ProcDNA Case Study Solution Report

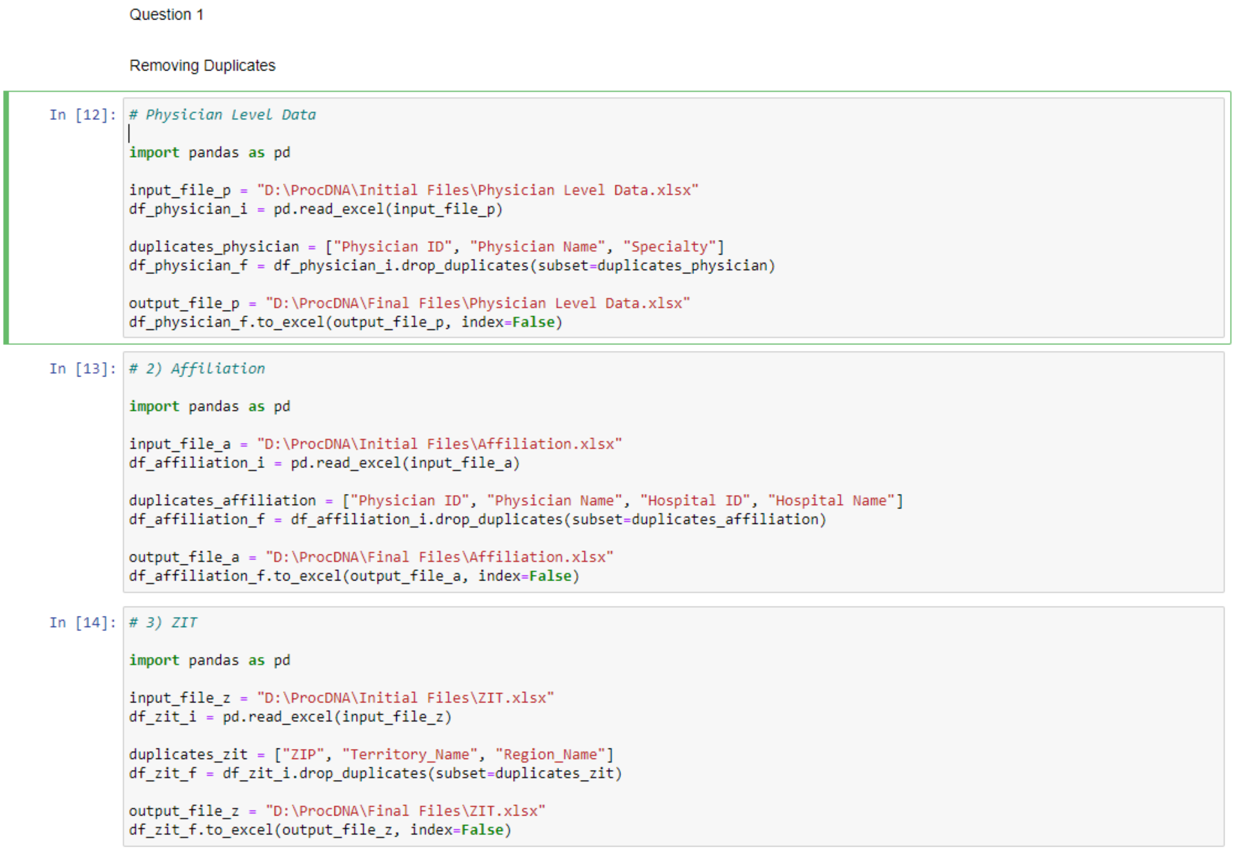
Group Number: 7
Group Members:

1. Ayushi Prasad
2. Sumeet Pavitrakar

Solutions(With their Key Insights):

1. What data checks will you apply on the given datasets?

- ➢ Checking for missing values in the dataset (There are some missing values in ‘Speciality’ column of Physician Level Data)
- ➢ Handling duplicates



Question 1

Removing Duplicates

In [12]:

```
# Physician Level Data

import pandas as pd

input_file_p = "D:\ProcDNA\Initial Files\Physician Level Data.xlsx"
df_physician_i = pd.read_excel(input_file_p)

duplicates_physician = ["Physician ID", "Physician Name", "Specialty"]
df_physician_f = df_physician_i.drop_duplicates(subset=duplicates_physician)

output_file_p = "D:\ProcDNA\Final Files\Physician Level Data.xlsx"
df_physician_f.to_excel(output_file_p, index=False)
```

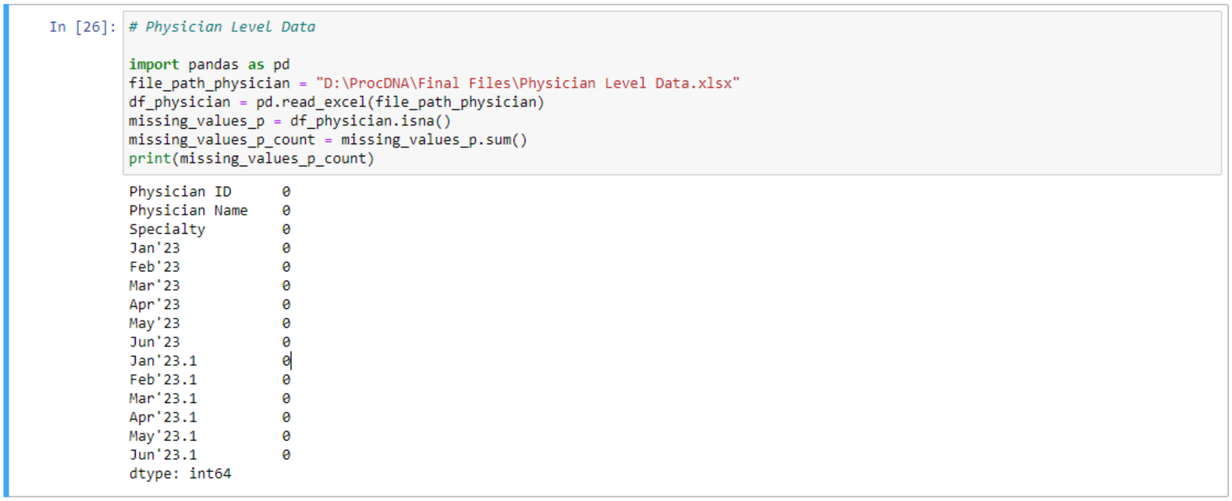
In [13]:

```
# 2) Affiliation

import pandas as pd

input_file_a = "D:\ProcDNA\Initial Files\Affiliation.xlsx"
df_affiliation_i = pd.read_excel(input_file_a)

duplicates_affiliation = ["Physician ID", "Physician Name", "Hospital ID", "Hospital Name"]
df_affiliation_f = df_affiliation_i.drop_duplicates(subset=duplicates_affiliation)

output_file_a = "D:\ProcDNA\Final Files\Affiliation.xlsx"
df_affiliation_f.to_excel(output_file_a, index=False)
```

In [14]:

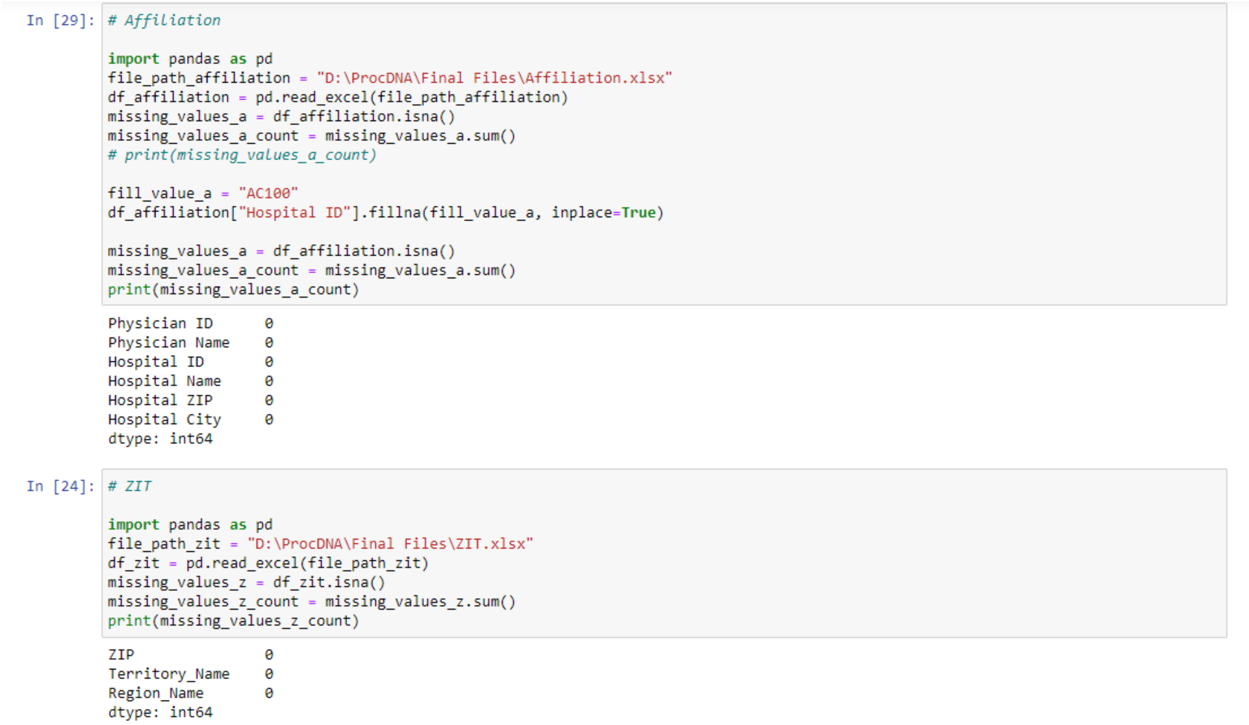
```
# 3) ZIT

import pandas as pd

input_file_z = "D:\ProcDNA\Initial Files\ZIT.xlsx"
df_zit_i = pd.read_excel(input_file_z)

duplicates_zit = ["ZIP", "Territory_Name", "Region_Name"]
df_zit_f = df_zit_i.drop_duplicates(subset=duplicates_zit)

output_file_z = "D:\ProcDNA\Final Files\ZIT.xlsx"
df_zit_f.to_excel(output_file_z, index=False)
```

Missing values



```
In [26]: # Physician Level Data

         import pandas as pd
         file_path_physician = "D:\ProcDNA\Final Files\Physician Level Data.xlsx"
         df_physician = pd.read_excel(file_path_physician)
         missing_values_p = df_physician.isna()
         missing_values_p_count = missing_values_p.sum()
         print(missing_values_p_count)
```

```
Physician ID      0
Physician Name    0
Specialty         0
Jan'23            0
Feb'23            0
Mar'23            0
Apr'23            0
May'23            0
Jun'23            0
Jan'23.1          0
Feb'23.1          0
Mar'23.1          0
Apr'23.1          0
May'23.1          0
Jun'23.1          0
dtype: int64
```



```
In [29]: # Affiliation

         import pandas as pd
         file_path_affiliation = "D:\ProcDNA\Final Files\Affiliation.xlsx"
         df_affiliation = pd.read_excel(file_path_affiliation)
         missing_values_a = df_affiliation.isna()
         missing_values_a_count = missing_values_a.sum()
         # print(missing_values_a_count)

         fill_value_a = "AC100"
         df_affiliation["Hospital ID"].fillna(fill_value_a, inplace=True)

         missing_values_a = df_affiliation.isna()
         missing_values_a_count = missing_values_a.sum()
         print(missing_values_a_count)
```

```
Physician ID      0
Physician Name    0
Hospital ID       0
Hospital Name     0
Hospital ZIP      0
Hospital City     0
dtype: int64
```

```
In [24]: # ZIT

         import pandas as pd
         file_path_zit = "D:\ProcDNA\Final Files\ZIT.xlsx"
         df_zit = pd.read_excel(file_path_zit)
         missing_values_z = df_zit.isna()
         missing_values_z_count = missing_values_z.sum()
         print(missing_values_z_count)
```

```
ZIP              0
Territory_Name   0
Region_Name      0
dtype: int64
```

2. Plot a graph showing the sales (# Total Prescriptions) of both the given products (Fludara and Mercapto) over months. Share key insights

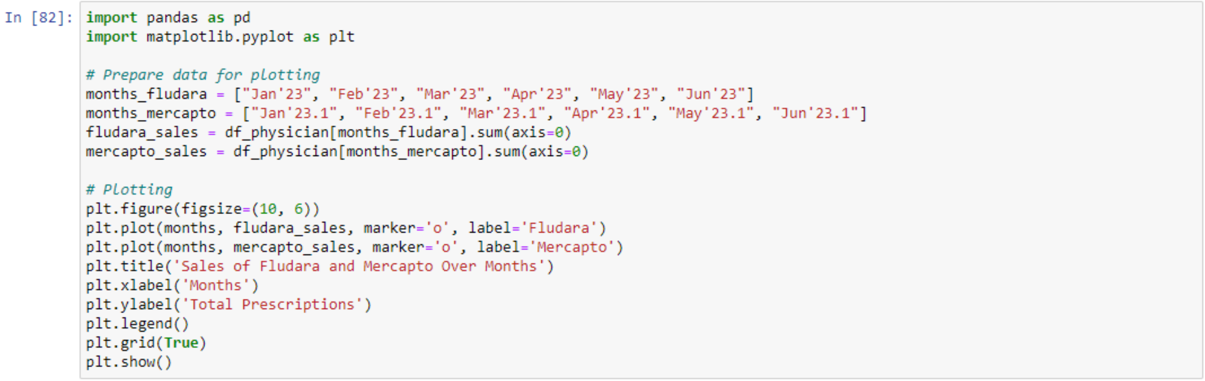
➢ **Key Insights:**

Observations	Possible Reasons
The sales of ‘Fludara’ were almost always lower than that of ‘Mercapto’ except for the month of February 2023.	This may happen due to marketing done with unanalysed and not-data-backed marketing strategies.
The need for these two products was the highest in June 2023.	Maybe the rise in cases of blood cancer arose in that period.
The sales of ‘Mercapto’ from January to February 2023 were consistent, while there was a significant increase in the sales of ‘Fludara’. But the sale of ‘Mercapto’ then increased significantly from February to March 2023.	It could be due to: 1. reduce in prices, 2. changes in marketing strategies 3. patient needs
The sales of ‘Mercapto’ have consistently been higher than that of ‘Fludara’ since March 2023.	It could be due to the fact that the aftereffects of ‘Mercapto’ might have been more favorable than that of ‘Fludara’ .
By mid-February, there was a continuous increase in the difference between the sales of both drug.	We assume that this happened due to the continuous research and innovation done by the rival company. Also, the researches might have reduced the curation period of the drug and hence there was an increase in number of physicians accepting ‘Mercapto’.

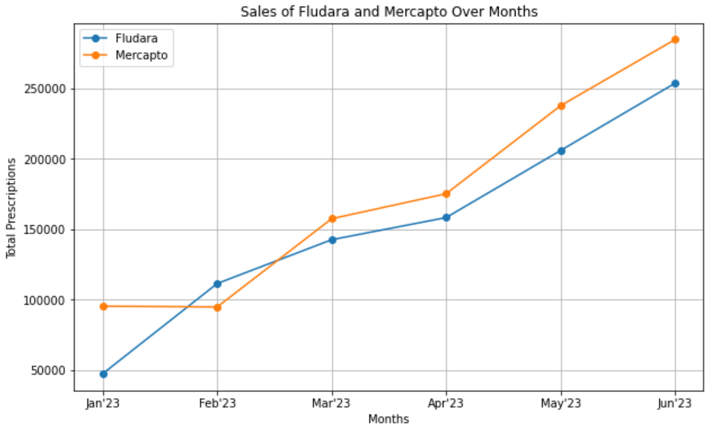
➢ The reason for rise in the sales of ‘Mercapto’ can be its increasing number of safety tests which gave positive results and hence marketing these results could have benefitted them in gaining trust of many physicians.

➢ Also, the rival company might have increased the production of the ‘Mercapto’, and hence, was able to reduce the price and make the drug comparatively more affordable.

Question 2

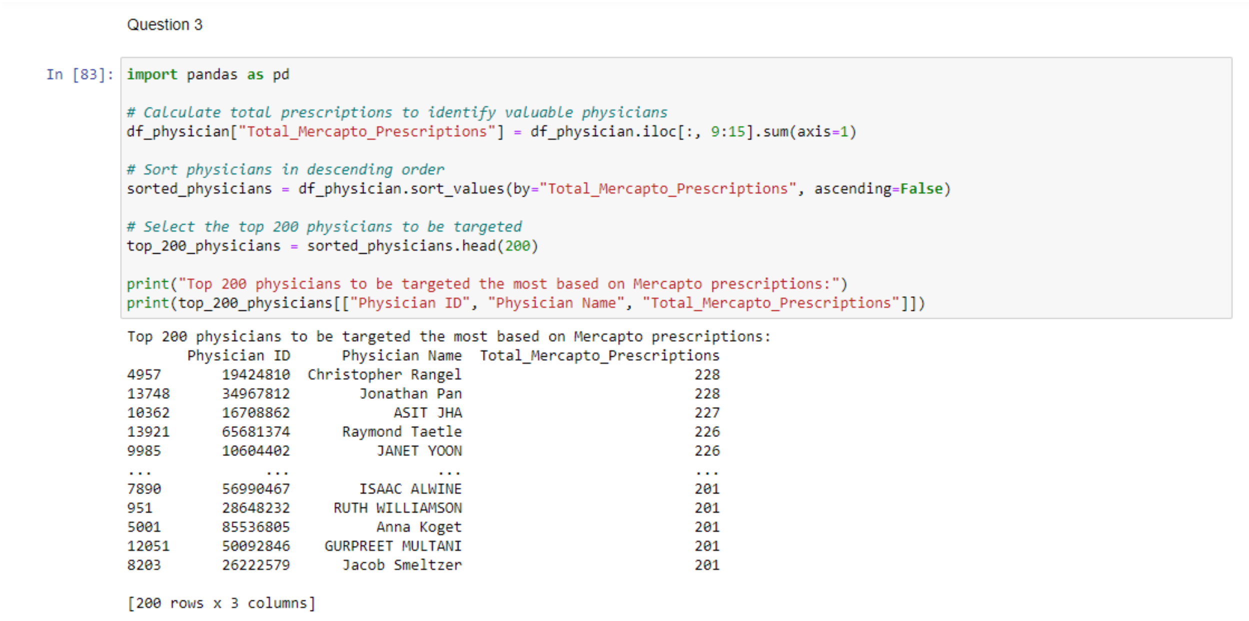


```
In [82]: import pandas as pd
         import matplotlib.pyplot as plt

         # Prepare data for plotting
         months_fludara = ["Jan'23", "Feb'23", "Mar'23", "Apr'23", "May'23", "Jun'23"]
         months_mercapto = ["Jan'23.1", "Feb'23.1", "Mar'23.1", "Apr'23.1", "May'23.1", "Jun'23.1"]
         fludara_sales = df_physician[months_fludara].sum(axis=0)
         mercapto_sales = df_physician[months_mercapto].sum(axis=0)

         # Plotting
         plt.figure(figsize=(10, 6))
         plt.plot(months, fludara_sales, marker='o', label='Fludara')
         plt.plot(months, mercapto_sales, marker='o', label='Mercapto')
         plt.title('Sales of Fludara and Mercapto Over Months')
         plt.xlabel('Months')
         plt.ylabel('Total Prescriptions')
         plt.legend()
         plt.grid(True)
         plt.show()
```



3. Who are the top 200 physicians that should be targeted the most? Explain the approach that you considered.

- ➢ First, we'll calculate the total sales of 'Fludara' and total sales of 'Mercapto' for each physician.
- ➢ We will target those physicians where the total sales of 'Mercapto' > total sales of 'Fludara'
- ➢ And sort the list in descending order where the difference will be high.
- ➢ **Key Insights:**
  - ○ With this data, it will help in prioritizing the targets where ProcDNA's Sales Representatives have to communicate more and market the product.
  - ○ This step will help in increasing the number of sales.



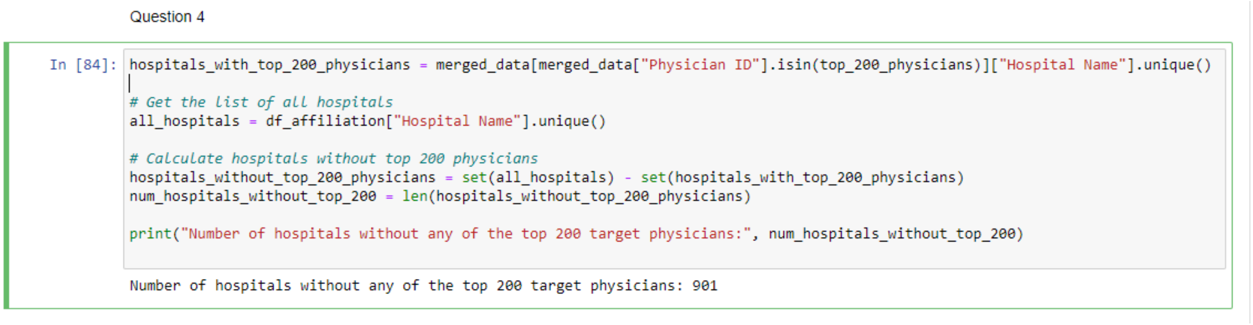
Question 3

```
In [83]: import pandas as pd

         # Calculate total prescriptions to identify valuable physicians
         df_physician["Total_Mercapto_Prescriptions"] = df_physician.iloc[:, 9:15].sum(axis=1)

         # Sort physicians in descending order
         sorted_physicians = df_physician.sort_values(by="Total_Mercapto_Prescriptions", ascending=False)

         # Select the top 200 physicians to be targeted
         top_200_physicians = sorted_physicians.head(200)

         print("Top 200 physicians to be targeted the most based on Mercapto prescriptions:")
         print(top_200_physicians[["Physician ID", "Physician Name", "Total_Mercapto_Prescriptions"]])
```

```
Top 200 physicians to be targeted the most based on Mercapto prescriptions:
       Physician ID      Physician Name  Total_Mercapto_Prescriptions
4957       19424810  Christopher Rangel                           228
13748      34967812        Jonathan Pan                           228
10362      16708862            ASIT JHA                           227
13921      65681374      Raymond Taetle                           226
9985       10604402          JANET YOON                           226
...             ...                 ...                           ...
7890       56990467        ISAAC ALWINE                           201
951        28648232     RUTH WILLIAMSON                           201
5001       85536805          Anna Koget                           201
12051      50092846    GURPREET MULTANI                           201
8203       26222579      Jacob Smeltzer                           201

[200 rows x 3 columns]
```

4. How many hospitals don't have any of the top 200 target physicians affiliated to them?

- ➢ First we evaluated the unique hospitals with top 200 physicians.
- ➢ Then we are evaluating all the unique hospitals.
- ➢ Then we'll remove the data of 'unique hospitals with top 200 physicians' from the data of 'all unique hospitals'.
- ➢ Then the remaining number of rows will be the number of unique hospitals without top 200 physicians.
- ➢ **Key Insights:**
  - ○ This will be consisting of hospitals which are not preferring 'Mercapto' as such.
  - ○ With this data, it will help ProcDNA in not prioritizing these hospital.
  - ○ Also, with this data we will be able to allocate less Sales Representatives and use the Sales Force more efficiently.



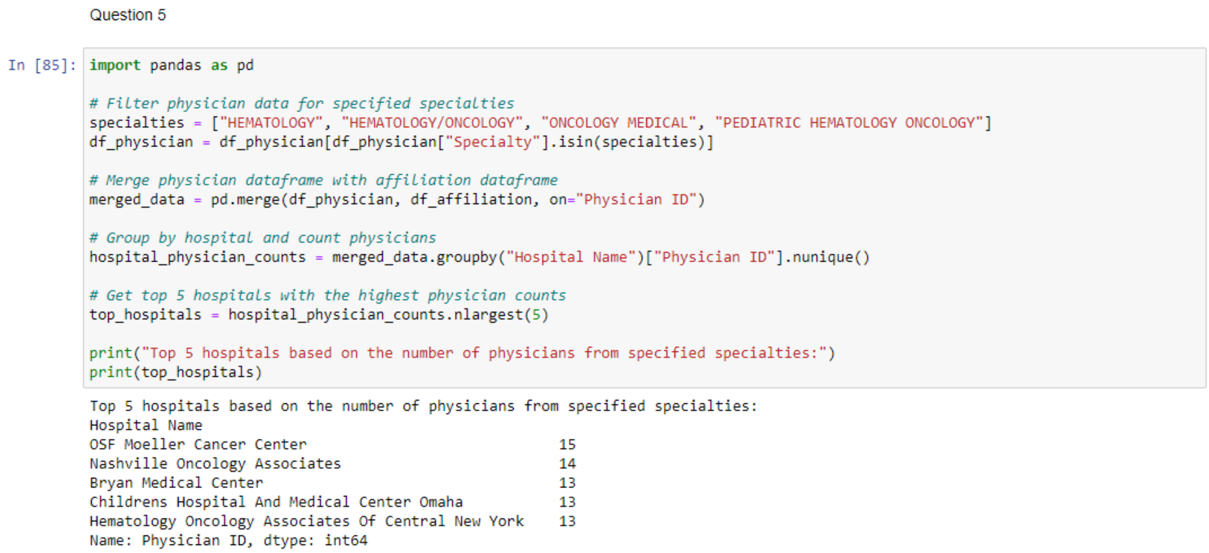
Question 4

```
In [84]: hospitals_with_top_200_physicians = merged_data[merged_data["Physician ID"].isin(top_200_physicians)]["Hospital Name"].unique()

         # Get the list of all hospitals
         all_hospitals = df_affiliation["Hospital Name"].unique()

         # Calculate hospitals without top 200 physicians
         hospitals_without_top_200_physicians = set(all_hospitals) - set(hospitals_with_top_200_physicians)
         num_hospitals_without_top_200 = len(hospitals_without_top_200_physicians)

         print("Number of hospitals without any of the top 200 target physicians:", num_hospitals_without_top_200)
```

```
Number of hospitals without any of the top 200 target physicians: 901
```

5. List the top 5 hospitals based on the # Physicians from the following 4 specialties affiliated to them: "Hematology", "Hematology/Oncology", "Oncology Medical" and "Pediatric Hematology Oncology".

- ➢ We'll replace physicians' dataframe with the physicians' dataframe whose speciality is either in 'Hematology', 'Hematology/Oncology', 'Oncology Medical' and 'Pediatric Hematology Oncology'
- ➢ We merge physician dataframe and affiliation dataframe on Physician ID.
- ➢ We then group the data on the basis of Hospital Name. We also take unique Physician ID in each group.
- ➢ And then, we're returning 5 Hospital Names with largest hospital physician count(which we calculated in above step).
- ➢ **<u>Key Insights:</u>**
    - ○ Hematology and Oncology are the study of blood and blood disorders.
    - ○ With this data we'll be able to target these hospitals who have most number of physicians in those particular fields.
    - ○ The top 5 hospitals having physicians specialized in these 4 specialities will lead to more patients coming in which will eventually lead to a greater need for 'Fludara' .
    - ○ These hospitals will be our top source of income.
    - ○ In these hospitals, we can prioritize on explaining the physicians about the after effects of the drug and that discuss its no-side-effect capability to gain trust.



Question 5

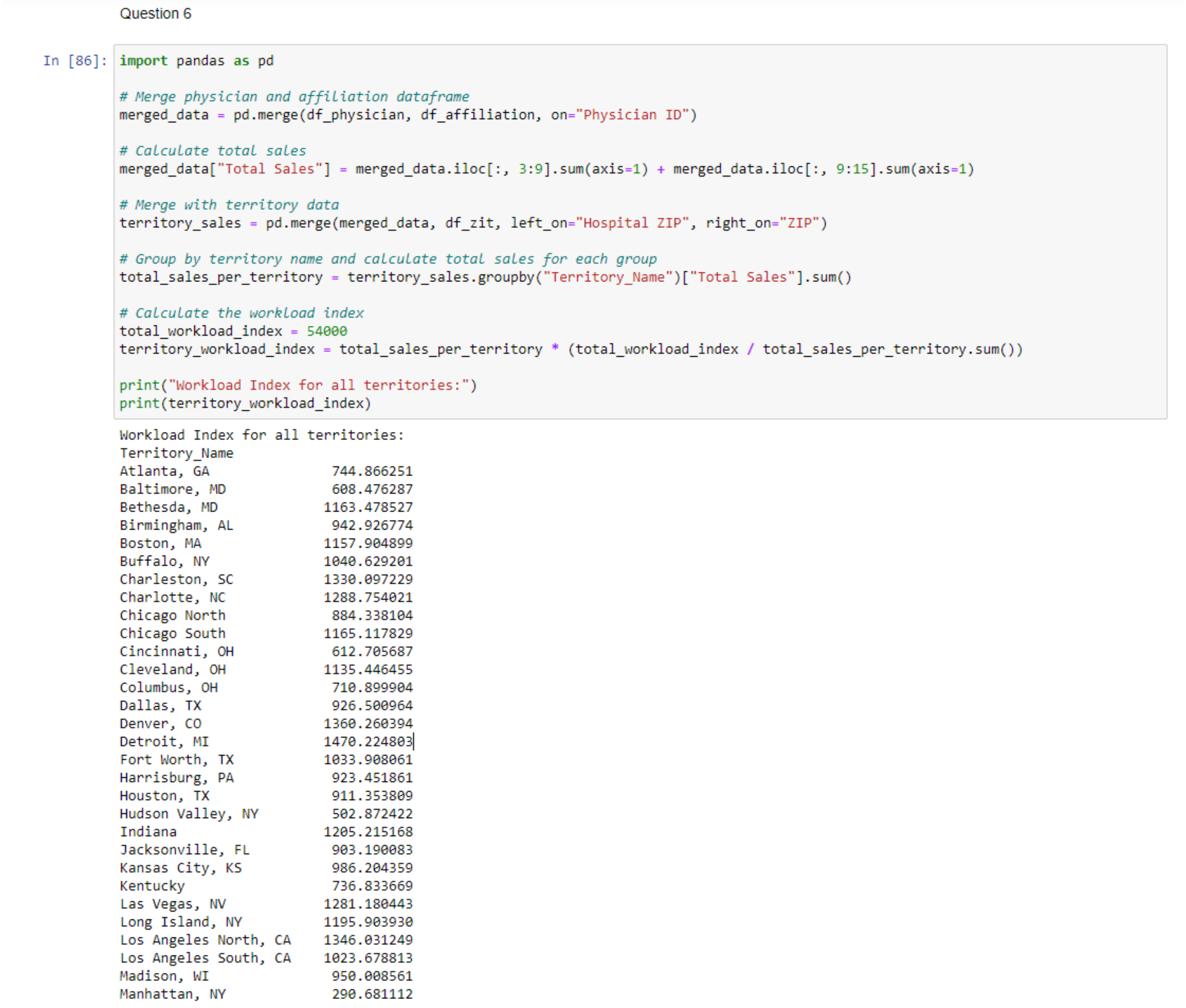
In [85]:

```
import pandas as pd

# Filter physician data for specified specialties
specialties = ["HEMATOLOGY", "HEMATOLOGY/ONCOLOGY", "ONCOLOGY MEDICAL", "PEDIATRIC HEMATOLOGY ONCOLOGY"]
df_physician = df_physician[df_physician["Specialty"].isin(specialties)]

# Merge physician dataframe with affiliation dataframe
merged_data = pd.merge(df_physician, df_affiliation, on="Physician ID")

# Group by hospital and count physicians
hospital_physician_counts = merged_data.groupby("Hospital Name")["Physician ID"].nunique()

# Get top 5 hospitals with the highest physician counts
top_hospitals = hospital_physician_counts.nlargest(5)

print("Top 5 hospitals based on the number of physicians from specified specialties:")
print(top_hospitals)
```

```
Top 5 hospitals based on the number of physicians from specified specialties:
Hospital Name
OSF Moeller Cancer Center                             15
Nashville Oncology Associates                         14
Bryan Medical Center                                  13
Childrens Hospital And Medical Center Omaha           13
Hematology Oncology Associates Of Central New York    13
Name: Physician ID, dtype: int64
```

6. Calculate the Workload index for all the territories.

- ➢ Workload calculation procedures:
- ➢ Add up all of the territories' combined sales of Fludara and Mercapto.
- ➢ To calculate the workload index for each territory, rescale the total sales to a value of 54,000 (# of territories* 1,000).
- ➢ It highlights which areas are experiencing higher demand and which ones have relatively lower demand, allowing for resource allocation and optimization in improving budget allocation and resource management.
- ➢ **Key insights:**
    - ○ We need to divide our budget of 54,000 among different territories based on their original ratios.
    - ○ In this way, each territory gets a fair share that considers their sales.
    - ○ This will help us to conclude how well each territory is doing according to their sizes.
    - ○ It helps in better resource allocation.

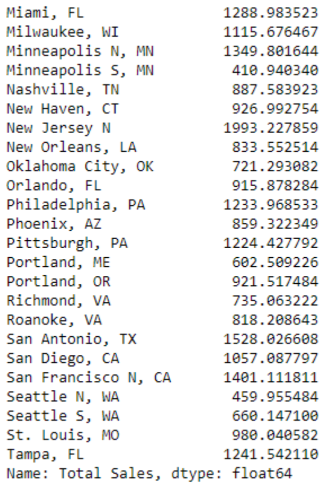


Question 6

In [86]:

```
import pandas as pd

# Merge physician and affiliation dataframe
merged_data = pd.merge(df_physician, df_affiliation, on="Physician ID")

# Calculate total sales
merged_data["Total Sales"] = merged_data.iloc[:, 3:9].sum(axis=1) + merged_data.iloc[:, 9:15].sum(axis=1)

# Merge with territory data
territory_sales = pd.merge(merged_data, df_zit, left_on="Hospital ZIP", right_on="ZIP")

# Group by territory name and calculate total sales for each group
total_sales_per_territory = territory_sales.groupby("Territory_Name")["Total Sales"].sum()

# Calculate the workload index
total_workload_index = 54000
territory_workload_index = total_sales_per_territory * (total_workload_index / total_sales_per_territory.sum())

print("Workload Index for all territories:")
print(territory_workload_index)
```

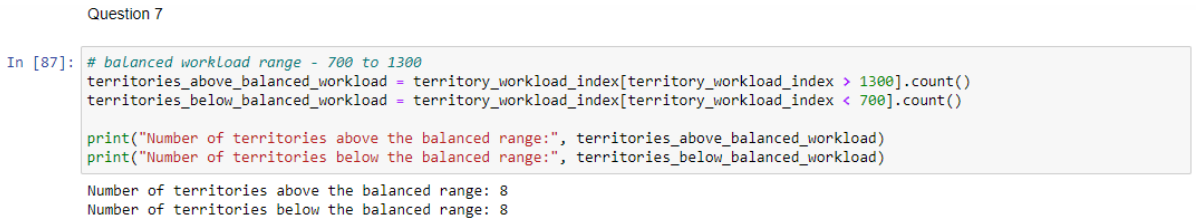
```
Workload Index for all territories:
Territory_Name
Atlanta, GA             744.866251
Baltimore, MD           608.476287
Bethesda, MD           1163.478527
Birmingham, AL          942.926774
Boston, MA             1157.904899
Buffalo, NY            1040.629201
Charleston, SC         1330.097229
Charlotte, NC          1288.754021
Chicago North           884.338104
Chicago South          1165.117829
Cincinnati, OH          612.705687
Cleveland, OH          1135.446455
Columbus, OH            710.899904
Dallas, TX              926.500964
Denver, CO             1360.260394
Detroit, MI            1470.224803
Fort Worth, TX         1033.908061
Harrisburg, PA          923.451861
Houston, TX             911.353809
Hudson Valley, NY       502.872422
Indiana                1205.215168
Jacksonville, FL        903.190083
Kansas City, KS         986.204359
Kentucky                736.833669
Las Vegas, NV          1281.180443
Long Island, NY        1195.903930
Los Angeles North, CA  1346.031249
Los Angeles South, CA  1023.678813
Madison, WI             950.008561
Manhattan, NY           290.681112
```



```
Miami, FL                1288.983523
Milwaukee, WI            1115.676467
Minneapolis N, MN        1349.801644
Minneapolis S, MN         410.940340
Nashville, TN             887.583923
New Haven, CT             926.992754
New Jersey N             1993.227859
New Orleans, LA           833.552514
Oklahoma City, OK         721.293082
Orlando, FL               915.878284
Philadelphia, PA         1233.968533
Phoenix, AZ               859.322349
Pittsburgh, PA           1224.427792
Portland, ME              602.509226
Portland, OR              921.517484
Richmond, VA              735.063222
Roanoke, VA               818.208643
San Antonio, TX          1528.026608
San Diego, CA            1057.087797
San Francisco N, CA      1401.111811
Seattle N, WA             459.955484
Seattle S, WA             660.147100
St. Louis, MO             980.040582
Tampa, FL                1241.542110
Name: Total Sales, dtype: float64
```

7. Calculate the # Territories above and below the balanced workload index range separately. (The territories having a workload index in the range of 700-1,300 (both inclusive) are considered to be balanced)

- As given, a balanced workload ranges between(700-1300)
- Now, to print the territories below balanced workload, we'll print the territories whose territory_workload_index is less than 700.
- And to print the territories above balanced workload, we'll print the territories whose territory_workload_index is greater than 1300.
- **Key Insights:**
    - This data will help us in balancing the budget in different territories based on workload index.
    - Our understanding is that 'below-workload' territories are deficient in budget and 'above-balanced' territories are surplus in budget.
    - This can be managed by shifting budget allocation of 'above-balanced' territories to that of 'below-balanced' territories.



Question 7

```
In [87]: # balanced workload range - 700 to 1300
         territories_above_balanced_workload = territory_workload_index[territory_workload_index > 1300].count()
         territories_below_balanced_workload = territory_workload_index[territory_workload_index < 700].count()

         print("Number of territories above the balanced range:", territories_above_balanced_workload)
         print("Number of territories below the balanced range:", territories_below_balanced_workload)

         Number of territories above the balanced range: 8
         Number of territories below the balanced range: 8
```

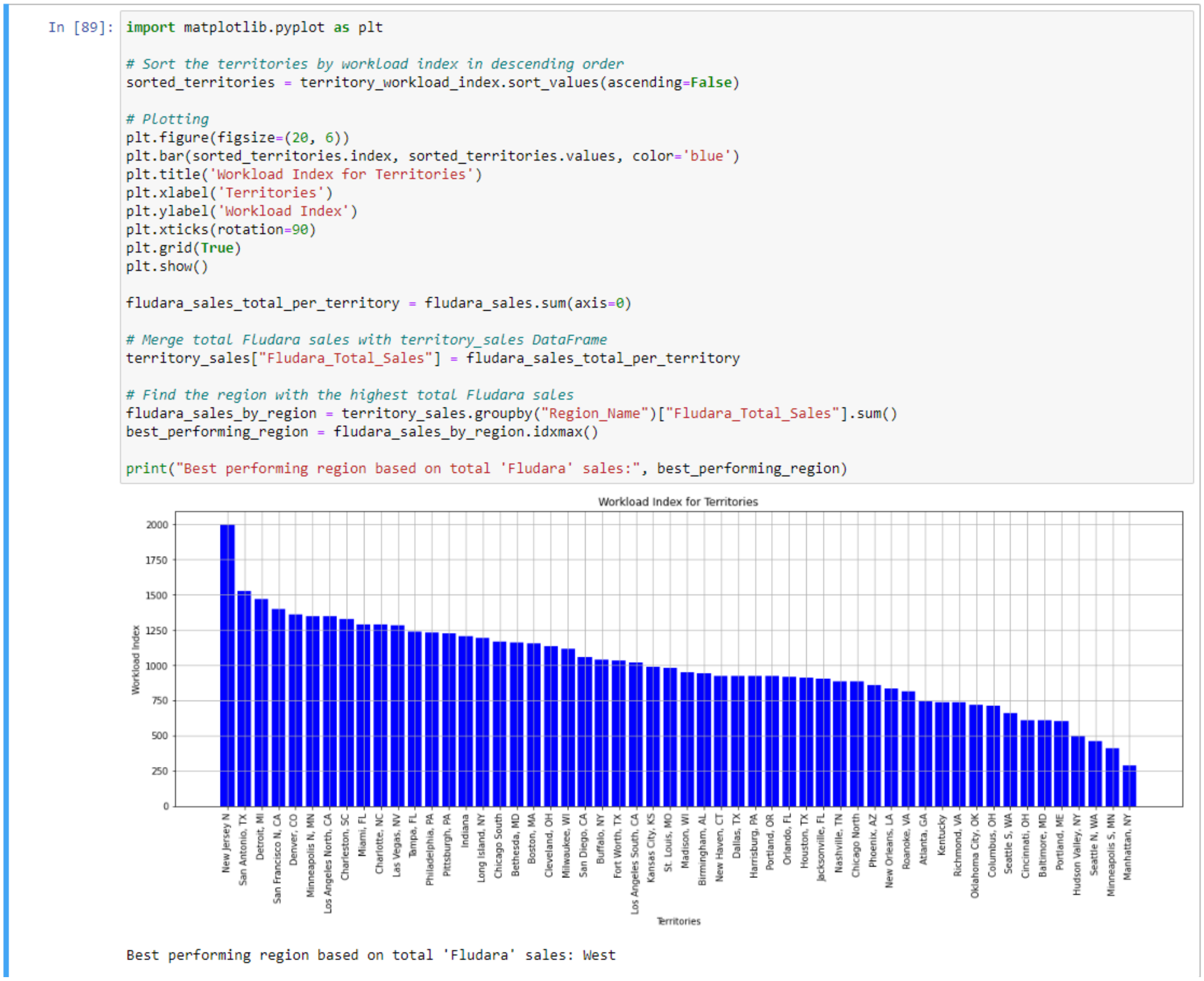
8. Plot a graph depicting the workload index for all the territories in descending order. Which region is performing best based on "Fludara" sales?

- We sort the territory_workload_index in descending order and then plot a bar chart.
- First, we take the sum of Fludara sales and merge it with the territory sales dataframe. We then group the merged dataframe by Region Name and calculate the sum of total fludara sales in each region.

- ➢ Then, we return the region corresponding to the maximum value in the dataframe.

- ➢ **Key Insights:**
  - ○ The bar chart will show us how the workload is spread across different territories, helping us see which areas have more or less work.
  - ○ Also, finding the region where Fludara sales are the highest tells us where the medicine is doing really well. This helps us decide where to focus our efforts to promote it even more effectively.
  - ○ Finding out where Fludara is selling the most tells us which areas really like the product. This helps us focus our advertising and marketing efforts on those places to make the product even more popular there.

Question 8



In [89]:

```
import matplotlib.pyplot as plt

# Sort the territories by workload index in descending order
sorted_territories = territory_workload_index.sort_values(ascending=False)

# Plotting
plt.figure(figsize=(20, 6))
plt.bar(sorted_territories.index, sorted_territories.values, color='blue')
plt.title('Workload Index for Territories')
plt.xlabel('Territories')
plt.ylabel('Workload Index')
plt.xticks(rotation=90)
plt.grid(True)
plt.show()

fludara_sales_total_per_territory = fludara_sales.sum(axis=0)

# Merge total Fludara sales with territory_sales DataFrame
territory_sales["Fludara_Total_Sales"] = fludara_sales_total_per_territory

# Find the region with the highest total Fludara sales
fludara_sales_by_region = territory_sales.groupby("Region_Name")["Fludara_Total_Sales"].sum()
best_performing_region = fludara_sales_by_region.idxmax()

print("Best performing region based on total 'Fludara' sales:", best_performing_region)
```

```
Best performing region based on total 'Fludara' sales: West
```

SUMMARY:

- We learnt how the business is influenced by the competitors.
- We discovered different strategies to control the product sales in a business.
- We learnt the importance of data analysis and how it impacts in making profitable data-backed decisions.
- We learnt how data analysis help
- We also learnt about the benefits of workload index which are helpful in understanding how the demand for specific products is distributed across different territories.
- According to Wilson’s Law “When you put Information and Intelligence first, the money keeps flowing in” and data analysis is the way to execute it.

REFERENCES:

- https://stackoverflow.com/
- https://www.procdna.com/case-studies/