MODULE – 7 ASSIGNMENT

Python for data analytics

Please implement Python coding for all the problems.

1) Please take care of missing data present in the "*Data.csv*" file using python module "sklearn.impute" and its methods, also collect all the data that has "Salary" less than "70,000".?

```
import pandas as pd
from sklearn.impute import SimpleImputer
data = pd.read_csv(r'Data.csv')
numeric_columns = data.select_dtypes(include=['number']).columns
imputer = SimpleImputer(strategy='mean')
data[numeric_columns] = imputer.fit_transform(data[numeric_columns])
filtered_data = data[data['Salaries'] < 70000]
filtered_data.to_csv('Filtered_Data.csv', index=False)
```

2) Subtracting dates:

Python date objects let us treat calendar dates as something similar to numbers: we can compare them, sort them, add, and even subtract them. Do math with dates in a way that would be a pain to do by hand. The 2007 Florida hurricane season was one of the busiest on record, with 8 hurricanes in one year. The first one hit on May 9th, 2007, and the last one hit on December 13th, 2007. How many days elapsed between the first and last hurricane in 2007?

Instructions:

Import date from datetime.

Create a date object for May 9th, 2007, and assign it to the start variable.

Create a date object for December 13th, 2007, and assign it to the end variable.

Subtract start from end, to print the number of days in the resulting timedelta object.

Sol:

```
from datetime import date

start = date(2007, 5, 9)

end = date(2007, 12, 13)

delta = end - start
```

```
print(f"Number of days between {start} and {end}: {delta.days}")
```

Output:-Number of days between 2007-05-09 and 2007-12-13: 218

3) Representing dates in different ways

Date objects in Python have a great number of ways they can be printed out as strings. In some cases, you want to know the date in a clear, language-agnostic format. In other cases, you want something which can fit into a paragraph and flow naturally.

Print out the same date, August 26, 1992 (the day that Hurricane Andrew made landfall in Florida), in a number of different ways, by using the “ .strftime() ” method. Store it in a variable called “Andrew”.

Instructions:

Print it in the format 'YYYY-MM', 'YYYY-DDD' and 'MONTH (YYYY)'

Sol:

```
from datetime import date

# Create a date object for August 26, 1992

Andrew = date(1992, 8, 26)

# Print the date in the format 'YYYY-MM'

print("Format 'YYYY-MM':", Andrew.strftime('%Y-%m'))

# Print the date in the format 'YYYY-DDD'

print("Format 'YYYY-DDD':", Andrew.strftime('%Y-%j'))

# Print the date in the format 'MONTH (YYYY)'

print("Format 'MONTH (YYYY)':", Andrew.strftime('%B (%Y)'))

#output:-

Format 'YYYY-MM': 1992-08

Format 'YYYY-DDD': 1992-239

Format 'MONTH (YYYY)': August (1992)
```

4) For the dataset “Indian_cities”,

a) Find out top 10 states in female-male sex ratio

```
import pandas as pd
file_path = (r'Indian_cities.csv')
data = pd.read_csv(file_path)
print(data.head())
top_10_states_sex_ratio = data.groupby('state_name')['sex_ratio'].mean().nlargest(10)
print("Top 10 states in female-male sex ratio:")
print(top_10_states_sex_ratio)
```

b) Find out top 10 cities in total number of graduates

```
top_10_cities_graduates = data[['name_of_city', 'total_graduates']].nlargest(10,
'total_graduates')
print("\nTop 10 cities in total number of graduates:")
print(top_10_cities_graduates)
```

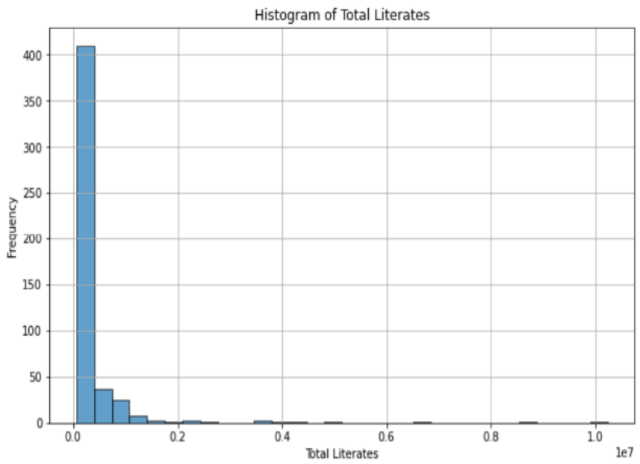
c) Find out top 10 cities and their locations in respect of total effective_literacy_rate.

```
top_10_cities_literacy = data[['name_of_city', 'location',
'effective_literacy_rate_total']].nlargest(10, 'effective_literacy_rate_total')
print("\nTop 10 cities and their locations in respect of total effective literacy rate:")
print(top_10_cities_literacy)
```

5) For the data set “Indian_cities”

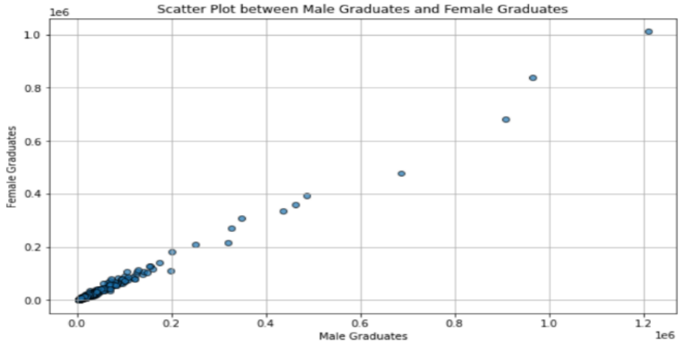
a) Construct histogram on literates_total and comment about the inferences

```
import pandas as pd
import matplotlib.pyplot as plt
file_path = (r'Indian_cities.csv')
data = pd.read_csv(file_path)
print(data.head())
plt.figure(figsize=(10, 6))
plt.hist(data['literates_total'], bins=30,
edgecolor='k', alpha=0.7)
plt.title('Histogram of Total Literates')
plt.xlabel('Total Literates')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```



b) Construct scatter plot between male graduates and female graduates

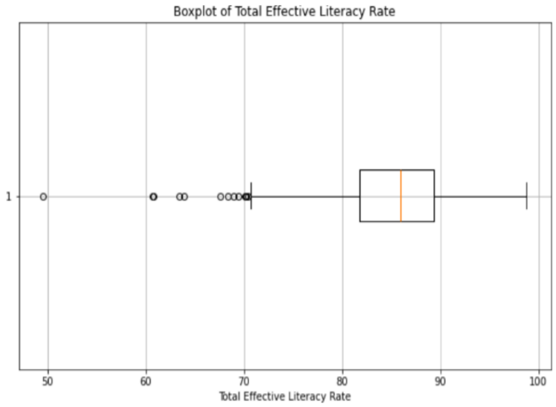
```
plt.figure(figsize=(10, 6))
plt.scatter(data['male_graduates'], data['female_graduates'], alpha=0.7, edgecolor='k')
plt.title('Scatter Plot between Male Graduates and Female Graduates')
plt.xlabel('Male Graduates')
plt.ylabel('Female Graduates')
plt.grid(True)
plt.show()
```



6) For the data set “Indian_cities”

a) Construct Boxplot on total effective literacy rate and draw inferences

```
import matplotlib.pyplot as plt
file_path = (r'Indian_cities.csv')
data = pd.read_csv(file_path)
print(data.head())
plt.figure(figsize=(10, 6))
plt.boxplot(data['effective_literacy_rate_total'].
dropna(), vert=False)
plt.title('Boxplot of Total Effective Literacy Rate')
plt.xlabel('Total Effective Literacy Rate')
plt.grid(True)
plt.show()
```



b) Find out the number of null values in each column of the dataset and delete them.

```
null_values = data.isnull().sum()

print("Number of null values in each column:")

print(null_values)

data_cleaned = data.dropna()

print(data_cleaned.head())
```