



Group of
Horribly
Optimistic
STatisticians

Data Manipulation pt.1

Intro to Data Science

Maksymilian Norkiewicz & Jędrzej Ogrodowski



File Formats



.CSV



```
1  "name","surname","email","phone"  
2  "Herminia","Marshall","herminia.marshall@example.com","678-313-8625"  
3  "Bernice","Richardson","bernice.richardson@example.com","406-640-0952"  
4  "Maeleachlainn","Albertson","maeleachlainn.albertson@example.com","936-514-5533"  
5  "Laloecen","Darwin","laloecen.darwin@example.com","772-216-4633"  
6  "Shib","Payton","shib.payton@example.com","817-657-1845"  
7  "Marcos","Bertolini","marcos.bertolini@example.com","209-977-7112"  
8  "Daniël","Teague","daniël.teague@example.com","773-750-0852"  
9  "Ikaros","Garber","ikaros.garber@example.com","615-289-1387"  
10 "Regulus","Cornell","regulus.cornell@example.com","832-572-5442"  
11 "Lars","Simen","lars.simen@example.com","843-441-7001"
```



.tsv



```
1  "name"  "surname"  "email"  "phone"
2  "Herminia" "Marshall" "herminia.marshall@example.com" "678-313-8625"
3  "Bernice"  "Richardson"  "bernice.richardson@example.com"  "406-640-0952"
4  "Maeleachlainn" "Albertson" "maeleachlainn.albertson@example.com" "936-514-5533"
5  "Laloecen"  "Darwin"  "laloecen.darwin@example.com"  "772-216-4633"
6  "Shib"  "Payton"  "shib.payton@example.com"  "817-657-1845"
7  "Marcos"  "Bertolini" "marcos.bertolini@example.com" "209-977-7112"
8  "Daniël"  "Teague"  "daniël.teague@example.com" "773-750-0852"
9  "Ikaros"  "Garber"  "ikaros.garber@example.com" "615-289-1387"
10 "Regulus"  "Cornell"  "regulus.cornell@example.com" "832-572-5442"
11 "Lars"  "Simen"  "lars.simen@example.com"  "843-441-7001"
```



.json



```
1 {  
2   "name": [  
3     "Herminia",  
4     "Bernice",  
5     "Maeleachlainn",  
6     "Laloeocen",  
7     "Shib",  
8     "Marcos",  
9     "Dani\u00e9l",  
10    "Ikaros",  
11    "Regulus",  
12    "Lars"  
13  ],  
14  "surname": [  
15    "Marshall",  
16    "Richardson",  
17    "Albertson",  
18    "Darwin",  
19    "Payton"
```



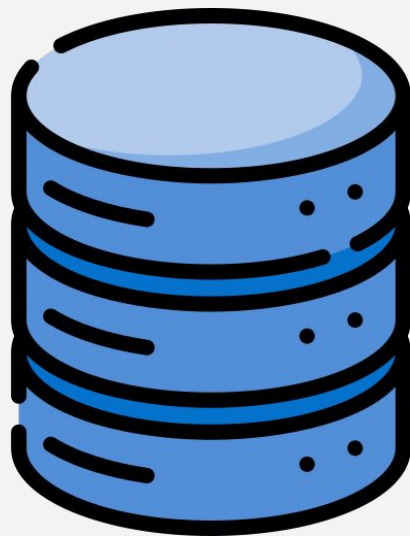
.xlsx



	A	B	C	D	
1	name	surname	email	phone	
2	Herminia	Marshall	herminia.marshall@example.com	678-313-8625	
3	Bernice	Richardson	bernice.richardson@example.com	406-640-0952	
4	Maeleachlainn	Albertson	maeleachlainn.albertson@example.com	936-514-5533	
5	Laloeocen	Darwin	laloeocen.darwin@example.com	772-216-4633	
6	Shib	Payton	shib.payton@example.com	817-657-1845	
7	Marcos	Bertolini	marcos.bertolini@example.com	209-977-7112	
8	Daniël	Teague	daniel.teague@example.com	773-750-0852	
9	Ikaros	Garber	ikaros.garber@example.com	615-289-1387	
10	Regulus	Cornell	regulus.cornell@example.com	832-572-5442	
11	Lars	Simen	lars.simen@example.com	843-441-7001	
12					



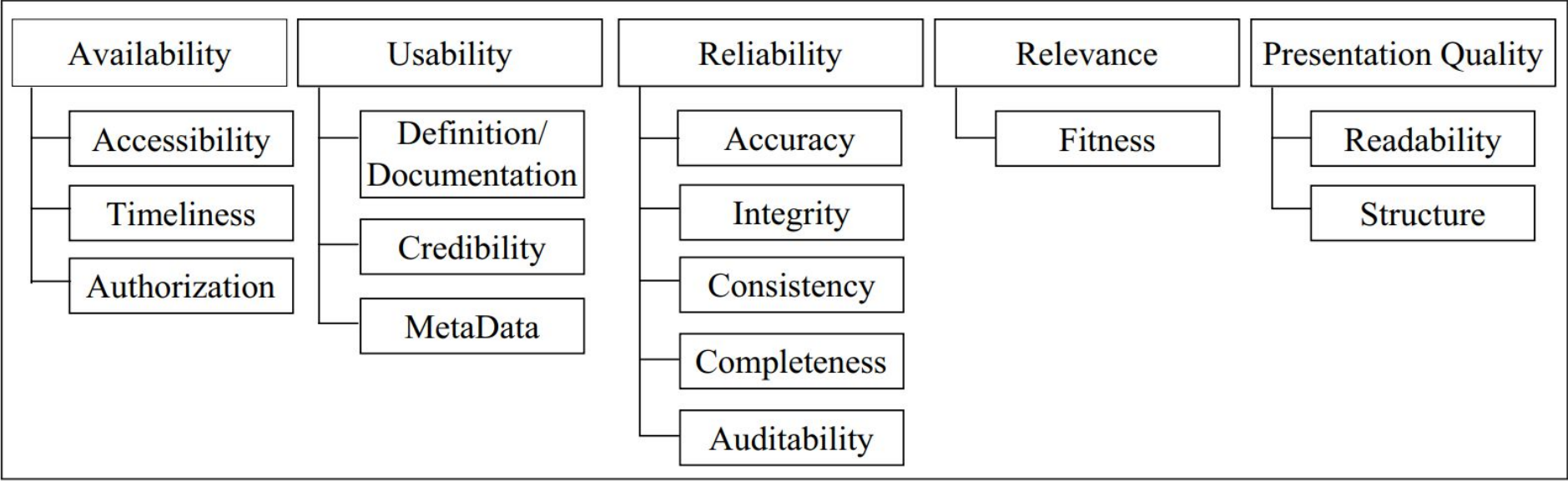
SQL



```
conn = sql.connect("path/to/file/users.db")
data_sql = pd.read_sql("SELECT * FROM users;", conn)
conn.close()
```



How to assess data quality










Availability







Accessibility

NY-House-Dataset.csv (1.33 MB)   

Detail Compact Column 10 of 17 columns 


About this file  **Add Suggestion**

Explore Trends, Prices, and Investment Opportunities with the NY Real Estate Insights Dataset

 BROKERTITLE	 TYPE	 PRICE
Title of the broker	Type of the house	Price of the house
Brokered by COM... 9%	Co-op for sale 30%	 2494
Brokered by Dougl... 2%	House for sale 21%	
Other (4235) 88%	Other (2339) 49%	
Brokered by Douglas Elliman -111 Fifth Ave	Condo for sale	315000


112 E 74th St
425 - 875 SF Medical Condo Units Offered at \$250,000 - \$595,000 Per Unit in New York, NY 10021

Apartment Buildings / New York / New York / 112 E 74th St, New York, NY 10021



PROPERTY FACTS

Price	\$250,000 - \$595,000	Property Subtype	Apartment
Unit Size	425 - 875 SF	Building Class	B
No. Units	2	Floors	9
Total Building Size	48,337 SF	Typical Floor Size	5,371 SF
Property Type	Multifamily	Year Built	1917





Timeliness





Timeliness





Timeliness





Authorization

otod^{em} Analytics



Authorization

otodom Analytics

1899 zł /30 dni

dostępny dla deweloperów
współpracujących z obido



Authorization

2. Postanowienia Ogólne

3. Treści publikowane w Serwisie, w tym w szczególności Ogłoszenia, niezależnie od ich formy, tj. materiały tekstowe, graficzne oraz wideo, są przedmiotem ochrony praw własności intelektualnej, w tym prawa autorskiego oraz praw własności przemysłowej, Grupy OLX, Użytkowników lub osób trzecich. Zabrania się w szczególności:
- a. jakiegokolwiek wykorzystywania tych treści bez pisemnej zgody uprawnionych;
 - b. jakiegokolwiek agregowania i przetwarzania danych oraz innych informacji dostępnych w Serwisie w celu ich dalszego udostępniania osobom trzecim w ramach innych serwisów internetowych jak i poza Internetem;**
 - c. wykorzystywania oznaczeń Serwisu oraz Grupy OLX, w tym charakterystycznych elementów grafiki bez zgody Grupy OLX.



Usability



Definition/ Documentation

About this Dataset

Updated

January 19, 2023

Data Last

Updated

December 5,
2022

Metadata Last

Updated

January 19,
2023

Date Created

April 28, 2020

Views

14.7K

Downloads

2,358

Data Collection

Data Collection

COVID-19 Health Data

Dataset Information

Agency

Department of Health and Mental Hygiene
(DOHMH)

Update

Update Frequency

Daily

Automation

Yes







Date Made Public

5/19/2020



Definition/ Documentation

Columns (6)

Column Name	Description	API Field Name	Data Type
 extract_date	Date of data extraction	extract_date	Floating Timestamp
 date	Date of emergency department visit	date	Floating Timestamp
 mod_zcta	Modified ZIP Code tabulation area (ZCTA) of patient residence	mod_zcta	Text
 total_ed_visits	Count of all emergency department visits	total_ed_visits	Number
 ili_pne_visits	Count of influenza-like illness and/or pneumonia visits	ili_pne_visits	Number
 ili_pne_admissions	Count of influenza-like illness and/or pneumonia visits admitted to the hospital	ili_pne_admissions	Number



Credibility



**CENTRALNA
KOMISJA
EGZAMINACYJNA**

Wyniki egzaminu maturalnego w 2024 roku

- [Wstępne informacje o wynikach egzaminu maturalnego 2024](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 CENTYLE](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 STANINY POL](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 STANINY ENG](#)
- [Wyniki egzaminu maturalnego 2024 PREZENTACJA](#)
- [Mapki z wynikami egzaminu maturalnego](#)



Credibility



Wyniki egzaminu maturalnego w 2024 roku

- [Wstępne informacje o wynikach egzaminu maturalnego 2024](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 CENTYLE](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 STANINY POL](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 STANINY ENG](#)
- [Wyniki egzaminu maturalnego 2024 PREZENTACJA](#)
- [Mapki z wynikami egzaminu maturalnego](#)



MAKSYMILIAN NORKIEWICZ ·
UPDATED 2 MONTHS AGO



14



Download (7 MB)



Matura exam results (2021-2024)

Result of matura exam in Poland from 2021 to 2024

Data Card

Code (1)

Discussion (0)

Suggestions (0)

About Dataset

The dataset contains the results of the secondary school leaving examination (matura) divided into schools. The data was downloaded from mapa.wyniki.edu.pl. The dataset contains as many as 420 columns, but this is because most of them are results from individual exams. Originally dataset was in polish, so matura_eng.csv file

Usability ⓘ

7.06

License

Unknown

Expected update frequency

Not specified



Credibility



Wyniki egzaminu maturalnego w 2024 roku

- [Wstępne informacje o wynikach egzaminu maturalnego 2024](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 CENTYLE](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 STANINY POL](#)
- [Wstępne informacje o wynikach egzaminu maturalnego 2024 STANINY ENG](#)
- [Wyniki egzaminu maturalnego 2024 PREZENTACJA](#)
- [Mapki z wynikami egzaminu maturalnego](#)



MAKSYMILIAN NORKIEWICZ ·
UPDATED 2 MONTHS AGO



14



Download (7 MB)



Matura exam results (2021-2024)

Result of matura exam in Poland from 2021 to 2024

Data Card

Code (1)

Discussion (0)

Suggestions (0)

About Dataset

The dataset contains the results of the secondary school leaving examination (matura) divided into schools. The data was downloaded from mapa.wyniki.edu.pl. The dataset contains as many as 420 columns, but this is because most of them are results from individual exams. Originally dataset was in polish, so matura_eng.csv file

Usability ⓘ

7.06

License

Unknown

Expected update frequency

Not specified



Reliability



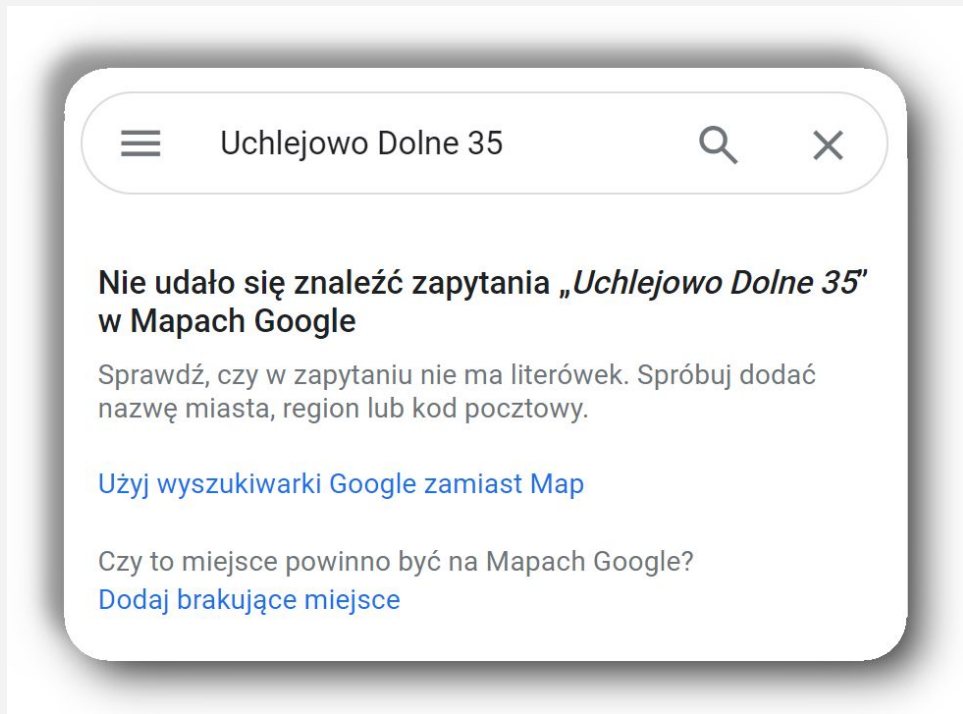
Accuracy

	name	surname	address
0	Kevin	McCallistera	Uchlejowo Dolne 35



Accuracy

	name	surname	address
0	Kevin	McCallistera	Uchlejowo Dolne 35





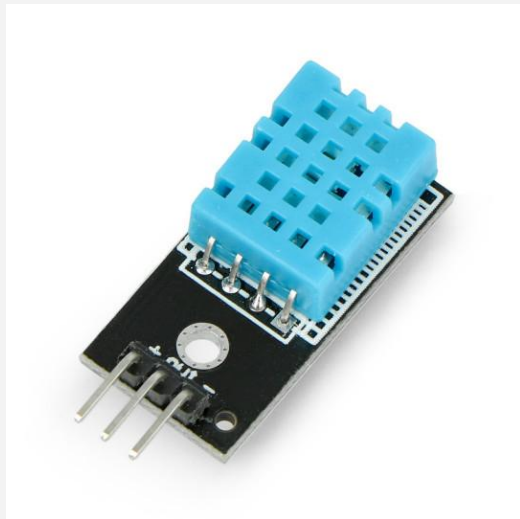
Accuracy

	name	surname	address
0	Kevin	McCallistera	Chicago, IL

	name	surname	address
0	Kevin	McCallistera	671 Lincoln Avenue, Chicago, IL 60614



Accuracy



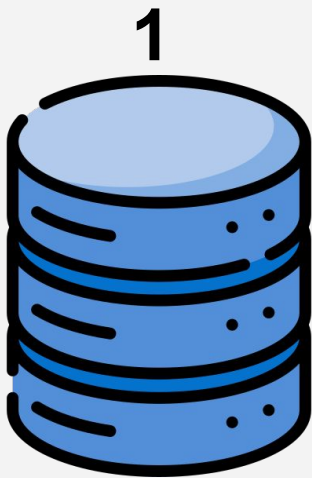
± 1°C



± 0.5°C



Consistency

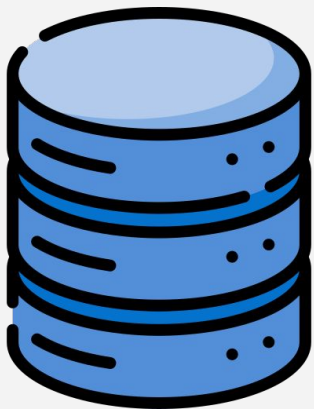


	name	surname	address
0	Kevin	McCallistera	671 Lincoln Avenue, Chicago



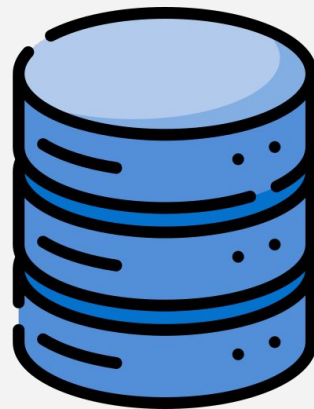
Consistency

1



	name	surname	address
0	Kevin	McCallistera	671 Lincoln Avenue, Chicago

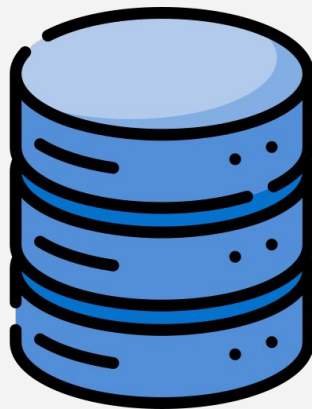
2



	name	surname	address
0	Kevin	McCallistera	352 Lincoln Avenue, Chicago



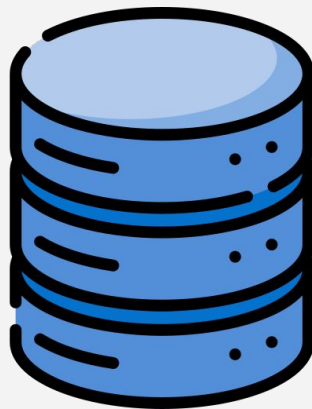
Consistency



	name	surname	age
0	Kevin	McCallistera	13



Consistency



	name	surname	age	marital_status
0	Kevin	McCallistera	13	divorced



Completeness

	price (PLN)	country	vintage	volume (liters)	kind	medals	wegan	natural	punctuation
0	49.0	Argentyna	2021.0	0.75	NaN	NaN	False	False	NaN
1	75.0	Węgry	2019.0	0.75	NaN	NaN	False	False	NaN
2	99.0	Hiszpania	2018.0	0.75	NaN	NaN	False	True	NaN
3	1163.0	Francja	2019.0	0.75	NaN	NaN	False	False	NaN
4	128.0	USA	2018.0	0.75	NaN	NaN	True	False	NaN



Presentation Quality



Readability

	name	weight	horsepower	model_year	col5
0	chevrolet chevelle malibu	3504	130.0	70	18.0
1	buick skylark 320	3693	165.0	70	15.0
2	plymouth satellite	3436	150.0	70	18.0
3	amc rebel sst	3433	150.0	70	16.0
4	ford torino	3449	140.0	70	17.0



Readability

	name	weight	horsepower	model_year	fuel_consumption
0	chevrolet chevelle malibu	3504	130.0	70	18.0
1	buick skylark 320	3693	165.0	70	15.0
2	plymouth satellite	3436	150.0	70	18.0
3	amc rebel sst	3433	150.0	70	16.0
4	ford torino	3449	140.0	70	17.0



Readability

l/100km (liters per 100km)

MPG (miles per gallon)



	name	weight	horsepower	model_year	fuel_consumption
0	chevrolet chevelle malibu	3504	130.0	70	18.0
1	buick skylark 320	3693	165.0	70	15.0
2	plymouth satellite	3436	150.0	70	18.0
3	amc rebel sst	3433	150.0	70	16.0
4	ford torino	3449	140.0	70	17.0



Structure

.csv, .tsv

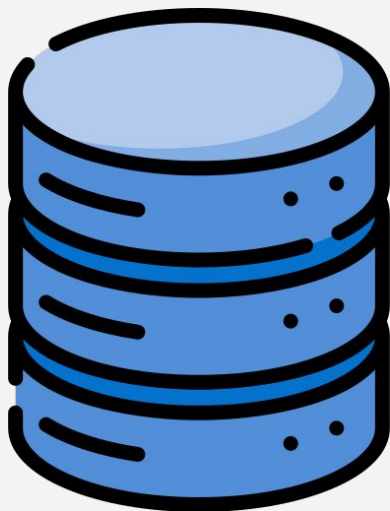
```
1 "name","surname","email"  
2 "Herminia","Marshall","herminia.marshall@example.com"  
3 "Bernice","Richardson","bernice.richardson@example.com"  
4 "Maeleachlainn","Albertson","maeleachlainn.albertson@example.com"  
5 "Laloecen","Darwin","laloecen.darwin@example.com"  
6 "Shib","Payton","shib.payton@example.com"  
7 "Marcos","Bertolini","marcos.bertolini@example.com"  
8 "Daniël","Teague","daniël.teague@example.com"  
9 "Ikaros","Garber","ikaros.garber@example.com"  
10 "Regulus","Cornell","regulus.cornell@example.com"  
11 "Lars","Simen","lars.simen@example.com"
```





Structure

SQL





Structure

.json

```
{  
  "name": {  
    "0": "Herminia",  
    "1": "Bernice",  
    "2": "Maeleachlainn",  
    "3": "Laloecen",  
    "4": "Shib",  
    "5": "Marcos",  
    "6": "Dani\u00e0",  
    "7": "Ikaros",  
    "8": "Regulus",  
    "9": "Lars"  
  },  
  "surname": {  
    "0": "Marshall",  
    "1": "Richardson",  
    "2": "Albertson",  
    "3": "Darwin",  
    "4": "Payton",  
    "5": "Bertolini",  
    "6": "Teague",  
  }  
}
```





Structure

Excel

The screenshot shows the Microsoft Excel interface. The ribbon is set to 'Narzędzi' (Tools) with the 'Style' group selected. Below the ribbon, the formula bar shows 'name' in cell A1. The main area displays a table with 11 rows of data. Each row has a green checkmark in the 'use_num' column, indicating data validation. The table columns are: name, surname, email, phone, use_num, street, city, state, and zipcode.

	A	B	C	D	E	F	G	H	I
1	name	surname	email	phone	use_num	street	city	state	zipcode
2	Hermينيا	Marshall	herminia.n	678-313-8	4893	Hanifan La	Atlanta	GA	30303
3	Bernice	Richardson	bernice.ric	406-640-0	4508	Coolidge S	Beach	MT	58621
4	Maeleachl	Albertson	maeleachl	936-514-5	1345	Norma Avi	Prairie Vie	TX	77445
5	Laloecen	Darwin	laloecen.d	772-216-4	470	Travis Stre	Fort Pierce	FL	34950
6	Shib	Payton	shib.payto	817-657-1	3075	Southside	Los Angele	CA	90017
7	Marcos	Bertolini	marcos.be	209-977-7	4090	New York	Fort Wort	TX	76147
8	Daniël	Teague	daniël.tea	773-750-0	1608	Center Ave	Fresno	CA	93704
9	Ikaros	Garber	ikaros.gar	615-289-1	788	Wilkinson	Nashville	TN	37211
10	Regulus	Cornell	regulus.co	832-572-5	3050	Chapel Str	Baytown	TX	77520
11	Lars	Simen	lars.simen	843-441-7	473	Jerry Dove	Beaufort	SC	29902
12									
13									





Structure

index	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6
1	1.474	-0.461	1.106	-0.559	-2.078	0.106
2	1.752	-0.688	0.21	-0.695	0.863	0.043
3	0.727	0.200	0.779	1.177	-1.702	-0.894
4	1.752	-0.688	0.21	-0.695	0.863	0.043
5	1.474	-0.461	1.106	-0.559	-2.078	0.106
6	1.752	-0.688	0.21	-0.695	0.863	0.043
7	0.727	0.200	0.779	1.177	-1.702	-0.894
8	1.474	-0.461	1.106	-0.559	-2.078	0.106
9	1.752	-0.688	0.21	-0.695	0.863	0.043
10	0.727	0.200	0.779	1.177	-1.702	-0.894
11	1.474	-0.461	1.106	-0.559	-2.078	0.106
12	1.752	-0.688	0.21	-0.695	0.863	0.043
13	0.727	0.200	0.779	1.177	-1.702	-0.894
14	0.727	0.200	0.779	1.177	-1.702	-0.894
15	1.474	-0.461	1.106	-0.559	-2.078	0.106
16	1.752	-0.688	0.21	-0.695	0.863	0.043

.pdf





Structure

index	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6
1	1.474	-0.461	1.106	-0.559	-2.078	0.106
2	1.752	-0.688	0.21	-0.695	0.863	0.043
3	0.727	0.200	0.779	1.177	-1.702	-0.894
4	1.752	-0.688	0.21	-0.695	0.863	0.043
5	1.474	-0.461	1.106	-0.559	-2.078	0.106
6	1.752	-0.688	0.21	-0.695	0.863	0.043
7	0.727	0.200	0.779	1.177	-1.702	-0.894
8	1.474	-0.461	1.106	-0.559	-2.078	0.106
9	1.752	-0.688	0.21	-0.695	0.863	0.043
10	0.727	0.200	0.779	1.177	-1.702	-0.894
11	1.474	-0.461	1.106	-0.559	-2.078	0.106
12	1.752	-0.688	0.21	-0.695	0.863	0.043
13	0.727	0.200	0.779	1.177	-1.702	-0.894
14	0.727	0.200	0.779	1.177	-1.702	-0.894
15	1.474	-0.461	1.106	-0.559	-2.078	0.106
16	1.752	-0.688	0.21	-0.695	0.863	0.043

.pdf





Structure

index	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6
1	1.474	-0.461	1.106	-0.559	-2.078	0.106
2	1.752	-0.688	0.21	-0.695	0.863	0.043
3	0.727	0.200	0.779	1.177	-1.702	-0.894
4	1.752	-0.688	0.21	-0.695	0.863	0.043
5	1.474	-0.461	1.106	-0.559	-2.078	0.106
6	1.752	-0.688	0.21	-0.695	0.863	0.043
7	0.727	0.200	0.779	1.177	-1.702	-0.894
8	1.474	-0.461	1.106	-0.559	-2.078	0.106
9	1.752	-0.688	0.21	-0.695	0.863	0.043
10	0.727	0.200	0.779	1.177	-1.702	-0.894
11	1.474	-0.461	1.106	-0.559	-2.078	0.106
12	1.752	-0.688	0.21	-0.695	0.863	0.043
13	0.727	0.200	0.779	1.177	-1.702	-0.894
14	0.727	0.200	0.779	1.177	-1.702	-0.894
15	1.474	-0.461	1.106	-0.559	-2.078	0.106
16	1.752	-0.688	0.21	-0.695	0.863	0.043

.pdf





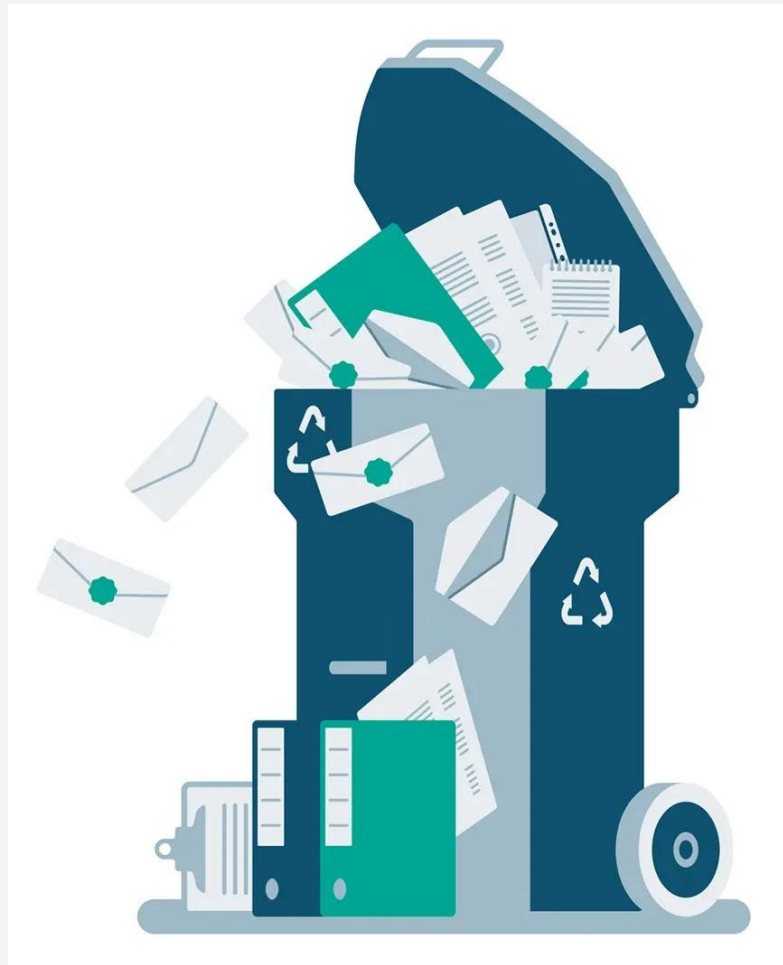
Data cleaning



Why data cleaning is important?

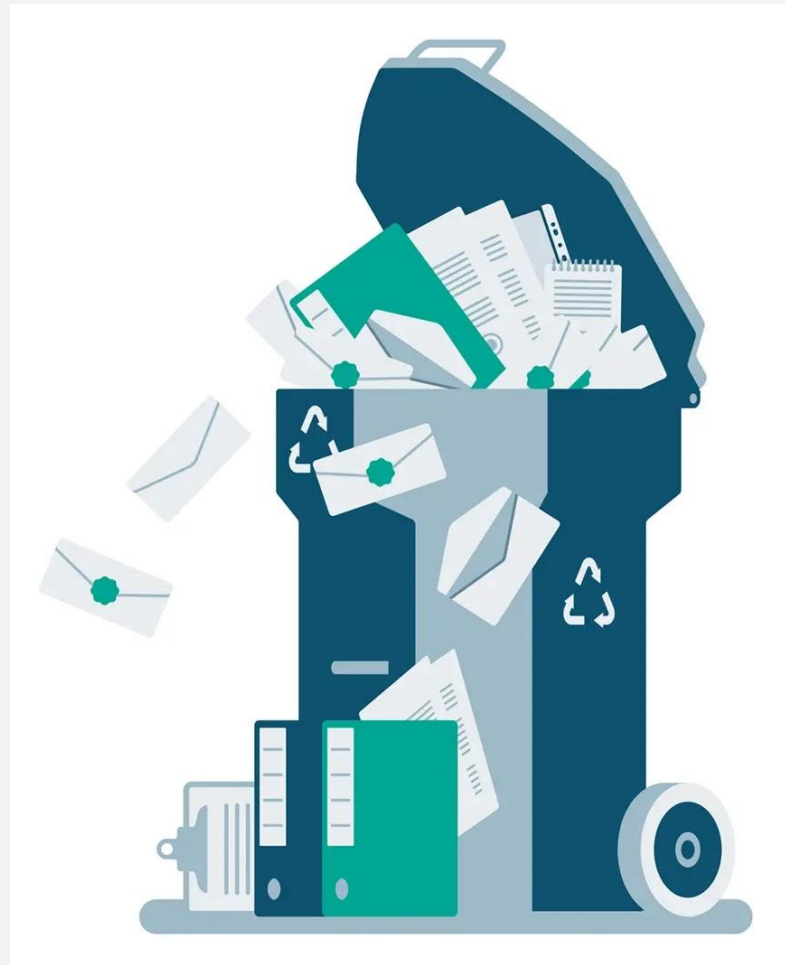


GIGO



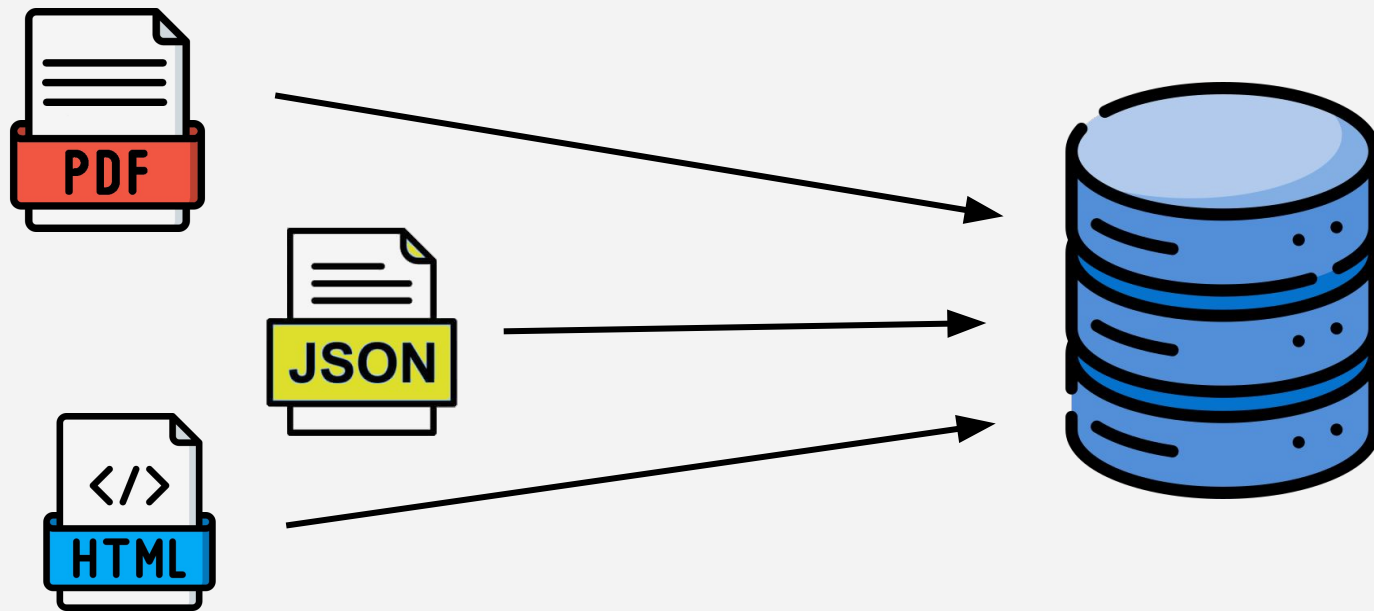


GARBAGE IN, GARBAGE OUT





From unstructured data to structured data





Removing irrelevant data

	name	weight (lbs)	horsepower	model_year	mpg	link
0	toyota corona	2702	96.0	75	24.0	https://pl.wikipedia.org/wiki/Toyota_Corona
1	chevrolet monte carlo s	4082	145.0	73	15.0	https://pl.wikipedia.org/wiki/Chevrolet_Monte_...
2	opel manta	2158	75.0	73	24.0	https://pl.wikipedia.org/wiki/Opel_Manta
3	honda civic 1500 gl	1850	NaN	80	44.6	https://pl.wikipedia.org/wiki/Honda_Civic
4	chevrolet vega	2401	72.0	73	21.0	https://pl.wikipedia.org/wiki/Chevrolet_Vega
5	chevrolet vega	2408	90.0	72	20.0	https://pl.wikipedia.org/wiki/Chevrolet_Vega



Removing irrelevant data

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	75	24.0
1	chevrolet monte carlo s	4082	145.0	73	15.0
2	opel manta	2158	75.0	73	24.0
3	honda civic 1500 gl	1850	NaN	80	44.6
4	chevrolet vega	2401	72.0	73	21.0
5	chevrolet vega	2408	90.0	72	20.0



Removing duplicates

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	75	24.0
1	chevrolet monte carlo s	4082	145.0	73	15.0
2	opel manta	2158	75.0	73	24.0
3	honda civic 1500 gl	1850	NaN	80	44.6
4	chevrolet vega	2401	72.0	73	21.0
5	chevrolet vega	2408	90.0	72	20.0



Removing duplicates

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	75	24.0
1	chevrolet monte carlo s	4082	145.0	73	15.0
2	opel manta	2158	75.0	73	24.0
3	honda civic 1500 gl	1850	NaN	80	44.6
4	chevrolet vega	2401	72.0	73	21.0
5	chevrolet vega	2408	90.0	72	20.0



Removing duplicates

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	75	24.0
1	chevrolet monte carlo s	4082	145.0	73	15.0
2	opel manta	2158	75.0	73	24.0
3	honda civic 1500 gl	1850	NaN	80	44.6
4	chevrolet vega	2401	72.0	73	21.0



Type conversions

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	75	24.0
1	chevrolet monte carlo s	4082	145.0	73	15.0
2	opel manta	2158	75.0	73	24.0
3	honda civic 1500 gl	1850	NaN	80	44.6
4	chevrolet vega	2401	72.0	73	21.0



`numpy.int64`



Type conversions

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	1975-01-01	24.0
1	chevrolet monte carlo s	4082	145.0	1973-01-01	15.0
2	opel manta	2158	75.0	1973-01-01	24.0
3	honda civic 1500 gl	1850	NaN	1980-01-01	44.6
4	chevrolet vega	2401	72.0	1973-01-01	21.0



```
pandas._libs.tslibs.timestamps.Timestamp
```




Units conversion

	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	1975-01-01	24.0
1	chevrolet monte carlo s	4082	145.0	1973-01-01	15.0
2	opel manta	2158	75.0	1973-01-01	24.0
3	honda civic 1500 gl	1850	NaN	1980-01-01	44.6
4	chevrolet vega	2401	72.0	1973-01-01	21.0



Units conversion

pounds



mile per gallon



	name	weight (lbs)	horsepower	model_year	mpg
0	toyota corona	2702	96.0	1975-01-01	24.0
1	chevrolet monte carlo s	4082	145.0	1973-01-01	15.0
2	opel manta	2158	75.0	1973-01-01	24.0
3	honda civic 1500 gl	1850	NaN	1980-01-01	44.6
4	chevrolet vega	2401	72.0	1973-01-01	21.0



Units conversion

kilograms



liters per 100 km



	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Dealing with missing values

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Dealing with missing values

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20





Dealing with missing values

Delete



Dealing with missing values

Delete or Impute



Deleting rows

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Deleting columns

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Imputing value

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Mean

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Mean

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Mean

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	97.0	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Median

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	NaN	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20

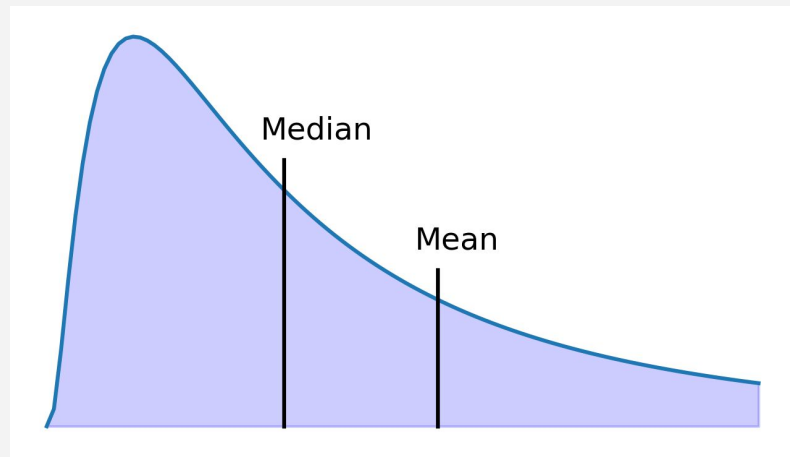
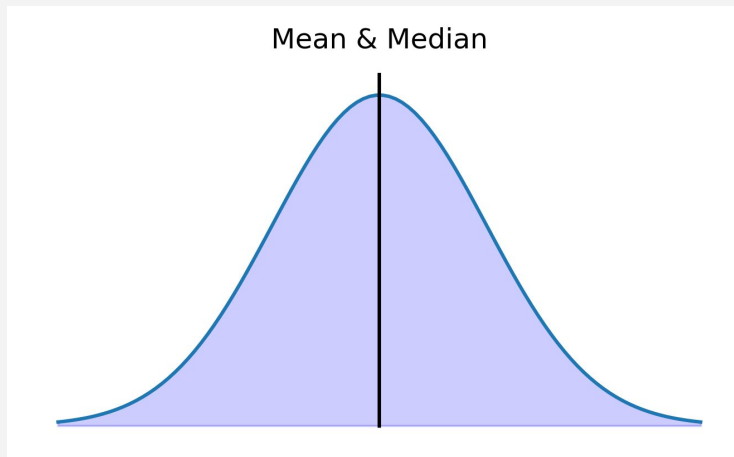


Median

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	85.5	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



Mean vs. Median





Generate random numbers

$\mu = 97$

horsepower
96.0
145.0
75.0
NaN
72.0

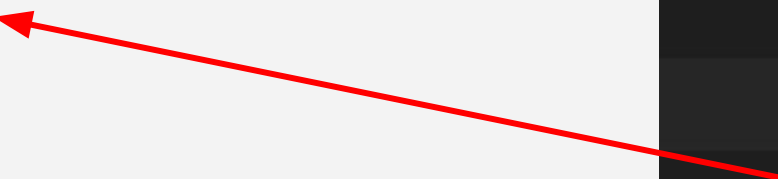


Generate random numbers

$$\mu = 97$$

$$\sigma = 34$$

horsepower	
	96.0
	145.0
	75.0
	NaN
	72.0

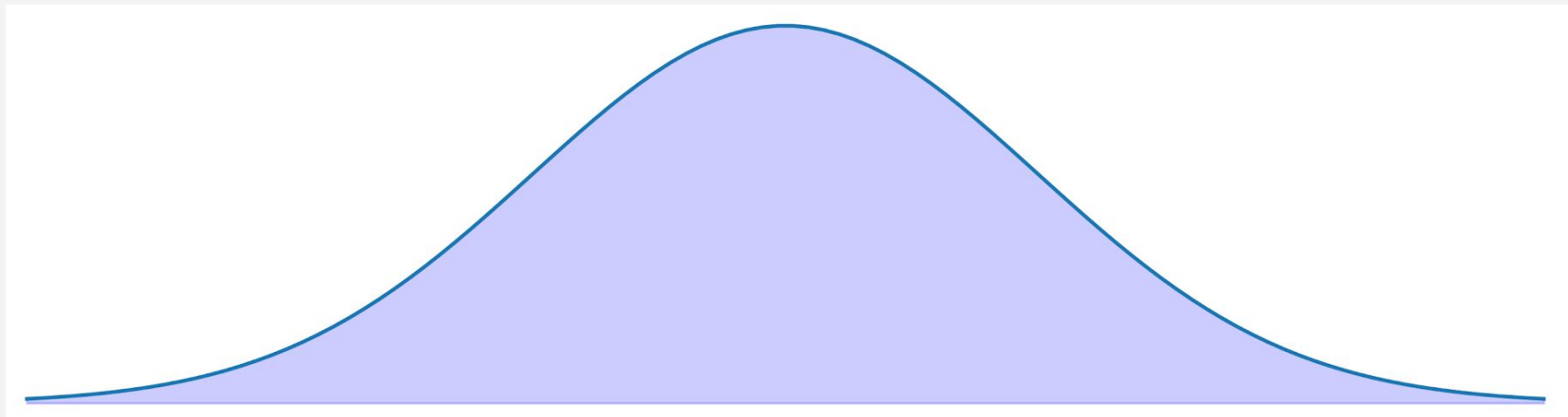




Generate random numbers

$$\mu = 97$$

$$\sigma = 34$$



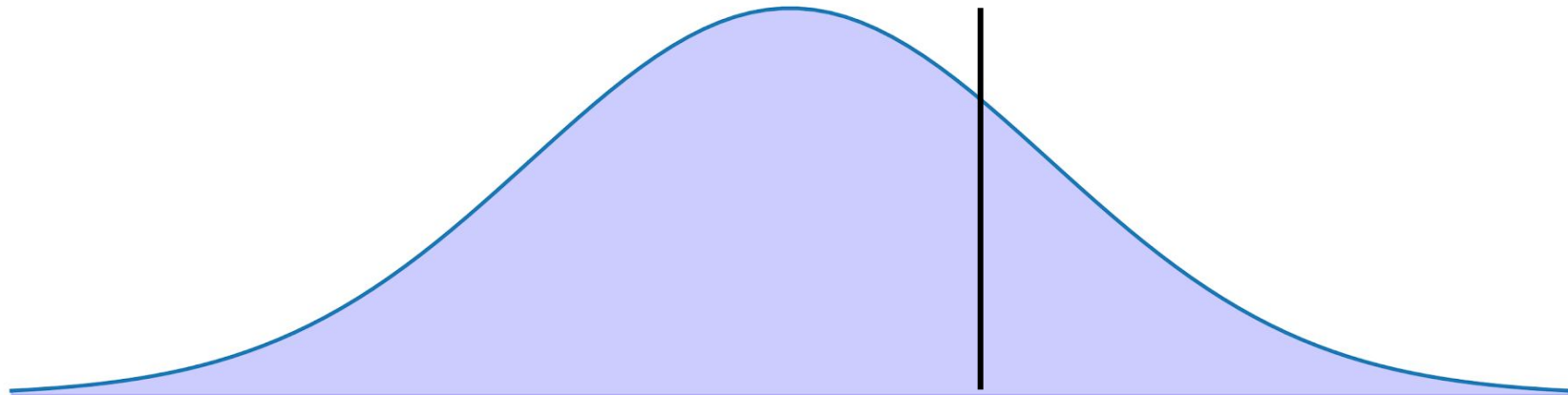


Generate random numbers

$$\mu = 97$$

$$\sigma = 34$$

131





Linear regression

weight (kg)	horsepower	model_year	l/100km
1225.6	96.0	1975-01-01	9.80
1851.6	145.0	1973-01-01	15.68
978.9	75.0	1973-01-01	9.80
839.1	NaN	1980-01-01	5.27
1089.1	72.0	1973-01-01	11.20



Linear regression

training features

weight (kg)	horsepower	model_year	l/100km
1225.6	96.0	1975-01-01	9.80
1851.6	145.0	1973-01-01	15.68
978.9	75.0	1973-01-01	9.80
839.1	NaN	1980-01-01	5.27
1089.1	72.0	1973-01-01	11.20



Linear regression

training features

training targets

weight (kg)	horsepower	model_year	l/100km
1225.6	96.0	1975-01-01	9.80
1851.6	145.0	1973-01-01	15.68
978.9	75.0	1973-01-01	9.80
839.1	NaN	1980-01-01	5.27
1089.1	72.0	1973-01-01	11.20



Linear regression

training features

training targets

features for prediction

weight (kg)	horsepower	model_year	l/100km
1225.6	96.0	1975-01-01	9.80
1851.6	145.0	1973-01-01	15.68
978.9	75.0	1973-01-01	9.80
839.1	NaN	1980-01-01	5.27
1089.1	72.0	1973-01-01	11.20



Linear regression

training features

training targets

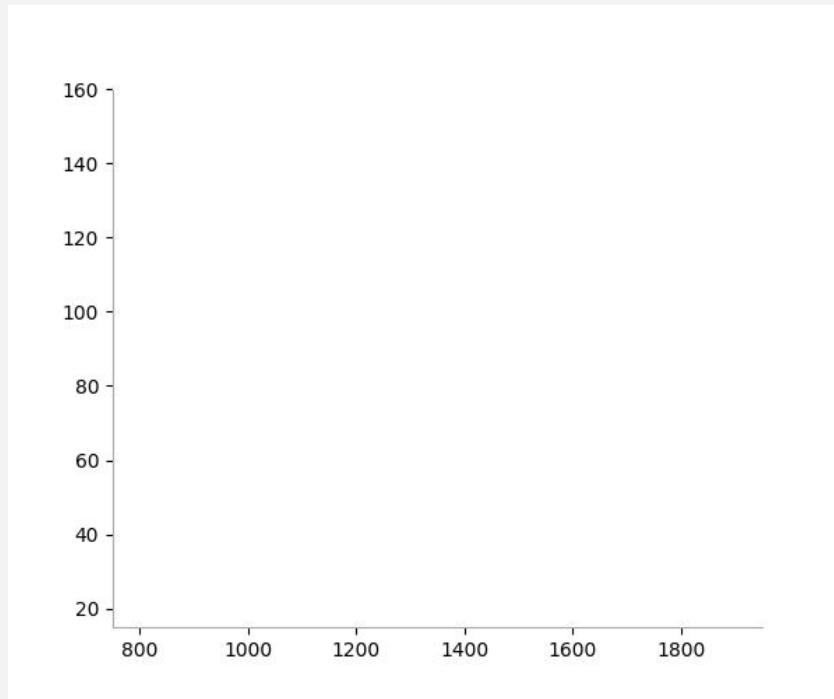
features for prediction

missing values

weight (kg)	horsepower	model_year	l/100km
1225.6	96.0	1975-01-01	9.80
1851.6	145.0	1973-01-01	15.68
978.9	75.0	1973-01-01	9.80
839.1	NaN	1980-01-01	5.27
1089.1	72.0	1973-01-01	11.20

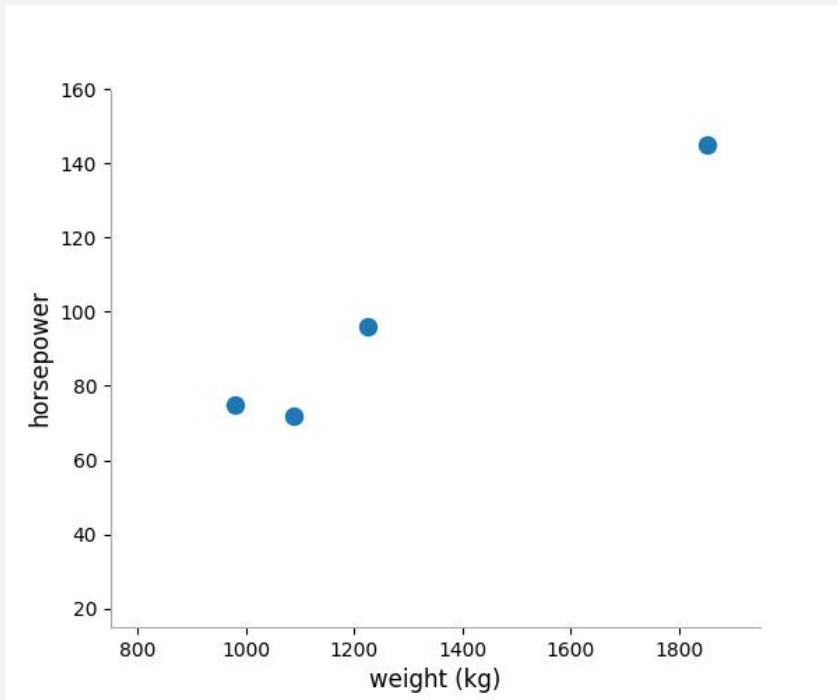


Linear regression



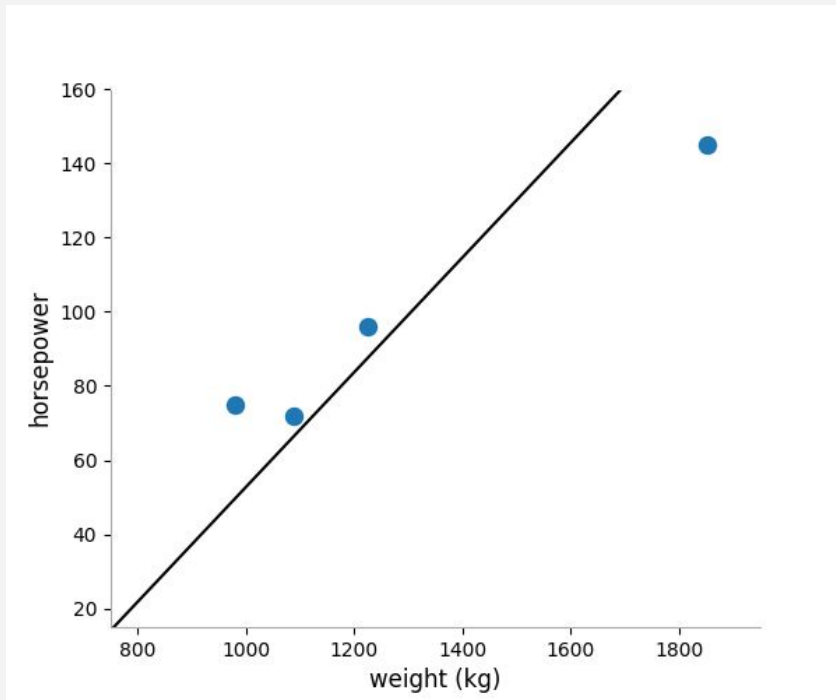


Linear regression



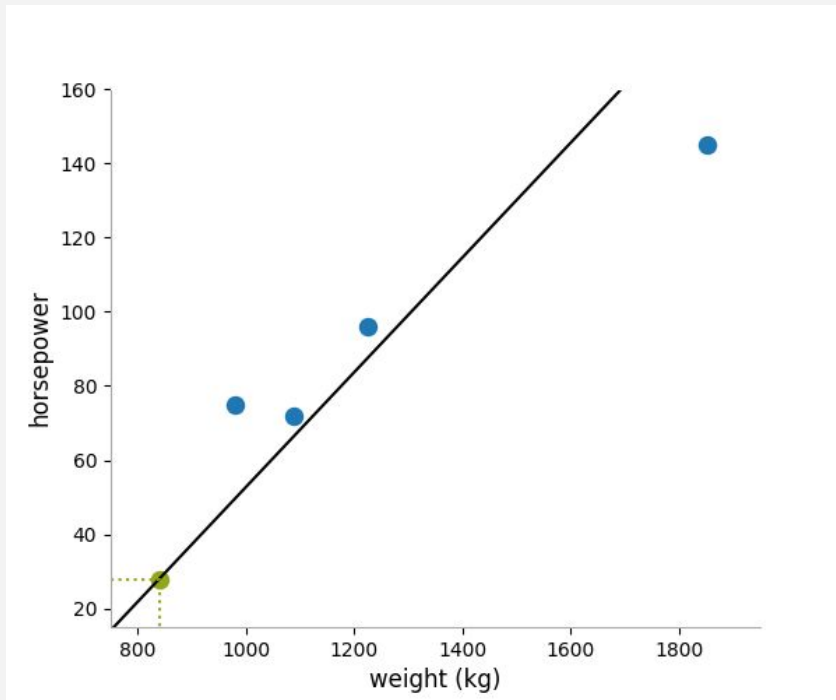


Linear regression





Linear regression





Linear regression

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	27.9	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20

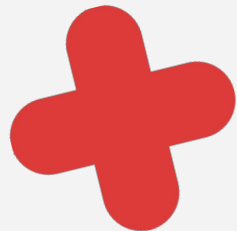


Flags

	name	weight (kg)	horsepower	model_year	l/100km
0	toyota corona	1225.6	96.0	1975-01-01	9.80
1	chevrolet monte carlo s	1851.6	145.0	1973-01-01	15.68
2	opel manta	978.9	75.0	1973-01-01	9.80
3	honda civic 1500 gl	839.1	missing	1980-01-01	5.27
4	chevrolet vega	1089.1	72.0	1973-01-01	11.20



References



- <https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>
 - <https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e>
 - <https://account.datascience.codata.org/index.php/up-j-dsj/article/view/dsj-2015-002>
 - <https://www.sbctc.edu/resources/documents/colleges-staff/commissions-councils/dgc/data-quality-d>
emissions.pdf
 - https://data.cityofnewyork.us/Health/Emergency-Department-Visits-and-Admissions-for-Inf/2nwg-uqyg/about_data
- 