DATA PREPROCESSING

Topics to be Covered

- 1. Categories of Data
- 2. How a Data Scientist Works
- 3. Data Pre-processing

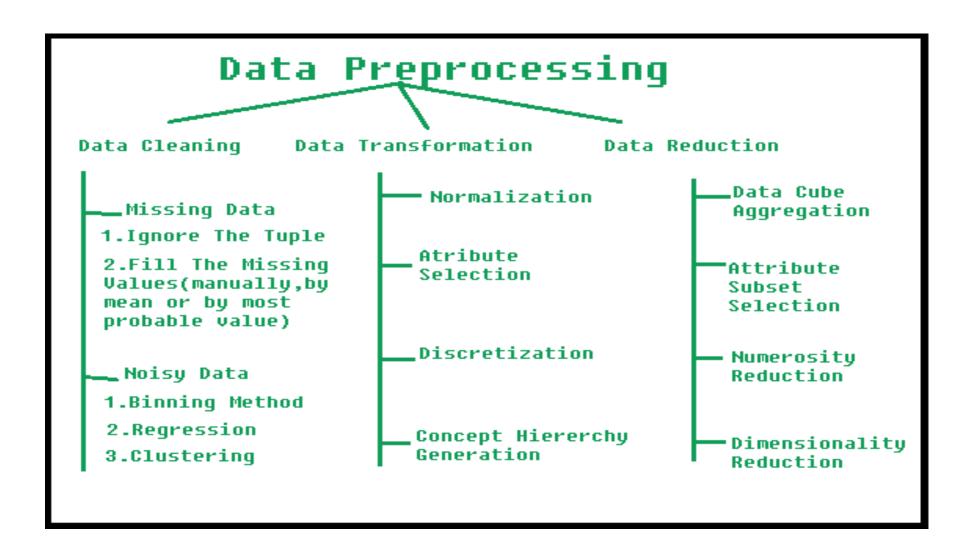
Categories of Data

- 1. Qualitative Data (no measurable meaning to the difference of number)
 - 1. Nominal (eg: vehicle number, computer model number, shirt number, etc)
 - 2. Ordinal (eg: highest degree, position, etc)
- 2. Quantitative Data (meaning to the difference)
 - 1. Interval (eg: year of graduation, year of experience, etc)
 - Ratio (eg: marks, no. Of children, no. Of country visited, no. Of companies worked)

How a Data Scientist Works

- 1. Ask right Question (to understand the business problem)
- 2. Explore and collect data (databases, web logs, customer feedbacks, etc)
- 3. Extract the Data (transform the data to a standard form)
- 4. Clean Data (remove erroneous, inconsitent values from data)
- 5. Replace Missing values (with 0, mean, median, mode, global constant)
- 6. Normalize Data (scale the value in a particular range)
- 7. Analyze data, Find Patterns, make future prediction
- 8. Represent the Result (present the result with useful insight in a way the company can understand)

Data Pre-processing



Data Cleaning

Data cleaning is done to handle missing data, wrong formats, wrong data, duplicate values

	Duration	Date	Pulse	Maxpulse	Calories
(2)	69	"2020/12/01"	110	130	489.1
=1.	68	"2020/12/02"	117	145	479.0
2	68	"2020/12/03"	1.83	135	340.0
3	45	"2020/12/04"	189	175	282.4
4	45	"2020/12/05"	1,17	148	406.0
5	68	"2020/12/06"	102	127	300.0
-6	68	12020/12/07	110	136	374.0
3	450	"2020/12/08"	104	134	253.3
23 4	30	"2020/12/09"	169	133	195.1
9	68	"2020/12/10"	98	124	269.0
1.0	68	"2020/12/11"	103	147	329.3
11.11	68	"2020/12/12"	100	120	250.7
1.2	68	"2020/12/12"	188	120	250.7
1.3	68	"2020/12/13"	186	128	345.3
1.4	68	"2020/12/14"	164	132	379.3
1.5	68	"2020/12/15"	98	123	275.0
11.65	68	"2020/12/16"	98	120	215.2
17	68	"2020/12/17"	100	120	300.0
18	45	"2020/12/18"	98	112	NaN
1.9	68	"2020/12/19"	1.83	123	323.0
28	45	"2020/12/20"	97	125	243.0
21	69	2020/12/21	108	131	364.2
22	45	NaN	1.88	119	282.0
23	69	"2020/12/23"	1.30	101	388.8
24	45	"2020/12/24"	1.05	132	246.9
25	69	"2020/12/25"	182	126	334.5
26	68	2020/12/26	100	120	250.0
27	69	"2020/12/27"	92	118	241.9
28	68	"2020/12/28"	1.93	132	NaN
29	68	"2020/12/29"	1.99	132	289.9
30	68	"2020/12/30"	1.92	129	389.3
31	68	12020/12/31	92	115	243.0

Missing Data

1. Remove rows containing missing data.

```
import pandas as pd

df = pd.read_csv('data.csv')

df.dropna(inplace = True)

print(df.to_string())
```

Rows 18, 22, and 28 are removed

	Duration	Date	Pulse	Maxpulse	Calories
8	68	'2020/12/01'	110	130	409.1
=1.	69	'2020/12/02'	117	145	479.0
2	69	'2020/12/03'	103	135	340.0
3	45	"2020/12/04"	109	175	282.4
4	45	"2020/12/05"	117	148	496.9
5	60	"2020/12/06"	182	127	300.0
6	69	"2020/12/07"	110	136	374.0
7	450	'2020/12/08'	184	134	253.3
23	30	'2020/12/09'	109	133	195.1
9	60	'2020/12/10'	98	124	269.0
1.8	60	"2020/12/11"	183	147	329.3
11	60	"2020/12/12"	188	120	250.7
12	60	"2020/12/12"	188	120	250.7
13	60	'2020/12/13'	106	128	345.3
14	60	"2020/12/14"	184	132	379.3
15	60	'2020/12/15'	98	123	275.0
16	60	"2020/12/16"	98	120	215.2
17	60	"2020/12/17"	188	120	300.0
1.9	60	"2020/12/19"	103	123	323.0
20	45	"2020/12/20"	97	125	243.0
21	60	"2020/12/21"	108	131	364.2
23	60	"2020/12/23"	139	101	300.0
24	45	"2020/12/24"	105	132	246.0
25	69	'2020/12/25'	182	126	334.5
26	69	2020/12/26	188	120	250.0
27	69	'2020/12/27'	92	118	241.0
29	69	'2020/12/29'	100	132	280.0
38	69	'2020/12/30'	102	129	380.3
3:1	68	'2020/12/31'	92	115	243.0

Missing Data

2. Replace missing data with values

```
import pandas as pd

df = pd.read_csv('data.csv')

df.fillna(130, inplace = True)
```

Replace missing data for specified column

```
import pandas as pd

df = pd.read_csv('data.csv')

df["Calories"].fillna(130, inplace = True)
```

```
Duration
                         Date
                                 Pulse
                                         Maxpulse
                                                     Calories
           60
                '2020/12/01'
                                   118
                                               138
                                                        409.1
           60
                '2020/12/02'
                                   117
                                               145
                                                        479.0
                "2020/12/03"
                                   103
                                               135
                                                        340.0
           68
           45
                '2020/12/04'
                                   109
                                               175
                                                        282.4
           45
                "2020/12/05"
                                   117
                                                        496.8
                                               148
           68
                '2020/12/06'
                                   102
                                               127
                                                        300.0
           68
                '2020/12/07'
                                                        374.8
                                   110
                                               136
          450
                '2020/12/08'
                                   184
                                                        253.3
                                               3 4
                '2020/12/09'
           30
                                   189
                                               133
                                                        195.1
           60
                '2020/12/10'
                                    98
                                               124
                                                        269.0
18
           60
                '2020/12/11'
                                   193
                                               147
                                                        329.3
60
                '2020/12/12'
                                   188
                                               1.28
                                                        250.7
12
           68
                "2020/12/12"
                                   100
                                                        259.7
                                               128
13
           68
                '2020/12/13'
                                   1.86
                                               128
                                                        345.3
14
           68
                '2020/12/14'
                                   184
                                               132
                                                        379.3
15
           60
                '2020/12/15'
                                               123
                                    98
                                                        275.0
16
                                                        215.2
           60
                '2020/12/16'
                                    98
                                               128
17
           60
                '2020/12/17'
                                   188
                                               128
                                                        300.0
18
           45
                '2020/12/18'
                                    90 (3)
                                               112
                                                        130.0
19
           60
                "2020/12/19"
                                   103
                                               123
                                                        323.0
28
           45
                "2020/12/20"
                                    97
                                               125
                                                        243.0
21
           60
                 '2020/12/21'
                                   364.2
22
           45
                           130
                                   188
                                               119
                                                        282.0
23
                                                        300.0
           60
                '2020/12/23'
                                   130
                                               24
           45
                <u>'2020/12/24'</u>
                                   185
                                               132
                                                        246.0
25
           60
                '2020/12/25'
                                   1.63.2
                                               126
                                                        334.5
26
           60
                                                        250.0
                  2020/12/26
                                   188
                                               128
27
           60
                '2020/12/27'
                                    92
                                               118
                                                        241.0
28
           60
                '2020/12/28'
                                   103
                                               132
                                                        130.0
29
           60
                '2020/12/29'
                                               1 3 2
                                   100
                                                        280.0
30
           68
                "2020/12/30"
                                   102
                                               129
                                                        380.3
```

Missing Data

3. Replace missing data with Attribute MEAN

```
import pandas as pd

df = pd.read_csv('data.csv')

x = df["Calories"].mean()

df["Calories"].fillna(x, inplace = True)
```

Missing data at column

"Calories", row 18 and 28 are replaced by MEAN, i.e.,

304.68

	Duration	Date	Pulse	Maxpulse	Calories
(3)	69	'2020/12/01'	110	130	409.10
1	60	"2020/12/02"	117	145	479.00
2.	60	'2020/12/03'	103	135	340.00
3	45	'2020/12/04'	109	175	282.40
4	45	'2020/12/05'	117	148	406.00
5	69	'2020/12/06'	102	127	300.00
6	69	'2020/12/07'	110	136	374.00
7	450	"2020/12/08"	184	134	253.30
8	30	"2020/12/09"	109	133	195.10
9	60	'2020/12/10'	98	124	269.00
10	60	'2020/12/11'	103	147	329.30
11	60	'2020/12/12'	100	120	250.70
12	69	'2020/12/12'	100	120	250.70
13	69	'2020/12/13'	106	128	345.30
14	69	'2020/12/14'	104	132	379.30
1.5	69	'2020/12/15'	98	123	275.00
16	69	'2020/12/16'	98	120	215.20
17	69	'2020/12/17'	100	120	300.00
18	45	'2020/12/18'	98	112	304.68
19	68	'2020/12/19'	103	123	323.00
28	45	"2020/12/20"	97	125	243.00
21	68	'2020/12/21'	108	131	364.20
22	45	NaN	100	119	282.00
23	68	"2020/12/23"	130	101	300.00
24	45	"2020/12/24"	105	132	246.00
25	68	"2020/12/25"	182	126	334.50
26	68	2020/12/26	100	120	250.00
27	60	'2020/12/27'	92	118	241.00
28	68	'2020/12/28'	103	132	304.68
29	68	'2020/12/29'	100	132	289.99
38	69	"2020/12/30"	102	129	380.30

Cleaning Data of Wrong Format

At Column "Date" and Row 26, the data format is wrong

```
import pandas as pd

df = pd.read_csv('data.csv')

df['Date'] = pd.to_datetime(df['Date'])

print(df.to_string())
```

	Duration	Date	Pulse	Maxpulse	Calories
(2)	68	'2020/12/01'	110	130	409.1
1	68	'2020/12/02'	117	145	479.0
2.	68	'2020/12/03'	193	135	340.0
3	45	'2020/12/04'	109	175	282.4
4	45	'2020/12/05'	117	148	406.0
5	68	'2020/12/06'	192	127	300.0
6	68	'2020/12/07'	110	136	374.0
7	450	'2020/12/08'	164	134	253.3
8	38	'2020/12/09'	109	133	195.1
9	68	'2020/12/10'	98	124	269.0
10	68	'2020/12/11'	193	147	329.3
1.1	68	'2020/12/12'	100	120	250.7
12	68	'2020/12/12'	100	120	250.7
13	68	'2020/12/13'	196	128	345.3
14	68	'2020/12/14'	164	132	379.3
15	68	'2020/12/15'	98	123	275.0
16	68	'2020/12/16'	98	120	215.2
17	68	'2020/12/17'	100	120	300.0
18	45	'2020/12/18'	98	112	NaN
19	68	'2020/12/19'	193	123	323.0
28	45	'2020/12/20'	97	125	243.0
21	68	'2020/12/21'	108	131	364.2
22	45	NaN	100	119	282.9
23	68	'2020/12/23'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	68	'2 <u>020/</u> 12/25'	102	126	334.5
26	68	2020/12/26	100	120	250.0
27	68	'2020/12/27'	92	118	241.0
28	68	'2020/12/28'	103	132	NaN
29	68	'2020/12/29'	100	132	280.0
30	68	'2020/12/30'	102	129	380.3
31	68	'2020/12/31'	92	115	243.0

Cleaning Wrong Data

At Column "Duration" and Row 7, the data is wrong

1. Replacing Value

2. Replacing with a Boundary Value

```
for x in df.index:
   if df.loc[x, "Duration"] > 120:
     df.loc[x, "Duration"] = 120
```

	Duration	Date	Pulse	Maxpulse	Calories
8	69	'2020/12/01'	110	130	409.1
1	60	'2020/12/02'	117	145	479.0
2	69	'2020/12/03'	103	135	340.0
3	45	'2020/12/04'	109	175	282.4
4	45	'2020/12/05'	117	148	406.0
5	69	'2020/12/06'	102	127	300.0
6	60	'2020/12/07'	110	136	374.0
7	450	'2020/12/08'	104	134	253.3
8	30	'2020/12/09'	109	133	195.1
9	69	'2020/12/10'	98	124	269.0
10	69	'2020/12/11'	103	147	329.3
11	69	'2020/12/12'	100	120	250.7
12	69	'2020/12/12'	100	120	250.7
13	69	'2020/12/13'	106	128	345.3
14	69	'2020/12/14'	104	132	379.3
15	69	'2020/12/15'	98	123	275.0
16	69	'2020/12/16'	98	120	215.2
17	69	'2020/12/17'	100	120	300.0
18	45	'2020/12/18'	90	112	NaN
19	69	'2020/12/19'	103	123	323.0
20	45	'2020/12/20'	97	125	243.0
21	69	'2020/12/21'	108	131	364.2
22	45	NaN	100	119	282.0
23	69	'2020/12/23'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	69	'2020/12/25'	102	126	334.5
26	69	2020/12/26	100	120	250.0
27	60	'2020/12/27'	92	118	241.0
28	69	'2020/12/28'	103	132	NaN
29	69	'2020/12/29'	100	132	280.0
30	69	'2020/12/30'	102	129	380.3
31	69	'2020/12/31'	92	115	243.0

Removing Duplicates

Rows 11 and 12 contain duplicate values

1. Discovering Duplicate Rows

print(df.duplicated())

2. Removing Duplicates

df.drop_duplicates(inplace = True)

	Duration	Date	Pulse	Maxpulse	Calories
0	60	'2020/12/01'	110	130	409.1
1.	60	'2020/12/02'	117	145	479.0
2.	60	'2020/12/03'	103	135	340.0
3	45	'2020/12/04'	109	175	282.4
4	45	'2020/12/05'	117	148	406.0
5	60	'2020/12/06'	102	127	300.0
6	60	'2020/12/07'	110	136	374.0
7	450	'2020/12/08'	104	134	253.3
8	30	'2020/12/09'	109	133	195.1
9	60	'2020/12/10'	98	124	269.0
10	60	'2020/12/11'	103	147	329.3
11	60	'2020/12/12'	100	120	250.7
12	60	'2020/12/12'	100	120	250.7
13	60	'2020/12/13'	106	128	345.3
14	60	'2020/12/14'	104	132	379.3
15	60	'2020/12/15'	98	123	275.0
16	69	'2020/12/16'	98	120	215.2
17	60	'2020/12/17'	100	120	300.0
18	45	'2020/12/18'	90	112	NaN
19	69	'2020/12/19'	103	123	323.0
20	45	'2020/12/20'	97	125	243.0
21	69	'2020/12/21'	108	131	364.2
22	45	NaN	100	119	282.0
23	69	'2020/12/23'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	69	'2020/12/25'	102	126	334.5
26	69	2020/12/26	100	120	250.0
27	69	'2020/12/27'	92	118	241.0
28	69	'2020/12/28'	103	132	NaN
29	69	'2020/12/29'	100	132	280.0
30	69	'2020/12/30'	102	129	380.3
31	60	'2020/12/31'	92	115	243.0