How to clean data in Python

data are not clean. Before pursuing any data analysis, cleaning data is the mandatory step. After cleaning, the data will be in a good shape and can be used for further analysis.

This tutorial will show the steps of cleaning data and then generate a donut chart like this.

Chart, sunburst chart

Description automatically generated

Importing libraries

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt

Import data

dataset = pd.read\_clipboard() #<https://en.wikipedia.org/wiki/List_of_largest_Internet_companies>dataset

Table

Description automatically generated

Above is the snapshot of the top 21 rows of the dataframe snapshot. The whole data frame has 57 rows and 9 columns.

*Rank: the rank of the company based on the revenue*

*Company: company name*

*Revenue: the annual revenue in the unit of billion US dollars*

*F.Y. :Fiscal Year*

*Employees: Number of employees*

*Market cap.($B) : market capital in the unit of billion US dollars*

*Headquarters: location of the company headquarter*

*Founded: the year company established*

*Refs: reference*

Step 1: drop column ‘Refs’

Since column ‘Refs’ has nothing to do with the following data cleaning and visualization, I will remove it from the dataset first.

dataset.drop(columns = ‘Refs’,inplace=True) #drop last column

Step 2: rename some columns

Columns ‘F.Y’ have ‘Market cap. ($B)’ have special character dot, I would like to remove those dot and make the column names easier to understand and process.

dataset.rename({‘F.Y.’: ‘FiscalYear’, ‘Market cap. ($B)’: ‘MarketCap($B)’}, axis=1,inplace=True)

Step 3: remove decimal in column ‘Founded’

Since there is a null value in the column ‘Founded’, we can not simply use round() to remove decimals. We need a lambda function here to process null value in a different way.

dataset[‘Founded’] = dataset[‘Founded’].apply(lambda x: None if pd.isnull(x) else ‘{0:.0f}’.format(pd.to\_numeric(x)))

Step 4: replace dollar sign

The data type of columns ‘Revenue ($B)’ and ‘MarketCap($B)’ is object. We need to convert the data type to string and then use str.replace function to remove dollar sign $.

#replace $  
dataset[‘Revenue ($B)’] =   
dataset[‘Revenue ($B)’].astype(str).str.replace(‘$’,’’)  
dataset[‘MarketCap($B)’] = dataset[‘MarketCap($B)’].astype(str).str.replace(‘$’,’’)

Step 5: replace comma in column ‘Employees’

Same as how we removed dollar sign above, We need to convert the data type to string and then use str.replace function to remove comma ,.

# replace ,  
dataset[‘Employees’] = dataset[‘Employees’].astype(str).str.replace(‘,’,’’)

Step 6: remove [39] in column ‘FiscalYear’ row 29**,**and [40] in column ‘MarketCap($B)’ row 29.

dataset[‘FiscalYear’] = dataset[‘FiscalYear’].str.replace(r”\[.\*\]”,””)  
dataset[‘MarketCap($B)’] = dataset[‘MarketCap($B)’].str.replace(r”\[.\*\]”,””)  
# \Used to escape a special character,   
# . Wildcard character, matches any character  
# \* Match zero, one or more of the previous

Step 6: change the datatype of column ‘Revenue ($B)’ from object to float and round the number to two decimals.

dataset[‘Revenue ($B)’] = dataset[‘Revenue ($B)’].astype(float).round(2)

Step 7: only keep columns ‘Rank’,’Company’ and ‘MarketCap($B)’.

top10 = dataset.drop([‘FiscalYear’, ‘Employees’,  
 ‘MarketCap($B)’, ‘Headquarters’, ‘Founded’], axis=1).head(10)top10

Table

Description automatically generated

Step 8: make the donut chart

# Donut chart  
labels = top10.Company  
sizes = top10[‘Revenue ($B)’]  
#explsion  
explode = (0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05) # explode every slice a little bitfig1, ax1 = plt.subplots(figsize=(10, 8))  
ax1.pie(sizes, explode = explode, labels=labels, autopct=’%1.1f%%’,  
 shadow=False, startangle=90,pctdistance=0.85)  
#draw circle  
centre\_circle = plt.Circle((0,0),0.70,fc=’white’)  
fig = plt.gcf()  
fig.gca().add\_artist(centre\_circle)  
fig.suptitle(‘Top10 Internet Company’, fontsize=20,y=1.1)  
ax1.axis(‘equal’) # Equal aspect ratio ensures that pie is drawn as a circle.  
plt.tight\_layout()  
plt.show()

Then, we will get this donut chart.

Chart, sunburst chart

Description automatically generated