2.​ Problem Statement

Read the dataset from the below link

https://raw.githubusercontent.com/guipsamora/pandas\_exercises/master/06\_Stats/US

\_Baby\_Names/US\_Baby\_Names\_right.csv

Code:

import numpy as np

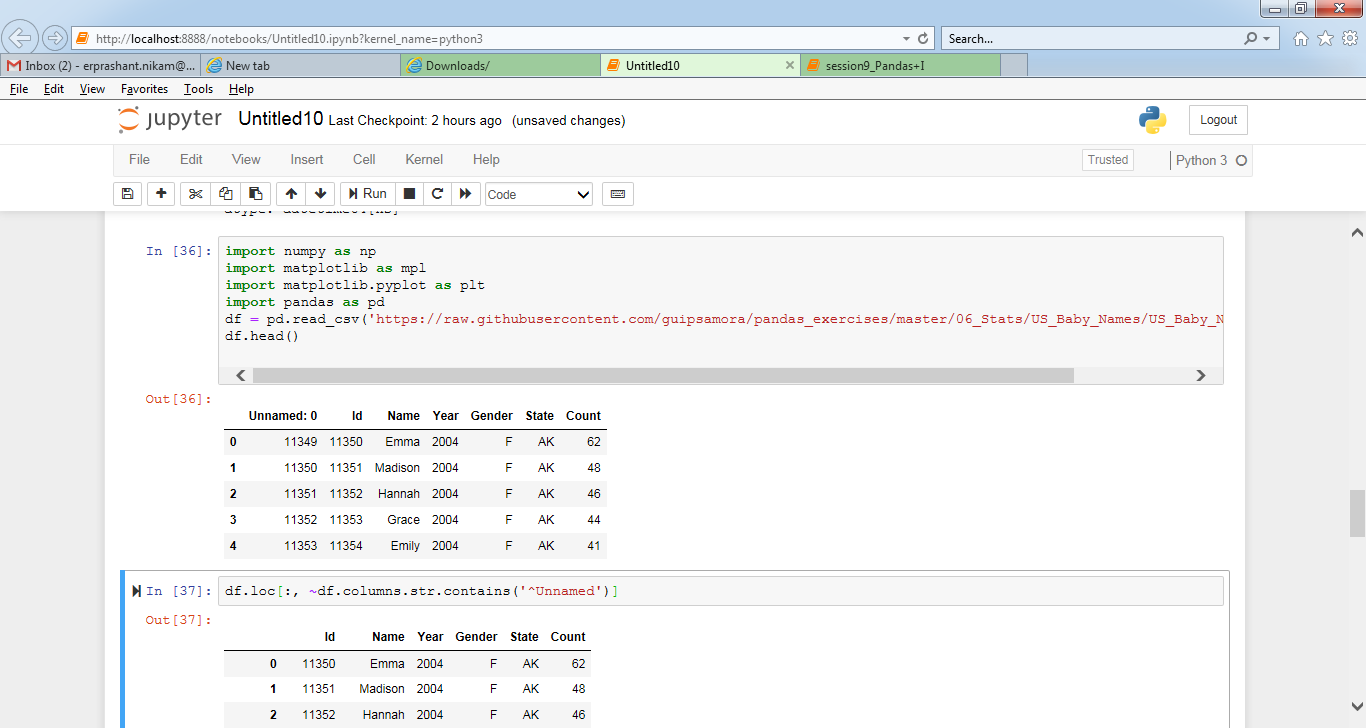
import matplotlib as mpl

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read\_csv('https://raw.githubusercontent.com/guipsamora/pandas\_exercises/master/06\_Stats/US\_Baby\_Names/US\_Baby\_Names\_right.csv')

df.head()



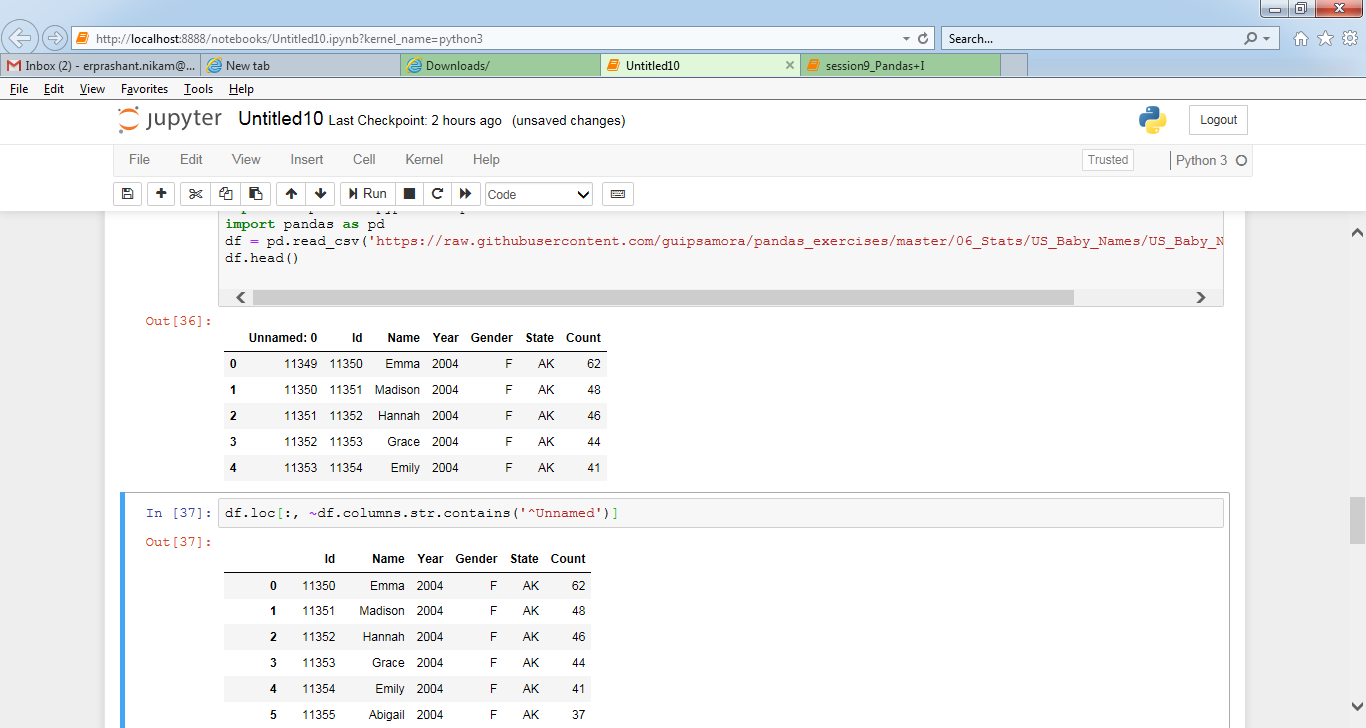
Questions:

1. Delete unnamed columns

Code:

df.loc[:, ~df.columns.str.contains('^Unnamed')]

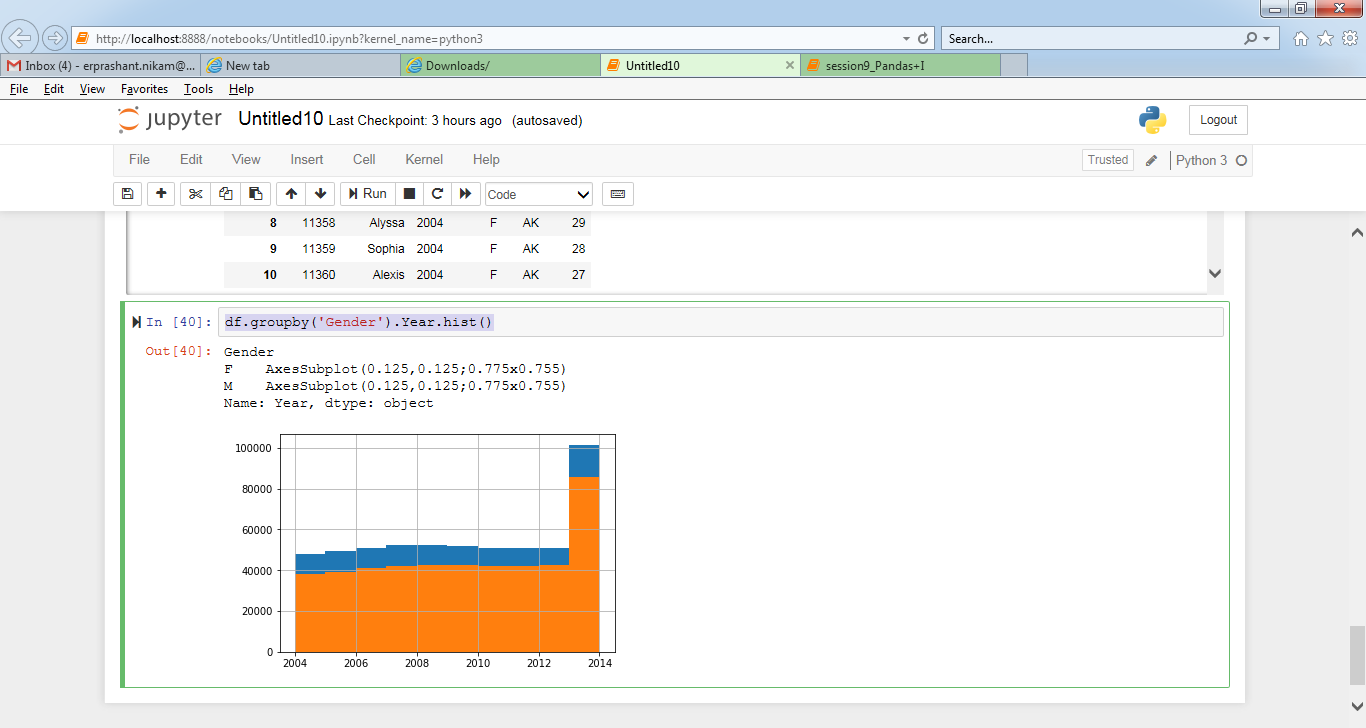
OutPut : << here i just removed from display , and not dropped from df >



1. Show the distribution of male and female

Code:

df.groupby('Gender').Year.hist()



1. Show the top 5 most preferred names

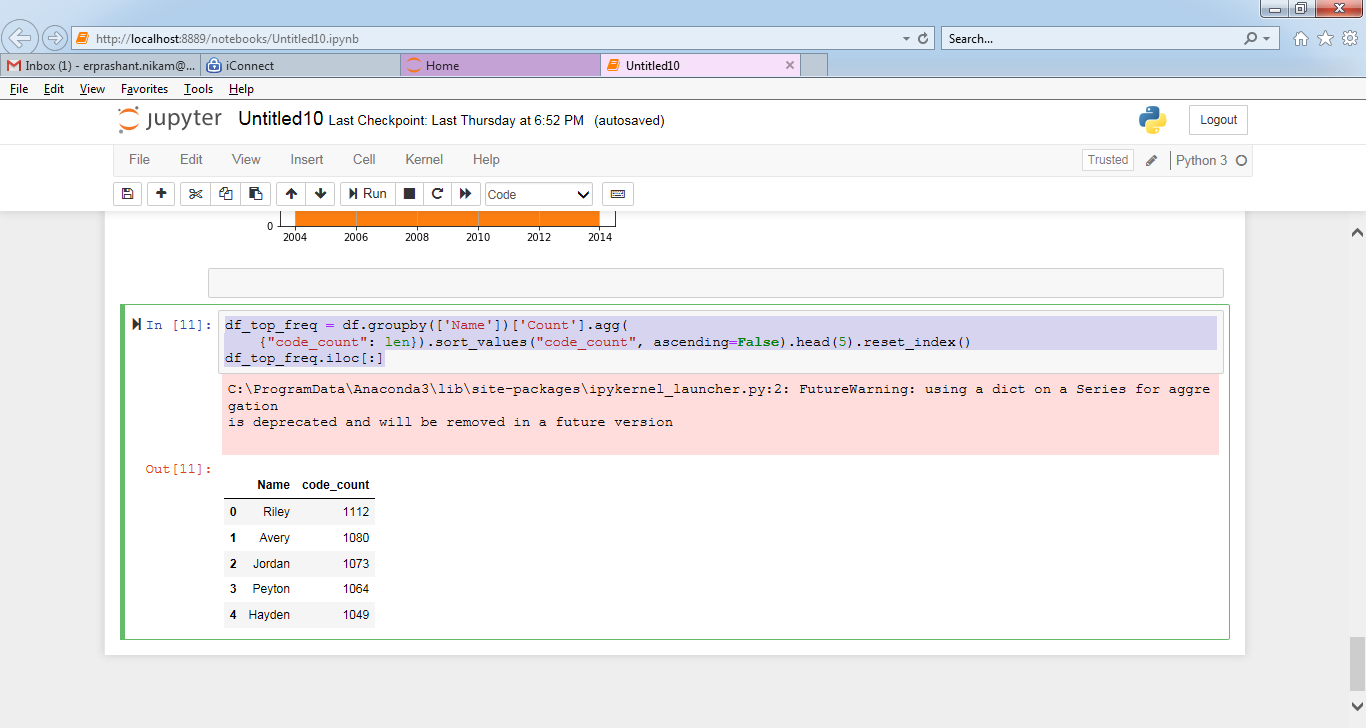
Code:

df\_top\_freq = df.groupby(['Name'])['Count'].agg(

{"code\_count": len}).sort\_values("code\_count", ascending=False).head(5).reset\_index()

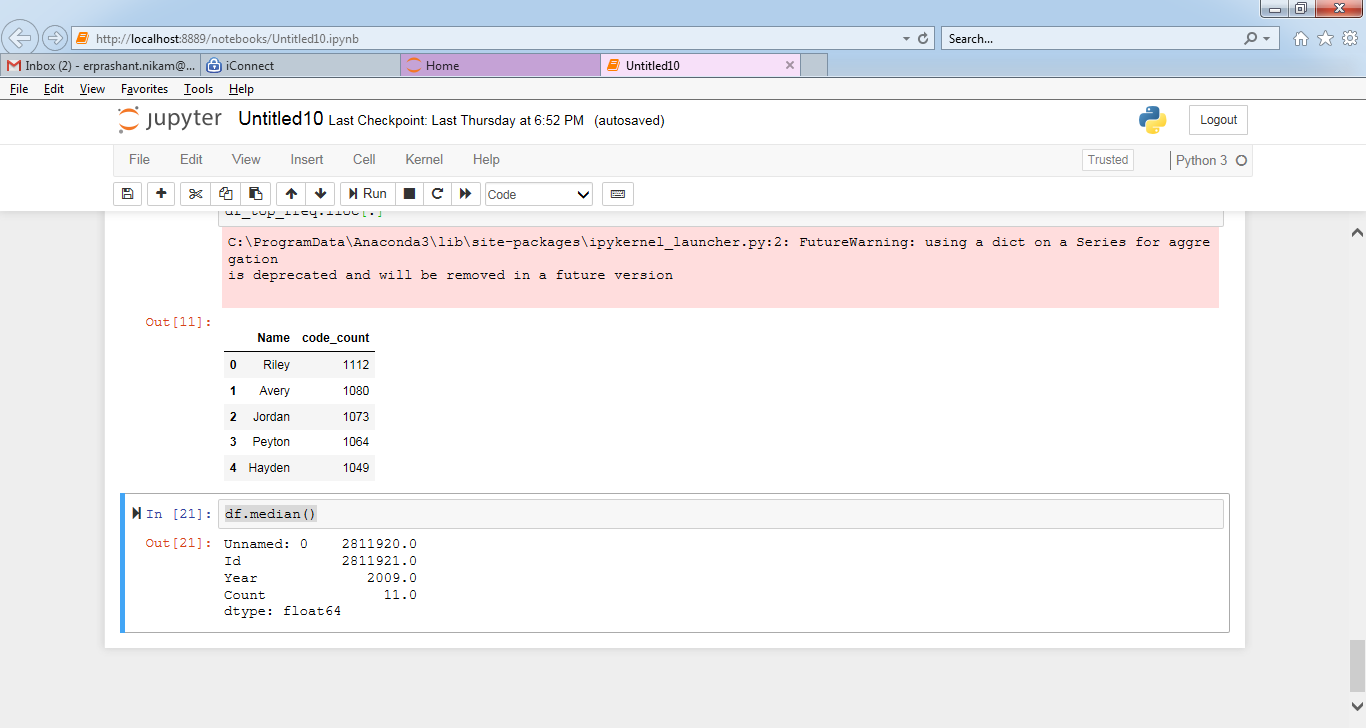
df\_top\_freq.iloc[:]

Output:



1. What is the median name occurence in the dataset

Code: df.median()

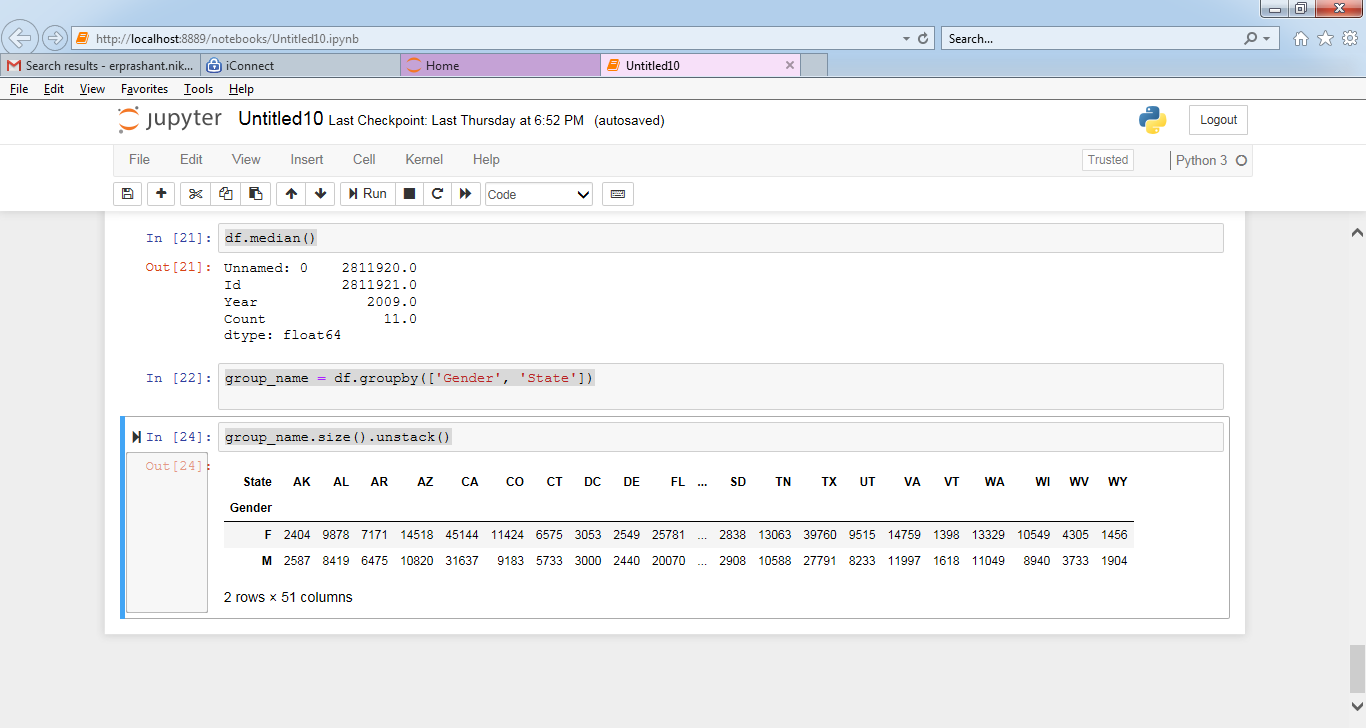


5. Distribution of male and female born count by states

Code:

group\_name = df.groupby(['Gender', 'State'])

group\_name.size().unstack()



3. Expected Output

Note: Solution submitted via github must contain all the source code and output.