تمرین پایانی تحلیل اکتشافی داده‌ها

محمدمهدی خانی

۹۷۲۴۳۰۲۷

مقدمه

در این تمرین محتوای یک پوشه که به عنوان آدرس ورودی به برنامه داده میشود بررسی میشود و تمامی فایل های زیر این پوشه در یک فایل csv قرار میگیرند. فایل csv مربوطه در مرحله بعد خوانده میشود و از روی آن دیتافریم تشکیل میشود و از روی دیتافریم مربوطه ۵ نمودار برای تحلیل داده ها رسم میشود.

Scan

در این بخش تک تک فایل هایی که زیر پوشه address هستند (چه به صورت مستقیم و چه به صورت غیر مستقیم) استخراج میشوند و داده های مورد نیاز از قبیل size، تاریخ ایجاد، تاریخ آخرین تغییر و ... از آن استخراج میشوند و در یک فایل csv ذخیره میشوند.

Scan files in given adderess

```
import os
import csv
import datetime
import time

address = "D:\SBU\Term 8"

with open('out.csv', mode='w' , newline = '') as out:
    writer = csv.writer(out, delimiter = ',',quotechar = '"', quoting = csv.QUOTE_MINIMAL)
    writer.writerow(['name', 'creation', 'format', 'modification', 'folder' ,'size'])
    for path,dirs,files in os.walk(address):
        for f in sorted(files):
            filename = os.path.splitext(f)[0]
            modification = datetime.datetime.fromtimestamp(os.path.getmtime(path + '/' + f))
            format = os.path.splitext(f)[1]
            creation = datetime.datetime.fromtimestamp(os.path.getctime(path + '/' + f))
            folder = path.split('\\')[-1]
            size = os.path.getsize(path + '/' + f)
            writer.writerow(["%s" % filename, "%s" % creation, "%s" % format, "%s" % modification, "%s" % folder, "%s" %  size])
```

Imports and set rcParams

Panads و ... را import میکنیم و برای اینکه هر دفعه متغییر های ویژگی های نمودار ها را ست نکنیم یکبار آن را در ابتدای کار انجام میدهیم.

Import Pandas, Numpy and set rcParams

```
In [10]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import matplotlib as mpl
         from cycler import cycler
         import matplotlib
         import seaborn as sns
         matplotlib.rcParams['font.size'] = 18
         matplotlib.rcParams['figure.dpi'] = 200
         matplotlib.rcParams['font.family'] = 'tahoma'
         matplotlib.rcParams['font.weight'] = 'normal'
         matplotlib.rcParams['figure.figsize'] = (80,60)
```

Read csv

در این قسمت فایل csv را میخوانیم و از روی آن دیتافریم df را تشکیل میدهیم.

Read CSV file and create Datagram

```
df = pd.read_csv('out.csv')
df = df.dropna()
```

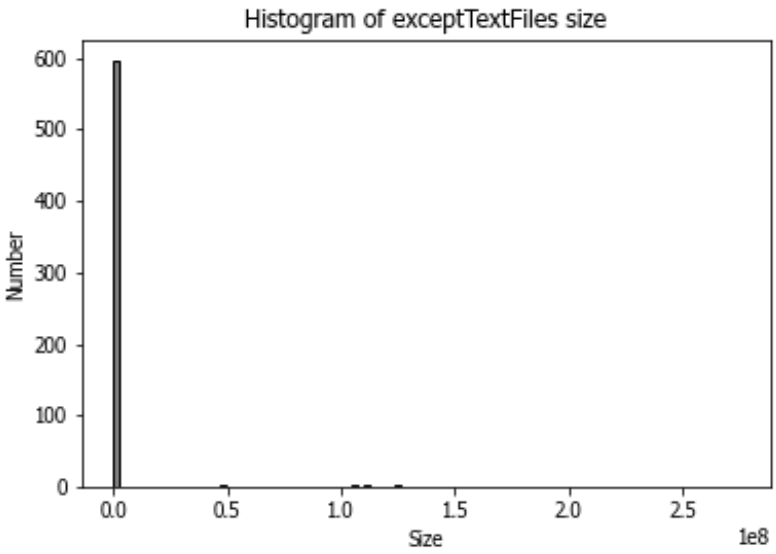
تشکیل دیتافریمی از فایل های با پسوند غیر txt.

make new dataframe with conditioning

```
exceptTextFiles = df[df.format != ".txt"]
exceptTextFiles
```

	name	creation	format	modification	folder	size
1	EXAM_DBLAB	2022-05-09 16:06:06.967873	.pdf	2021-12-28 18:43:58.664631	exam	226171
2	EXAM_DBLAB_97243081	2022-05-09 16:06:07.079806	.docx	2021-12-29 23:56:50.241467	exam	22050
3	EXAM_DBLAB_97243081	2022-05-09 16:06:07.197753	.pdf	2021-12-29 23:57:26.642171	exam	697852
4	Homework_01	2022-05-09 16:06:07.412600	.pdf	2021-10-27 10:02:20	HW1	160724
5	dblab_hw1_97243081	2022-05-09 16:06:07.301669	.docx	2021-11-01 23:39:07.201535	HW1	15497
...	...	...	...	...	...	...
665	OS_project	2022-05-09 16:08:13.505031	.pptx	2021-12-29 16:32:49.198159	q1	1086035
666	azos	2022-05-09 16:08:08.531090	.docx	2021-12-29 16:59:51.543673	q1	12701
667	danny_detection	2022-05-09 16:08:08.535088	.mp4	2021-12-31 15:33:00.269259	q1	242210066
668	eraeye bachcheha	2022-05-09 16:08:10.742731	.rar	2021-12-26 12:30:37.984761	q1	183245654

نمودار هیستوگرام این سایز دیتافریم



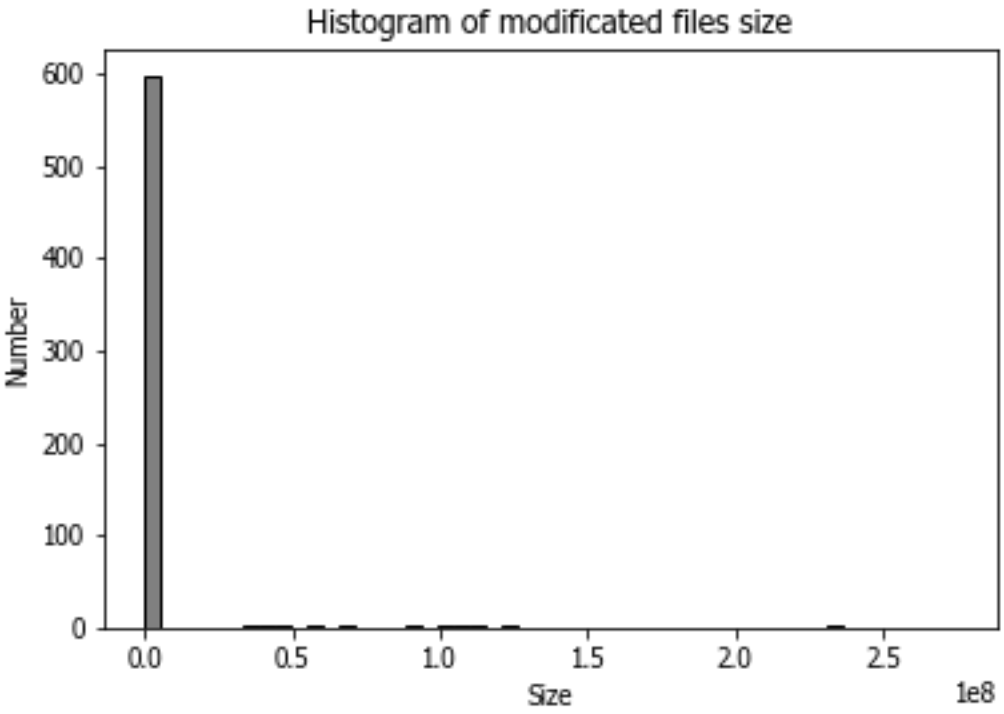
تشکیل یک سری جدید در دیتافریم به نام isChanged

این سری نشان میدهد که آیا فایل ها از زمان ایجاد تغییر کرده اند یا خیر.

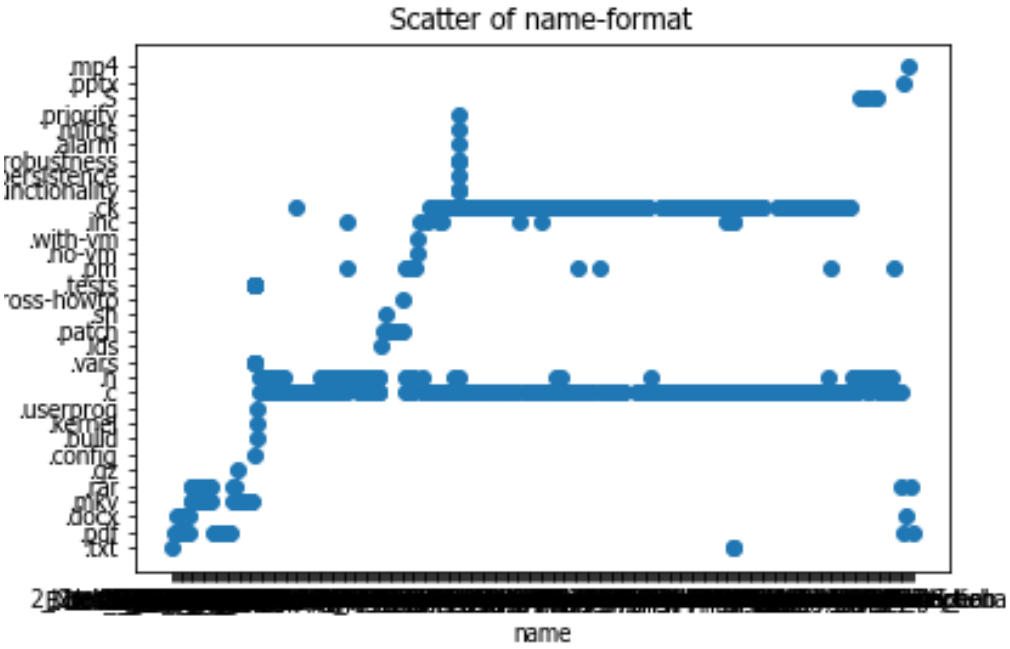
Create new Col ¶

```
df["isChanged"] = df["modification"] != df["creation"]
```

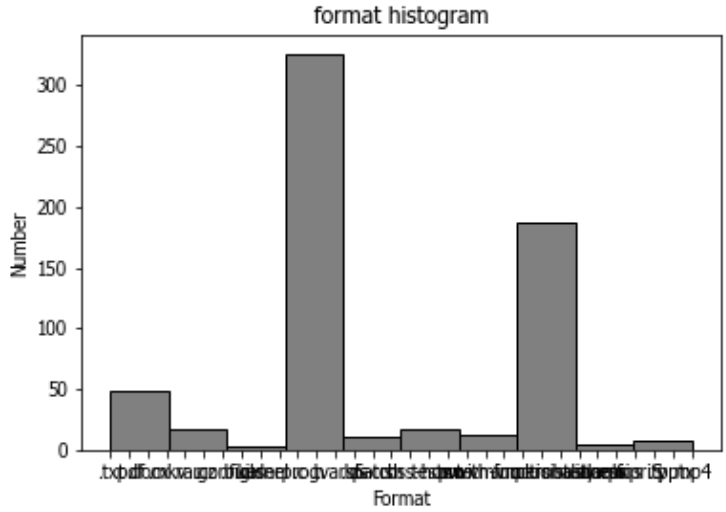
نمودار هیستوگرام سایز داده هایی که از زمان ایجاد تغییر کرده اند



نمودار scatter برای name-format



نمودار هیستوگرام فرمت



نمودار density برای size