Day5_Exc_Key

Question 1

Write Pandas program to read the given “employees.csv” file, then accomplish the following functions:

- Fill all the null values in Gender column with “No Gender”
- Drop all the records have no “Name” Value
- Use z-score to normalize the “Salary” column.
- Detect any outlier in the given “Bonus” data.

```
#Start by uploding the file to CoLab

from google.colab import files
uploaded = files.upload()
```

```
# Import the needed libraries
import pandas as pd
import numpy as np

# Read the uploaded file to a new data frame
dataset = pd.read_csv("employees.csv")

# To see information about the dataset
print(dataset.columns)

print("\n The dataset has ", len(dataset), " recoreds")
```

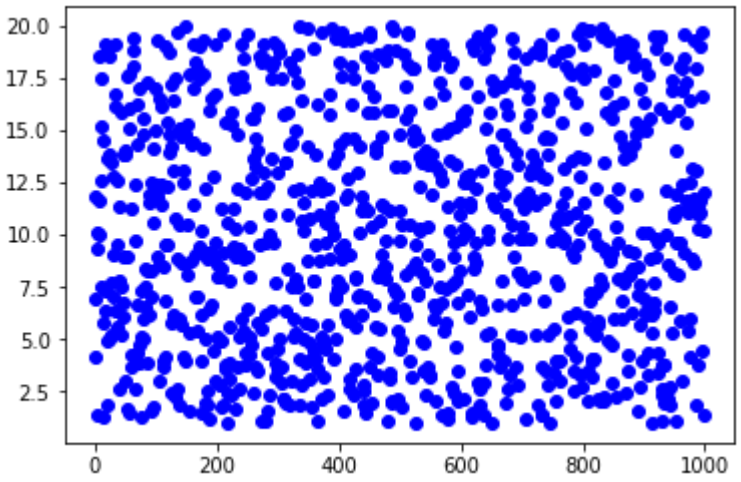
```
Index(['Name', 'Gender', 'Start Date', 'Last Login Time', 'Salary', 'Bonus %',
       'Senior Management', 'Team', 'Position'],
      dtype='object')

 The dataset has  1000  recoreds
```

```
# Fill all the null values in Gender column with “No Gender”
dataset['Gender'].replace(np.nan, 'No Gender', inplace=True)
```

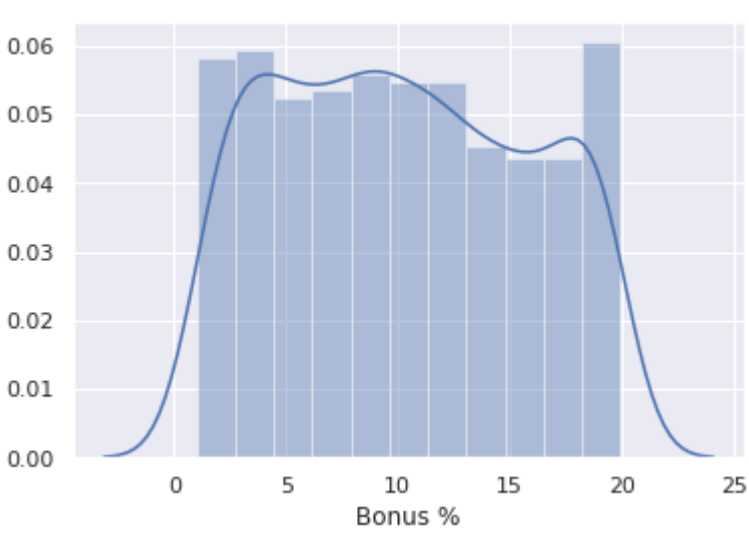
```
# Drop all the records have no “Name” Value
dataset['Name'].dropna(inplace=True)
```

```
# Use z-score to normalize the “Salary” column.

from scipy.stats import zscore

dataset['Salary'] = zscore(dataset['Salary'])
```

```
#Detect any outlier in the given “Bonus” data.

# The easiest way is to plot the given values
import matplotlib.pyplot as plt

plt.plot(dataset['Bonus %'], 'o', color='blue');
plt.show()
```




```
# Another way

import seaborn as sns
from scipy import stats

sns.set(color_codes=True)
sns.distplot(dataset['Bonus %']);
```



Question 2:

This study analyzes gun deaths in the United States of America between 2012 and 2014.

The data set for this study comes from GitHub and can be accessed here:
https://github.com/fivethirtyeight/guns-data.git

Load and clean the dataset and prepare it for processing.

```
# First bring the data folder from GITHUB

!git clone https://github.com/fivethirtyeight/guns-data.git
```

```
Cloning into 'guns-data'...
remote: Enumerating objects: 43, done.
remote: Total 43 (delta 0), reused 0 (delta 0), pack-reused 43
```

```
# to access the path of the CSV file
cd guns-data
```

```
/content/guns-data
```

```
import pandas as pd
import numpy as np

dataq2= pd.read_csv("full_data.csv")
print(" A new dataframe has been created... ")

all_records=len(dataq2)
print("Total number of recorrds = ", all_records)
print("the columns are :", dataq2.columns)
```

```
 A new dataframe has been created...
Total number of recorrds =  100798
the columns are : Index(['Unnamed: 0', 'year', 'month', 'intent', 'police', 'sex', 'ag
       'hispanic', 'place', 'education'],
      dtype='object')
```

```
# drop all the records have ,issing data
dataq2.dropna(inplace=True)
rem_records=len(dataq2)
print(" Records have been dropped = ", all_records- rem_records )
print(" Remaining records = ",rem_records )
```

```
 Records have been dropped =  2783
 Remaining records =  98015
```

```
# Check for possible outliers in the data (hispanic column)

import seaborn as sns
from scipy import stats

sns.set(color_codes=True)
sns.distplot(dataq2['hispanic']);
```


```
# To normalize the (hispanic column)
from scipy.stats import zscore
dataq2['hispanic'] = zscore(dataq2['hispanic'])
```


Case Study:

This study analyses the leading causes of death in the United States of America between 1999 and 2015.

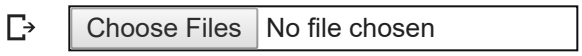
The data set in this case study comes from open data from the U.S. government, which can be accessed through https://data.gov.

You can download it from here:

https://catalog.data.gov/dataset/age-adjusted-death-rates-for-the-top-10-leading-causes-of-death-united-states-2013

- What is the total number of records in the dataset?
- Drop all records with NA cases
- Check the size of your dataset again.
- What were the causes of death in this data set?
- What was the total number of deaths in the United States from 1999 to 2015?
- What is the number of deaths per each year from 1999 to 2015?
- Which ten states had the highest number of deaths overall?
- What were the top causes of deaths in the United States during this period?

```
# First bring the data

from google.colab import files
uploaded = files.upload()
```



Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving NCHS_US.csv to NCHS_US.csv
```

```
import pandas as pd
import numpy as np

studydata= pd.read_csv("NCHS_US.csv")
print(" A new dataframe has been created... ")
print("the columns are :", studydata.columns)
```

```
 A new dataframe has been created...
the columns are : Index(['Year', '113 Cause Name', 'Cause Name', 'State', 'Deaths',
       'Age-adjusted Death Rate'],
      dtype='object')
```

```
# What is the total number of records in the dataset?
print("Total number of recorrds = ", len(studydata))
```

```
Total number of recorrds =  10296
```

```
# Drop all records with NA cases
studydata.dropna(inplace=True)
print("Total number of recorrds after dropping = ", len(studydata))
```

```
Total number of recorrds after dropping =  10296
```

```
#What were the causes of death in this data set?
causes=studydata['Cause Name'].unique()
print('The number of death causes ', len(causes))
print(' The list of causes include:', causes)
```

```
The number of death causes  11
 The list of causes include: ['Kidney disease' 'Suicide' "Alzheimer's disease"
 'Influenza and pneumonia' 'Diabetes' 'CLRD' 'Unintentional injuries'
 'Stroke' 'Heart disease' 'Cancer' 'All causes']
```

```
# notice that the last cause is callaed "All causes", to exclude this one:

studydata=studydata[studydata['Cause Name'] != 'All causes']
causes=studydata['Cause Name'].unique()
print('The number of death causes ', len(causes))
print(' The list of causes include:', causes)
```

```
The number of death causes  10
 The list of causes include: ['Kidney disease' 'Suicide' "Alzheimer's disease"
 'Influenza and pneumonia' 'Diabetes' 'CLRD' 'Unintentional injuries'
 'Stroke' 'Heart disease' 'Cancer']
```

```
# What was the total number of deaths in the United States from 1999 to 2015?
partial_data=studydata[studydata['Year'] != 2016]
num_deaths= partial_data['Deaths'].sum()
print("The total number of deaths in the United States from 1999 to 2015 = ", num_deaths)
```

```
The total number of deaths in the United States from 1999 to 2015 =  64329866
```

```
# What is the number of deaths per each year from 1999 to 2015?
dyear = partial_data.groupby(['Year']).sum()
print('The number of deaths per each year from 1999 to 2015?: \n')
dyear
```

```
#* Which ten states had the highest number of deaths overall?
states = studydata.groupby(['State'])
sumbystate = states['Deaths'].agg(np.sum).reset_index()
sumbystate.nlargest(10,'Deaths')
```

```
#* What were the top causes of deaths in the United States during this period?
causes_groups = studydata.groupby('Cause Name')
sumbycause=causes_groups['Deaths'].agg(np.sum).reset_index()
sumbycause.nlargest(10,'Deaths')
```