Homework 5

import numpy as np

import pandas as pd

df=pd.read\_csv("D:/Old\_Data/math/Data science toseeh/Files/digi2.csv",encoding='UTF-8')

#1

df.columns

output: 'ID\_Order', 'ID\_Customer', 'ID\_Item', 'DateTime\_CartFinalize', 'Amount\_Gross\_Order','city\_name\_fa', 'Quantity\_item', 'Year'],dtype='object'

#2

np.sum(df.isna())

Out[4]:

ID\_Order 0

ID\_Customer 0

ID\_Item 0

DateTime\_CartFinalize 0

Amount\_Gross\_Order 0

city\_name\_fa 0

Quantity\_item 0

dtype: int64

#3

df.drop(labels="Amount\_Gross\_Order",axis=1,inplace=True)

df.columns

output:

[ID\_Order', 'ID\_Customer', 'ID\_Item', 'DateTime\_CartFinalize',

'city\_name\_fa', 'Quantity\_item'],dtype='object')

#4

df.drop\_duplicates(inplace=True)

len(df.index)

output: 199993

#5

df["DateTime\_CartFinalize"]=df["DateTime\_CartFinalize"].agg(pd.Timestamp)

#JalaliDate

from persiantools.jdatetime import JalaliDate

JalaliDate(pd.Timestamp("1/17/2020"))

Output: JalaliDate(1398, 10, 27, Jomeh)

JalaliDate(pd.Timestamp('1/6/2020')).year

Output:1398

JalaliDate(pd.Timestamp('1/6/2020')).month

Output:10

JalaliDate(pd.Timestamp('1/6/2020')).day

Output:16

df["DateTime\_CartFinalize"].agg(JalaliDate)

اگر بخواهیم این ستون در داده کاملا تغییر کند باید بنویسیم

df["DateTime\_CartFinalize"]=df["DateTime\_CartFinalize"].agg(JalaliDate)

#6

def ghadim\_jadid(x):

x=pd.Timestamp(x)

baze=(x.now()-x).days

return(baze)

fasele=df["DateTime\_CartFinalize"].agg(ghadim\_jadid)

df["DateTime\_CartFinalize"][fasele==fasele.max()]

output: 2013-09-27

df["DateTime\_CartFinalize"][fasele==fasele.min()]

output: 2018-12-10

#7

len(np.unique(df["ID\_Item"]))

or (len(df["ID\_Item"].unique()))

output:95232

len(np.unique(df["ID\_Customer"]))

Output: 151634

#8

len(np.unique(df["city\_name\_fa"]))

output: 906

#9

city=df.groupby("city\_name\_fa").agg({"ID\_Order":["count"]})

sort=city.sort\_values(("ID\_Order","count"),ascending=False)

sort.head(20)

output:



#10

def yearfunction(x):

y=pd.Timestamp(x).year

return(y)

df["Year"]=df["DateTime\_CartFinalize"].agg(yearfunction)

yeargroup=df.groupby("Year").agg({"ID\_Order":["count"]})

sortyear=yeargroup.sort\_values(("ID\_Order","count"),ascending=False)

sortyear.head()

output:

2018 75936

2017 61018

2016 39208

2015 19263

2014 4283

2013 285

#11

year\_city=df.groupby(["city\_name\_fa","Year"]).count()["ID\_Order"]

year\_city.get("تهران")

Out[73]:

Year

2013 140

2014 2274

2015 10545

2016 23121

2017 34130

2018 38093

Name: ID\_Order, dtype: int64

year\_city.get("اصفهان")

2013 6

2014 99

2015 622

2016 1156

2017 1689

2018 2420

#12

np.unique(df.iloc[:,5])

df.iloc[:,5].value\_counts().head(20).index

output:

Int64Index([1, 2, 3, 4, 5, 10, 6, 8, 7, 30, 20, 15, 12, 9, 11, 14, 13, 16, 25,

17],

dtype='int64')

#13

s=df.groupby("city\_name\_fa")["Quantity\_item"]

np.unique(s.get\_group("مشهد"))

output:

array([ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 15, 20, 22, 26, 30,

47], dtype=int64)

np.unique(s.get\_group("رشت"))

Out[85]:

array([ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 17, 20, 30],

dtype=int64)

#14

syear=df.groupby("Year")["Quantity\_item"]

np.unique(syear.get\_group(2018))

Out[87]:

array([ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,

14, 15, 16, 17, 18, 20, 22, 24, 25, 26, 27, 30, 40,

50, 200], dtype=int64)

#15

customer\_order=df.groupby("ID\_Customer").agg({"ID\_Order":["count"]})

customer\_order.get\_value(466132,("ID\_Order","count"))

output: 11

#16

df["ID\_Item"].value\_counts(ascending=False).head(20).index

output:

Int64Index([294942, 36871, 51778, 45121, 8289, 416448, 19890, 22839,

294943, 153067, 7225, 42124, 294923, 118375, 129574, 12532,

77707, 245223, 20215, 143130],

dtype='int64')

#17

Top=df.groupby('city\_name\_fa')["ID\_Item"].value\_counts(ascending=False).head()

چون به ازای یک شهر ممکن است چندین آیدی آیتم داشته باشیم نمیتوان دیتا فریم تعریف کرد و از دیکشنری استفاده میکنیم که نظیر هر کلید چندین مقدار میتواند داشته باشد.

b=dict()

for i in np.unique(df['city\_name\_fa']):

b[i]=df.groupby('city\_name\_fa').get\_group(i)["ID\_Item"].value\_counts(ascending=False).index[:5]

b["تهران"]

Int64Index([294942, 36871, 51778, 45121, 8289], dtype='int64')

b["محمود آباد"]

Int64Index([66484, 242685, 381617, 1042626, 67262], dtype='int64')