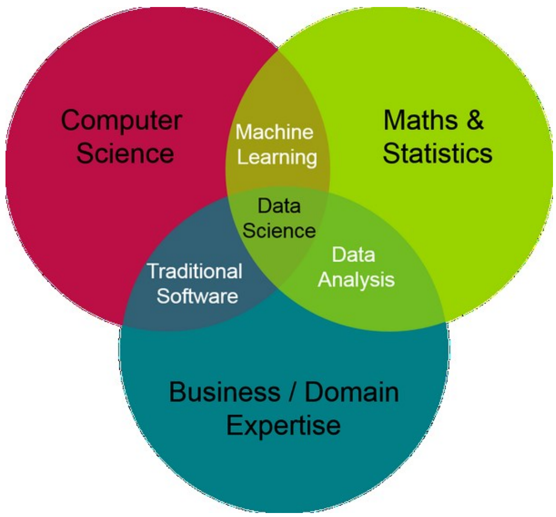
Introduction to Data Science

Data Science Lifecycle

What is Data Science?

What is Data Science?

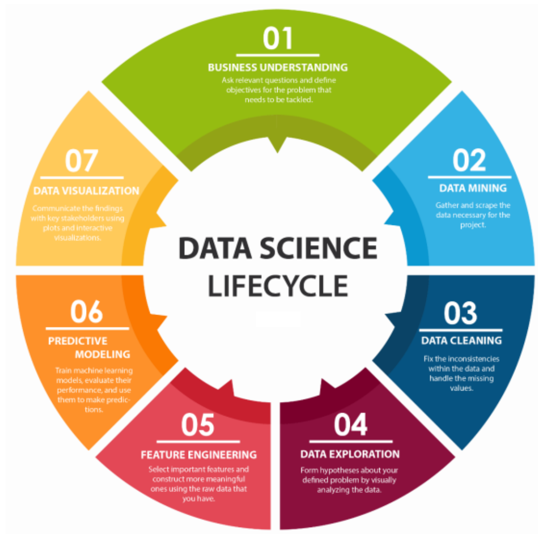


3

Data Science Lifecycle

4

Data Science Lifecycle



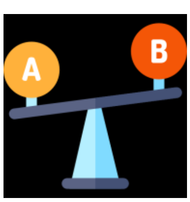
1. BUSINESS UNDERSTANDING

- A project starts by understanding the *what*, the *why*, and the *how* of your project.
- The outcome of this phase:
  - clear research goal
  - a good understanding of the context
  - well-defined deliverables
  - a plan of action with a timetable and cost estimate
- The design team should think carefully about the use scenario
  - The business problem will be mapped to data science tasks.

Problem Definition

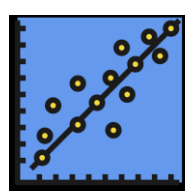
- **Define objectives**: work with your customer to understand and identify the business problems.
- **Formulate questions**: convert the business goals into questions that the data science techniques can target.
- **Define the success metrics**: look for specific, measurable, achievable, relevant, and time-bound metrics.
- **Identify data sources**: look for the data that is relevant to the question.

Formulate Questions



Comparison


Description


Regression

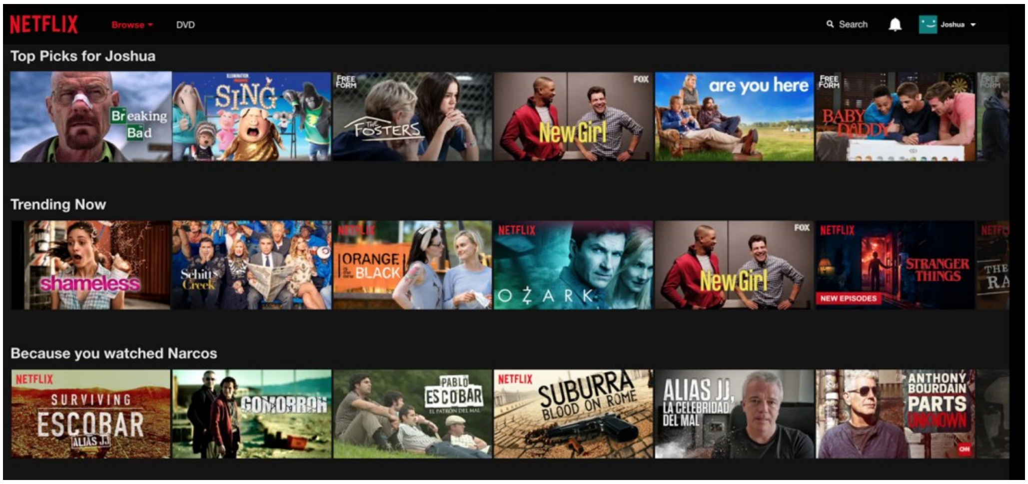

Classification


Clustering

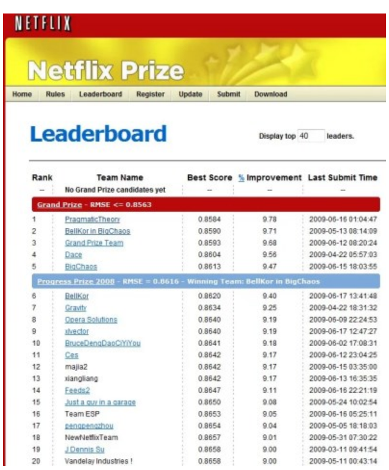

Anomaly Detection


Recommendation

Netflix Recommender System



Netflix Recommender System


NETFLIX

Netflix Prize

Home | Rules | Leaderboard | Register | Update | Submit | Download

Leaderboard

Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
1	PragmaticTheory	0.8584	9.78	2009-06-16 01:04:47
2	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
3	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
4	Dace	0.8604	9.56	2009-04-22 05:57:03
5	BigChaos	0.8613	9.47	2009-06-15 18:03:55
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
6	BellKor	0.8620	9.40	2009-06-17 13:41:48
7	Gravity	0.8634	9.25	2009-04-22 18:31:32
8	Opera Solutions	0.8640	9.19	2009-06-09 22:24:53
9	xlvector	0.8640	9.19	2009-06-17 12:47:27
10	BruceDengDaoCiYiYou	0.8641	9.18	2009-06-02 17:08:31
11	Ces	0.8642	9.17	2009-06-12 23:04:25
12	majia2	0.8642	9.17	2009-06-15 03:35:00
13	xiangliang	0.8642	9.17	2009-06-13 16:35:35
14	Feeds2	0.8647	9.11	2009-06-16 22:21:19
15	Just a guy in a garage	0.8650	9.08	2009-05-24 10:02:54
16	Team ESP	0.8653	9.05	2009-06-16 05:25:11
17	pengpengzhou	0.8654	9.04	2009-05-05 18:18:03
18	NewNetflixTeam	0.8657	9.01	2009-05-31 07:30:22
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	Vandelay Industries !	0.8658	9.00	2009-05-11 00:43:14

Define Success Metric

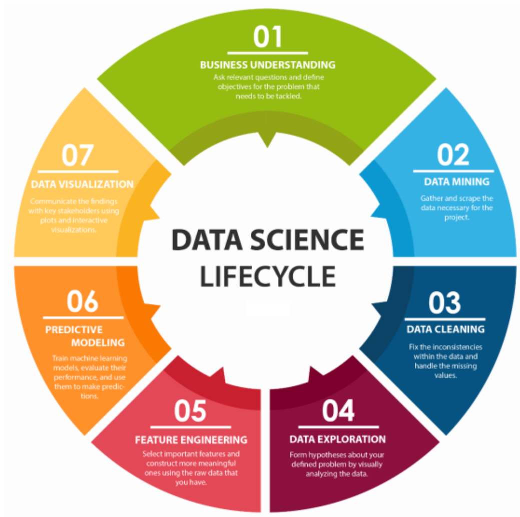
- Most companies don’t care about the fancy ML metrics.
- The sole purpose of businesses: maximize profits.
- In case of Netflix:
  - The objective is to increase revenue by 5%.
  - To increase revenue, we need to increase the customer retention rate by 8%.
  - To increase the customer retention rate, we need to increase the accuracy of the recommender system by 10%.
- Look for specific, measurable, achievable, relevant, and time-bound metrics.

Identify Data Sources

- Internal Data: many companies will have already collected and stored the data for you.

- External Data: the data outside your organization that needs to be bought from third parties or collected.

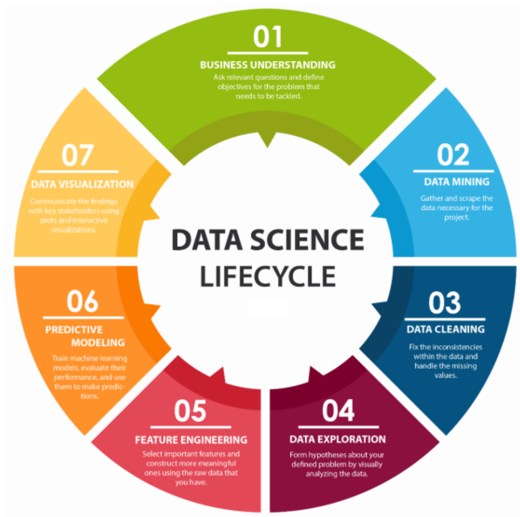
2. DATA MINING



Data Collection

- **Data collection** is the process of gathering and measuring information of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.
  - What data do I need for my project?
  - Where does it live?
  - How can I obtain it?
  - What is the most efficient way to store and access all of it?

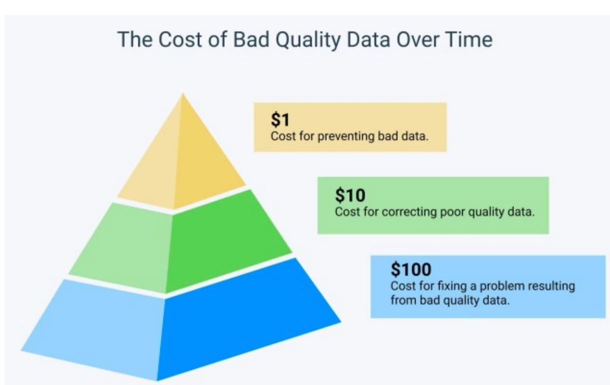
3. DATA CLEANING



15

Data Cleaning

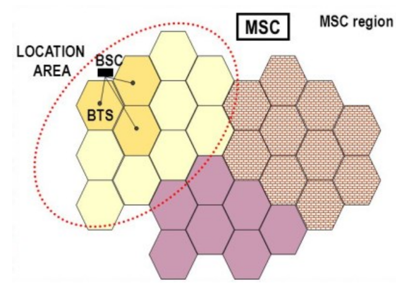
- Data cleaning is the process of editing, correcting, and structuring data within a data set so that it's generally uniform and prepared for analysis.



16

Scrub for Duplicate

- Duplicates: repeated data entries.
  - It usually happens when data is coming from different sources or users.

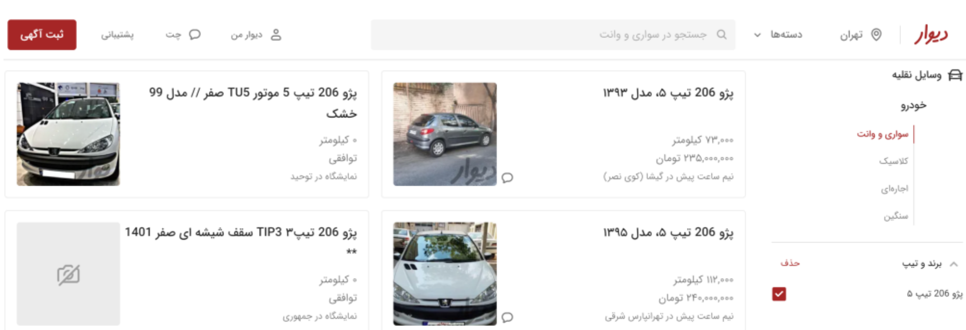


CALLING	CALLED	TARIKH	SAAT	MODDAT	LAC	CELL	IMSI	IMEI	OWNER	SOURCE	TYPE
9129348134	9104438695	1395/03/03	22:16	58	2223	32008	432112007113296	35206601728831	CALLING	mci	voice
9129348134	9104438695	1395/03/03	22:16	58					CALLED	mci	voice
9129348134	9104438695	1395/03/03	22:16	58	2223	32008	432112007113296	35206601728831	CALLING	mci	voice
9129348134	9104438695	1395/03/03	22:16	58					CALLED	mci	voice

CALLING	CALLED	TARIKH	SAAT	MODDAT	LAC	CELL	IMSI	IMEI	OWNER	SOURCE	TYPE
9906045127	22801240	1395/08/27	15:10:33	555	1264	30003	4.3212E+14	3.5645E+13	CALLING	mci	voice
9906045127	22801240	1395/08/27	15:10:20	554					CALLED	tci	voice
9906045127	22801240	1395/08/27	16:08:42	554					CALLED	tci	voice

Scrub for Irrelevant Data

- Irrelevant data is the type of information that doesn’t have any formal errors but is just not useful for your project.



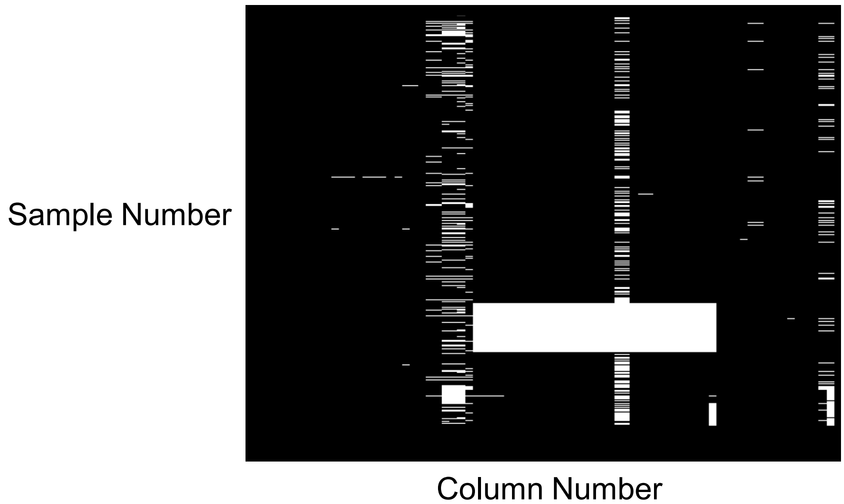
Scrub for Incorrect Data

- Incorrect data is often easy to spot, as it's just illogical.
  - Example: you're preparing a report about the app users' average age, and you see entries like -1 or 420.
- The reason for incorrect data lies within the processing stage, be it preparation or cleaning.
  - It is usually attributed to imprecisely defined functions, and transformations data went through.
- Amend the functions that caused the wrong calculations.
  - If not possible, then remove the data.

Handle Missing Data

- Missing data is just unavoidable. You're likely to find even whole rows and columns of missing values in your datasets.
- There three main methods of dealing with missing data:
  - **Drop**: When the missing values in a column are few and far between, the easiest way to handle them is to drop the missing data rows.
  - **Impute:** Calculate the missing values based on other observations.
    - Statistical techniques like median, mean, or linear regression.
    - Replacing missing data with entries from another "similar" database.
  - **Flag:** Missing data can be informative, especially if there is a pattern in play. Flagging the data can help you with those subtle insights.

Visualizing Missing Values



Check the Outliers

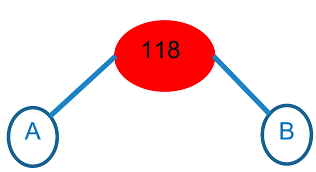
- Outliers are values that stand out and are significantly different from the others.
- They are not necessarily mistakes, but they can be.
- So how do you differentiate?
  - What you need to watch out for is the context.
  - Example: you're researching your app users' age, and you find entries like 72 and 2.
- Don't remove an outlier unless you know for a fact that it's a mistake.

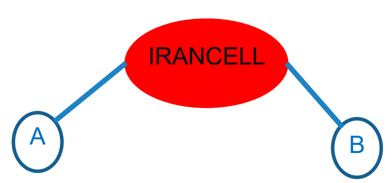
Standardize + Normalize

- Standardization and normalization make data ripe for statistical analysis and easy to compare and analyze.
- **Standardization** is a process during which you're making sure all your values adhere to a specific standard:
  - Deciding whether to go with kilos or grams, upper or lower case, etc.
  - Example: +989121234567, 00989121234567, 989121234567, 09121234567 → 9121234567
- **Normalization** is the process of adjusting the values to a common scale.
  - Example: rescale values into the 0-1 range.

Important Note!

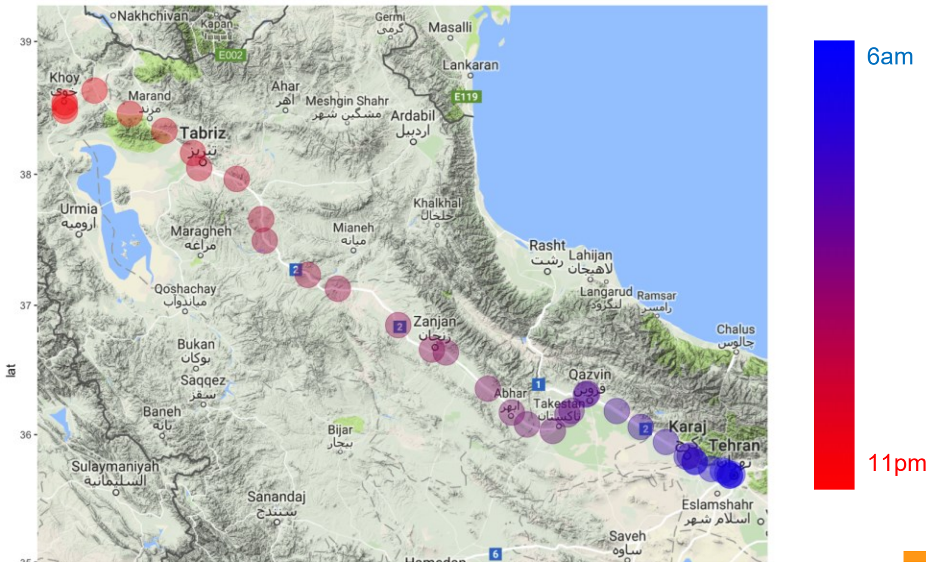
- In data cleaning we removed numbers that had less than 8 characters and numbers that contained letters:



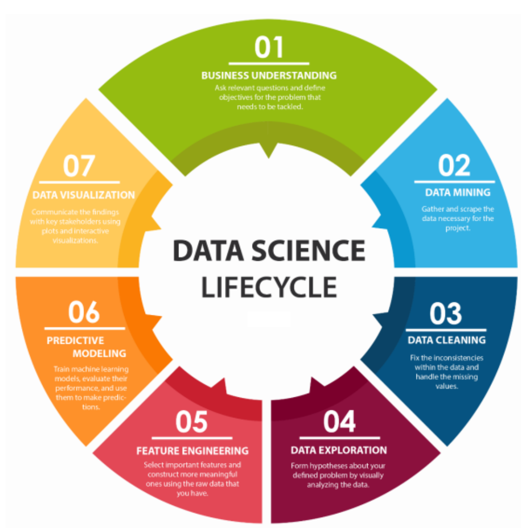


- Removing records that contain special numbers (e.g. 118, *8#, IRANCELL) may help social network analysis but it damages mobility analysis.
- Data preparation should be tailored to the specific analysis.

Mobility Analysis



4. DATA EXPLORATION



Data Exploration

- Data exploration is an approach to analyze the dataset using **visual** techniques, in order to better understand the nature of the data.

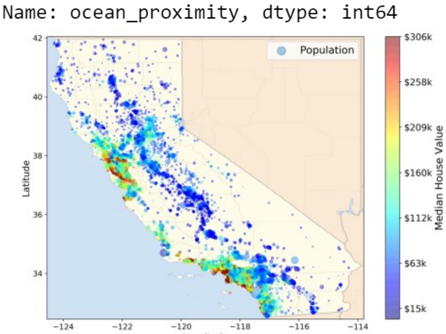
Variable Identification

```
housing.info()
```

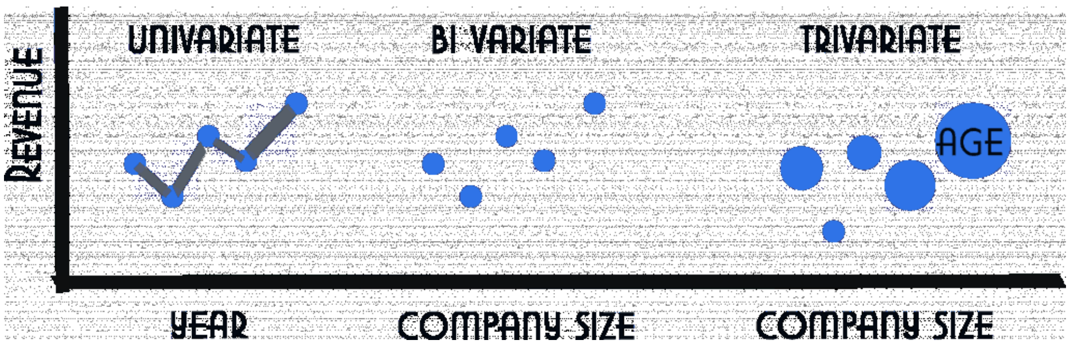
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

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

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



Exploratory Data Analysis



29

Anscombe's Quartet

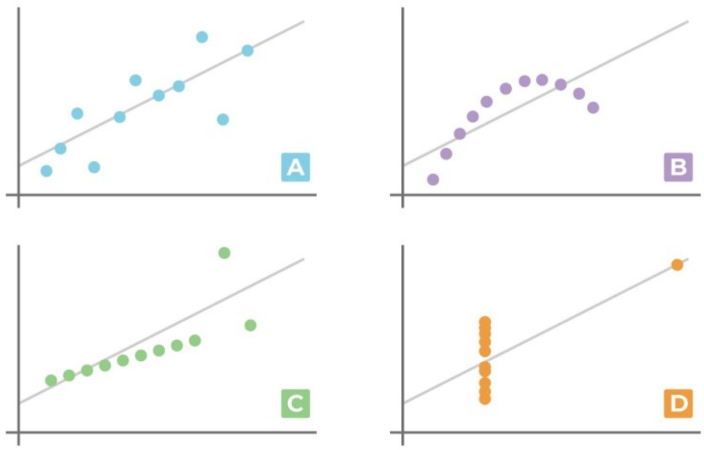
- For all four datasets:

Property	Value
Mean of *x*	9
Sample variance of *x*	11
Mean of *y*	7.50
Sample variance of *y*	4.125
Correlation between *x* and *y*	0.816
Linear regression line	$y = 3.00 + 0.500x$

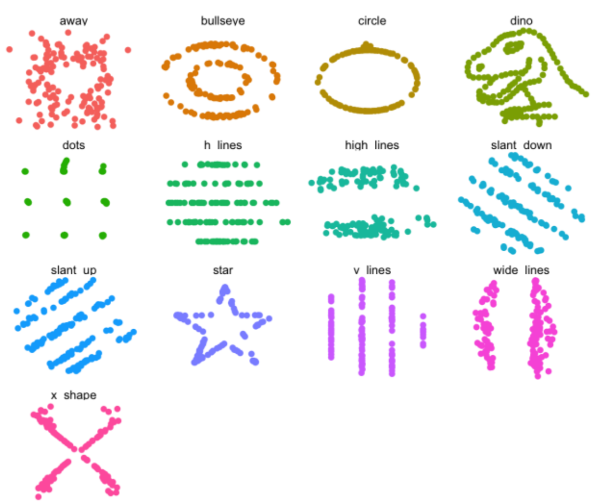
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

30

Anscombe's Quartet

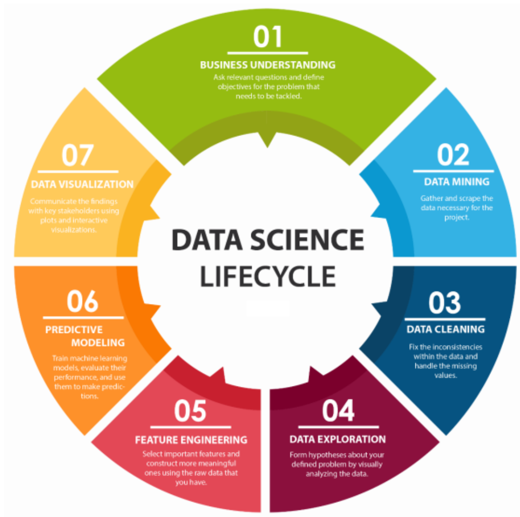


DataSaurus



dataset	mean(x)	mean(y)	var(x)	var(y)	cor(x, y)
away	54.266	47.835	281.227	725.750	−0.064
bullseye	54.269	47.831	281.207	725.533	−0.069
circle	54.267	47.838	280.898	725.227	−0.068
dino	54.263	47.832	281.070	725.516	−0.064
dots	54.260	47.840	281.157	725.235	−0.060
h_lines	54.261	47.830	281.095	725.757	−0.062
high_lines	54.269	47.835	281.122	725.763	−0.069
slant_down	54.268	47.836	281.124	725.554	−0.069
slant_up	54.266	47.831	281.194	725.689	−0.069
star	54.267	47.840	281.198	725.240	−0.063
v_lines	54.270	47.837	281.232	725.639	−0.069
wide_lines	54.267	47.832	281.233	725.651	−0.067
x_shape	54.260	47.840	281.231	725.225	−0.066

5. FEATURE ENGINEERING



Feature Engineering

- Feature engineering is the process of using domain knowledge to transform your raw data into informative features.
- This step requires a creative combination of domain expertise and the insights obtained from the data exploration step.
- This stage will directly influence the accuracy of the predictive model you construct in the next stage.

Feature Engineering

- Feature selection: is the process of cutting down the features that add more noise than information.
  - **Filter methods**: apply statistical measure to assign scoring to each feature
  - **Wrapper methods**: frame the selection of features as a search problem and use a heuristic to perform the search
  - **Embedded methods**: use machine learning to figure out which features contribute best to the accuracy
- Feature construction: involves creating new features from the ones that you already have.

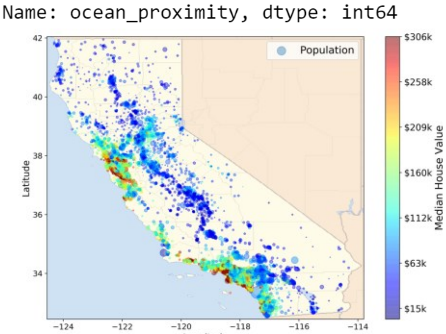
Housing Dataset

```
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

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

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



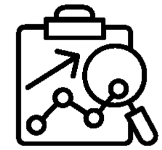
Feature Combinations

- Try out various feature combinations.
- Example: the total number of rooms in a district is not very useful if you don’t know how many households there are.
  - The number of rooms per household is more informative.
- Create new attributes:

```
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"]
```


6. PREDICTIVE MODELING

Predictive Modeling



- Predictive modeling is where the machine learning finally comes into your data science project.
- Depending on the type of question that you're trying to answer, there are many modeling algorithms available.
- The models that you train will be dependent on:
  - the size, type and quality of your data
  - how much time and computational resources you are willing to invest
  - the type of output you intend to derive.

7. DATA VISUALIZATION

Data Visualization

- Data visualization combines the fields of communication, psychology, statistics, and art, with an ultimate goal of communicating the data in a simple yet effective and visually pleasing way.

- Present your solution:
  - Highlighting what you have learned
  - Expose the model with an interface
    - Data Dashboards