Colab: https://colab.research.google.com/drive/1wccJMr8n7Hw8IDhzxdWO9Lqu2-xXxmXH?usp=sharing

+ Code + Text

```
import pandas as pd
import numpy as np
```

```
!gdown 173A59xh2mnpmljCCB9bhC4C5eP2IS6qZ
```

```
    Downloading...
    From: https://drive.google.com/uc?id=173A59xh2mnpmljCCB9bhC4C5eP2IS6qZ
    To: /content/Pfizer_1.csv
    100% 1.51k/1.51k [00:00<00:00, 2.13MB/s]
```

```
data = pd.read_csv("Pfizer_1.csv")
```

```
data.info()
```

```
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 18 entries, 0 to 17
    Data columns (total 15 columns):
     #   Column     Non-Null Count  Dtype  
    ---  ------     --------------  -----  
     0   Date       18 non-null     object 
     1   Drug_Name  18 non-null     object 
     2   Parameter  18 non-null     object 
     3   1:30:00    16 non-null     float64
     4   2:30:00    16 non-null     float64
     5   3:30:00    12 non-null     float64
     6   4:30:00    14 non-null     float64
     7   5:30:00    16 non-null     float64
     8   6:30:00    18 non-null     int64  
     9   7:30:00    16 non-null     float64
     10  8:30:00    14 non-null     float64
     11  9:30:00    16 non-null     float64
     12  10:30:00   18 non-null     int64  
     13  11:30:00   16 non-null     float64
     14  12:30:00   18 non-null     int64  
    dtypes: float64(9), int64(3), object(3)
    memory usage: 2.2+ KB
```

```
data.head()
```

	Date	Drug_Name	Parameter	1:30:00	2:30:00	3:30:00	4:30:00	5:30:00	6:30:00	7:30:00	8:30:00	9:30:00
0	15-10-2020	diltiazem hydrochloride	Temperature	23.0	22.0	NaN	21.0	21.0	22	23.0	21.0	22.0
1	15-10-2020	diltiazem hydrochloride	Pressure	12.0	13.0	NaN	11.0	13.0	14	16.0	16.0	24.0
2	15-10-2020	docetaxel injection	Temperature	NaN	17.0	18.0	NaN	17.0	18	NaN	NaN	23.0
3	15-10-2020	docetaxel injection	Pressure	NaN	22.0	22.0	NaN	22.0	23	NaN	NaN	27.0
4	15-10-2020	ketamine hydrochloride	Temperature	24.0	NaN	NaN	27.0	NaN	26	25.0	24.0	23.0

```
# Can I restructure my dataset to turn it into a long format dataset?
# Melting
```

```
data_melt = pd.melt(data, id_vars = ["Date", "Drug_Name", "Parameter"],
              var_name = "time", value_name = "reading")
```

```
18*12
```

```
    216
```

```
data_melt.head()
```

	Date	Drug_Name	Parameter	time	reading
0	15-10-2020	diltiazem hydrochloride	Temperature	1:30:00	23.0
1	15-10-2020	diltiazem hydrochloride	Pressure	1:30:00	12.0
2	15-10-2020	docetaxel injection	Temperature	1:30:00	NaN
3	15-10-2020	docetaxel injection	Pressure	1:30:00	NaN
4	15-10-2020	ketamine hydrochloride	Temperature	1:30:00	24.0

```
data_tidy = data_melt.pivot(index=["Date", "Drug_Name", "time"],
                 columns = "Parameter",
                 values="reading").reset_index()
data_tidy
```

Parameter	Date	Drug_Name	time	Pressure	Temperature
0	15-10-2020	diltiazem hydrochloride	10:30:00	18.0	20.0
1	15-10-2020	diltiazem hydrochloride	11:30:00	19.0	20.0
2	15-10-2020	diltiazem hydrochloride	12:30:00	20.0	21.0
3	15-10-2020	diltiazem hydrochloride	1:30:00	12.0	23.0
4	15-10-2020	diltiazem hydrochloride	2:30:00	13.0	22.0
...	...	...	...	...	...
103	17-10-2020	ketamine hydrochloride	5:30:00	11.0	17.0
104	17-10-2020	ketamine hydrochloride	6:30:00	12.0	18.0
105	17-10-2020	ketamine hydrochloride	7:30:00	12.0	19.0
106	17-10-2020	ketamine hydrochloride	8:30:00	11.0	20.0
107	17-10-2020	ketamine hydrochloride	9:30:00	12.0	21.0

108 rows × 5 columns

```
data_melt.pivot(index=["Date", "Drug_Name", "Parameter"],
                 columns = "time",
                 values="reading").reset_index()
```

time	Date	Drug_Name	Parameter	10:30:00	11:30:00	12:30:00	1:30:00	2:30:00	3:30:00	4:30:00	5:30:00	6:
0	15-10-2020	diltiazem hydrochloride	Pressure	18.0	19.0	20.0	12.0	13.0	NaN	11.0	13.0	
1	15-10-2020	diltiazem hydrochloride	Temperature	20.0	20.0	21.0	23.0	22.0	NaN	21.0	21.0	
2	15-10-2020	docetaxel injection	Pressure	26.0	29.0	28.0	NaN	22.0	22.0	NaN	22.0	
3	15-10-2020	docetaxel injection	Temperature	23.0	25.0	25.0	NaN	17.0	18.0	NaN	17.0	
4	15-10-2020	ketamine hydrochloride	Pressure	9.0	9.0	11.0	8.0	NaN	NaN	7.0	NaN	
5	15-10-2020	ketamine hydrochloride	Temperature	22.0	21.0	20.0	24.0	NaN	NaN	27.0	NaN	
6	16-10-2020	diltiazem hydrochloride	Pressure	24.0	NaN	27.0	18.0	19.0	20.0	21.0	22.0	
7	16-10-2020	diltiazem hydrochloride	Temperature	40.0	NaN	42.0	34.0	35.0	36.0	36.0	37.0	
8	16-10-2020	docetaxel injection	Pressure	28.0	29.0	30.0	23.0	24.0	NaN	25.0	26.0	
9	16-10-2020	docetaxel injection	Temperature	56.0	57.0	58.0	46.0	47.0	NaN	48.0	48.0	
10	16-10-2020	ketamine hydrochloride	Pressure	16.0	17.0	18.0	12.0	12.0	13.0	NaN	15.0	
11	16-10-2020	ketamine hydrochloride	Temperature	13.0	14.0	15.0	8.0	9.0	10.0	NaN	11.0	
12	17-10-2020	diltiazem hydrochloride	Pressure	11.0	13.0	14.0	3.0	4.0	4.0	4.0	6.0	
13	17-10-2020	diltiazem hydrochloride	Temperature	14.0	11.0	10.0	20.0	19.0	19.0	18.0	17.0	
14	17-10-2020	docetaxel injection	Pressure	28.0	29.0	28.0	20.0	22.0	22.0	22.0	22.0	
15	17-10-2020	docetaxel injection	Temperature	21.0	22.0	23.0	12.0	13.0	14.0	15.0	16.0	

```
# Pivot --> Opposite of Melting
data_tidy.columns.name = None
```

```
data_tidy
```

	Date	Drug_Name	time	Pressure	Temperature
0	15-10-2020	diltiazem hydrochloride	10:30:00	18.0	20.0
1	15-10-2020	diltiazem hydrochloride	11:30:00	19.0	20.0
2	15-10-2020	diltiazem hydrochloride	12:30:00	20.0	21.0
3	15-10-2020	diltiazem hydrochloride	1:30:00	12.0	23.0
4	15-10-2020	diltiazem hydrochloride	2:30:00	13.0	22.0
...	...	...	...	...	...
103	17-10-2020	ketamine hydrochloride	5:30:00	11.0	17.0
104	17-10-2020	ketamine hydrochloride	6:30:00	12.0	18.0
105	17-10-2020	ketamine hydrochloride	7:30:00	12.0	19.0
106	17-10-2020	ketamine hydrochloride	8:30:00	11.0	20.0
107	17-10-2020	ketamine hydrochloride	9:30:00	12.0	21.0

108 rows × 5 columns

```
# Missing Values - NaN, None
```

```
type(None)
```

```
NoneType
```

```
type(np.nan)
```

```
float
```

```
pd.Series([1, np.nan, 2, None])
```

```
0    1.0
1    NaN
2    2.0
3    NaN
dtype: float64
```

```
pd.Series(["1", "np.nan", "2", None])
```

```
0         1
1    np.nan
2         2
```

```
3      None
dtype: object
```

```
pd.Series(["1", "np.nan", "2", "Anant", np.nan])
```

```
0         1
1    np.nan
2         2
3     Anant
4       NaN
dtype: object
```

```
data_tidy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 108 entries, 0 to 107
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Date         108 non-null    object 
 1   Drug_Name    108 non-null    object 
 2   time         108 non-null    object 
 3   Pressure     95 non-null     float64
 4   Temperature  95 non-null     float64
dtypes: float64(2), object(3)
memory usage: 4.3+ KB
```

```
data_tidy.isna().sum(axis=0)
```

```
Date            0
Drug_Name       0
time            0
Pressure       13
Temperature    13
dtype: int64
```

```
data_tidy.isna().sum(axis=1)
```

```
0      0
1      0
2      0
3      0
4      0
      ..
103    0
104    0
105    0
106    0
107    0
Length: 108, dtype: int64
```

```
data_tidy.isnull().sum(axis=0)
```

```
Date            0
Drug_Name       0
time            0
Pressure       13
Temperature    13
dtype: int64
```

```
pd.isna
```

```
<function pandas.core.dtypes.missing.isna(obj)>
```

```
pd.isnull
```

```
<function pandas.core.dtypes.missing.isna(obj)>
```

```
# handle missing values
# 1. Simply remove the rows/columns having missing values
# 2. Replace it with some values (Imputation)
#.   - Either fill it up with some placeholder -> 0, 999999999
#    - Either replace it with some estimator (mean, median for numeric) (mode for categorical)
#.   - If data is a time-series (R2 --> R1 seq fashion) - fill-up with the last values
```

```
data_tidy.dropna(axis=0)
```

	Date	Drug_Name	time	Pressure	Temperature
0	15-10-2020	diltiazem hydrochloride	10:30:00	18.0	20.0
1	15-10-2020	diltiazem hydrochloride	11:30:00	19.0	20.0
2	15-10-2020	diltiazem hydrochloride	12:30:00	20.0	21.0
3	15-10-2020	diltiazem hydrochloride	1:30:00	12.0	23.0
4	15-10-2020	diltiazem hydrochloride	2:30:00	13.0	22.0
...	...	...	...	...	...
103	17-10-2020	ketamine hydrochloride	5:30:00	11.0	17.0
104	17-10-2020	ketamine hydrochloride	6:30:00	12.0	18.0
105	17-10-2020	ketamine hydrochloride	7:30:00	12.0	19.0
106	17-10-2020	ketamine hydrochloride	8:30:00	11.0	20.0

```
data_tidy.dropna(axis=1)
```

	Date	Drug_Name	time
0	15-10-2020	diltiazem hydrochloride	10:30:00
1	15-10-2020	diltiazem hydrochloride	11:30:00
2	15-10-2020	diltiazem hydrochloride	12:30:00
3	15-10-2020	diltiazem hydrochloride	1:30:00
4	15-10-2020	diltiazem hydrochloride	2:30:00
...	...	...	...
103	17-10-2020	ketamine hydrochloride	5:30:00
104	17-10-2020	ketamine hydrochloride	6:30:00
105	17-10-2020	ketamine hydrochloride	7:30:00
106	17-10-2020	ketamine hydrochloride	8:30:00
107	17-10-2020	ketamine hydrochloride	9:30:00

108 rows × 3 columns

```
data_tidy.fillna(999999).head(20)
```

	Date	Drug_Name	time	Pressure	Temperature
0	15-10-2020	diltiazem hydrochloride	10:30:00	18.0	20.0
1	15-10-2020	diltiazem hydrochloride	11:30:00	19.0	20.0
2	15-10-2020	diltiazem hydrochloride	12:30:00	20.0	21.0
3	15-10-2020	diltiazem hydrochloride	1:30:00	12.0	23.0
4	15-10-2020	diltiazem hydrochloride	2:30:00	13.0	22.0
5	15-10-2020	diltiazem hydrochloride	3:30:00	999999.0	999999.0
6	15-10-2020	diltiazem hydrochloride	4:30:00	11.0	21.0
7	15-10-2020	diltiazem hydrochloride	5:30:00	13.0	21.0
8	15-10-2020	diltiazem hydrochloride	6:30:00	14.0	22.0
9	15-10-2020	diltiazem hydrochloride	7:30:00	16.0	23.0
10	15-10-2020	diltiazem hydrochloride	8:30:00	16.0	21.0
11	15-10-2020	diltiazem hydrochloride	9:30:00	24.0	22.0
12	15-10-2020	docetaxel injection	10:30:00	26.0	23.0
13	15-10-2020	docetaxel injection	11:30:00	29.0	25.0
14	15-10-2020	docetaxel injection	12:30:00	28.0	25.0
15	15-10-2020	docetaxel injection	1:30:00	999999.0	999999.0
16	15-10-2020	docetaxel injection	2:30:00	22.0	17.0
17	15-10-2020	docetaxel injection	3:30:00	22.0	18.0
18	15-10-2020	docetaxel injection	4:30:00	999999.0	999999.0
19	15-10-2020	docetaxel injection	5:30:00	22.0	17.0

```
data_tidy["Temperature"].mean()
```

```
24.326315789473686
```

```
data_tidy["Temperature"].fillna(data_tidy["Temperature"].mean()).head(20)
```

```
0     20.000000
1     20.000000
2     21.000000
3     23.000000
4     22.000000
5     24.326316
6     21.000000
7     21.000000
8     22.000000
9     23.000000
10    21.000000
11    22.000000
12    23.000000
13    25.000000
14    25.000000
15    24.326316
16    17.000000
17    18.000000
18    24.326316
19    17.000000
Name: Temperature, dtype: float64
```

```
def temp_mean(x):
  x["Avg_Temperature"] = x["Temperature"].mean()
  return x

data_tidy = data_tidy.groupby("Drug_Name").apply(temp_mean)
```

```
data_tidy.head(20)
```

	Date	Drug_Name	time	Pressure	Temperature	Avg_Temperature
0	15-10-2020	diltiazem hydrochloride	10:30:00	18.0	20.0	24.848485
1	15-10-2020	diltiazem hydrochloride	11:30:00	19.0	20.0	24.848485
2	15-10-2020	diltiazem hydrochloride	12:30:00	20.0	21.0	24.848485
3	15-10-2020	diltiazem hydrochloride	1:30:00	12.0	23.0	24.848485
4	15-10-2020	diltiazem hydrochloride	2:30:00	13.0	22.0	24.848485
5	15-10-2020	diltiazem hydrochloride	3:30:00	NaN	NaN	24.848485
6	15-10-2020	diltiazem hydrochloride	4:30:00	11.0	21.0	24.848485
7	15-10-2020	diltiazem hydrochloride	5:30:00	13.0	21.0	24.848485
8	15-10-2020	diltiazem hydrochloride	6:30:00	14.0	22.0	24.848485
9	15-10-2020	diltiazem hydrochloride	7:30:00	16.0	23.0	24.848485
10	15-10-2020	diltiazem hydrochloride	8:30:00	16.0	21.0	24.848485
11	15-10-2020	diltiazem hydrochloride	9:30:00	24.0	22.0	24.848485
12	15-10-2020	docetaxel injection	10:30:00	26.0	23.0	30.387097
13	15-10-2020	docetaxel injection	11:30:00	29.0	25.0	30.387097
14	15-10-2020	docetaxel injection	12:30:00	28.0	25.0	30.387097
15	15-10-2020	docetaxel injection	1:30:00	NaN	NaN	30.387097
16	15-10-2020	docetaxel injection	2:30:00	22.0	17.0	30.387097
17	15-10-2020	docetaxel injection	3:30:00	22.0	18.0	30.387097
18	15-10-2020	docetaxel injection	4:30:00	NaN	NaN	30.387097
19	15-10-2020	docetaxel injection	5:30:00	22.0	17.0	30.387097

```
def pressure_mean(x):
  x["Avg_Pressure"] = x["Pressure"].mean()
  return x

data_tidy = data_tidy.groupby("Drug_Name").apply(pressure_mean)
```

```
data_tidy
```

	Date	Drug_Name	time	Pressure	Temperature	Avg_Temperature	Avg_Pressure
0	15-10-2020	diltiazem hydrochloride	10:30:00	18.0	20.0	24.848485	15.424242
1	15-10-2020	diltiazem hydrochloride	11:30:00	19.0	20.0	24.848485	15.424242
2	15-10-2020	diltiazem hydrochloride	12:30:00	20.0	21.0	24.848485	15.424242
3	15-10-2020	diltiazem hydrochloride	1:30:00	12.0	23.0	24.848485	15.424242
4	15-10-2020	diltiazem hydrochloride	2:30:00	13.0	22.0	24.848485	15.424242
...	...	...	...	...	...	...	...
103	17-10-2020	ketamine hydrochloride	5:30:00	11.0	17.0	17.709677	11.935484
104	17-10-2020	ketamine hydrochloride	6:30:00	12.0	18.0	17.709677	11.935484
105	17-10-2020	ketamine hydrochloride	7:30:00	12.0	19.0	17.709677	11.935484
106	17-10-2020	ketamine hydrochloride	8:30:00	11.0	20.0	17.709677	11.935484
107	17-10-2020	ketamine hydrochloride	9:30:00	12.0	21.0	17.709677	11.935484

108 rows × 7 columns

```
data_tidy["Temperature"].fillna(data_tidy["Avg_Temperature"])
```

```
0      20.0
1      20.0
2      21.0
3      23.0
4      22.0
       ...
103    17.0
104    18.0
105    19.0
106    20.0
107    21.0
Name: Temperature, Length: 108, dtype: float64
```

```
# NPS - #detractors [0-6], # neutrals [7-8], #promoters [9-10]
# Numerical Data --> Categorical Data
# Temperature
```

```
data_tidy["Temperature"].min()
```

```
8.0
```

```
data_tidy["Temperature"].max()
```

```
58.0
```

```
# Temperature 5-60 usually
# low, medium, high, very high
# bucketisation
temp_labels = ["low", "medium", "high", "very high"]
temp_edges = [5, 20, 35, 50, 60]
pd.cut(data_tidy["Temperature"], bins=temp_edges, labels=temp_labels)
```

```
0         low
1         low
2      medium
3      medium
4      medium
        ...
103       low
104       low
105       low
106       low
107    medium
Name: Temperature, Length: 108, dtype: category
Categories (4, object): ['low' < 'medium' < 'high' < 'very high']
```

```
# string methods, datetime --> Revision Notes
```