session-2

August 25, 2024

1 Data Cleaning

```
[1]: import pandas as pd
     import csv
```

```
[2]: #pd.read_csv('laptops.csv', encoding = 'utf-8')
     #pd.read_csv('laptops.csv', encoding = 'windowm-1251')
     laptops = pd.read_csv('laptops.csv', encoding = 'latin-1')
```

```
[3]: laptops
```

```
[3]:      Manufacturer                           Model Name            Category  \
     0           Apple                          MacBook Pro           Ultrabook
     1           Apple                          Macbook Air           Ultrabook
     2              HP                               250 G6            Notebook
     3           Apple                          MacBook Pro           Ultrabook
     4           Apple                          MacBook Pro           Ultrabook
     …           …                                  …                 …
     1298       Lenovo                       Yoga 500-14ISK  2 in 1 Convertible
     1299       Lenovo                       Yoga 900-13ISK  2 in 1 Convertible
     1300       Lenovo                   IdeaPad 100S-14IBR            Notebook
     1301             HP  15-AC110nv (i7-6500U/6GB/1TB/Radeon            Notebook
     1302           Asus  X553SA-XX031T (N3050/4GB/500GB/W10)            Notebook

          Screen Size                                      Screen  \
     0          13.3"          IPS Panel Retina Display 2560x1600
     1          13.3"                                    1440x900
     2          15.6"                           Full HD 1920x1080
     3          15.4"          IPS Panel Retina Display 2880x1800
     4          13.3"          IPS Panel Retina Display 2560x1600
     …          …                                          …
     1298       14.0"   IPS Panel Full HD / Touchscreen 1920x1080
     1299       13.3"  IPS Panel Quad HD+ / Touchscreen 3200x1800
     1300       14.0"                                    1366x768
     1301       15.6"                                    1366x768
     1302       15.6"                                    1366x768
```

```
                                      CPU   RAM              Storage  \
0                    Intel Core i5 2.3GHz   8GB            128GB SSD   
1                    Intel Core i5 1.8GHz   8GB  128GB Flash Storage   
2              Intel Core i5 7200U 2.5GHz   8GB            256GB SSD   
3                    Intel Core i7 2.7GHz  16GB            512GB SSD   
4                    Intel Core i5 3.1GHz   8GB            256GB SSD   
…                                    …   …                   …   
1298           Intel Core i7 6500U 2.5GHz   4GB            128GB SSD   
1299           Intel Core i7 6500U 2.5GHz  16GB            512GB SSD   
1300  Intel Celeron Dual Core N3050 1.6GHz   2GB   64GB Flash Storage   
1301           Intel Core i7 6500U 2.5GHz   6GB              1TB HDD   
1302  Intel Celeron Dual Core N3050 1.6GHz   4GB            500GB HDD   

                               GPU Operating System Operating System Version  \
0     Intel Iris Plus Graphics 640            macOS                      NaN   
1           Intel HD Graphics 6000            macOS                      NaN   
2            Intel HD Graphics 620            No OS                      NaN   
3               AMD Radeon Pro 455            macOS                      NaN   
4     Intel Iris Plus Graphics 650            macOS                      NaN   
…                             …              …                      …   
1298         Intel HD Graphics 520          Windows                       10   
1299         Intel HD Graphics 520          Windows                       10   
1300             Intel HD Graphics          Windows                       10   
1301            AMD Radeon R5 M330          Windows                       10   
1302             Intel HD Graphics          Windows                       10   

      Weight Price (Euros)  
0     1.37kg       1339,69  
1     1.34kg        898,94  
2     1.86kg        575,00  
3     1.83kg       2537,45  
4     1.37kg       1803,60  
…       …            …  
1298   1.8kg        638,00  
1299   1.3kg       1499,00  
1300   1.5kg        229,00  
1301  2.19kg        764,00  
1302   2.2kg        369,00  

[1303 rows x 13 columns]
```

[4]:

```
laptops.head()
```

[4]:

```
  Manufacturer   Model Name   Category Screen Size  \
0        Apple  MacBook Pro  Ultrabook       13.3"   
1        Apple  Macbook Air  Ultrabook       13.3"   
2           HP       250 G6   Notebook       15.6"   
```

```
3        Apple  MacBook Pro  Ultrabook       15.4"
4        Apple  MacBook Pro  Ultrabook       13.3"

                               Screen                         CPU   RAM  \
0  IPS Panel Retina Display 2560x1600        Intel Core i5 2.3GHz   8GB
1                            1440x900        Intel Core i5 1.8GHz   8GB
2                   Full HD 1920x1080  Intel Core i5 7200U 2.5GHz   8GB
3  IPS Panel Retina Display 2880x1800        Intel Core i7 2.7GHz  16GB
4  IPS Panel Retina Display 2560x1600        Intel Core i5 3.1GHz   8GB

               Storage                           GPU Operating System  \
0            128GB SSD  Intel Iris Plus Graphics 640            macOS
1  128GB Flash Storage        Intel HD Graphics 6000            macOS
2            256GB SSD         Intel HD Graphics 620            No OS
3            512GB SSD            AMD Radeon Pro 455            macOS
4            256GB SSD  Intel Iris Plus Graphics 650            macOS

  Operating System Version  Weight Price (Euros)
0                      NaN  1.37kg       1339,69
1                      NaN  1.34kg        898,94
2                      NaN  1.86kg        575,00
3                      NaN  1.83kg       2537,45
4                      NaN  1.37kg       1803,60
```

[5]:
```
laptops.shape
```

[5]:
```
(1303, 13)
```

[6]:
```
laptops.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1303 entries, 0 to 1302
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Manufacturer              1303 non-null   object
 1   Model Name                1303 non-null   object
 2   Category                  1303 non-null   object
 3   Screen Size               1303 non-null   object
 4   Screen                    1303 non-null   object
 5   CPU                       1303 non-null   object
 6   RAM                       1303 non-null   object
 7    Storage                  1303 non-null   object
 8   GPU                       1303 non-null   object
 9   Operating System          1303 non-null   object
 10  Operating System Version  1133 non-null   object
 11  Weight                    1303 non-null   object
 12  Price (Euros)             1303 non-null   object
dtypes: object(13)
memory usage: 132.5+ KB
```

2 str.strip() method

```
[7]: #for removing spaces in string.
     " Techma Zone ".strip()
```

```
[7]: 'Techma Zone'
```

```
[8]: "-Techma-Zone-".strip('-')
```

```
[8]: 'Techma-Zone'
```

```
[9]: "(-  Techma-Zone-())".strip(' -()')
```

```
[9]: 'Techma-Zone'
```

```
[10]: laptops.columns
```

```
[10]: Index(['Manufacturer', 'Model Name', 'Category', 'Screen Size', 'Screen',
              'CPU', 'RAM', ' Storage', 'GPU', 'Operating System',
              'Operating System Version', 'Weight', 'Price (Euros)'],
             dtype='object')
```

```
[11]: laptops[' Storage']
```

```
[11]: 0                 128GB SSD
      1       128GB Flash Storage
      2                 256GB SSD
      3                 512GB SSD
      4                 256GB SSD
                     …
      1298              128GB SSD
      1299              512GB SSD
      1300     64GB Flash Storage
      1301                1TB HDD
      1302              500GB HDD
      Name:  Storage, Length: 1303, dtype: object
```

3 str.replace()

```
[12]: 'Techma Zone *'.replace(' *', '')
```

```
[12]: 'Techma Zone'
```

```
[13]: 'Techma Zone'.replace('Z', 'T')
```

```
[13]: 'Techma Tone'
```

4 str.lower() & str.upper()

```
[14]: 'techma zone'.upper()
```

```
[14]: 'TECHMA ZONE'
```

```
[15]: 'TECHMA ZONE'.lower()
```

```
[15]: 'techma zone'
```

5 Function

```
[16]: def clean_label(label):
          label = label.strip()
          label = label.replace('(', '')
          label = label.replace(')', '')
          label = label.lower()
          label = label.replace(' ', '_')
          label = label.replace('operating_system', 'os')
          return label
```

```
[17]: clean_label('(Column Name)')
```

```
[17]: 'column_name'
```

```
[18]: type(laptops.columns)
```

```
[18]: pandas.core.indexes.base.Index
```

6 pd.Series()

```
[19]: pd.Series(laptops.columns)
```

```
[19]: 0                  Manufacturer
      1                    Model Name
      2                      Category
      3                   Screen Size
      4                        Screen
```

```
5                          CPU
6                          RAM
7                      Storage
8                          GPU
9             Operating System
10    Operating System Version
11                      Weight
12               Price (Euros)
dtype: object
```

7 .apply()

[20]:
```
#difference between .apply() or .map()
```

[21]:
```
laptops.columns = pd.Series(laptops.columns).apply(clean_label)
```

[22]:
```
laptops.columns
```

[22]:
```
Index(['manufacturer', 'model_name', 'category', 'screen_size', 'screen',
       'cpu', 'ram', 'storage', 'gpu', 'os', 'os_version', 'weight',
       'price_euros'],
      dtype='object')
```

8 .isnull()

[23]:
```
laptops.manufacturer.isnull()
```

[23]:
```
0       False
1       False
2       False
3       False
4       False
        …
1298    False
1299    False
1300    False
1301    False
1302    False
Name: manufacturer, Length: 1303, dtype: bool
```

[24]:
```
laptops['os_version'].isnull()
```

[24]:
```
0        True
1        True
2        True
```

```
3         True
4         True
        …
1298    False
1299    False
1300    False
1301    False
1302    False
Name: os_version, Length: 1303, dtype: bool
```

9 Boolean Addition

- True -> 1
- False -> 0

```
[25]: True + True
```

```
[25]: 2
```

```
[26]: True + False
```

```
[26]: 1
```

```
[27]: False + False
```

```
[27]: 0
```

```
[28]: laptops['os_version'].isnull().sum()
```

```
[28]: 170
```

```
[29]: laptops.isnull().sum()
```

```
[29]: manufacturer      0
      model_name        0
      category          0
      screen_size       0
      screen            0
      cpu               0
      ram               0
      storage           0
      gpu               0
      os                0
      os_version      170
      weight            0
      price_euros       0
      dtype: int64
```

10 Skill Test - 1

- Extract complete records containing null values in os_version column

[30]:
```
laptops[laptops['os_version'].isnull()].head()
```

[30]:
```
  manufacturer  model_name   category screen_size  \
0        Apple  MacBook Pro  Ultrabook       13.3"
1        Apple  Macbook Air  Ultrabook       13.3"
2           HP       250 G6   Notebook       15.6"
3        Apple  MacBook Pro  Ultrabook       15.4"
4        Apple  MacBook Pro  Ultrabook       13.3"

                               screen                         cpu   ram  \
0  IPS Panel Retina Display 2560x1600        Intel Core i5 2.3GHz   8GB
1                            1440x900        Intel Core i5 1.8GHz   8GB
2                   Full HD 1920x1080  Intel Core i5 7200U 2.5GHz   8GB
3  IPS Panel Retina Display 2880x1800        Intel Core i7 2.7GHz  16GB
4  IPS Panel Retina Display 2560x1600        Intel Core i5 3.1GHz   8GB

               storage                           gpu     os os_version  \
0            128GB SSD  Intel Iris Plus Graphics 640  macOS        NaN
1  128GB Flash Storage        Intel HD Graphics 6000  macOS        NaN
2            256GB SSD         Intel HD Graphics 620  No OS        NaN
3            512GB SSD            AMD Radeon Pro 455  macOS        NaN
4            256GB SSD  Intel Iris Plus Graphics 650  macOS        NaN

   weight price_euros
0  1.37kg     1339,69
1  1.34kg      898,94
2  1.86kg      575,00
3  1.83kg     2537,45
4  1.37kg     1803,60
```

11 .fillna()

[31]:
```
laptops['os_version'] = laptops['os_version'].fillna('Version Unknown')
```

[32]:
```
laptops.isnull().sum()
```

[32]:
```
manufacturer    0
model_name      0
category        0
screen_size     0
screen          0
cpu             0
ram             0
```

```
storage         0
gpu             0
os              0
os_version      0
weight          0
price_euros     0
dtype: int64
```

12 Skill Test - 2

- Clean the RAM column i.e. remove GBs from each entry
- Clean the screen_size column i.e. remove inches from entry

```
[37]: laptops['ram'] = laptops['ram'].str.replace('GB', '').astype(int)
```

```
[41]: laptops['ram']
```

```
[41]: 0        8
      1        8
      2        8
      3       16
      4        8
              ..
      1298     4
      1299    16
      1300     2
      1301     6
      1302     4
      Name: ram, Length: 1303, dtype: int32
```

```
[40]: laptops['ram'].value_counts()
```

```
[40]: 8     619
      4     375
      16    200
      6      41
      12     25
      2      22
      32     17
      24      3
      64      1
      Name: ram, dtype: int64
```

```
[42]: laptops['screen_size'] = laptops['screen_size'].str.replace('"', '').
        ↪astype(float)
```

```
[43]: laptops['screen_size']
```

```
[43]: 0       13.3
      1       13.3
      2       15.6
      3       15.4
      4       13.3
              …
      1298    14.0
      1299    13.3
      1300    14.0
      1301    15.6
      1302    15.6
      Name: screen_size, Length: 1303, dtype: float64
```

```
[44]: laptops['screen_size'].value_counts()
```

```
[44]: 15.6    665
      14.0    197
      13.3    164
      17.3    164
      12.5     39
      11.6     33
      12.0      6
      13.5      6
      13.9      6
      12.3      5
      10.1      4
      15.4      4
      15.0      4
      13.0      2
      18.4      1
      17.0      1
      14.1      1
      11.3      1
      Name: screen_size, dtype: int64
```

13 Changing column name

```
[48]: laptops.columns
```

```
[48]: Index(['manufacturer', 'model_name', 'category', 'screen_size', 'screen',
             'cpu', 'ram', 'storage', 'gpu', 'os', 'os_version', 'weight',
             'price_euros'],
            dtype='object')
```

```
[49]: laptops.rename({'screen_size':'screen_size_inches', 'ram':'ram_gb'}, axis=1,␣
       ↪inplace=True)
```

```
[50]: laptops.columns
```

```
[50]: Index(['manufacturer', 'model_name', 'category', 'screen_size_inches',
              'screen', 'cpu', 'ram_gb', 'storage', 'gpu', 'os', 'os_version',
              'weight', 'price_euros'],
             dtype='object')
```

14 .astype()

- used to change datatype of a column

15 .split() method

```
[56]: name = 'Techma Zone Danial fastTrack'
```

```
[57]: name.split()
```

```
[57]: ['Techma', 'Zone', 'Danial', 'fastTrack']
```

```
[59]: first_name, last_name = name.split()[0], name.split()[1]
```

```
[60]: first_name
```

```
[60]: 'Techma'
```

```
[61]: last_name
```

```
[61]: 'Zone'
```

16 Skill Test - 3

- Find out the number of GPU manufactured by each manufacturer

```
[77]: laptops.gpu
```

```
[77]: 0        Intel Iris Plus Graphics 640
      1              Intel HD Graphics 6000
      2               Intel HD Graphics 620
      3                  AMD Radeon Pro 455
      4        Intel Iris Plus Graphics 650
                           …
      1298            Intel HD Graphics 520
```

```
1299           Intel HD Graphics 520
1300               Intel HD Graphics
1301              AMD Radeon R5 M330
1302               Intel HD Graphics
Name: gpu, Length: 1303, dtype: object
```

[78]:

```
laptops.gpu.value_counts()
```

[78]:

```
Intel HD Graphics 620      281
Intel HD Graphics 520      185
Intel UHD Graphics 620      68
Nvidia GeForce GTX 1050     66
Nvidia GeForce GTX 1060     48
                          …
AMD Radeon R5 520            1
AMD Radeon R7                1
Intel HD Graphics 540        1
AMD Radeon 540               1
ARM Mali T860 MP4            1
Name: gpu, Length: 110, dtype: int64
```

[79]:

```
laptops['gpu'].str.split() #laptops['gpu'].str.split(n=0)
```

[79]:

```
0       [Intel, Iris, Plus, Graphics, 640]
1              [Intel, HD, Graphics, 6000]
2               [Intel, HD, Graphics, 620]
3                  [AMD, Radeon, Pro, 455]
4       [Intel, Iris, Plus, Graphics, 650]
                       …
1298            [Intel, HD, Graphics, 520]
1299            [Intel, HD, Graphics, 520]
1300                 [Intel, HD, Graphics]
1301               [AMD, Radeon, R5, M330]
1302                 [Intel, HD, Graphics]
Name: gpu, Length: 1303, dtype: object
```

[80]:

```
laptops['gpu'].str.split(n=1, expand=False) #this is not good because we want data in dataframe pattern.
```

[80]:

```
0       [Intel, Iris Plus Graphics 640]
1             [Intel, HD Graphics 6000]
2              [Intel, HD Graphics 620]
3                 [AMD, Radeon Pro 455]
4       [Intel, Iris Plus Graphics 650]
                     …
1298           [Intel, HD Graphics 520]
1299           [Intel, HD Graphics 520]
```

```
1300               [Intel, HD Graphics]
1301             [AMD, Radeon R5 M330]
1302               [Intel, HD Graphics]
Name: gpu, Length: 1303, dtype: object
```

[81]:
```python
laptops['gpu'].str.split(n=1, expand=True)
```

[81]:
```
          0                       1
0     Intel  Iris Plus Graphics 640
1     Intel        HD Graphics 6000
2     Intel         HD Graphics 620
3       AMD          Radeon Pro 455
4     Intel  Iris Plus Graphics 650
…       …                      …
1298  Intel         HD Graphics 520
1299  Intel         HD Graphics 520
1300  Intel             HD Graphics
1301    AMD          Radeon R5 M330
1302  Intel             HD Graphics

[1303 rows x 2 columns]
```

[86]:
```python
laptops['gpu'] = laptops['gpu'].str.split(n=1, expand=True)[1]
```

[87]:
```python
laptops['gpu']
```

[87]:
```
0       Iris Plus Graphics 640
1             HD Graphics 6000
2              HD Graphics 620
3               Radeon Pro 455
4       Iris Plus Graphics 650
                 …
1298           HD Graphics 520
1299           HD Graphics 520
1300               HD Graphics
1301            Radeon R5 M330
1302               HD Graphics
Name: gpu, Length: 1303, dtype: object
```

[88]:
```python
laptops['gpu_company'] = laptops['gpu'].str.split(n=1, expand=True)[0]
```

[89]:
```python
laptops['gpu_company']
```

[89]:
```
0         Iris
1           HD
2           HD
3       Radeon
```

```
4         Iris
         …
1298        HD
1299        HD
1300        HD
1301    Radeon
1302        HD
Name: gpu_company, Length: 1303, dtype: object
```

[90]:
```
laptops['gpu_company'].value_counts()
```

[90]:
```
HD            639
GeForce       368
Radeon        173
UHD            68
Quadro         31
Iris           14
FirePro         5
R4              1
GTX             1
R17M-M1-70      1
Graphics        1
Mali            1
Name: gpu_company, dtype: int64
```

[91]:
```
laptops.head(3)
```

[91]:
```
  manufacturer   model_name   category  screen_size_inches  \
0        Apple  MacBook Pro  Ultrabook                13.3
1        Apple  Macbook Air  Ultrabook                13.3
2           HP       250 G6   Notebook                15.6

                               screen                         cpu  ram_gb  \
0  IPS Panel Retina Display 2560x1600        Intel Core i5 2.3GHz       8
1                            1440x900        Intel Core i5 1.8GHz       8
2                   Full HD 1920x1080  Intel Core i5 7200U 2.5GHz       8

               storage                     gpu     os       os_version  \
0            128GB SSD  Iris Plus Graphics 640  macOS  Version Unknown
1  128GB Flash Storage        HD Graphics 6000  macOS  Version Unknown
2            256GB SSD         HD Graphics 620  No OS  Version Unknown

   weight price_euros gpu_company
0  1.37kg     1339,69        Iris
1  1.34kg      898,94          HD
2  1.86kg      575,00          HD
```

17 Complete your Task

- Find out the number of CPU manufactured by each manufacturer

[94]:

```
laptops['cpu']
```

[94]:

```
0                       Intel Core i5 2.3GHz
1                       Intel Core i5 1.8GHz
2                 Intel Core i5 7200U 2.5GHz
3                       Intel Core i7 2.7GHz
4                       Intel Core i5 3.1GHz
                        …
1298              Intel Core i7 6500U 2.5GHz
1299              Intel Core i7 6500U 2.5GHz
1300    Intel Celeron Dual Core N3050 1.6GHz
1301              Intel Core i7 6500U 2.5GHz
1302    Intel Celeron Dual Core N3050 1.6GHz
Name: cpu, Length: 1303, dtype: object
```

[95]:

```
laptops['cpu'].value_counts()
```

[95]:

```
Intel Core i5 7200U 2.5GHz       190
Intel Core i7 7700HQ 2.8GHz      146
Intel Core i7 7500U 2.7GHz       134
Intel Core i7 8550U 1.8GHz        73
Intel Core i5 8250U 1.6GHz        72
                                …
Intel Core M M3-6Y30 0.9GHz        1
AMD A9-Series 9420 2.9GHz          1
Intel Core i3 6006U 2.2GHz         1
AMD A6-Series 7310 2GHz            1
Intel Xeon E3-1535M v6 3.1GHz      1
Name: cpu, Length: 118, dtype: int64
```

[101]:

```
laptops['cpu'].str.split(n=1, expand=True)
```

[101]:

```
          0                                1
0     Intel               Core i5 2.3GHz
1     Intel               Core i5 1.8GHz
2     Intel         Core i5 7200U 2.5GHz
3     Intel               Core i7 2.7GHz
4     Intel               Core i5 3.1GHz
…     …                              …
1298  Intel         Core i7 6500U 2.5GHz
1299  Intel         Core i7 6500U 2.5GHz
1300  Intel  Celeron Dual Core N3050 1.6GHz
1301  Intel         Core i7 6500U 2.5GHz
1302  Intel  Celeron Dual Core N3050 1.6GHz
```

```
[1303 rows x 2 columns]
```

```
[104]: laptops['cpu_company'] = laptops['cpu'].str.split(n=1, expand=True)[0]
```

```
[105]: laptops['cpu'] = laptops['cpu'].str.split(n=1, expand=True)[1]
```

```
[107]: laptops['cpu_company'].value_counts()
```

```
[107]: Intel      1240
       AMD          62
       Samsung       1
       Name: cpu_company, dtype: int64
```

```
[109]: laptops['cpu'].value_counts()
```

```
[109]: Core i5 7200U 2.5GHz       190
       Core i7 7700HQ 2.8GHz      146
       Core i7 7500U 2.7GHz       134
       Core i7 8550U 1.8GHz        73
       Core i5 8250U 1.6GHz        72
                                 …
       Core M M3-6Y30 0.9GHz        1
       A9-Series 9420 2.9GHz        1
       Core i3 6006U 2.2GHz         1
       A6-Series 7310 2GHz          1
       Xeon E3-1535M v6 3.1GHz      1
       Name: cpu, Length: 118, dtype: int64
```

```
[106]: laptops.head(3)
```

```
[106]:   manufacturer   model_name   category  screen_size_inches  \
       0        Apple  MacBook Pro  Ultrabook                13.3
       1        Apple  Macbook Air  Ultrabook                13.3
       2           HP       250 G6   Notebook                15.6

                                     screen                    cpu  ram_gb  \
       0  IPS Panel Retina Display 2560x1600        Core i5 2.3GHz       8
       1                            1440x900        Core i5 1.8GHz       8
       2                  Full HD 1920x1080   Core i5 7200U 2.5GHz       8

                       storage                    gpu     os       os_version  \
       0             128GB SSD  Iris Plus Graphics 640  macOS  Version Unknown
       1  128GB Flash Storage        HD Graphics 6000  macOS  Version Unknown
       2             256GB SSD         HD Graphics 620  No OS  Version Unknown

          weight price_euros gpu_company cpu_company
```

```
0  1.37kg    1339,69        Iris       Intel
1  1.34kg     898,94          HD       Intel
2  1.86kg     575,00          HD       Intel
```

18 Saving the File

```
[110]: laptops.to_csv('laptops_clean.csv', index=False)
```

```
[ ]:
```