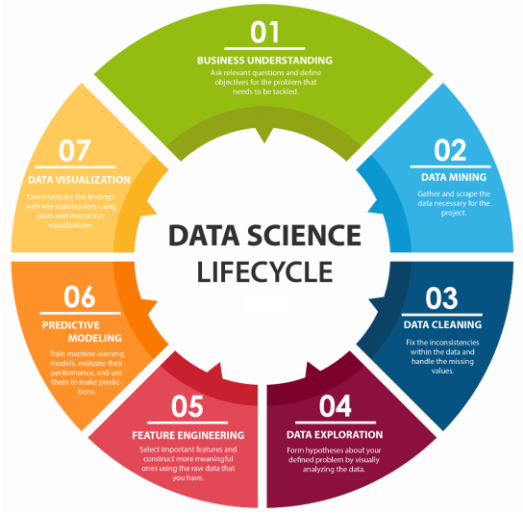
Data Visualization

Data Science Lifecycle

1. BUSINESS UNDERSTANDING



Problem Definition

- 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.

3

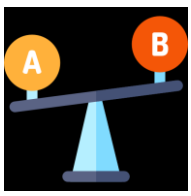
Problem Definition

- **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.

4

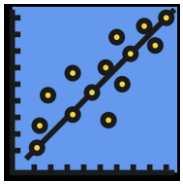
Formulate Questions

Typical Questions Answered by Data Science



Comparison

Description



Regression

Classification

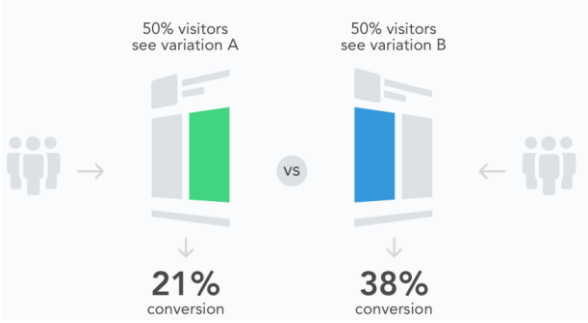
Clustering

Anomaly Detection

Recommendation

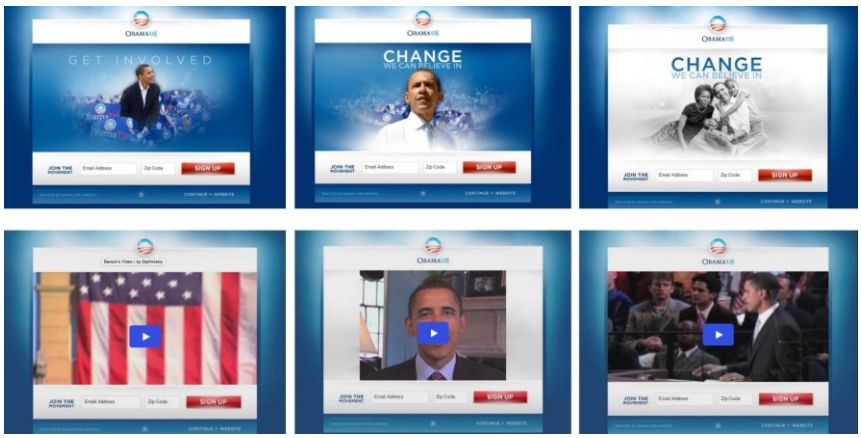
Comparison

- Is A better in some way than B?
  - Do users click on a green button more than a blue button?
  - Are males more inclined to buy our products than females?



7

A/B Testing for US Presidential Campaign



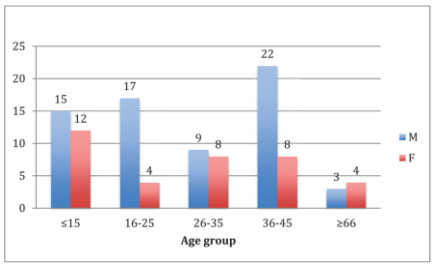
8

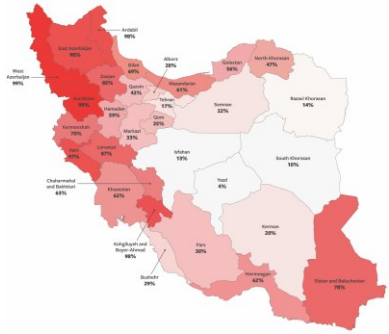
The Winner

9

Description

- What are the characteristics of the data?
  - What is the mean/variance of the monthly sales volume of a product in different regions?
  - What is the age distribution of the respondents?

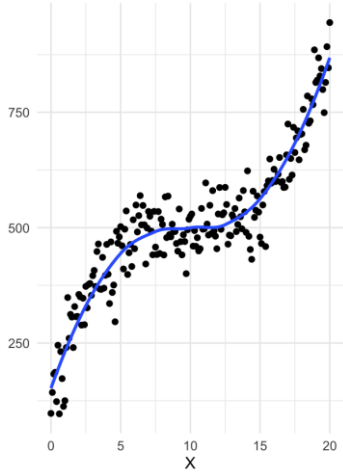




10

Regression

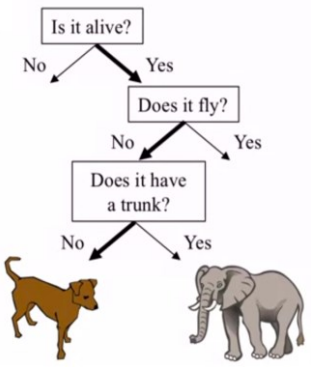
- How much or how many?
- *Regression* attempts to estimate or predict the numerical value of some variable for an entity.
  - How much demand a company will have for a given service?
  - How much will be the price of a certain stock tomorrow?
  - How many item **A** will be sold by store **S**?



11

Classification

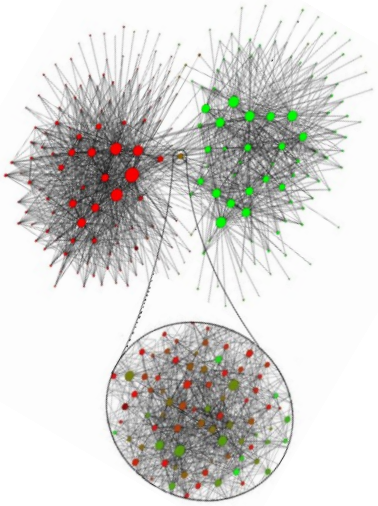
- Which category?
- *Classification* attempts to predict, for each individual in a population, which of a (small) set of classes this individual belongs to.
  - Is an email is spam or not?
  - Is this picture belongs to a mouse, a cat, or a dog?
  - Is a certain transaction fraudulent or not?
  - Is a website user male or female?
  - Is this customer going to buy our product or not?



12

Clustering

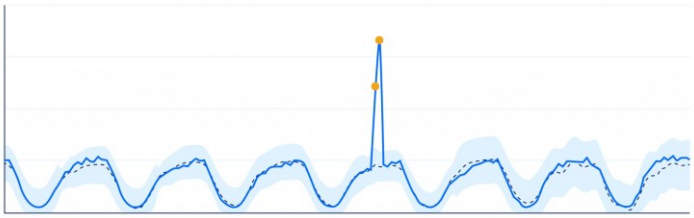
- Which group?
- *Clustering* attempts to *group* individuals in a population together by their similarity.
  - Which consumers have similar product preferences?
  - Which server performs similar pattern to the broken ones?
  - How many different kinds of employees are there in the company?



13

Anomaly Detection

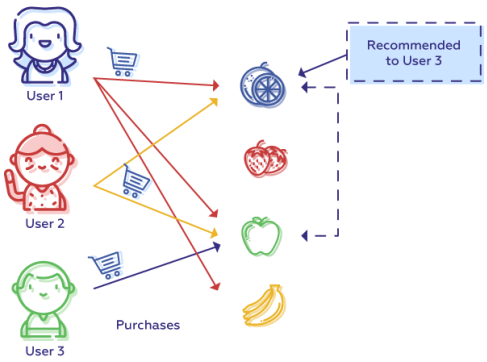
- Is this weird?
- Anomaly detection attempts to identify abnormal observations which are inconsistent or deviate significantly from the data.
  - Is the packet sizes of a specific IP address in our network normal?
  - Is this insurance claim look normal?
  - Is this tissue normal or cancerous?



14

Recommendation

- Which option should be taken?
- Recommender systems provide suggestions for items that are most suitable for a particular user.
  - Who to follow on Twitter?
  - Which items are most likely to be bought by a Digikala user?



15

Success Metric

16

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.

17

Identify Data Sources

18

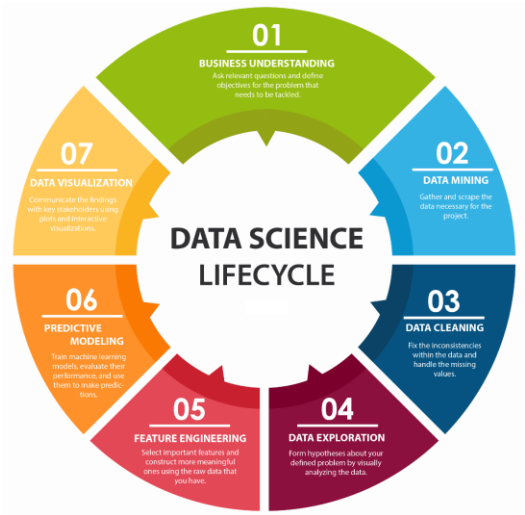
Identify Data Sources

- 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.

19

2. DATA MINING

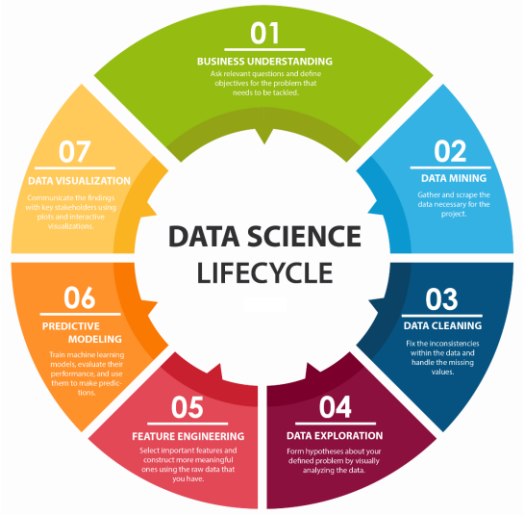


20

Data Collection

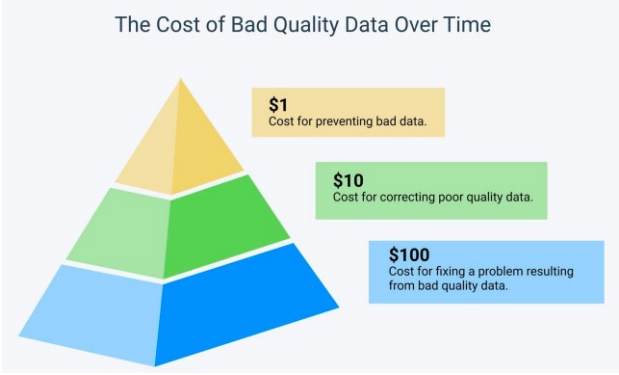
- **Data collection** is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.
  - What data do I need for my project?
  - Where does it live?
  -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

3. DATA CLEANING



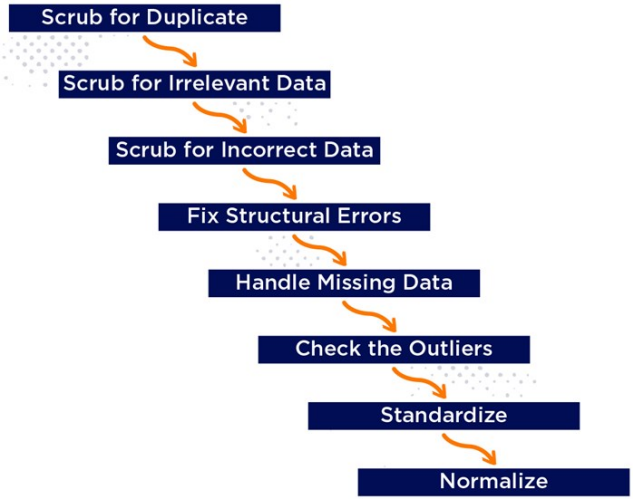
Data Cleaning

- 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.



23

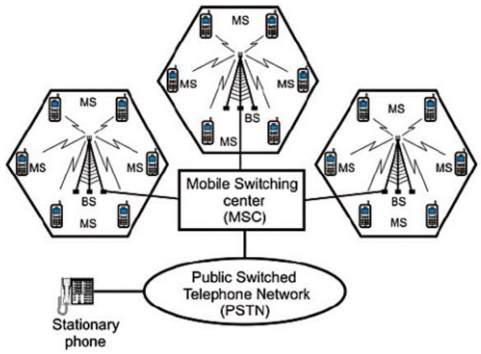
Data Cleaning Workflow



24

Scrub for Duplicate

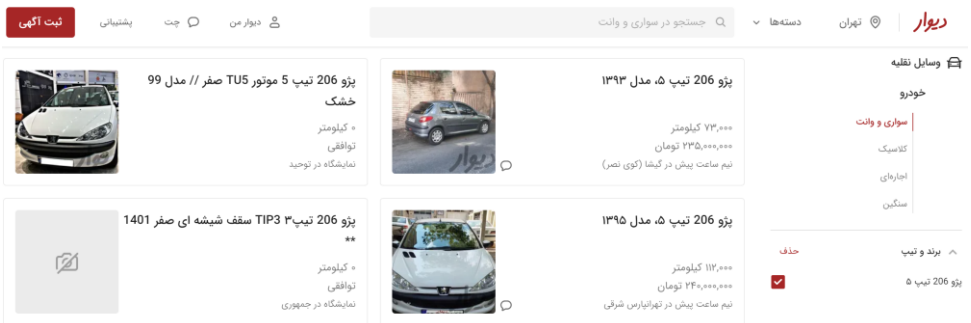
- Duplicates: repeated data entries.
  - It happens when data is coming from different sources or users, for any reason, submit their entry more than once.
- You should usually remove duplicates.



25

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.



26

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.

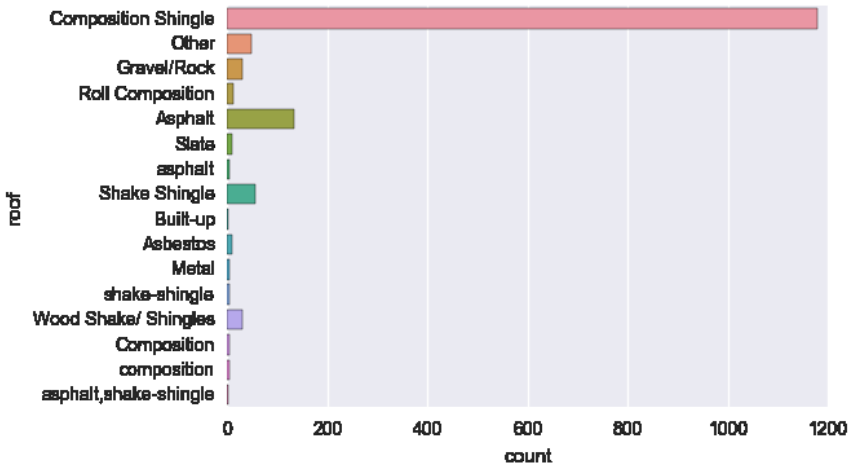
27

Fix Structural Errors

- Structural errors occur during measurement, data transfer, or maintenance activities.
- They include odd naming, typos, inconsistent capitalization.
  - While they may be obvious to humans, most machine learning applications wouldn't recognize the mistakes.
- Structural errors lead to inconsistent data, duplicates, and mislabeled categories.
- Review your data collection and data transformation process to prevent structural errors.

28

Fix Structural Errors



29

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.

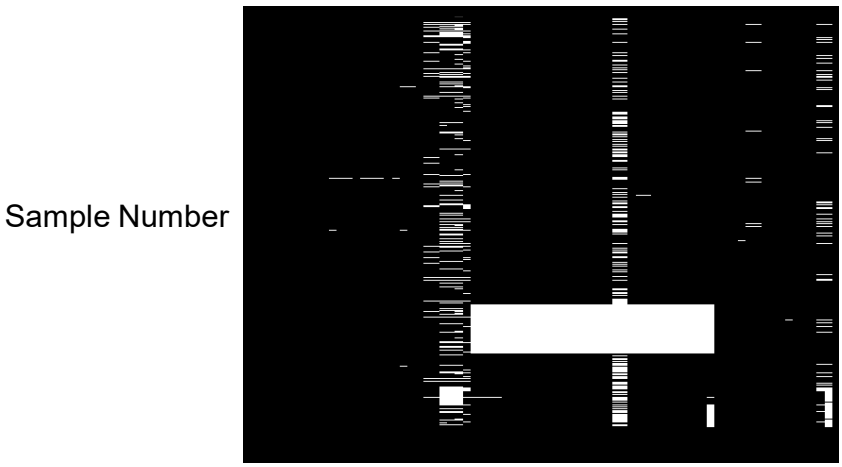
30

Handling Missing Data: Example

ST_NUM	ST_NAME	NUM_BEDROOMS	OWN_OCCUPIED
104	PUTNAM	3	Y
197	LEXINGTON	3	N
	LEXINGTON	n/a	N
201	BERKELEY	1	12
203	BERKELEY	3	Y
207	BERKELEY	NA	Y
NA	WASHINGTON	2	
213	TREMONT	--	Y
215	TREMONT	na	Y

31

Visualizing Missing Values



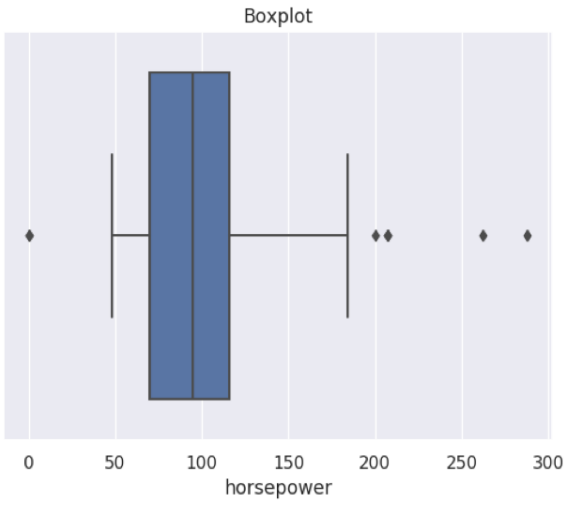
32

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.

33

Outlier Analysis for One Variable



34

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.

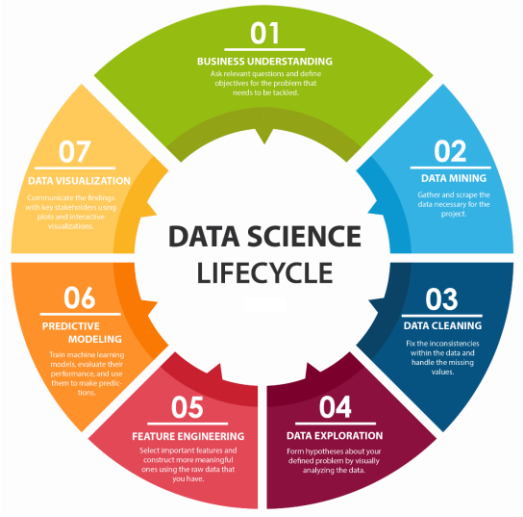
35

Data Cleaning Tips

- Validate your data after cleaning.
  - Confirm that it's high quality, consistent, and properly formatted for downstream processes.
- Create the right process and use it consistently.
- Use data cleaning tools.
- Pay attention to errors and track where dirty data comes from.

36

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.

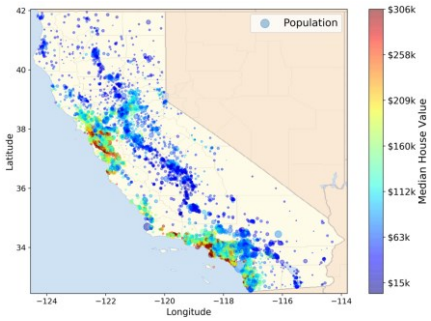
Variable Identification

```
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
```

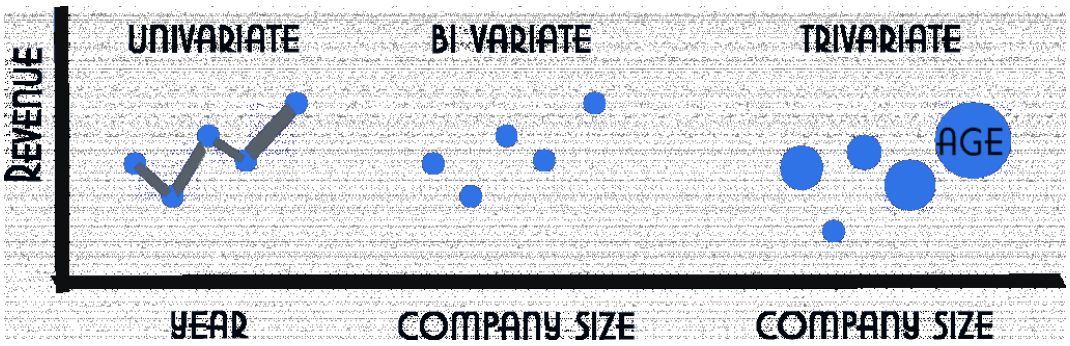
```
housing["ocean_proximity"].value_counts()
```

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



39

Exploratory Data Analysis



40

Anscombe's Quartet

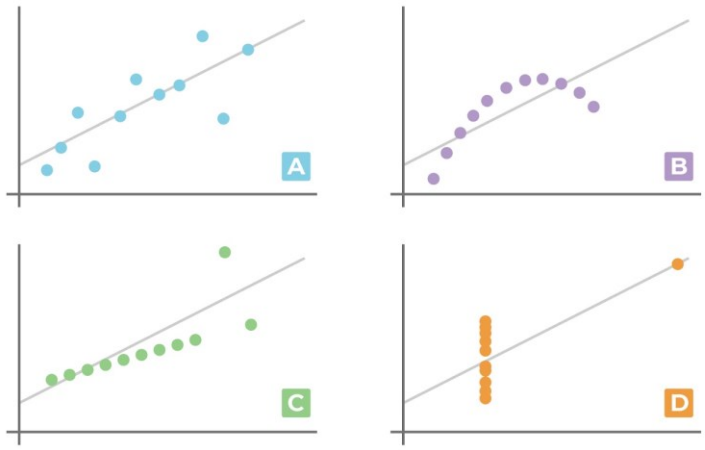
- For all four datasets:

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

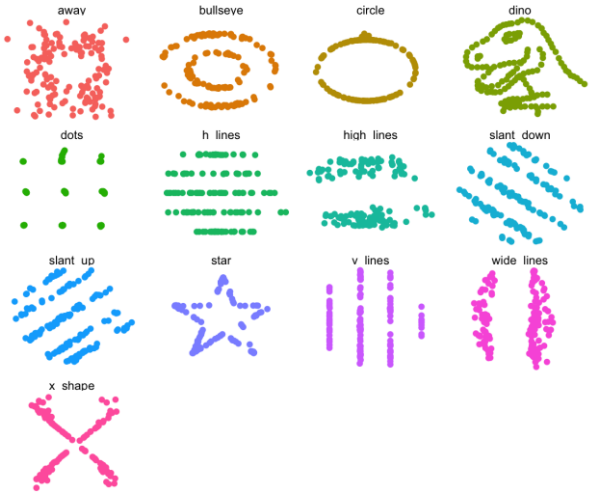
41

Anscombe's Quartet



42

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

43

5. FEATURE ENGINEERING

44

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.

Feature Construction

Converting between Feature Types

- Construct binary feature from numerical feature.
  - Use a single threshold:
  - Example: Age[<18] : 0 ， Age[≥18] : 1
- Construct categorical feature from numerical feature.
  - Use multiple thresholds:
  - Example: Age[<18] : 0 ， Age[from 18 to 35] : 1 ， Age[≥35] : 2
- Construct binary feature from categorical feature.
  - Use one-hot encoding
  - Example: { short or medium or long } → [1, 0, 0] or [0, 1, 0] or [0, 0, 1]

Feature Transformation

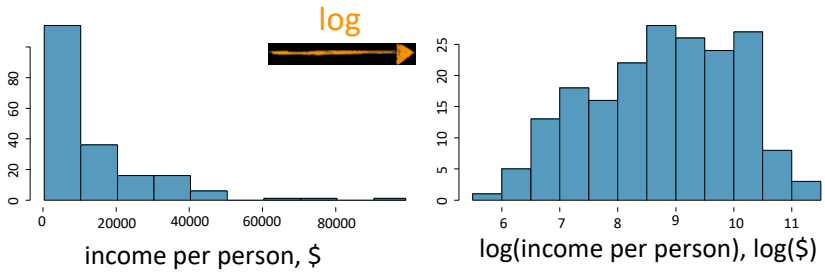
- A transformation is a rescaling of the data using a function.
  - Log transformation
  - Square root transformation
  - Inverse transformation

- Goals of transformation:
  - To see the data structure differently
  - To reduce skew and assist in modeling
  - To straighten a nonlinear relationship in a scatterplot
  - To model the relationship with simpler method

49

Log Transformation

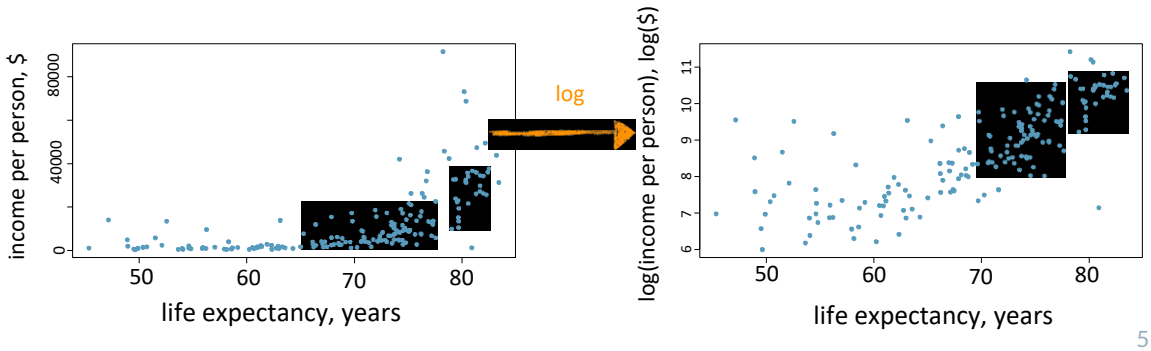
- Often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive.



50

Log Transformation

- To make the relationship between the variables more linear, and hence easier to model with simple methods



51

Feature Scaling

- Feature scaling is done owing to the sensitivity of some machine learning algorithms to the scale of the input values.
  - Min-Max Scaling: This process involves the rescaling of all values in a feature in the range 0 to 1:

$$x_i \rightarrow \frac{x_i - x_{Min}}{x_{Max} - x_{Min}}$$

  - Variance Scaling: All the data points are subtracted by their mean and the result divided by the distribution's standard variation to arrive at a distribution with a 0 mean and variance of 1:

$$x_i \rightarrow \frac{x_i - \mu_x}{\sigma_x}$$

52

Feature Combinations

- Feature combination involves deriving new features from existing ones.
- This can be done by simple mathematical operations such as aggregations to obtain the mean, median, mode, sum, or difference and even product of two values.
- These features, although derived directly from the given data, when carefully chosen to relate to the target can have an impact on the performance.

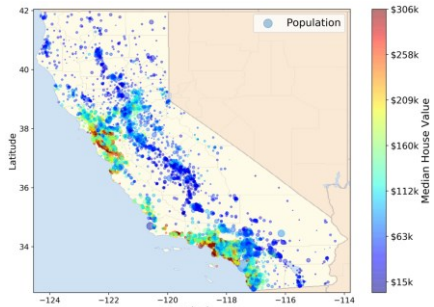
Housing Dataset

```
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
```



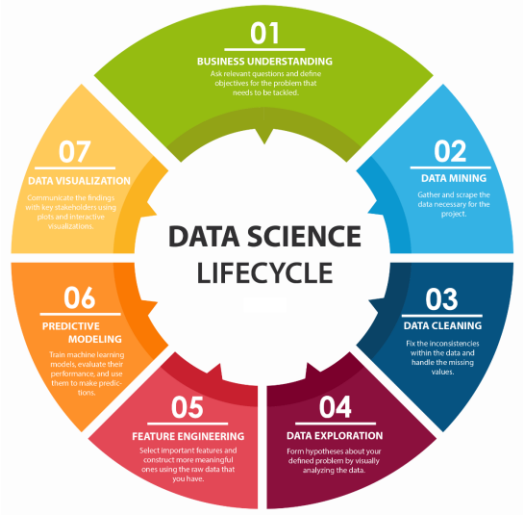
Feature Combinations

- 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"]
```

55

6. PREDICTIVE MODELING



56

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.

Data Visualization in Machine Learning

- Identify trends and patterns in data
- Communicate insights to stakeholders
- Monitor machine learning models
- Improve data quality

7. DATA VISUALIZATION

59

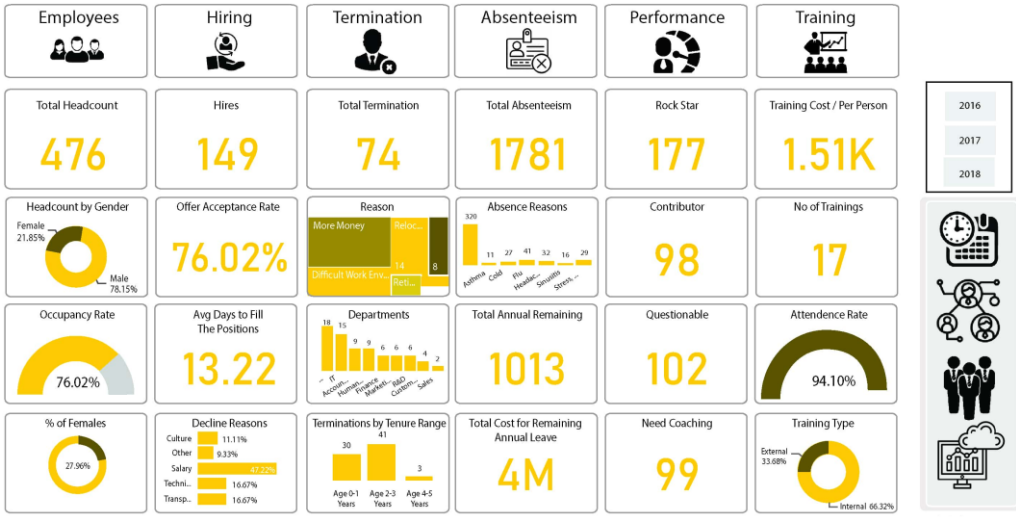
Data Visualization

- 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

60

Example: HR Analytics Dashboard



Example: Bad Dashboard Design