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)

   2. **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
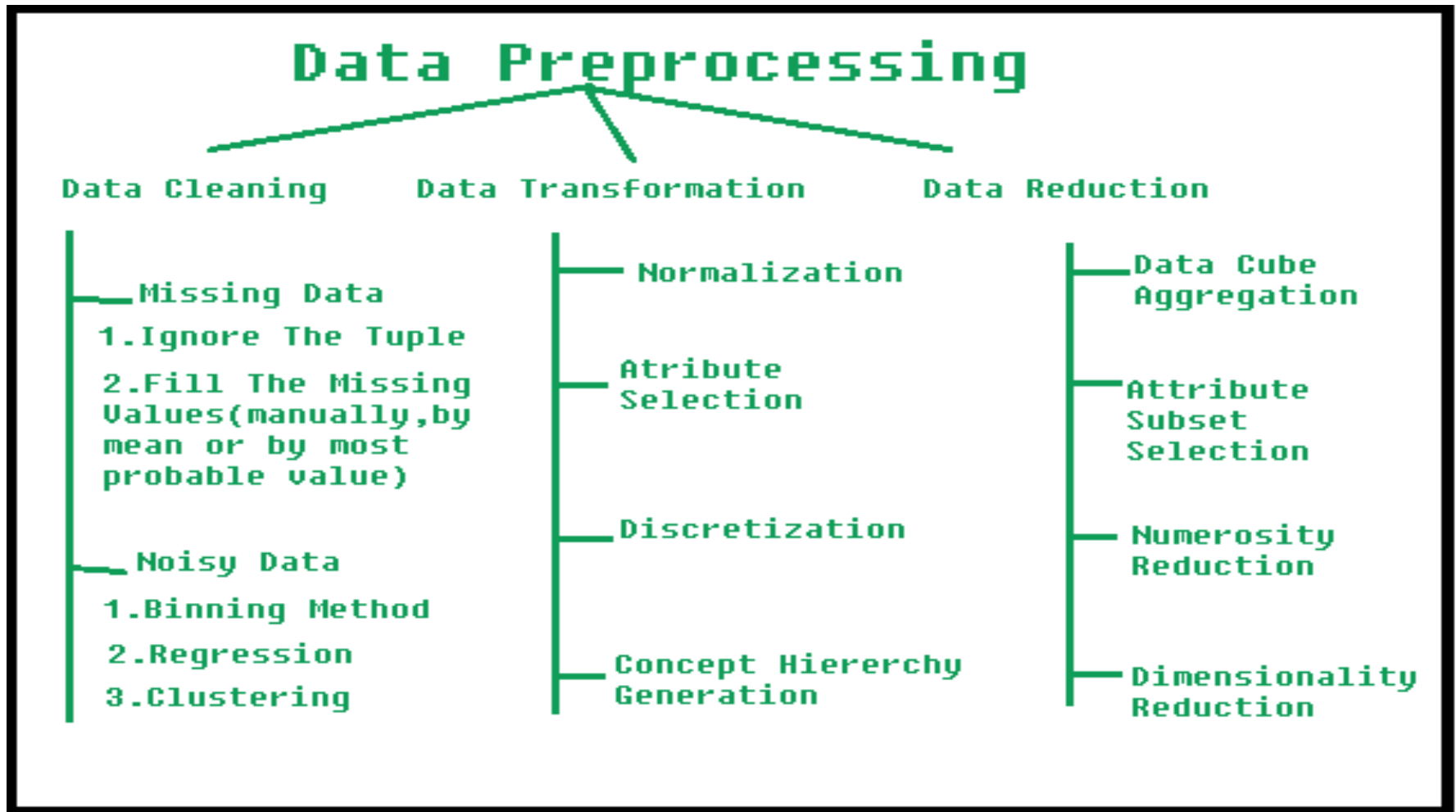
Data Cleaning

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

	Duration	Date	Pulse	Maxpulse	Calories
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	60	"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	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
22	45	NaN	100	119	282.0
23	60	"2020/12/23"	130	101	300.0
24	45	"2020/12/24"	105	132	246.0
25	60	"2020/12/25"	102	126	334.5
26	60	2020/12/26	100	120	250.0
27	60	"2020/12/27"	92	118	241.0
28	60	"2020/12/28"	103	132	NaN
29	60	"2020/12/29"	100	132	280.0
30	60	"2020/12/30"	102	129	380.3
31	60	"2020/12/31"	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
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	60	'2020/12/16'	98	120	215.2
17	60	'2020/12/17'	100	120	300.0
19	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'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	60	'2020/12/25'	102	126	334.5
26	60	2020/12/26	100	120	250.0
27	60	'2020/12/27'	92	118	241.0
29	60	'2020/12/29'	100	132	280.0
30	60	'2020/12/30'	102	129	380.3
31	60	'2020/12/31'	92	115	243.0

Missing Data

2. Replace missing data with values

```python
import pandas as pd

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

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

Replace missing data for specified column

```python
import pandas as pd

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

df["Calories"].fillna(130, 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	60	'2020/12/16'	98	120	215.2
17	60	'2020/12/17'	100	120	300.0
18	45	'2020/12/18'	90	112	130.0
19	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
22	45	130	100	119	282.0
23	60	'2020/12/23'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	60	'2020/12/25'	102	126	334.5
26	60	2020/12/26	100	120	250.0
27	60	'2020/12/27'	92	118	241.0
28	60	'2020/12/28'	103	132	130.0
29	60	'2020/12/29'	100	132	280.0
30	60	'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
0	60	'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	60	'2020/12/06'	102	127	300.00
6	60	'2020/12/07'	110	136	374.00
7	450	'2020/12/08'	104	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	60	'2020/12/12'	100	120	250.70
13	60	'2020/12/13'	106	128	345.30
14	60	'2020/12/14'	104	132	379.30
15	60	'2020/12/15'	98	123	275.00
16	60	'2020/12/16'	98	120	215.20
17	60	'2020/12/17'	100	120	300.00
18	45	'2020/12/18'	90	112	304.68
19	60	'2020/12/19'	103	123	323.00
20	45	'2020/12/20'	97	125	243.00
21	60	'2020/12/21'	108	131	364.20
22	45	NaN	100	119	282.00
23	60	'2020/12/23'	130	101	300.00
24	45	'2020/12/24'	105	132	246.00
25	60	'2020/12/25'	102	126	334.50
26	60	2020/12/26	100	120	250.00
27	60	'2020/12/27'	92	118	241.00
28	60	'2020/12/28'	103	132	304.68
29	60	'2020/12/29'	100	132	280.00
30	60	'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
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         60    '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         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
22         45             NaN    100       119     282.0
23         60    '2020/12/23'    130       101     300.0
24         45    '2020/12/24'    105       132     246.0
25         60    '2020/12/25'    102       126     334.5
26         60      2020/12/26    100       120     250.0
27         60    '2020/12/27'     92       118     241.0
28         60    '2020/12/28'    103       132       NaN
29         60    '2020/12/29'    100       132     280.0
30         60    '2020/12/30'    102       129     380.3
31         60    '2020/12/31'     92       115     243.0
```

Cleaning Wrong Data

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

1. Replacing Value

```
df.loc[7, 'Duration'] = 45
```

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
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	60	'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	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
22	45	NaN	100	119	282.0
23	60	'2020/12/23'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	60	'2020/12/25'	102	126	334.5
26	60	2020/12/26	100	120	250.0
27	60	'2020/12/27'	92	118	241.0
28	60	'2020/12/28'	103	132	NaN
29	60	'2020/12/29'	100	132	280.0
30	60	'2020/12/30'	102	129	380.3
31	60	'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	60	'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	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
22	45	NaN	100	119	282.0
23	60	'2020/12/23'	130	101	300.0
24	45	'2020/12/24'	105	132	246.0
25	60	'2020/12/25'	102	126	334.5
26	60	2020/12/26	100	120	250.0
27	60	'2020/12/27'	92	118	241.0
28	60	'2020/12/28'	103	132	NaN
29	60	'2020/12/29'	100	132	280.0
30	60	'2020/12/30'	102	129	380.3
31	60	'2020/12/31'	92	115	243.0