Assignment 9

**Problem Statement 1:**

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

Questions:

1. Delete unnamed columns

2. Show the distribution of male and female

3. Show the top 5 most preferred names

4. What is the median name occurrence in the dataset

5. Distribution of male and female born count by states

**Code:**

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

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# Read in small data from .csv file

# Filepath

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

print(df)

print('Shape', df.shape)

print('-------------------------')

print('Number of rows', len(df))

print('-------------------------')

print('Column headers', df.columns)

print('-------------------------')

print('Data types', df.dtypes)

print('-------------------------')

print('Index', df.index)

print('-------------------------')

1. Delete unnamed columns

# Drop Unnamed columns - Column 0

#df = df.drop('Unnamed: 0', axis =1, inplace = true)

df = df.drop('Unnamed: 0', 1)

print(df)

2. Show the distribution of male and female

# Show the distribution of male and female

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

print(df)

#Show the distribution of male and female

df.hist(column="Year",by="Gender",bins=30)

3. Show the top 5 most preferred names

#Show the top 5 most preferred names

df1= df['Name'].value\_counts()

df1.head(5)

4. What is the median name occurrence in the dataset

# What is the median name occurrence in the dataset

df1= df['Name'].value\_counts()

print(df1)

print("Median name occurrence in the dataset:",df1.median())

5. Distribution of male and female born count by states

# Distribution of male and female born count by states

impute\_grps = df.pivot\_table(values=["Count"], index=["Gender","State"], aggfunc=np.mean)

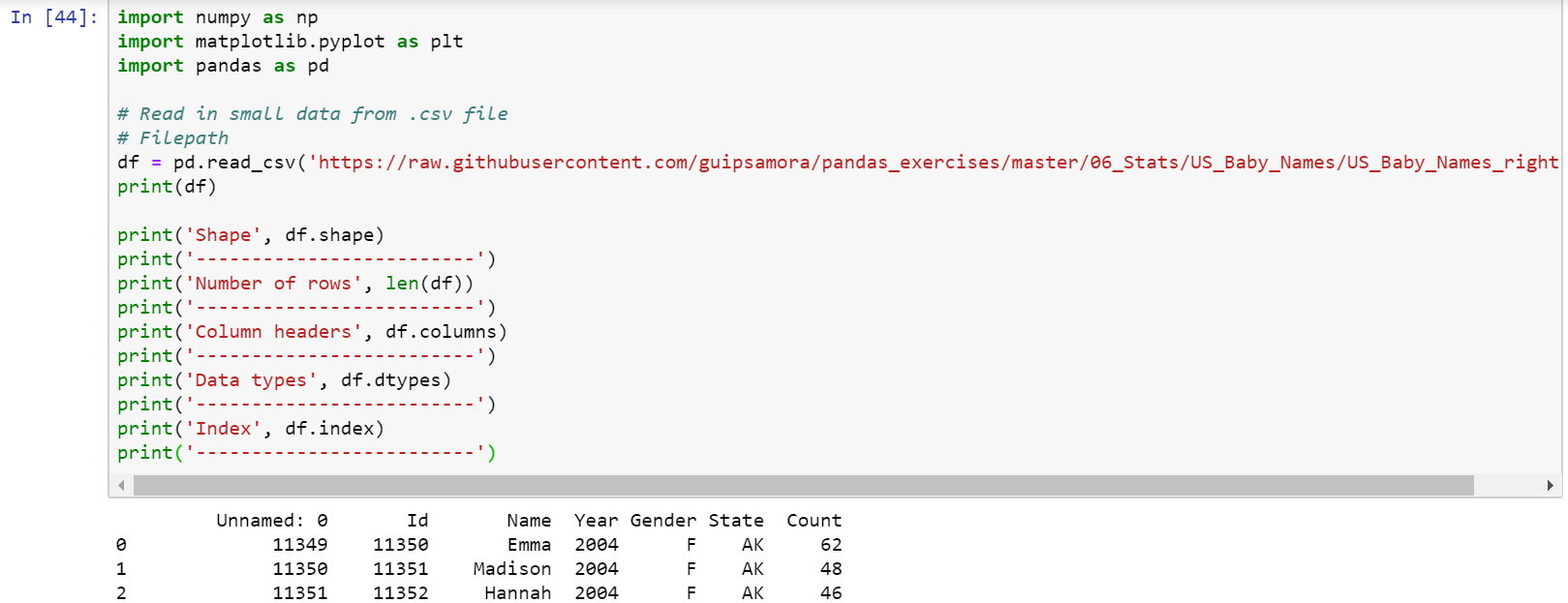
print(impute\_grps)

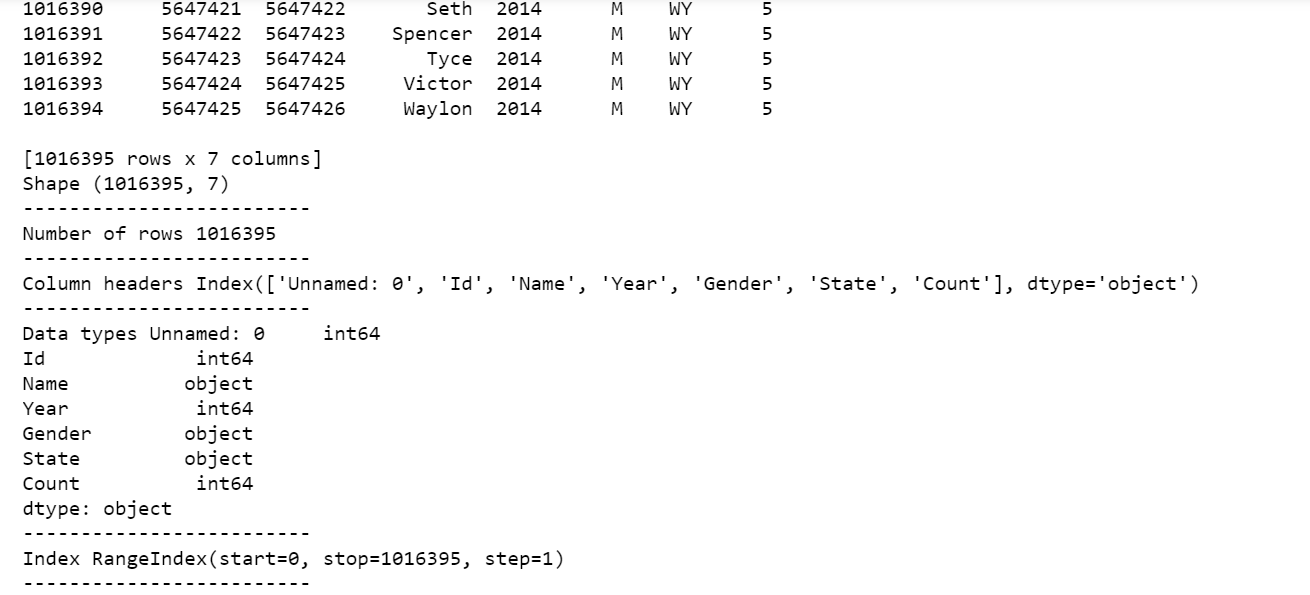
**Screenshot**

Read the dataset from the below link

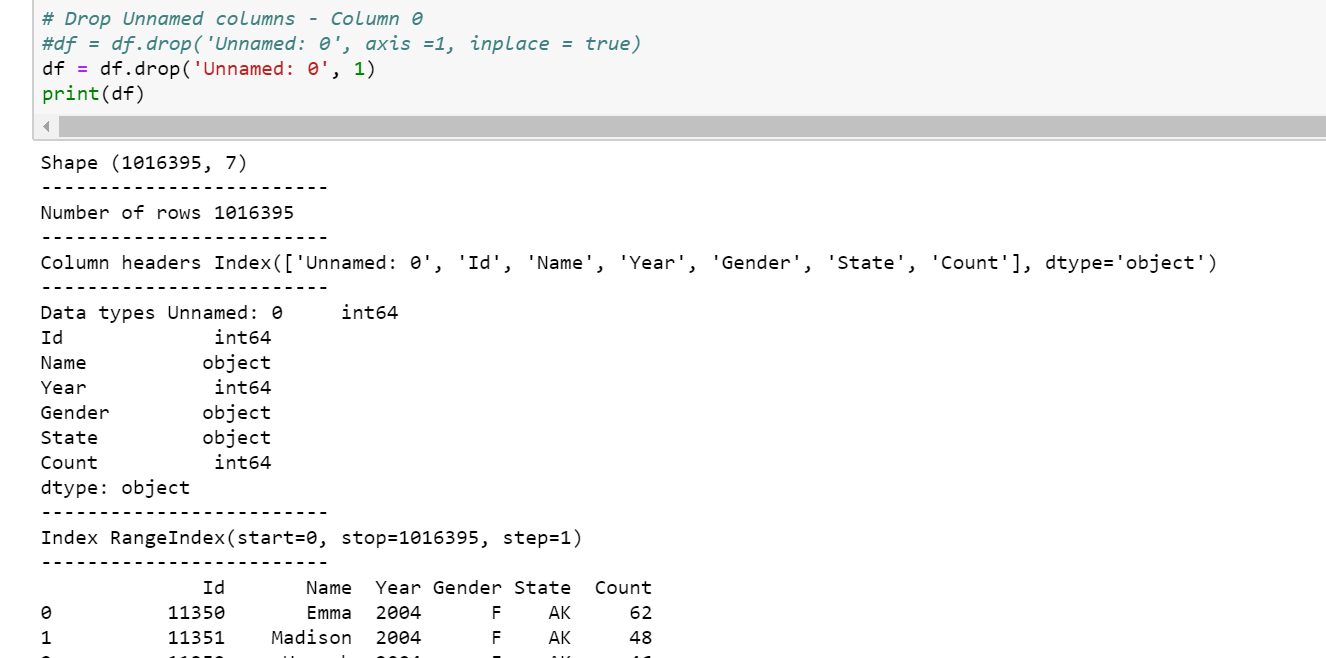
https://raw.githubusercontent.com/guipsamora/pandas\_exercises/master/06\_Stats/US

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

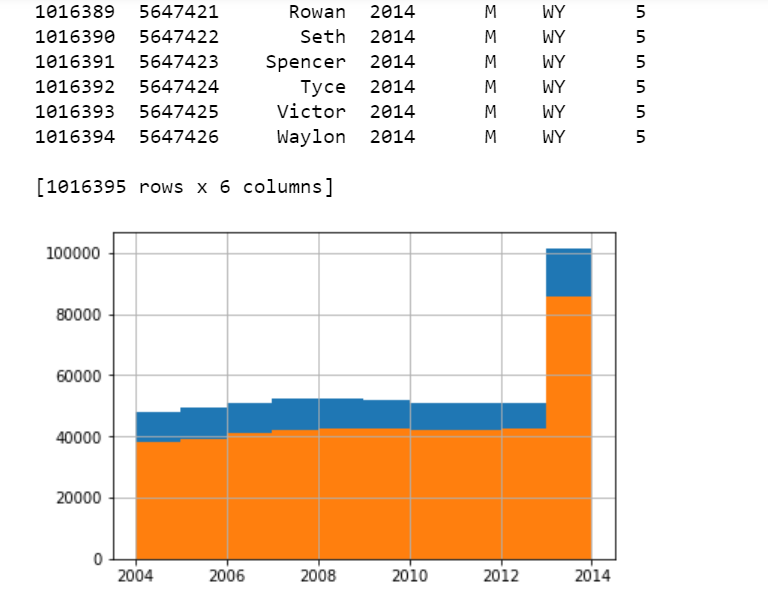


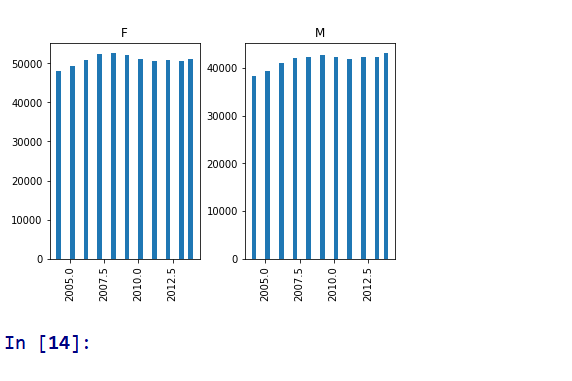


1. Delete unnamed columns

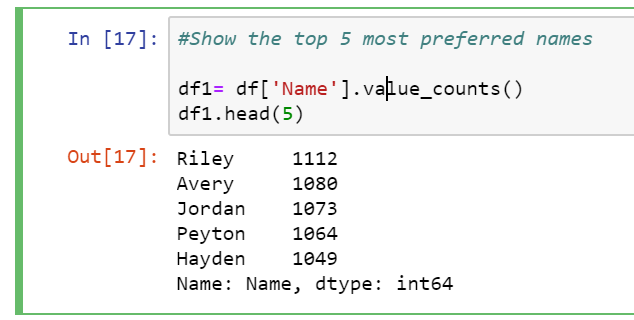


2. Show the distribution of male and female

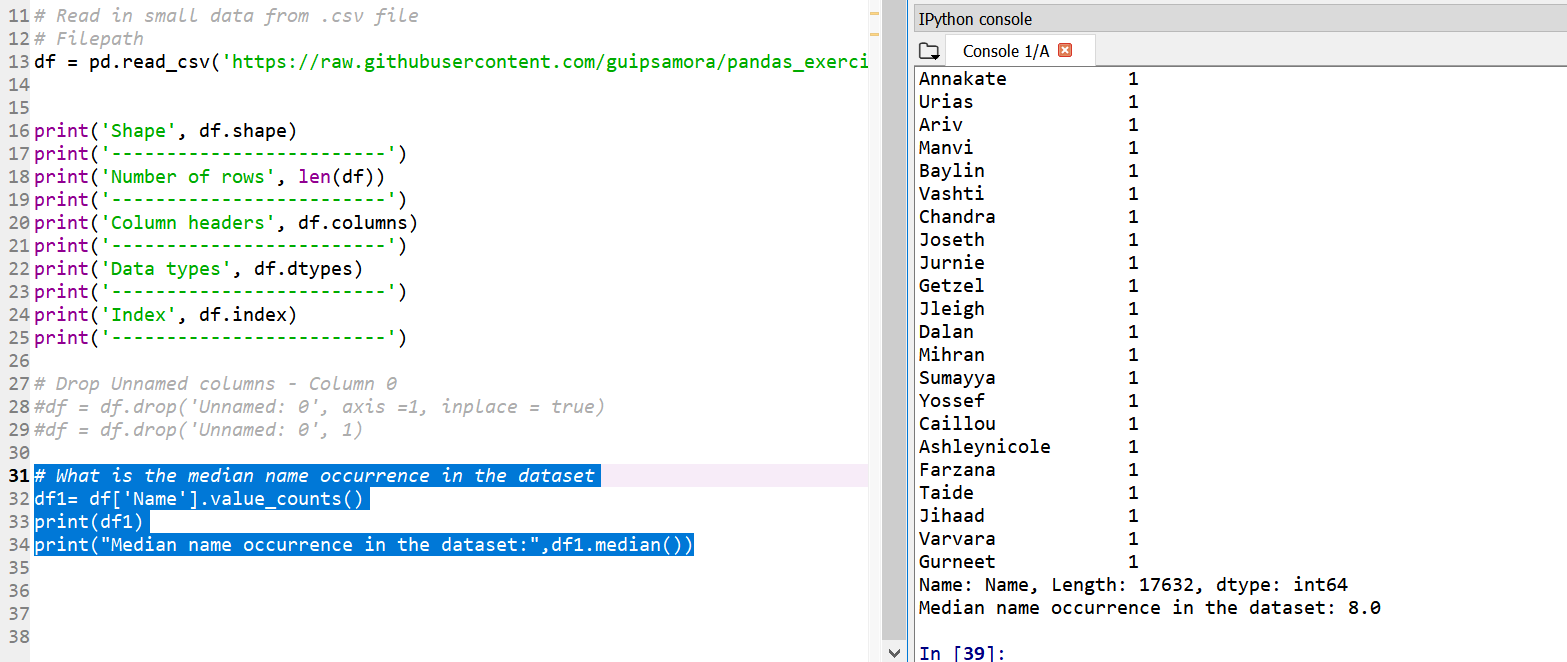




3. Show the top 5 most preferred names



4. What is the median name occurrence in the dataset



5. Distribution of male and female born count by states

