1

create following data frames

```
import pandas as pd

import numpy as np

d1=pd.DataFrame({'v1':np.random.random(size=50),'v2':np.random.choice(range(10
0),size=50)})

d2=pd.DataFrame({'v1':np.random.random(size=50),'v2':np.random.choice(range(2,
300),size=50)})
```

combine these dataframes to create a larger dataframe d3 with 100 observations and then sort the dataframe with column v1.

Hints :

- use function pd.concat for combining, chose appropriate value for option axis for combining them by rows
- use function sort_values for sorting

2

using data frame d3, calculate mean of column v2, ensuring values from only dataframe d1 are used

Hints :

- before combining dataframes d1 and d2, add a column name 'data' to them d1['data']='d1' and d2['data']='d2' . You can use this column to differentiate between observations from d1 and d2 in the larger dataframe d3
- use .loc with the dataframe to conditionally filter and refer to column v2 before applying function mean to calculate mean

3

add a column v3 to dataframe d3 such that it takes value 0 when v1>0.5 and value log(v2) othwerwise

Hints :

- make use of function np.where

4

Separate dataframe d3 into d1 and d2 again

Hints :

- make use of column 'data' and revise how to conditionally filter the data

5

Read file rg_train.csv as pandas data frame. Extract names of all categorical columns in the file

Hints :

- use function pd.read_csv
- use function select_dtypes on the dataframe

6

For the data frame that you read in exercise 6, find out categories in column Region which have frequency higher than 5000.

Hints :

- calculate frequencies using function value_counts on the column apply condition on the index of the result from value_counts

7

Find out names of variables in the dataframe that you read in exercise 6 which have less than 10 unique values

Hints :

- use function nunique on the data frame
- apply condition on the index of the above result

8

Find out percentage of values of Revenue.Grid across categories of TVarea . Hints :

- use pd.crosstab for calculating raw counts
- experiment with values for argument normalize in function crosstab . it takes three values True,'columns','index'. see what these options do