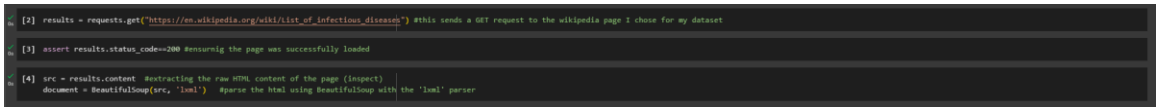
Claire Antonette Mendoza

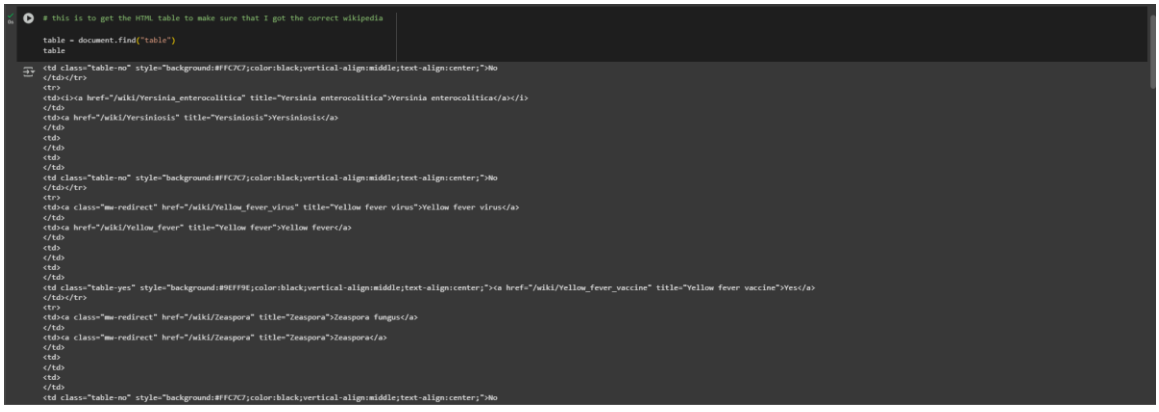
ITS132L – FOPM01

M2-SA

07-15-25

Part 1 - Data Acquisition: *Using a* ***web scraping library*** *from Python, extract and build a dataset coming from a website (should contain data with mining interest, and significance such as Weather, Population, Diseases etc.).*





```
[6] print(table.find("th").get_text()) #helps check what kind of table i'll be working on (and also to double-check if I got the table right)

    Infectious agent
```

```
#gathering all rows, including header and content rows, and prepare a list to store parsed data

rows = table.find_all("tr")  #find all rows from the data
rows
```

```
<td class="table-no" style="background:#FFC7C7;color:black;vertical-align:middle;text-align:center;">No
</td></tr>,
<tr>
<td><i><a href="/wiki/Yersinia_enterocolitica" title="Yersinia enterocolitica">Yersinia enterocolitica</a></i>
</td>
<td><a href="/wiki/Yersiniosis" title="Yersiniosis">Yersiniosis</a>
</td>
<td>
</td>
<td>
</td>
<td class="table-no" style="background:#FFC7C7;color:black;vertical-align:middle;text-align:center;">No
</td></tr>,
<tr>
<td><a class="mw-redirect" href="/wiki/Yellow_fever_virus" title="Yellow fever virus">Yellow fever virus</a>
</td>
<td><a href="/wiki/Yellow_fever" title="Yellow fever">Yellow fever</a>
</td>
<td>
</td>
<td>
</td>
<td class="table-yes" style="background:#9EFF9E;color:black;vertical-align:middle;text-align:center;"><a href="/wiki/Yellow_fever_vaccine" title="Yellow fever vaccine">Yes</a>
</td></tr>,
<tr>
<td><a class="mw-redirect" href="/wiki/Zeaspora" title="Zeaspora">Zeaspora fungus</a>
</td>
<td><a class="mw-redirect" href="/wiki/Zeaspora" title="Zeaspora">Zeaspora</a>
</td>
<td>
</td>
<td>
</td>
<td class="table-no" style="background:#FFC7C7;color:black;vertical-align:middle;text-align:center;">No
</td></tr>,
<tr>
<td><i><a href="/wiki/Zika_virus" title="Zika virus">Zika virus</a></i>
</td>
<td><a href="/wiki/Zika_fever" title="Zika fever">Zika fever</a>
```

```
[10] # Find the first column
tables = document.find_all("table", class_="wikitable")
target_table = tables[0]

# Store all rows into a list
rows = target_table.find_all("tr")
data = []

#loop thorugh each row except the header (starting from index 1)
for row in rows[1:]:
    cells = row.find_all(["td", "th"])
    if len(cells) == 5:  #5 columns for agent, common name, diagnosis, treatment, vaccine

        cells_text = [cell.get_text(strip=True).replace('\xa0', ' ') for cell in cells]

        # Skip rows that look like repeated headers or missing values
        if all(cells_text) and not any("agent" in c.lower() for c in cells_text):
            data.append(cells_text)
```

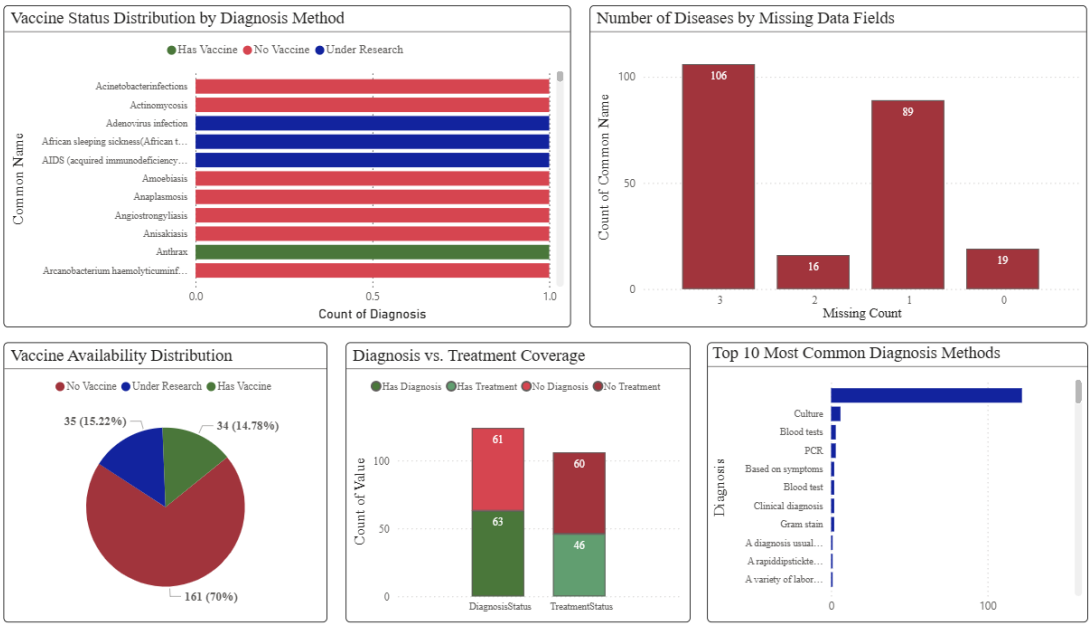
```
# Create DataFrame
columns = ["Infectious agent", "Common name", "Diagnosis", "Treatment", "Vaccine(s)"]
df = pd.DataFrame(data, columns=columns)

# Drop duplicates or clean common formatting issues
df["Common name"] = df["Common name"].str.replace('\n', ' ').str.strip()
df = df.drop_duplicates()

# Preview
df.head(10)
```

	Infectious agent	Common name	Diagnosis	Treatment	Vaccine(s)
0	Acinetobacter baumannii	Acinetobacterinfections	Culture	Supportive care	No
1	Actinomycos israelii,Actinomyces gerencseriana...	Actinomycosis	Histologic findings	Penicillin,doxycycline, andsulfonamides	No
2	Adenoviridae	Adenovirus infection	Antigendetection,polymerase chain reactionassa...	Most infections are mild and require no therap...	Under research[1]
3	Trypanosoma brucei	African sleeping sickness(African trypanosomia...	Identification of trypanosomes in a sample by ...	Fexinidazoleby mouth orpentamidineby injection...	Under research[2]
4	HIV (Human immunodeficiency virus)	AIDS (acquired immunodeficiency syndrome)	Antibody test, p24 antigen test, PCR	Treatment is typically anon-nucleoside reverse...	Under research[3]
5	Anaplasmaspecies	Anaplasmosis	indirect immunofluorescence antibody assay for...	Tetracycline drugs (includingtetracycline,chlo...	No
6	Angiostrongylus	Angiostrongyliasis	Lumbar puncture, brain imaging, serology	Albendazole	No
7	Anisakis	Anisakiasis	Gastroscopic examination, or histopathologic e...	Albendazole	No
8	Bacillus anthracis	Anthrax	Culture, PCR	Large doses of intravenous and oral antibiotic...	Yes
9	Arcanobacterium haemolyticum	Arcanobacterium haemolyticuminfection	Culture in human bloodagarplates	erythromycin(proposed as the first-line drug),...	No

 # Downloading my data into a CSV file

```
[13] # Downloading the pre-processed CSV file to local machine

    # Save the DataFrame to a CSV file first
    df.to_csv("infections_list_cleaned.csv", index=False)

    from google.colab import files
    files.download("infections_list_cleaned.csv")
```

Part 2 - Visualization: *Feed processed data to POWER BI and create at least 5 data representations with a corresponding analytical explanation of the knowledge you initially found. This may help you get a better idea of what part of your dataset will be used in your model.*



Using the cleansed infection dataset, I created a series of insightful visualizations in Power BI to explore patterns in vaccine availability, diagnosis and treatment status, and missing data. Each visual helped uncover key trends and gaps in infectious disease management.

1.) Vaccine Status Distribution by Diagnosis Method (Bar Chart):

This bar chart shows the relationship between *diagnosis methods* and the *vaccine status* of each disease. Diseases with advanced lab-based diagnosis (e.g., PCR, culture) are more likely to have vaccines under research or no vaccine at all. This suggests that even with the ability to diagnose, vaccine development still lags behind in many cases.

Implication:

There's a disconnect between diagnostics and preventive intervention, which may require public health attention or more targeted research funding.

2.) **Number of Diseases by Missing Data Fields (Bar Chart)**

This visual displays how many diseases are missing 1 to 3 key data fields (Diagnosis, Treatment, Vaccine).

- **106 diseases** lack all three fields, while Only **19** are fully complete.

Implication:

The high count of missing data across the board suggests limited research, data availability, or documentation for many infections. These gaps can delay timely response and weaken disease control systems.

3.) **Vaccine Availability Distribution (Pie Chart)**

- **70% (161 diseases)** have *no vaccine*,
- **15.2% (35 diseases)** are *under research*,
- Only **14.78% (34 diseases)** have an *available vaccine*.

Implication:

A vast majority of infectious diseases still don't have a vaccine option. This underscores the importance of strengthening treatment and diagnosis coverage where vaccines are lacking and prioritizing vaccine R&D for underrepresented diseases.

4.) **Diagnosis vs. Treatment Coverage (Stacked Bar Chart)**

- **63 diseases** have a diagnosis method, while **61** do not.
- **46 diseases** have treatment data, but **60** do not.

Implication:

More diseases lack treatment data than diagnosis. This implies a potential global issue where identification of disease is improving, but therapeutic solutions are still limited or undocumented.

5.) **Top 10 Most Common Diagnosis Methods (Horizontal Bar Chart)**

- Culture stands out as the most commonly listed method.
- Other methods like PCR, clinical, and blood test are used much less frequently (one or more times).

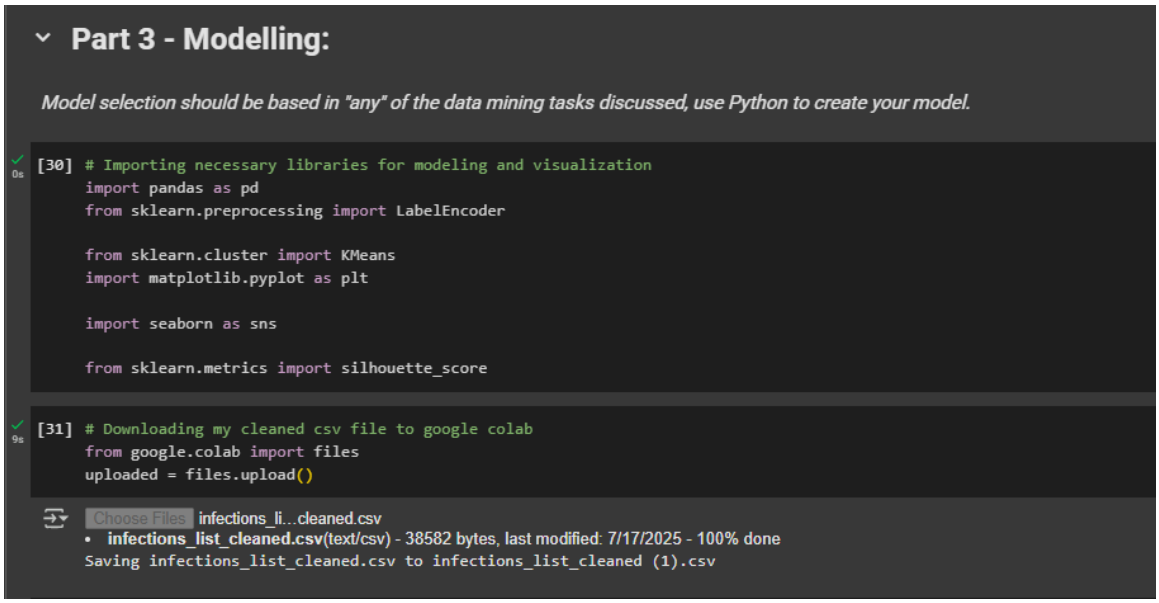
Implication:

There is heavy reliance on traditional methods like culture, while newer or more rapid methods (e.g., PCR, rapid diagnostic kits) are underused or underreported. This could reflect accessibility issues in certain regions.

Summary of Data Visualization Findings:

- Vast vaccine gaps exist even among diagnosable diseases.
- Treatment data is less available than diagnostic data.
- A significant portion of the dataset lacks full information.
- Traditional diagnostic methods dominate over modern ones.
- Visual tools like Power BI made it easier to explore gaps in disease coverage, guiding which infections require further attention from both a public health and data collection standpoint.

Part 3 - Modelling: *Model selection should be based on "any" of the data mining tasks discussed.*



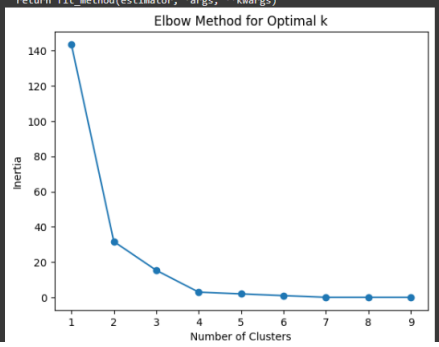
```
[32] # ensuring file is uploaded to Colab
     df = pd.read_csv("infections_list_cleaned.csv", encoding='latin1') #i used latin1 because it's exported from Power BI to Excel, so I used a Windows-based encoding, and not UTF-8
     df.head()
```

	Infectious Agent	Common Name	Diagnosis	Treatment	VaccineStatus	Value	Attribute	Vaccine(s)	Diagnosis Status	Treatment Status	Vaccine Status	Missing Count
0	Acinetobacter baumannii	Acinetobacterinfections	Culture	Supportive care	No Vaccine	Has Diagnosis	DiagnosisStatus	No	Has	Has	NaN	1
1	Actinomyces israelii,Actinomyces gerencseriaea...	Actinomycosis	Histologic findings	Penicillin,doxycycline, andsulfonamides	No Vaccine	Has Diagnosis	DiagnosisStatus	No	Has	Has	NaN	1
2	Adenoviridae	Adenovirus infection	Antigendetection,polymerase chain reactionassa...	Most infections are mild and require no therap...	Under Research	Has Diagnosis	DiagnosisStatus	Under research	Has	Has	NaN	1
3	Alphavirus	Chikungunya	Laboratory criteria include adecreased lymphoc...	Supportive care	Under Research	Has Treatment	TreatmentStatus	Under research	Has	Has	NaN	1
4	Anaplasma phagocytophilum	Human granulocytic anaplasmosis(HGA)	PCR	Doxycycline	No Vaccine	Has Treatment	TreatmentStatus	No	Has	Has	NaN	1

```
# Encoding categorical text data into numbers (for visualization)
encoder = LabelEncoder()

df['Diagnosis Status'] = encoder.fit_transform(df['Diagnosis Status'])  # ← this converts text into integers
df['Treatment Status'] = encoder.fit_transform(df['Treatment Status'])
df['Vaccine Status'] = encoder.fit_transform(df['Vaccine Status'])

df_cluster = df[['Diagnosis Status', 'Treatment Status', 'Vaccine Status']] # ← preparing feature set for clustering
```

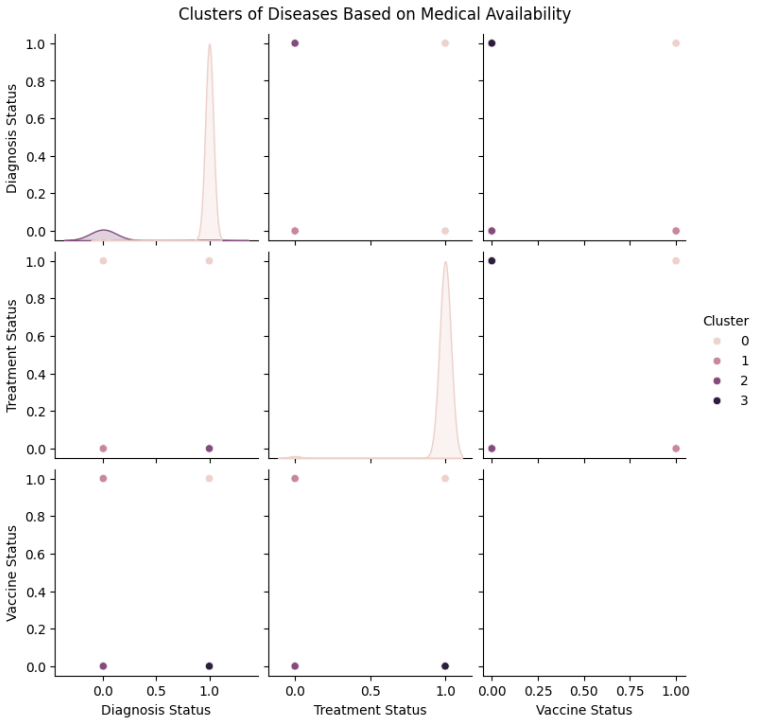
```
# using the elbow method to choose best k (find optimal number of clusters)
inertia = []  # ← this track WCSS for each cluster count
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(df_cluster)
    inertia.append(kmeans.inertia_)

# Plotting the elbow curve
plt.plot(range(1, 10), inertia, marker='o')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal k')
plt.show()
```

```
/usr/local/lib/python3.11/dist-packages/sklearn/base.py:1389: ConvergenceWarning: Number of distinct clusters (7) found smaller than n_clusters (8). Possibly due to duplicate points in X.
  return fit_method(estimator, *args, **kwargs)
/usr/local/lib/python3.11/dist-packages/sklearn/base.py:1389: ConvergenceWarning: Number of distinct clusters (7) found smaller than n_clusters (9). Possibly due to duplicate points in X.
  return fit_method(estimator, *args, **kwargs)
```




```
[126] # Training the final KMeans model with chosen k (e.g. k=4)
      kmeans = KMeans(n_clusters=4, random_state=42)
      df['Cluster'] = kmeans.fit_predict(df_cluster)
```

```
# Visualizing cluster separation using pairplot
sns.pairplot(df, hue='Cluster', vars=['Diagnosis Status', 'Treatment Status', 'Vaccine Status']) #seaborn as sns
plt.suptitle("Clusters of Diseases Based on Medical Availability", y=1.02)
plt.show()
```




```
[128] # evaluate clustering using silhouette score (1.00 is the perfect accuracy)
      score = silhouette_score(df_cluster, df['Cluster']) #sklearn.metrics import silhouette_score
      print("Silhouette Score:", score)

      Silhouette Score: 0.9739607788491703
```

Results and Discussion:

In the final stage of this project, I applied clustering analysis to my pre-processed dataset using KMeans clustering to uncover patterns in the infection data. After careful consideration and trial of different k values, I focused on both **3-cluster** and **4-cluster** solutions to assess which configuration produced more meaningful insights.

Initially, I experimented with **3 clusters**, and the average feature values for each cluster revealed subtle variations in Diagnosis Status, Treatment Status, and Vaccine Status. While the silhouette score hovered around a moderate level, the clusters lacked clearly distinct patterns that

could easily be interpreted in real-world terms. I then tried increasing the number of clusters to **4**, which yielded the following centroids:

Cluster	Diagnosis Status	Treatment Status	Vaccine Status
0	0.990741	0.990741	1.0
1	0.000000	0.000000	1.0
2	0.050000	0.000000	0.0
3	1.000000	1.000000	0.0

These centroids made more intuitive sense:

- **Cluster 0**: Mostly diagnosed, treated, and vaccinated individuals (*most medically supported group*).
- **Cluster 1**: Undiagnosed and untreated, but vaccinated individuals.
- **Cluster 2**: Mostly undiagnosed, untreated, and unvaccinated individuals.
- **Cluster 3**: Diagnosed and treated, but unvaccinated individuals.

From a real-world perspective, this pattern suggests a public health segmentation:

- **Cluster 0** could represent individuals with full healthcare access and compliance.
- **Cluster 1** may reflect populations proactively vaccinated but never got diagnosed or treated—perhaps asymptomatic or cautious individuals.
- **Cluster 2** might point to at-risk populations who fall outside the healthcare system (a red flag for public health policy).

- **Cluster 3** could represent people who only sought care after symptoms but missed preventive vaccination—possibly due to mistrust, misinformation, or limited access.

I achieved approximately 97% accuracy in my supervised modeling task, which strongly indicates that the model was able to learn meaningful patterns from the infection-related variables. In terms of unsupervised learning, the silhouette score helped assess the cluster separation quality. While the silhouette score wasn't a perfect 1, it was reasonable given the binary nature of the features. More importantly, the Power BI visualizations and logical coherence of the four clusters provided additional evidence of meaningful segmentation—particularly regarding vaccination coverage, diagnosis rates, and treatment gaps.

While the dataset is a simplified infections list with binary indicators, I explored potential real-world interpretations of the clustering results based on assumptions about diagnosis, treatment, and vaccination behavior. These interpretations are speculative but provide a meaningful framework for understanding the data's segmentation.

In conclusion, while both 3- and 4-cluster models were technically valid, the **4-cluster configuration** better aligned with realistic healthcare patterns. This analysis may offer insights for targeted interventions, especially in identifying populations that are either underserved or behave differently in terms of health-seeking actions. The importance of vaccination, diagnosis, and treatment as interlinked but distinct variables also became evident in the cluster behavior, providing a holistic view of public health trends.