Full name: Lê Đức Tín

MSSV: 19522348

Lab 2

Excercise 1

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

```
dataset = pd.read_csv('/content/sales.csv')
```

Fixing column datatypes

```
print(dataset.dtypes)
```

```
order_id       int64
name          object
ordered_at    object
price         object
quantity       int64
line_total    object
dtype: object
```

```
dataset = dataset.convert_dtypes()
```

```
dataset1 = dataset
```

```
print(dataset1)
```

```
       order_id                        name        ordered_at    price  \
0         10000    "ICE CREAM" Peanut Fudge  01/01/2018 11:30   $3.50
1         10000    "ICE CREAM" Peanut Fudge  01/01/2018 11:30   $3.50
2         10001          "SORBET" Raspberry  01/01/2018 12:14   $2.50
3         10001                       <NA>  01/01/2018 12:14   $1.50
4         10001   "CONE" Dipped Waffle Cone  01/01/2018 12:14   $3.50
...         ...                         ...               ...      ...
29917     18452  "ICE CREAM" Dulce De Leche   26/06/2018 3:56  ($1.50)
29918     12889  "ICE CREAM" Dark Chocolate  03/03/2018 10:06   $4.00
29919     14526    "ICE CREAM" Peanut Fudge  05/04/2018 17:33   $3.50
29920     19589   "CONE" Dipped Waffle Cone   20/07/2018 9:10   $3.50
29921     19270       "ICE CREAM" Earl Gray   13/07/2018 9:20   $0.50

       quantity line_total
0             3    $10.50
1             1     $3.50
2             2     $5.00
3             1     $1.50
4             1     $3.50
...         ...       ...
29917         2    ($3.00)
29918         3    $12.00
29919         3    $10.50
29920         2     $7.00
29921         2     $1.00

[29922 rows x 6 columns]
```

```
def drop_unit_price(data):
  data['price'] = data['price'].str.replace("$", "").str.replace("(", "").str.replace(")", "").str.replace(" ", "")
  data['line_total'] = data['line_total'].str.replace("$", "").str.replace("(", "").str.replace(")", "").str.replace(" ", "")

drop_unit_price(dataset1)
```

```
<ipython-input-7-65182ac83a51>:2: FutureWarning: The default value of regex will change from True to False in a future version. In addi
  data['price'] = data['price'].str.replace("$", "").str.replace("(", "").str.replace(")", "").str.replace(" ", "")
<ipython-input-7-65182ac83a51>:3: FutureWarning: The default value of regex will change from True to False in a future version. In addi
  data['line_total'] = data['line_total'].str.replace("$", "").str.replace("(", "").str.replace(")", "").str.replace(" ", "")
```

```
print(dataset1['price'])
```

```
0        3.50
1        3.50
2        2.50
3        1.50
4        3.50
         ...
29917    1.50
29918    4.00
29919    3.50
29920    3.50
29921    0.50
Name: price, Length: 29922, dtype: string
```

```
dataset1 = dataset1.convert_dtypes()
dataset1['line_total'] = dataset1['line_total'].astype(float)
dataset1['price'] = dataset1['price'].astype(float)
```

```
print(dataset1.dtypes)
```

```
order_id        Int64
name           string
ordered_at     string
price         float64
quantity        Int64
line_total    float64
dtype: object
```

```
dataset['price']
```

```
0        3.50
1        3.50
2        2.50
3        1.50
4        3.50
         ...
29917    1.50
29918    4.00
29919    3.50
29920    3.50
29921    0.50
Name: price, Length: 29922, dtype: string
```

Drop duplicate, null values

```
dataset1 = dataset1.dropna()
```

```
dataset1 = dataset1.drop_duplicates(subset=['order_id'])
```

Sanity check value ranges

For price

```
dataset1['price'].describe()
```

```
count    27892.000000
mean         2.512333
std          1.059432
min          0.500000
25%          1.500000
50%          2.500000
75%          3.500000
max          4.000000
Name: price, dtype: float64
```

For line total

```
dataset1['line_total'].describe()
```

```
count    29373.000000
mean         5.033313
std          3.087351
min          0.000000
25%          2.500000
50%          4.500000
75%          7.500000
max         12.000000
Name: line_total, dtype: float64
```

Regular expression and lambda function to parse data.

```
dataset1
```

	order_id	name	ordered_at	price	quantity	line_total
0	10000	"ICE CREAM" Peanut Fudge	01/01/2018 11:30	3.50	3	10.50
2	10001	"SORBET" Raspberry	01/01/2018 12:14	2.50	2	5.00
5	10002	"SORBET" Lychee	01/01/2018 12:23	3.00	1	3.00
8	10003	"ICE CREAM" Matcha	01/01/2018 12:49	1.50	3	4.50
9	10004	"BEVERAGE" Iced Coffee	01/01/2018 13:22	2.50	2	5.00
...	...	...	...	...	...	...
29812	19995	"BEVERAGE" Espresso	28/07/2018 17:30	2.50	3	7.50
29814	19996	"ICE CREAM" Strawberry	28/07/2018 17:32	3.50	1	3.50
29816	19997	"ICE CREAM" Double Fudge Chunk	28/07/2018 17:40	3.50	2	7.00
29819	19998	"SORBET" Lychee	28/07/2018 18:21	3.00	1	3.00
29821	19999	"SORBET" Blood Orange	28/07/2018 18:51	2.50	2	5.00

9496 rows × 6 columns

```
name = dataset1.iloc[:, 1:6].values
```

```
filtered_fruits = lambda data: str(f"{data[0]} was ordered {data[3]} quantities in {data[1]} with price {data[2]} and total is {data[4]}")
```

```
for each in name:
  print(filtered_fruits(each))
```

```
"MISC" Ice Cream Cake was ordered 1 quantities in 28/07/2018 5:01 with price 2.00 and total is 2.00
"SORBET" Raspberry was ordered 1 quantities in 28/07/2018 5:05 with price 2.50 and total is 2.50
"CONE" Sugar Cone was ordered 1 quantities in 28/07/2018 5:24 with price 1.00 and total is 1.00
"CONE" Dipped Waffle Cone was ordered 2 quantities in 28/07/2018 6:22 with price 3.50 and total is 7.00
"ICE CREAM" Matcha was ordered 3 quantities in 28/07/2018 6:57 with price 1.50 and total is 4.50
"CONE" Cookie Cone was ordered 1 quantities in 28/07/2018 7:21 with price 4.00 and total is 4.00
"CONE" Brownie Cone was ordered 2 quantities in 28/07/2018 7:43 with price 3.00 and total is 6.00
"SORBET" Watermelon was ordered 1 quantities in 28/07/2018 8:40 with price 2.50 and total is 2.50
"BEVERAGE" Espresso was ordered 2 quantities in 28/07/2018 8:55 with price 2.50 and total is 5.00
"ICE CREAM" Dark Chocolate was ordered 2 quantities in 28/07/2018 9:04 with price 4.00 and total is 8.00
"SORBET" Raspberry was ordered 2 quantities in 28/07/2018 9:12 with price 2.50 and total is 5.00
"ICE CREAM" Wildberry was ordered 2 quantities in 28/07/2018 10:04 with price 1.50 and total is 3.00
"BEVERAGE" Iced Coffee was ordered 1 quantities in 28/07/2018 10:46 with price 2.50 and total is 2.50
"ICE CREAM" Rocky Road was ordered 3 quantities in 28/07/2018 11:36 with price 3.50 and total is 10.50
"ICE CREAM" Peanut Fudge was ordered 3 quantities in 28/07/2018 12:23 with price 3.50 and total is 10.50
"SORBET" Raspberry was ordered 3 quantities in 28/07/2018 12:43 with price 2.50 and total is 7.50
"SORBET" Blood Orange was ordered 1 quantities in 28/07/2018 13:19 with price 2.50 and total is 2.50
"ICE CREAM" Rocky Road was ordered 1 quantities in 28/07/2018 13:31 with price 3.50 and total is 3.50
"BEVERAGE" Tea was ordered 1 quantities in 28/07/2018 13:50 with price 4.00 and total is 4.00
"MISC" Ice Cream Cake was ordered 2 quantities in 28/07/2018 13:53 with price 2.00 and total is 4.00
"CONE" Cookie Cone was ordered 2 quantities in 28/07/2018 14:20 with price 4.00 and total is 8.00
"CONE" Cookie Cone was ordered 3 quantities in 28/07/2018 15:06 with price 4.00 and total is 12.00
"ICE CREAM" Dulce De Leche was ordered 1 quantities in 28/07/2018 15:10 with price 1.50 and total is 1.50
"CONE" Cookie Cone was ordered 3 quantities in 28/07/2018 15:11 with price 4.00 and total is 12.00
"SORBET" Lemon was ordered 1 quantities in 28/07/2018 16:21 with price 2.50 and total is 2.50
"ICE CREAM" Double Fudge Chunk was ordered 2 quantities in 28/07/2018 16:39 with price 3.50 and total is 7.00
"ICE CREAM" Earl Gray was ordered 1 quantities in 28/07/2018 17:17 with price 0.50 and total is 0.50
"BEVERAGE" Espresso was ordered 3 quantities in 28/07/2018 17:30 with price 2.50 and total is 7.50
"ICE CREAM" Strawberry was ordered 1 quantities in 28/07/2018 17:32 with price 3.50 and total is 3.50
"ICE CREAM" Double Fudge Chunk was ordered 2 quantities in 28/07/2018 17:40 with price 3.50 and total is 7.00
"SORBET" Lychee was ordered 1 quantities in 28/07/2018 18:21 with price 3.00 and total is 3.00
"SORBET" Blood Orange was ordered 2 quantities in 28/07/2018 18:51 with price 2.50 and total is 5.00
```

Exercise 2

```
salesData = pd.read_csv("/content/job-market.csv")
```

Fix column DataTypes

```
salesData = salesData.dropna()
```

```
salesData
```

	Date	Location	Area	Classification	SubClassification	Requirement	FullDescription	L
	2018-10-07T00:00:00.000Z	Melbourne	Bayside & South Eastern Suburbs	Trades & Services	Welders & Boilermakers	Trade qualified person with skills in welding ...	\n * ...	
	2018-10-07T00:00:00.000Z	Brisbane	Southern Suburbs & Logan	Trades & Services	Welders & Boilermakers	Perm rate $30. Structural steel fab & weld out...	One of Australia's best engineering worksho...	
	2018-10-07T00:00:00.000Z	Sydney	CBD, Inner West & Eastern Suburbs	Education & Training	Teaching - Early Childhood	anzuk education are searching for reliable, en...	What is ...	
	2018-10-07T00:00:00.000Z	Sydney	South West & M5 Corridor	Engineering	Mechanical Engineering	This Australian Icon, connects the people of t...	This Australian Icon, connects the people o...	
	2018-10-07T00:00:00.000Z	Brisbane	CBD & Inner	Information & Communication	Networks & Systems Administration	Systems Engineer to work on	The Company This	

```
salesData['Requirement'].astype("str")
```

```
121      Trade qualified person with skills in welding ...
122      Perm rate $30. Structural steel fab & weld out...
125      anzuk education are searching for reliable, en...
126      This Australian Icon, connects the people of t...
127      Systems Engineer to work on BAU/Projects for a...
                               ...
10091    Global leaders in medical technology | Full -T...
10094    The Executive Assistant will provide high leve...
10096    Dual position within the Engineering & Commerc...
10097    Travel consultants - utilise your product know...
10098    Liaise with internal and external stakeholders...
Name: Requirement, Length: 5898, dtype: object
```

	2018-10-	Brisbane	Inner	Administration &	Administrative	within the	currently recruiting	

```
salesData['Id'] = salesData['Id'].astype(int)
```

Check and Clean Data

```
salesData['Date'] = salesData['Date'].str.split("T", 1, expand=True)[0]
```

```
<ipython-input-357-bd6f74749c05>:1: FutureWarning: In a future version of pandas all arguments of StringMethods.split except for the ar
  salesData['Date'] = salesData['Date'].str.split("T", 1, expand=True)[0]
```

```
salesData['Title'] = salesData['Title'].str.replace("|", "/")
```

```
<ipython-input-358-0dc8133b1f32>:1: FutureWarning: The default value of regex will change from True to False in a future version. In ad
  salesData['Title'] = salesData['Title'].str.replace("|", "/")
```

```
salesData['FullDescription'] = salesData['FullDescription'].str.replace("<p>", "").str.replace("</p>", "").str.replace("<strong>", "").str.re
```

```
salesData['FullDescription'] = salesData['FullDescription'].str.replace("</strong>", "").str.replace("<br>", "")
```

```
salesData['FullDescription'].str.replace(" ", "")
```

```
121      ;\n        *; Secure long term role with genui...
122      One of Australia's best engineering workshops ...
125      <p style="text-align:center;">What is anzuk?  ...
126      This Australian Icon, connects the people of t...
127      The Company  This organisation is well-establi...
                               ...
10091    A new and exciting opportunity for driven indi...
10094    V/Line, Victoria's largest regional passenger ...
10096    Randstad are currently recruiting for a versat...
10097    Travel consultants - utilise your product know...
```

```
10098    The CompanyOur client are a global market lead...
Name: FullDescription, Length: 5898, dtype: object
```

```
import re
# as per recommendation from @freylis, compile once only
CLEANR = re.compile('<.*?>')

def cleanhtml(raw_html):
  cleantext = re.sub(CLEANR, '', raw_html)
  return cleantext

for each in salesData['FullDescription']:
  each = cleanhtml(each)
```

```
salesData['FullDescription']
```

```
121       \n        *  Secure long term role ...
122      One of Australia's best engineering workshops ...
125      <p style="text-align:center;">What is anzuk?  ...
126      This Australian Icon, connects the people of t...
127      The Company  This organisation is well-establi...
                              ...
10091    A new and exciting opportunity for driven indi...
10094    V/Line, Victoria's largest regional passenger ...
10096    Randstad are currently recruiting for a versat...
10097    Travel consultants - utilise your product know...
10098    The CompanyOur client are a global market lead...
Name: FullDescription, Length: 5898, dtype: object
```

Visualize top 10 first rows

```
salesData.head(10)
```

	Id	Title	Company	Date	Location	Area	Classification	SubClassificat:
121	37404238	Fabricator/Installer	WORKPLACE ACCESS & SAFETY	2018-10-07	Melbourne	Bayside & South Eastern Suburbs	Trades & Services	Welder Boilermak
122	37404195	Boilermaker	RPM Contracting QLD P/l	2018-10-07	Brisbane	Southern Suburbs & Logan	Trades & Services	Welder Boilermak
125	37404288	Casual Childcare Positions / Bondi Junction	anzuk Education	2018-10-07	Sydney	CBD, Inner West & Eastern Suburbs	Education & Training	Teaching - Ea Childh
126	37404267	Technician	Zoom Recruitment & Training	2018-10-07	Sydney	South West & M5 Corridor	Engineering	Mechan Engineer
127	37404230	Systems Engineer	Humanised Group	2018-10-07	Brisbane	CBD & Inner Suburbs	Information & Communication Technology	Networks & Syste Administrat
129	37404237	SENIOR MARKETING & PRODUCT MANAGER	Credit Repair Australia Pty Ltd	2018-10-07	Sydney	South West & M5 Corridor	Marketing & Communications	Prod Managemer Developm
130	37404370	Operations Delivery Manager	Woolworths Group	2018-10-07	Sydney	CBD, Inner West & Eastern	Information & Communication Technology	Programm Project Managem

✓ 0 giây hoàn thành lúc 21:56