine Learning

Swati Mishra

Applications of Machine Learning (4AL3)

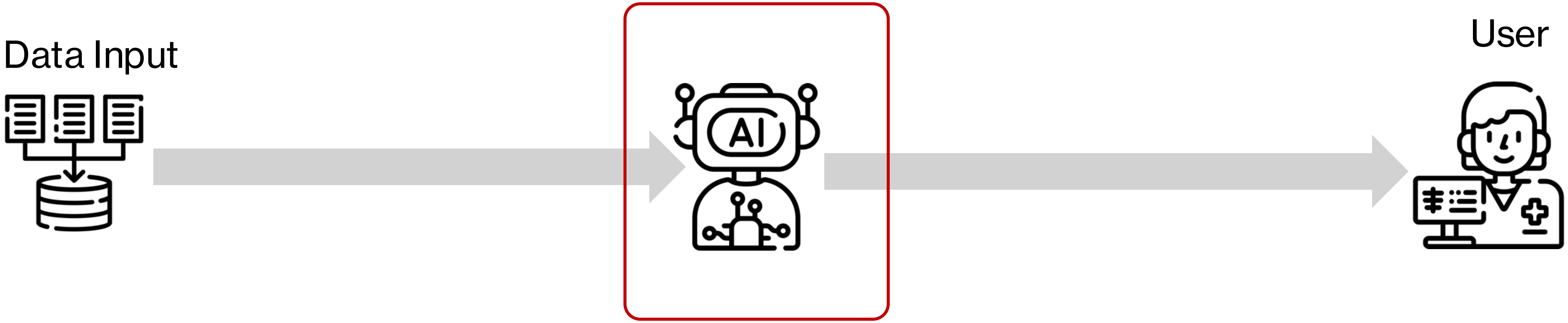
Fall 2024

McMaster University
ENGINEERING

Review

- Classification
- Logistic Regression
- Log Likelihood, Cross Entropy Loss
- Stochastic Gradient Descent

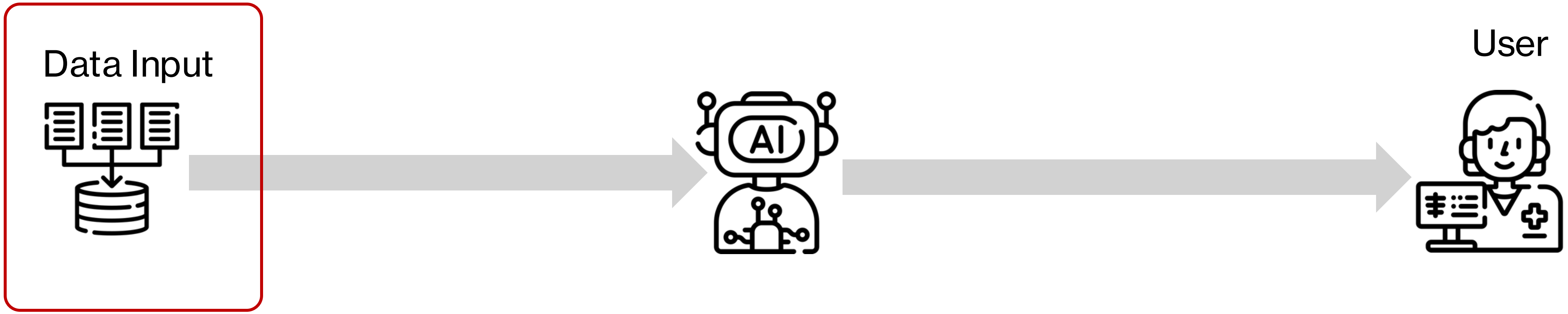
Review



So far we focused on ML model design

Data + Algorithm ⟶ Model

Review



Let’s pivot today and understand how data is input

Data + Algorithm ⟶ Model

What is Data

- Lens of Data Model – “raw” data
  - What the data *is* ? (i.e operations, statistical relations)
    - Example: 3-dimensional floating-point vector

f1	f2	f3
70	4	1
120	3	5
70	4	1
50	4	0
110	2	2
110	2	0
97.7	2.3	1.5

What is Data

- Lens of Conceptual Model – constructs
  - What does data *mean* ? semanticity
    - Example: temperature, grade point, dollars

f1	f2	f3
70	4	1
120	3	5
70	4	1
50	4	0
110	2	2
110	2	0
97.7	2.3	1.5

What is Data

- Lens of Data Model – “raw” data
  - A data model can capture the complex relationships between these numbers in a very statistical sense, e.g one might say that a higher f3 leads to higher f1.

  - What data *is* ? Statistical Relationship, Operations
    - Example: temperature, grade point, dollars

- Lens of Conceptual Model – conceptual
  - A conceptual model captures the constructs in data and adds meaning to attributes, e.g. one might say that if f1, f2, and f3, are constructs like Calories, Proteins and Fats.

  - What data *means* ? Semanticity
    - Example: temperature, grade point, dollars

What is the use of Data

- Lens of Data Model – “raw” data
  - Analyze, process, clean
  - e.g. what is the relationship

- Lens of Conceptual Model – conceptual
  - Provide meaning, direct hypotheses
  - e.g. why is it increasing/decreasing

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.

Types of Data

- Nominal (labels, categories, groups)
    - Nominal data is a type of qualitative data that represents categories or labels.
    - They are primarily mutually exclusive and cannot be ranked.
    - Example: Dog Breeds, what else?

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.
- Ordered (ordinal, rankings)
  - Ordinal data is a type of qualitative or quantitative data that represents categories or values
  - They have inherent order, but differences between values are not informative.

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.
- Ordered (ordinal, rankings)
  - Ordinal data is a type of qualitative or quantitative data that represents categories or values
  - They have inherent order, but differences between values are not informative.
  - Example: Class grade, what else?

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.
- Ordered (ordinal, rankings)
  - Ordinal data is a type of qualitative or quantitative data that represents categories or values
  - They have inherent order, but differences between values are not informative.
- Interval (no true zero)
  - Interval data is a type of quantitative data that represents measurements on a scale
  - The intervals between the values are equal, but there is no true zero point.

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.
- Ordered (ordinal, rankings)
  - Ordinal data is a type of qualitative or quantitative data that represents categories or values
  - They have inherent order, but differences between values are not informative.
- Interval (no true zero)
  - Interval data is a type of quantitative data that represents measurements on a scale
  - The intervals between the values are equal, but there is no true zero point.
  - Example: Date, what else?

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.
- Ordered (ordinal, rankings)
  - Ordinal data is a type of qualitative or quantitative data that represents categories or values
  - They have inherent order, but differences between values are not informative.
- Interval (no true zero)
  - Interval data is a type of quantitative data that represents measurements on a scale
  - The intervals between the values are equal, but there is no true zero point.
- Ratio (with true zero)
  - Ratio data is a type of quantitative data that represents measurements on a scale
  - There is an inherent order, equal intervals between the values, and differences are meaningful.

Types of Data

- Nominal (labels, categories, groups)
  - Nominal data is a type of qualitative data that represents categories or labels.
  - They are primarily mutually exclusive and cannot be ranked.
- Ordered (ordinal, rankings)
  - Ordinal data is a type of qualitative or quantitative data that represents categories or values
  - They have inherent order, but differences between values are not informative.
- Interval (no true zero)
  - Interval data is a type of quantitative data that represents measurements on a scale
  - The intervals between the values are equal, but there is no true zero point.
- Ratio (with true zero)
  - Ratio data is a type of quantitative data that represents measurements on a scale
  - There is an inherent order, equal intervals between the values, and differences are meaningful.
  - Example: Height, what else?

Types of Data

- Nominal (labels, categories, groups)
  - Example: Dog breeds, gender, nationality
- Ordered (ordinal, rankings)
  - Example: Class grade, Likert scale (like, neutral, dislike)
- Interval (no true zero)
  - Example: Date, time
- Ratio (with true zero)
  - Example: Height, weight, income

Types of Data: Example

- Data : 44.0, 54.2, 78.4 , 42.1, 102.3
- Concept : Temperature in Celsius
- Kinds
  - Nominal
    - ?
  - Ordinal
    - ?
  - Interval / Ratio
    - ?

What kinds of data can it represent?

McMaster University

Types of Data: Example

- Data : 44.0, 54.2, 78.4 , 42.1, 102.3
- Concept : Temperature in Celsius
- Kinds
  - Nominal
    - Water freezes / water doesn’t freeze
  - Ordinal
    - Warm, cold, freezing
  - Interval / Ratio
    - Celsius or Kelvin

Source: S. S. Stevens, On the theory of scales of measurements, 1946

Types of Data: Example

What are Nominal, Ordered, Quantitative (Interval, Ratio) attributes here?

name	Manufacturer	Calories	Protein	Fat	Sodium	Fiber	Carbohydrates
100% Bran	Nabisco	70	4	1	130	10	5
100% Natural Bran	Quaker Oats	120	3	5	15	2	8
All-Bran	Kelloggs	70	4	1	260	9	7
All-Bran with Extra Fiber	Kelloggs	50	4	0	140	14	8
Apple Cinnamon Cheerio	General Mills	110	2	2	180	1.5	10.5
Apple Jacks	Kelloggs	110	2	0	125	1	11
Basic 4	General Mills	97.7	2.3	1.5	157.9	1.5	13.5
Bran Chex	Ralston Purina	90	2	1	200	4	15
Bran Flakes	Post	90	3	0	210	5	13
Cap'n'Crunch	Quaker Oats	120	1	2	220	0	12
Cheerios	General Mills	110	6	2	290	2	17
Cinnamon Toast Crunch	General Mills	120	1	3	210	0	13
Clusters	General Mills	110	3	2	140	2	13
Cocoa Puffs	General Mills	110	1	1	180	0	12
Corn Chex	Ralston Purina	110	2	0	280	0	22
Corn Flakes	Kelloggs	100	2	0	290	1	21
Corn Pops	Kelloggs	110	1	0	90	1	13
Count Chocula	General Mills	110	1	1	180	0	12
Cracklin' Oat Bran	Kelloggs	110	3	3	140	4	10
Crispix	Kelloggs	110	2	0	220	1	21

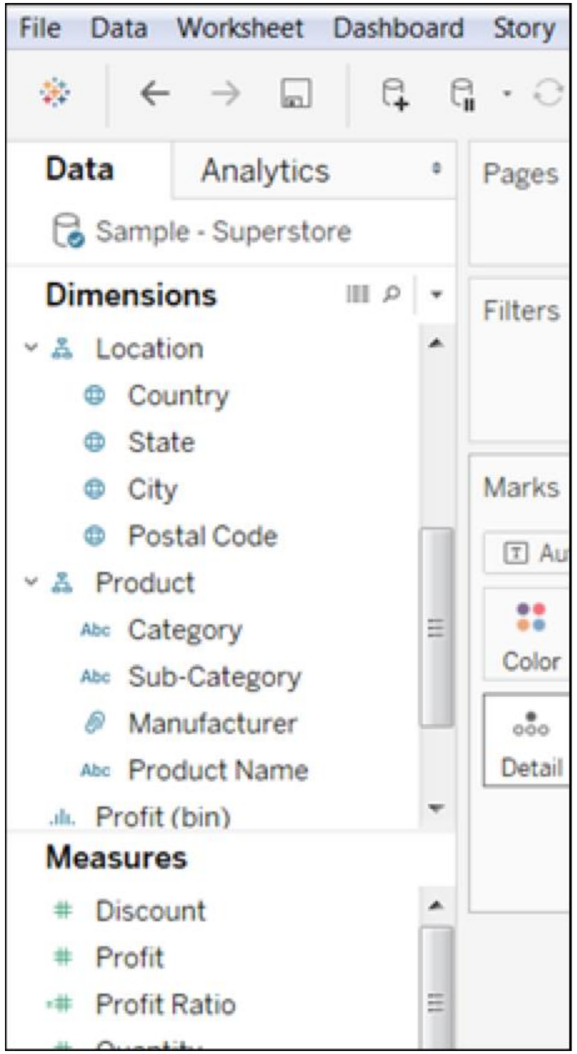
Empirical Operations

- Nominal (labels, categories, groups) - (Determination of Equality)
  - $= \neq \in \notin$
- Ordered (ordinal, rankings) - (Determination of greater or less)
  - $= \neq \in \notin < >$
- Interval - (Determination of equality of intervals or differences)
  - $= \neq > < + -$
- Ratio - (Determination of equality of ratios)
  - $= \neq > < + - \times \div \%$

Common Conventions

- Dimensions
  - Qualitative values
  - Categorize, segment
  - Dates, Names

- Measures
  - “Mathematical” data
  - Numeric, able to be aggregated w/functions
  - Petal width, height, temperature, grade point



Types of Data: Example

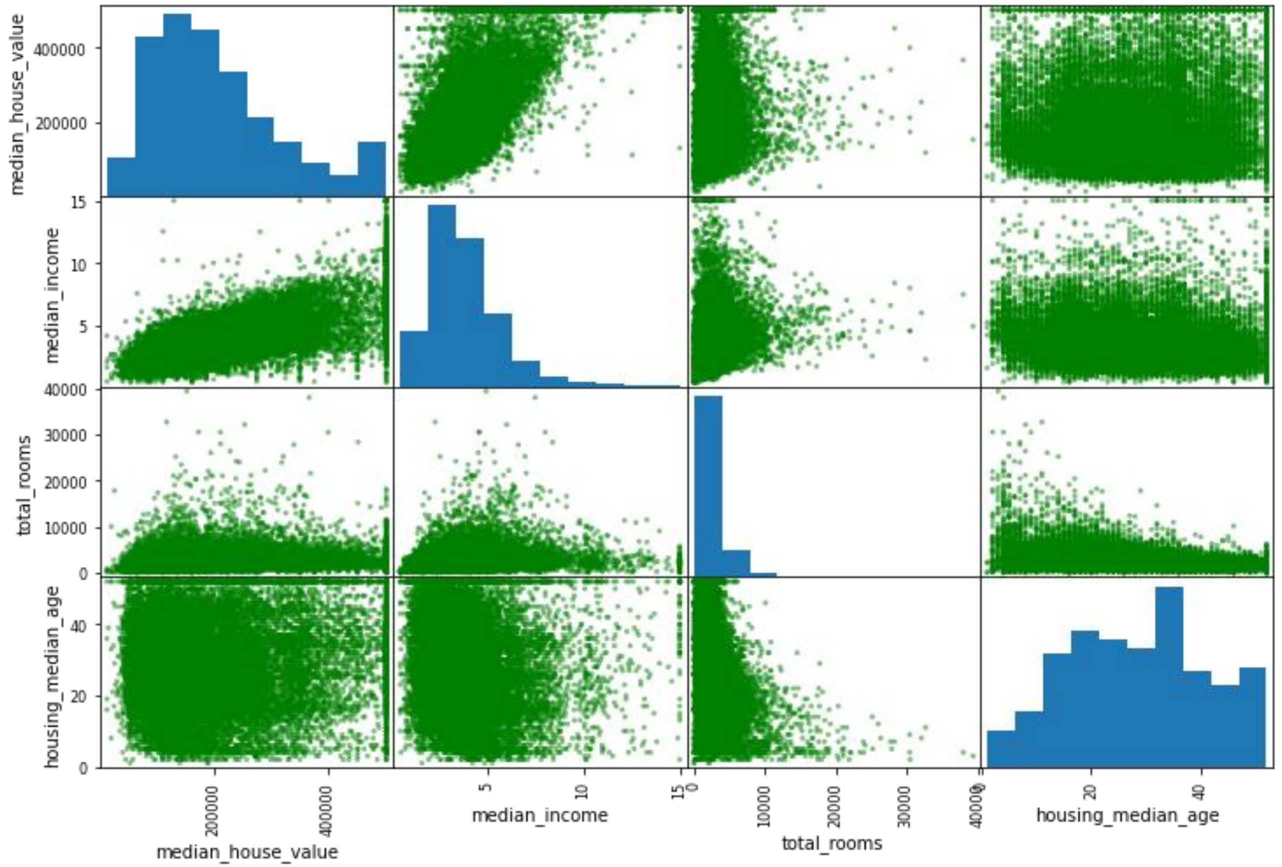
What are Dimensions and Measures here?

name	Manufacturer	Calories	Protein	Fat	Sodium	Fiber	Carbohydrates
100% Bran	Nabisco	70	4	1	130	10	5
100% Natural Bran	Quaker Oats	120	3	5	15	2	8
All-Bran	Kelloggs	70	4	1	260	9	7
All-Bran with Extra Fiber	Kelloggs	50	4	0	140	14	8
Apple Cinnamon Cheerio	General Mills	110	2	2	180	1.5	10.5
Apple Jacks	Kelloggs	110	2	0	125	1	11
Basic 4	General Mills	97.7	2.3	1.5	157.9	1.5	13.5
Bran Chex	Ralston Purina	90	2	1	200	4	15
Bran Flakes	Post	90	3	0	210	5	13
Cap'n'Crunch	Quaker Oats	120	1	2	220	0	12
Cheerios	General Mills	110	6	2	290	2	17
Cinnamon Toast Crunch	General Mills	120	1	3	210	0	13
Clusters	General Mills	110	3	2	140	2	13
Cocoa Puffs	General Mills	110	1	1	180	0	12
Corn Chex	Ralston Purina	110	2	0	280	0	22
Corn Flakes	Kelloggs	100	2	0	290	1	21
Corn Pops	Kelloggs	110	1	0	90	1	13
Count Chocula	General Mills	110	1	1	180	0	12
Cracklin' Oat Bran	Kelloggs	110	3	3	140	4	10
Crispix	Kelloggs	110	2	0	220	1	21

Finding Correlations

Index	Longitude	Latitude	Housing Median Age	Total Rooms	Total Bedrooms	Population	Households	Median Income	Median House value	Ocean Proximity	Rooms Per Household	Bedrooms	Population per household
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY	6.984127	0.146591	2.555556
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY	6.238137	0.155797	2.109842
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY	8.288136	0.129516	2.802260
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY	5.817352	0.184458	2.547945
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY	6.281853	0.172096	2.181467
...	...	...	...	...	...	...	...	...	...	...	...	...	...
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	78100.0	INLAND	5.045455	0.224625	2.560606
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	77100.0	INLAND	6.114035	0.215208	3.122807
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	92300.0	INLAND	5.205543	0.215173	2.325635
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	84700.0	INLAND	5.329513	0.219892	2.123209
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	89400.0	INLAND	5.254717	0.221185	2.616981

Finding Correlations

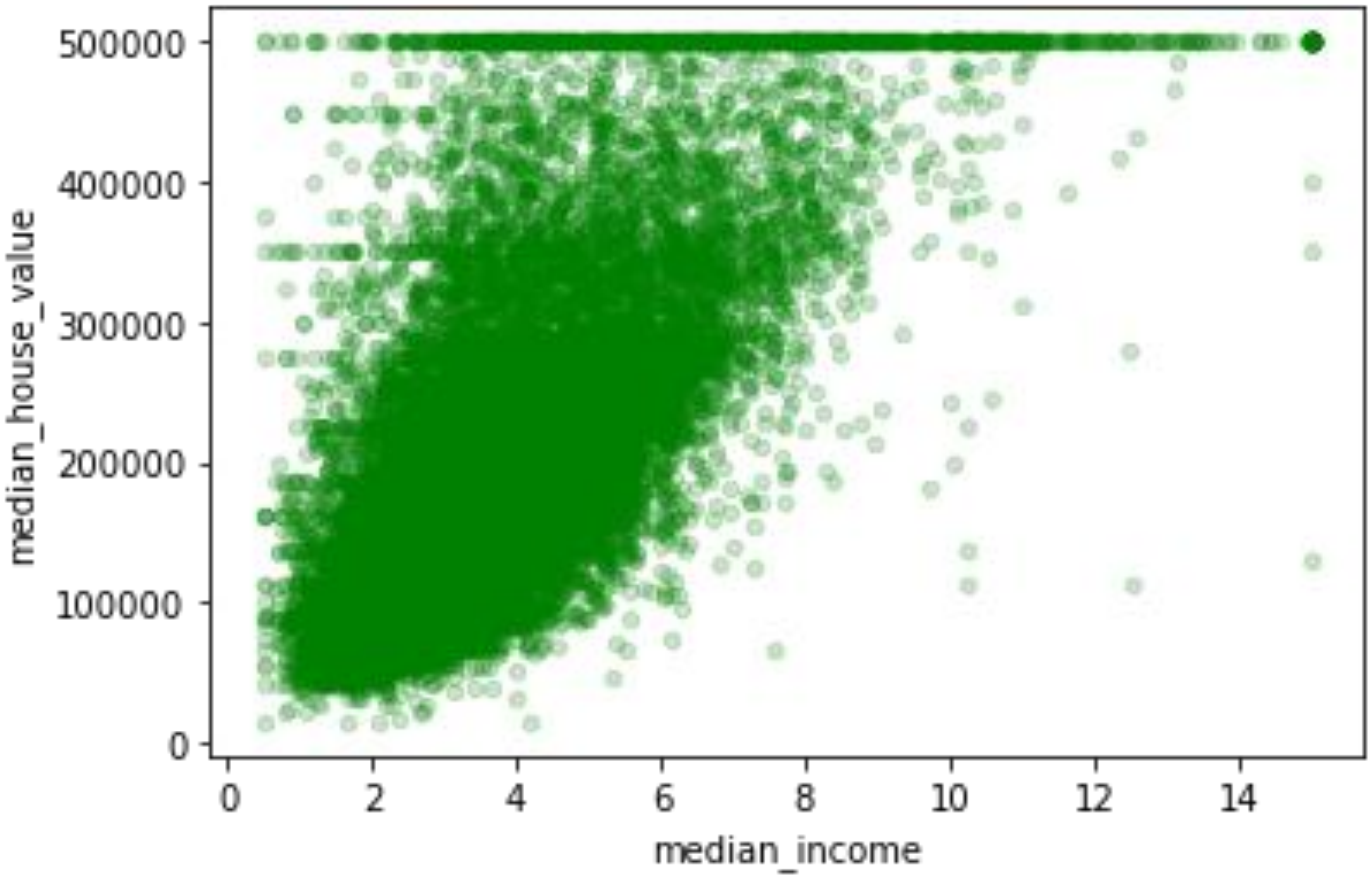



```python
corr_matrix = housing.corr()
corr_matrix["median_house_value"].sort_values(ascending=False)
```

```
median_house_value    1.000000
median_income         0.688075
total_rooms           0.134153
housing_median_age    0.105623
households            0.065843
total_bedrooms        0.049686
population           -0.024650
longitude            -0.045967
latitude             -0.144160
Name: median_house_value, dtype: float64
```

- Correlation close to 1 = strong positive
- Correlation close to -1 =strong negative

Finding Correlations



McMaster University

Data Model vs Concept Model

Before adding correlation feature

```
corr_matrix = housing.corr()
corr_matrix["median_house_value"].sort_values(ascending=False)
```

[6] 0.0s

```
median_house_value    1.000000
median_income         0.688075
total_rooms           0.134153
housing_median_age    0.105623
households            0.065843
total_bedrooms        0.049686
population           -0.024650
longitude            -0.045967
latitude             -0.144160
Name: median_house_value, dtype: float64
```

After adding correlation feature

```
housing["rooms_per_household"] = housing["total_rooms"]/housing["households"]
housing["bedrooms_per_room"] = housing["total_bedrooms"]/housing["total_rooms"]
housing["population_per_household"]=housing["population"]/housing["households"]
```

8] 0.0s

```
corr_matrix = housing.corr()
corr_matrix["median_house_value"].sort_values(ascending=False)
```

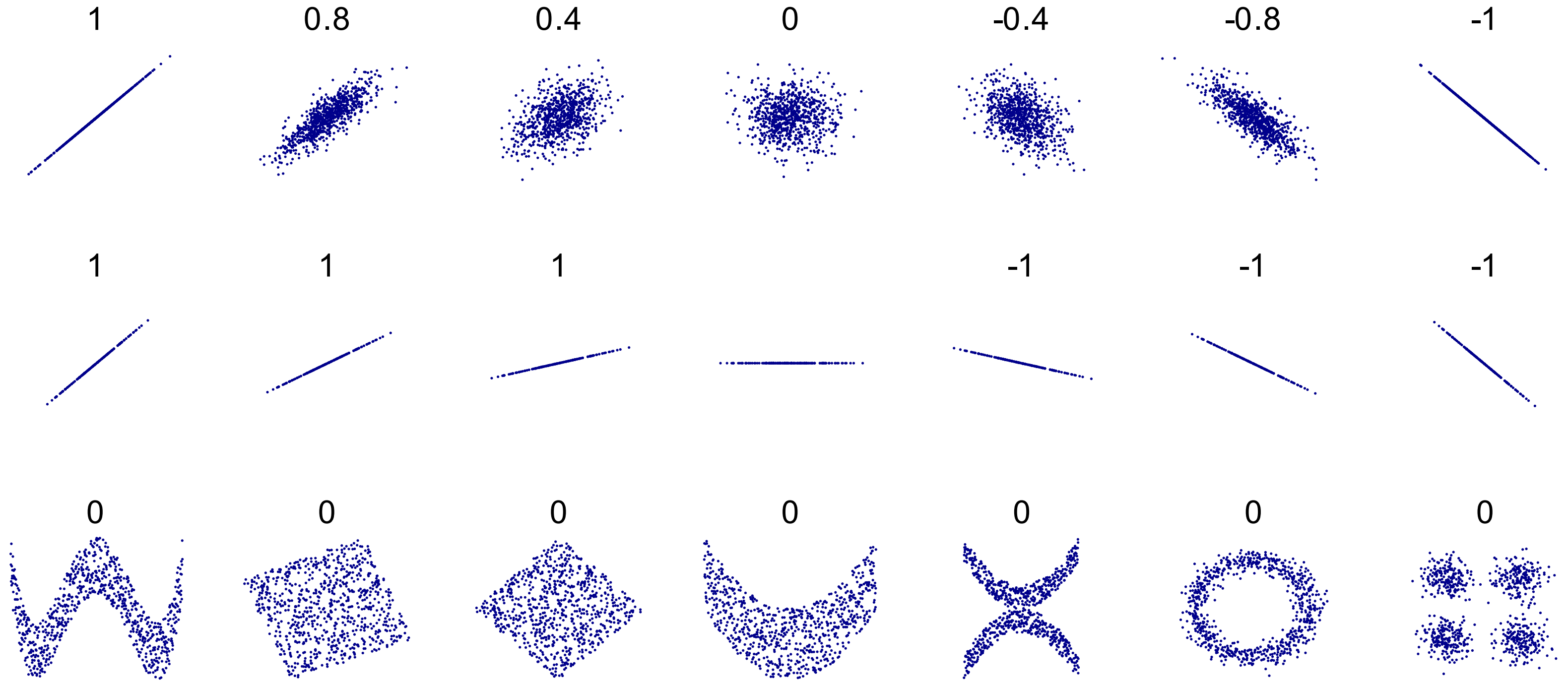
[20] 0.0s

```
median_house_value          1.000000
median_income               0.688075
rooms_per_household         0.151948
total_rooms                 0.134153
housing_median_age          0.105623
households                  0.065843
total_bedrooms              0.049686
population_per_household   -0.023737
population                 -0.024650
longitude                  -0.045967
latitude                   -0.144160
bedrooms_per_room          -0.255880
Name: median_house_value, dtype: float64
```

McMaster University

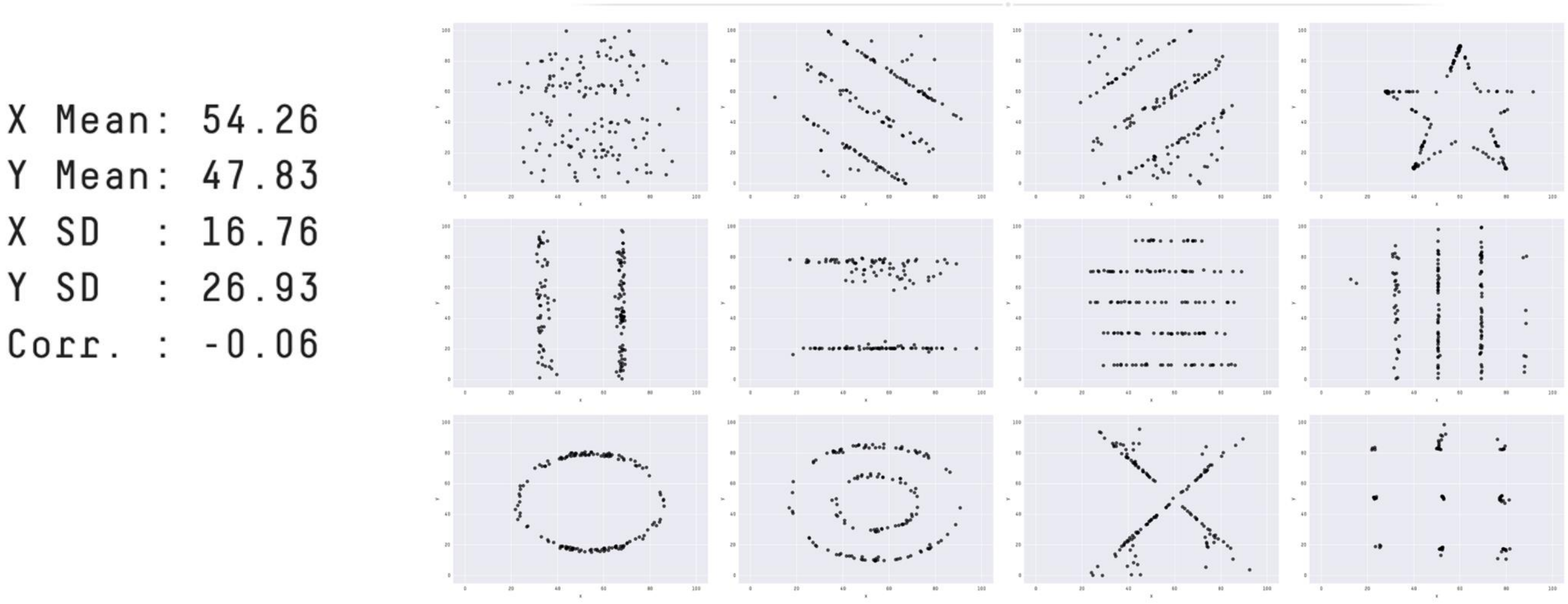
Why data model is not sufficient

- Different distributions, different correlations



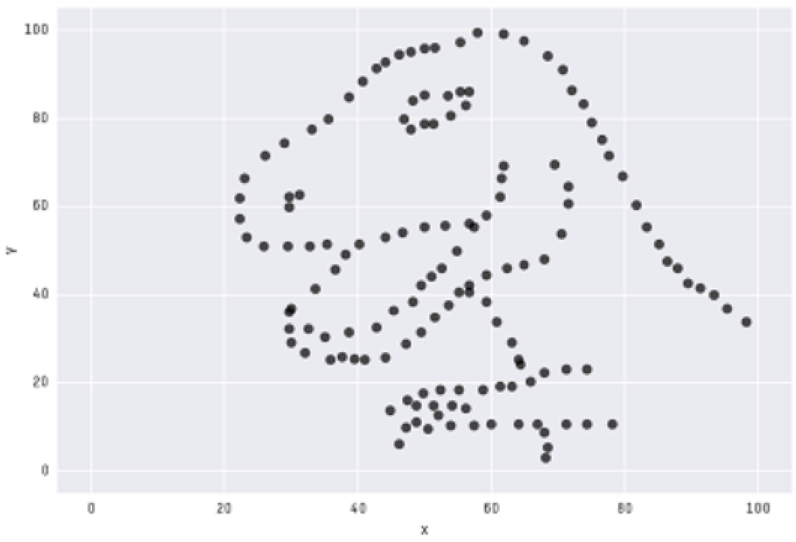
Why data model is not sufficient

- Same Statistics, Different Data



Why data model is not sufficient

- Same Statistics, Different Data

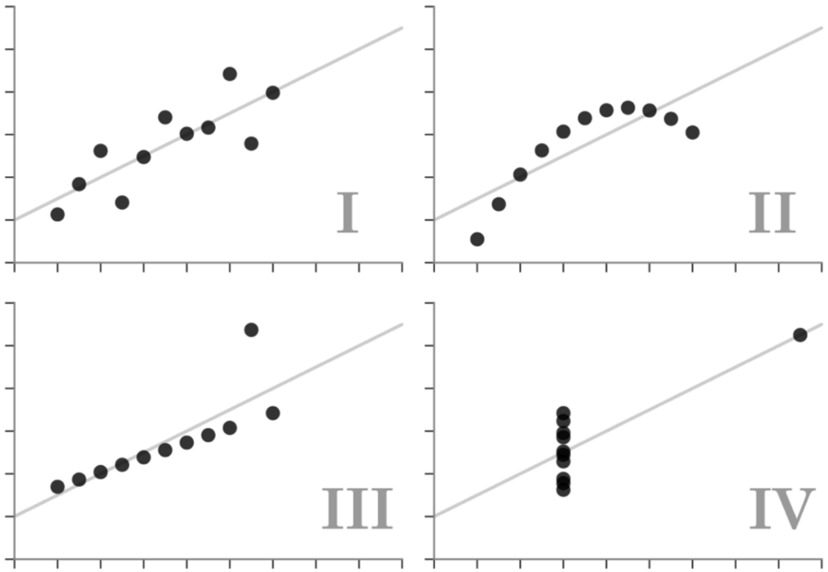


X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

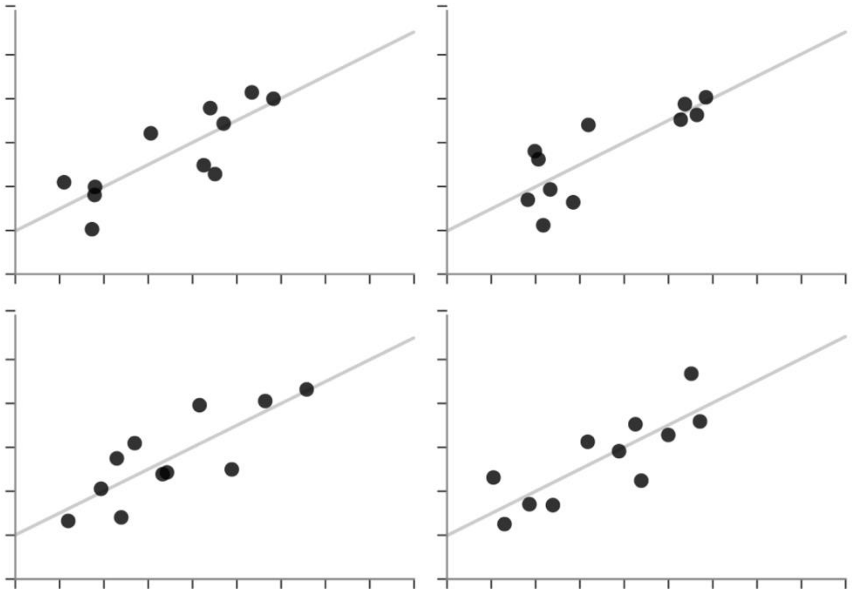
Why data model is not sufficient

- Same Statistics, Different Data

Anscombe's Quartet
Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



Unstructured Quartet
Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



Preparing Data for Machine Learning

- Data Cleaning
- Categorical Variables
- Feature Scaling
- Creating Test Sets

Removing missing data

- Delete corresponding row.
- Delete entire column.
- Set missing values to some value
  - (zero, the mean, the median, etc.).

Preparing Data for Machine Learning

- Data Cleaning
- Categorical Variables
- Feature Scaling
- Creating Test Sets

```
housing_cat = housing[["ocean_proximity"]]
housing_cat.value_counts()
```

✓ 0.0s

```
ocean_proximity
<1H OCEAN         9136
INLAND            6551
NEAR OCEAN        2658
NEAR BAY          2290
ISLAND               5
dtype: int64
```

```
ordinal_encoder = OrdinalEncoder()
housing_cat_encoded = ordinal_encoder.fit_transform(housing_cat)
print(ordinal_encoder.categories_)
housing_cat_encoded[:10]
```

[20] ✓ 0.0s

```
[array(['<1H OCEAN', 'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN'],
      dtype=object)]
```

```
array([[3.],
       [3.],
       [3.],
       [3.],
       [3.],
       [3.],
       [3.],
       [3.],
       [3.],
       [3.]])
```

```
housing_cat.head(10)
```

[18] ✓ 0.0s

	ocean_proximity
0	NEAR BAY
1	NEAR BAY
2	NEAR BAY
3	NEAR BAY
4	NEAR BAY
5	NEAR BAY
6	NEAR BAY
7	NEAR BAY
8	NEAR BAY
9	NEAR BAY

Converting text to numbers

Preparing Data for Machine Learning

- Data Cleaning
- Categorical Variables
- Feature Scaling
- Creating Test Sets

```
from sklearn.preprocessing import OneHotEncoder
cat_encoder = OneHotEncoder()
housing_cat_1hot = cat_encoder.fit_transform(housing_cat)
housing_cat_1hot.toarray()
```

[23] ✓ 0.0s

```
array([[0., 0., 0., 1., 0.],
       [0., 0., 0., 1., 0.],
       [0., 0., 0., 1., 0.],
       ...,
       [0., 1., 0., 0., 0.],
       [0., 1., 0., 0., 0.],
       [0., 1., 0., 0., 0.]])
```

Converting numbers to arrays

Preparing Data for Machine Learning

- Data Cleaning
- Categorical Variables
- Feature Scaling
- Creating Test Sets

Min-Max Scaler

```
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
X_scaled = X_std * (max - min) + min
```

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

Standard Scaler

```
x_scaled = [(x - x.mean)/std_dev]
```

$$x' = \frac{x - \bar{x}}{\sigma}$$

Preparing Data for Machine Learning

- Data Cleaning
- Categorical Variables
- Feature Scaling
- Creating Test Sets

How to split the test dataset:

- 80-20 split
- Stratified split

Readings

***Reference Material*:**

1. Feature Scaling
2. Stratified Split
3. Categorical Vairables
4. Datasaurus

Thank You