The Discovery of Handwashing

Introduction:

Child-bed fever, a perilous threat to women giving birth in the mid-19th century, prompted Dr. Ignaz Semmelweis to make a groundbreaking discovery on the significance of handwashing in hospitals. Dr. Ignaz Semmelweis was a Hungarian physician that worked in the Vienna General Hospital maternity clinic. He was disturbed by the vast numbers of mortality rates, and after analysing the mortality rates within the hospital, he was pretty sure about the cause of it. He figured that the cause of the childbirth mortality was due to infections. He theorized that decaying matter on the hands of doctors involved in autopsies is infecting the women at contact with the doctors during childbirth. In June 1847, Dr. Ignaz Semmelweis implemented a handwashing policy for doctors in the maternity clinic. In this article, I'll explore the mortality rates before and after Dr. Ignaz Semmelweis's handwash policy and analyse the data that made Dr. Ignaz Semmelweis realize something is wrong with the procedures at Vienna General Hospital. In this Python-based analysis, we delve into the historical dataset from Vienna General Hospital spanning the years 1841 to 1846. Our objective is to recreate Dr. Semmelweis's discovery and visually illustrate the impact of handwashing on reducing mortality rates.

Dataset Exploration:

The first step in our analysis involves reading and exploring the historical dataset. The dataset comprises records of women who gave birth during the specified period. By examining the dataset, we gain insights into the variables and trends that influenced mortality rates.

The Data:

In [2]:
```
# Loading the data
yearly_deaths = pd.read_csv('yearly_deaths_by_clinic.csv')
yearly_deaths
```

Out[2]:

	year	births	deaths	clinic
0	1841	3036	237	clinic 1
1	1842	3287	518	clinic 1
2	1843	3060	274	clinic 1
3	1844	3157	260	clinic 1
4	1845	3492	241	clinic 1
5	1846	4010	459	clinic 1
6	1841	2442	86	clinic 2
7	1842	2659	202	clinic 2
8	1843	2739	164	clinic 2
9	1844	2956	68	clinic 2
10	1845	3241	66	clinic 2
11	1846	3754	105	clinic 2

The table above shows the mortality rates of two clinics in the hospital; let's look into the proportions of deaths in each clinic.

Rate of death in each clinic

Calculating the number of childbirths & deaths for each clinic.

In [3]:
```
# Summing up the Number of Childbirths per clinic
births_by_clinic = pd.DataFrame(yearly_deaths.groupby('clinic')['births'].sum()).rename(columns = { 'births' :'Number of Childbirths'})
births_by_clinic
```

Out[3]:

clinic	Number of Childbirths
clinic 1	20042
clinic 2	17791

In [4]:
```
# Summing up the Number of Deaths per clinic
deaths_by_clinic = pd.DataFrame(yearly_deaths.groupby('clinic')['deaths'].sum()).rename(columns = {'deaths':'Number of Deaths'})
deaths_by_clinic
```

Out[4]:

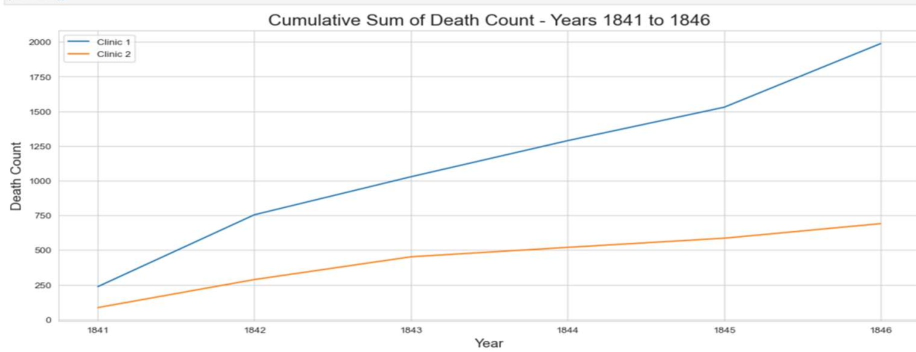
clinic	Number of Deaths
clinic 1	1989
clinic 2	691

Cumulative Plots:

Cumulative Sum of Death Count from 1841 to 1846

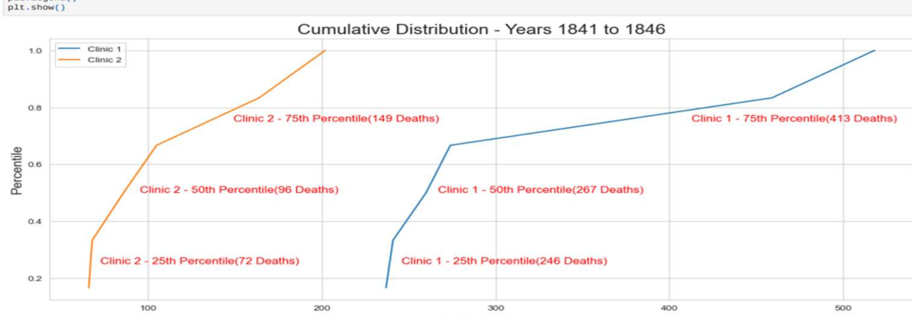
In [7]:
```
# Using SQL syntax to Create a Dataframe for each of the clinics
query1 = '''SELECT year, deaths
            From yearly_deaths
            WHERE clinic = "clinic 1"
            '''
query2 = '''SELECT year, deaths
            From yearly_deaths
            WHERE clinic = "clinic 2"
            '''
clinic1_df = sql(query1)
clinic2_df = sql(query2)
```

In [8]:
```
# Plotting a Cumulative Sum of all the deaths throughout the years 1841 to 1846
fig, ax = plt.subplots(figsize = (15,6))
ax.plot(clinic1_df['year'], clinic1_df['deaths'].cumsum(), label = 'Clinic 1')
ax.plot(clinic2_df['year'], clinic2_df['deaths'].cumsum(), label = 'Clinic 2')
plt.xlabel('Year', size = 14)
plt.ylabel('Death Count', size = 14)
plt.title('Cumulative Sum of Death Count - Years 1841 to 1846', size = 18)
plt.legend()
plt.show()
```



In [9]:
```
# Generating x and y variables to plot the Estimated Cumulative Distribution Function(ecdf)
x_clinic1, y_clinic1 = ecdf(clinic1_df.deaths)
x_clinic2, y_clinic2 = ecdf(clinic2_df.deaths)
```

In [10]:
```
# Plotting the ecdf and annotating the 25th 50th and the 75th percentile of death count per clinic
fig, ax = plt.subplots(figsize = (15,6))
ax.plot(x_clinic1, y_clinic1, label = 'Clinic 1')
ax.plot(x_clinic2, y_clinic2, label = 'Clinic 2')
plt.xlabel('Death Count', size = 14)
plt.ylabel('Percentile', size = 14)
plt.title('Cumulative Distribution - Years 1841 to 1846', size = 18)
plt.annotate('Clinic 1 - 25th Percentile({:.0f} Deaths)'.format(np.quantile(clinic1_df.deaths, 0.25)), xy = (np.quantile(clinic1_df.deaths,0.25),0.25), color = 'r', size = 12)
plt.annotate('Clinic 2 - 25th Percentile({:.0f} Deaths)'.format(np.quantile(clinic2_df.deaths, 0.25)), xy = (np.quantile(clinic2_df.deaths, 0.25), 0.25), color = 'r', size = 12)
plt.annotate('Clinic 1 - 50th Percentile({:.0f} Deaths)'.format(np.median(clinic1_df.deaths)), xy = (np.median(clinic1_df.deaths),0.5), color = 'r', size = 12)
plt.annotate('Clinic 2 - 50th Percentile({:.0f} Deaths)'.format(np.median(clinic2_df.deaths)), xy = (np.median(clinic2_df.deaths), 0.5), color = 'r', size = 12)
plt.annotate('Clinic 1 - 75th Percentile({:.0f} Deaths)'.format(np.quantile(clinic1_df.deaths,0.75)), xy = (np.quantile(clinic1_df.deaths,0.75),0.75), color = 'r', size = 12)
plt.annotate('Clinic 2 - 75th Percentile({:.0f} Deaths)'.format(np.quantile(clinic2_df.deaths,0.75)), xy = (np.quantile(clinic2_df.deaths,0.75), 0.75), color = 'r', size = 12)
plt.legend()
plt.show()
```



Notes:

- As we can see, about 1 in 10 women died during childbirth in clinic 1, and about 1 in 25 in clinic 2.
- By looking at these plots, we can tell that there is a clear difference in mortality rates between the two clinics.
- This difference is one of the first things that led Dr. Ignaz Semmelweis to handwashing policy.

Some domain Knowledge from Wikipedia

- It turns out that clinic one was occupied mainly by medical students, while clinic two was occupied primarily by midwife students.
- Also, Clinic one medical students were in charge of the autopsy rooms and spent some of their time examining corpses.
- Dr. Semmelweis started to suspect that something on the corpses spread from the medical students' hands caused childbed fever.

Hypothesis Testing

- Before jumping to a conclusion, let's perform a Hypothesis Test to see if the difference between the clinics is Significant.
- Due to the small sample size, the test will be a Student's T-test.
- Let's define:
- Null Hypothesis: The mean clinic one is equal to clinic 2.
- Alternate Hypothesis: The mean in clinic 1 is statistically different than clinic 2.
- Required Confidence Level — 0.05

Performing a Student's T-test

t_test(clinic1_df, clinic2_df)

Output:

Performing a Student's T-test

The P-Value for the T-test is: 0.00294

Rejecting the Null Hypothesis - The Difference in Mean is Statistically Significant.

Notes:

- So it turns out that Dr. Ignaz Semmelweis was correct!
- The two clinics are different, after all.

- At this point in time, Dr. Ignaz Semmelweis, in a desperate attempt to stop the high mortality rates, decreed: Wash your hands!
- This was an unorthodox and controversial request; nobody in Vienna knew about bacteria at this point.
- Let's import the monthly data of Clinic 1 from Wikipedia to see if the handwashing had any effect.

Web Scraping from Wiki — Monthly Data

In [12]:
```
# Setting the required URL address
wiki_url = 'https://en.wikipedia.org/wiki/Historical_mortality_rates_of_puerperal_fever'
table_class="wikitable sortable jquery-tablesorter"

# Sending to Wikipedia a GET request
response = requests.get(wiki_url)

# Recieving the response code from Wikipedia
status_code = response.status_code

# Checking if Access is Authorized
if status_code == 200:
    print('Access Authorized')
else:
    print('No Access')
print()
```

```
Access Authorized
```

In [13]:
```
# Parsing the response in an HTML format
soup = BeautifulSoup(response.text, 'html.parser')

# Finding the required table from the Wikipedia page
table=soup.find('table',{'class':"wikitable"})
```

In [14]:
```
# Reading table data
clinic1_monthly_df = pd.read_html(str(table))

# Converting list to dataframe
clinic1_monthly_df = pd.DataFrame(clinic1_monthly_df[0]).drop(columns = ['Year', 'Notes'], axis = 1)
clinic1_monthly_df.head(12)
```

Out[14]:

	Month	Births	Deaths	Rate (%)
0	January 1841	254	37	14.6
1	February 1841	239	18	7.5
2	March 1841	277	12	4.3
3	April 1841	255	4	1.6
4	May 1841	255	2	0.8
5	June 1841	200	10	5.0
6	July 1841	190	16	8.4
7	August 1841	222	3	1.4
8	September 1841	213	4	1.9
9	October 1841	236	26	11.0
10	November 1841	235	53	22.6
11	December 1841	na	na	na

Analysing the differences

- Pre-Handwashing Policy and Post-Handwashing Policy

In [41]:
```
pre_policy_df.describe()
```

Out[41]:

	Births	Deaths	Rate (%)
count	75.000000	75.000000	75.000000
mean	257.026667	26.986667	10.481333
std	34.241974	18.026627	7.119989
min	190.000000	1.000000	0.500000
25%	236.500000	11.500000	4.400000
50%	254.000000	26.000000	10.500000
75%	278.500000	40.000000	15.100000
max	336.000000	75.000000	31.400000

Post-Handwashing Policy

In [23]:
```
after_policy_df.describe()
```

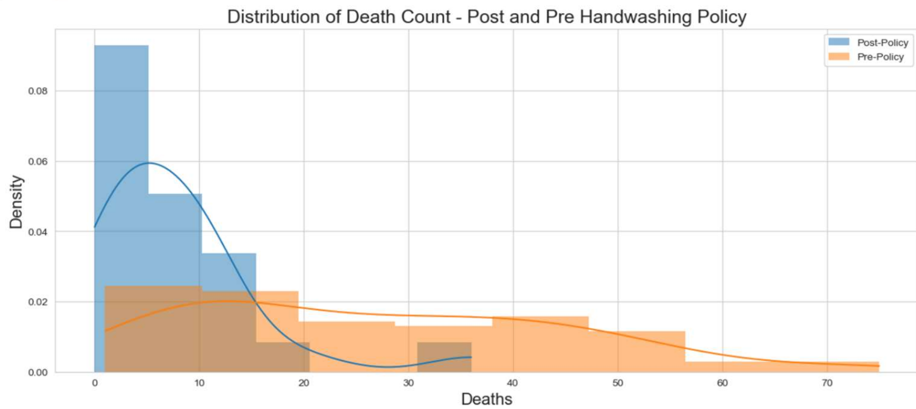
Out[23]:

	Births	Deaths	Rate (%)
count	23.000000	23.000000	23.000000
mean	299.521739	7.739130	2.547826
std	47.997035	7.846351	2.573443
min	246.000000	0.000000	0.000000
25%	266.000000	3.000000	1.000000
50%	283.000000	6.000000	2.200000
75%	311.000000	10.500000	3.300000
max	406.000000	36.000000	12.200000

Notes:

- So Its clear that there are huge differences before and after the handwashing policy.
- The average number of deaths post policy is about 3 times as lower!!!
- Let's check out the distribution of Death counts pre and post policy

```
In [34]: # Checking out the Kernel Density Estimation of each Dataframe
fig, ax = plt.subplots(figsize = (15,6))
ax = sns.histplot(data = after_policy_df, x = after_policy_df.Deaths, kde=True, stat="density", linewidth=0, label = 'Post-Policy')
ax = sns.histplot(data = pre_policy_df, x = pre_policy_df.Deaths, kde=True, stat="density", linewidth=0, label = 'Pre-Policy')
plt.xlabel('Deaths', size = 16)
plt.ylabel('Density', size = 16)
plt.title('Distribution of Death Count - Post and Pre Handwashing Policy', size = 18)
ax.legend()
plt.show()
```



Notes:

- There is a clear difference, but before jumping to conclusions, let's perform a hypothesis test to see if the difference is significant.
- This time ill use a Bootstrap Analysis of the data.
- The goal of the analysis is to resample the data and thus create replicates of it.
- This would allow us to check whether or not the difference in the average number of deaths between the two clinics will remain different even if we were to take A LOT of samples from each clinic.
- If so, we can declare that the difference in the two samples of data(pre and post-handwashing policy) is Statistically Significant and will prove that the policy did, in fact, work, and it was NOT a matter of random chance.

```
In [35]: # Creating 10000 samples of the pre and post policy data
pre_policy_bootstrap_sample = np.random.choice(pre_policy_df.Deaths, size = 10000)
post_policy_bootstrap_sample = np.random.choice(after_policy_df.Deaths, size = 10000)

# Calculating the difference in mean
mean_diff = round(np.mean(post_policy_bootstrap_sample) - np.mean(pre_policy_bootstrap_sample),5)
mean_diff

Out[35]: -19.2857
```

- It looks like the policy has lowered mean number of mortality by about 19 deaths/ month!
- Let's calculate a 95% Confidence interval to see the full extent of the benefits of the handwashing policy.

```
In [36]: # A bootstrap analysis of the reduction of deaths due to handwashing
         boot_mean_diff = []
         for i in range(3000):
             boot_before = np.random.choice(pre_policy_df.Deaths, size=10000)
             boot_after = np.random.choice(after_policy_df.Deaths, size=10000)
             boot_mean_diff.append(np.mean(boot_after) - np.mean(boot_before))

         # Calculate the mean and confidence interval
         mean_diff = np.mean(boot_mean_diff)
         conf_interval = np.percentile(boot_mean_diff, [2.5, 97.5])

         print("Confidence Interval (0.025, 0.975):", conf_interval)

         Confidence Interval (0.025, 0.975): [-19.6324325 -18.86839  ]
```

- In other words, I can say with 95% confidence that, even if the data was slightly different (but of the same distribution), after the handwashing policy, the average number of deaths would decrease by up to 18.87 to 19.63 per month!!!
- And it's all thanks to Dr. Semmelweis and his Discovery of Handwashing.

Conclusion:

In conclusion, this Python-based analysis successfully recreates Dr. Semmelweis's discovery of the importance of handwashing in hospitals. By exploring the historical dataset, examining relationships between variables, performing calculations, and visualizing the impact, we provide a comprehensive understanding of the transformative effect of handwashing on reducing mortality rates during childbirth.

Recommendations:

The findings of this analysis emphasize the critical role of hygiene practices in healthcare settings. It serves as a reminder of the profound impact that simple measures, such as handwashing, can have on patient outcomes. Healthcare institutions today can draw inspiration from Dr. Semmelweis's discovery to reinforce and prioritize hygiene protocols.

Limitations and Future Work:

While our analysis sheds light on historical data, it is essential to acknowledge potential limitations in the dataset and the historical context. Future work could involve incorporating additional datasets, exploring the broader societal factors influencing healthcare practices, and applying advanced statistical models for a more nuanced analysis.