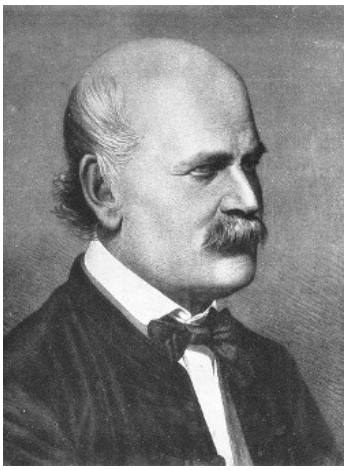
DR. SEMMELWEIS AND THE DISCOVERY OF HANDWASHING

datalab

1. Meet Dr. Ignaz Semmelweis



This is Dr. Ignaz Semmelweis, a Hungarian physician born in 1818 and active at the Vienna General Hospital. If Dr. Semmelweis looks troubled it's probably because he's thinking about *childbed fever*: A deadly disease affecting women that just have given birth. He is thinking about it because in the early 1840s at the Vienna General Hospital as many as 10% of the women giving birth die from it. He is thinking about it because he knows the cause of childbed fever: It's the contaminated hands of the doctors delivering the babies. And they won't listen to him and *wash their hands*!

In this notebook, we're going to reanalyze the data that made Semmelweis discover the importance of *handwashing*. Let's start by looking at the data that made Semmelweis realize that something was wrong with the procedures at Vienna General Hospital.

```
# Load in the tidyverse package
# .... YOUR CODE FOR TASK 1 ....
library(tidyverse)
# Read datasets/yearly_deaths_by_clinic.csv into yearly
yearly <- read_csv("datasets/yearly_deaths_by_clinic.csv")

# Print out yearly
# .... YOUR CODE FOR TASK 1 ....
print(yearly)
```

```
Rows: 12 Columns: 4
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): clinic
dbl (3): year, births, deaths

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# A tibble: 12 × 4
    year births deaths clinic
   <dbl>  <dbl>  <dbl> <chr>
 1  1841   3036    237 clinic 1
 2  1842   3287    518 clinic 1
 3  1843   3060    274 clinic 1
 4  1844   3157    260 clinic 1
 5  1845   3492    241 clinic 1
 6  1846   4010    459 clinic 1
 7  1841   2442     86 clinic 2
 8  1842   2659    202 clinic 2
 9  1843   2739    164 clinic 2
10  1844   2956     68 clinic 2
11  1845   3241     66 clinic 2
12  1846   3754    105 clinic 2
```

2. The alarming number of deaths

The table above shows the number of women giving birth at the two clinics at the Vienna General Hospital for the years 1841 to 1846. You'll notice that giving birth was very dangerous; an *alarming* number of women died as the result of childbirth, most of them from childbed fever.

We see this more clearly if we look at the *proportion of deaths* out of the number of women giving birth.

```
# Adding a new column to yearly with proportion of deaths per no. births
# .... YOUR CODE FOR TASK 1 ....
yearly <- yearly %>%
    mutate(proportion_deaths= deaths/births)
# Print out yearly
yearly
```

index	year	births	deaths	clinic	proportion_deaths
1	1841	3036	237	clinic 1	0.0781
2	1842	3287	518	clinic 1	0.1576
3	1843	3060	274	clinic 1	0.0895
4	1844	3157	260	clinic 1	0.0824
5	1845	3492	241	clinic 1	0.069
6	1846	4010	459	clinic 1	0.1145
7	1841	2442	86	clinic 2	0.0352
8	1842	2659	202	clinic 2	0.076
9	1843	2739	164	clinic 2	0.0599
10	1844	2956	68	clinic 2	0.023
11	1845	3241	66	clinic 2	0.0204
12	1846	3754	105	clinic 2	0.028

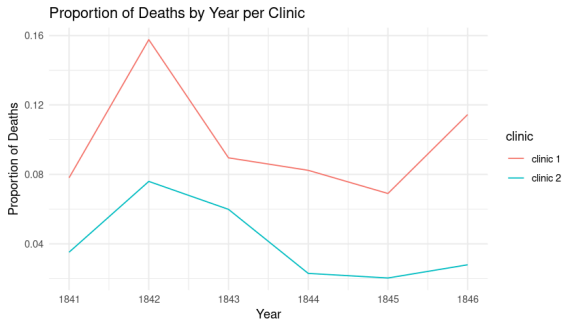
Rows: 12 Expand

3. Death at the clinics

If we now plot the proportion of deaths at both clinic 1 and clinic 2 we'll see a curious pattern...

```
# Setting the size of plots in this notebook
options(repr.plot.width=7, repr.plot.height=4)

# Plot yearly proportion of deaths at the two clinics
# .... YOUR CODE FOR TASK 3 ....
ggplot(yearly, aes(x = year, y = proportion_deaths, color = clinic, group = clinic)) +
  geom_line() +
  labs(title = "Proportion of Deaths by Year per Clinic",
       x = "Year",
       y = "Proportion of Deaths") +
  theme_minimal()
```



4. The handwashing begins

Why is the proportion of deaths constantly so much higher in Clinic 1? Semmelweis saw the same pattern and was puzzled and distressed. The only difference between the clinics was that many medical students served at Clinic 1, while mostly midwife students served at Clinic 2. While the midwives only tended to the women giving birth, the medical students also spent time in the autopsy rooms examining corpses.

Semmelweis started to suspect that something on the corpses, spread from the hands of the medical students, caused childbed fever. So in a desperate attempt to stop the high mortality rates, he decreed: *Wash your hands!* This was an unorthodox and controversial request, nobody in Vienna knew about bacteria at this point in time.

Let's load in monthly data from Clinic 1 to see if the handwashing had any effect.

```r
# Read datasets/monthly_deaths.csv into monthly
monthly <- read_csv("datasets/monthly_deaths.csv")

# Adding a new column with proportion of deaths per no. births
# .... YOUR CODE FOR TASK 4 ....
monthly$proportion_deaths <- monthly$deaths / monthly$births
# Print out the first rows in monthly
# .... YOUR CODE FOR TASK 4 ....
head(monthly)
```

```
Rows: 98 Columns: 3
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl  (2): births, deaths
date (1): date

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

	d…			proportion_d…
1	1841-01-01	254	37	0.1457
2	1841-02-01	239	18	0.0753
3	1841-03-01	277	12	0.0433
4	1841-04-01	255	4	0.0157
5	1841-05-01	255	2	0.0078
6	1841-06-01	200	10	0.05

Rows: 6 Expand

5. The effect of handwashing

With the data loaded we can now look at the proportion of deaths over time. In the plot below we haven't marked where obligatory handwashing started, but it reduced the proportion of deaths to such a degree that you should be able to spot it!

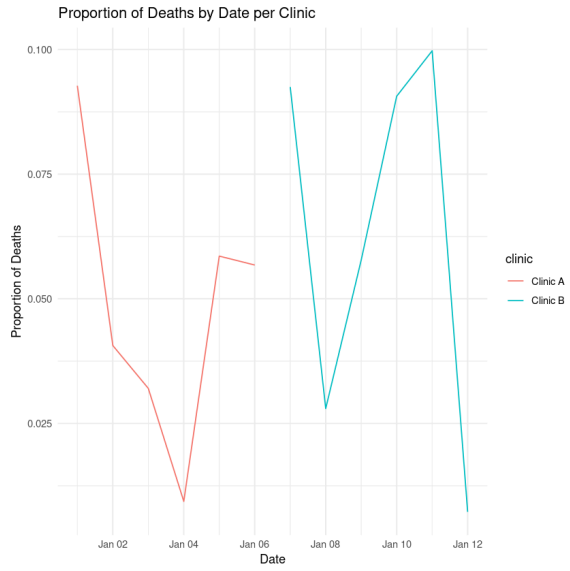
```r
# Load the ggplot2 package
library(ggplot2)

# Assuming 'monthly' is a data frame that needs to be created or loaded
# Here is an example of how you might create a sample 'monthly' data frame
monthly <- data.frame(
  date = as.Date('2021-01-01') + 0:11,  # Example dates
  proportion_deaths = runif(12, 0, 0.1),  # Example proportions
  clinic = rep(c("Clinic A", "Clinic B"), each = 6)  # Example clinic names
)

# Print the 'monthly' data frame to ensure it is created correctly
print(monthly)

# Plot monthly proportion of deaths
ggplot(monthly, aes(x = date, y = proportion_deaths, color = clinic, group = clinic)) +
  geom_line() +
  labs(title = "Proportion of Deaths by Date per Clinic",
       x = "Date",
       y = "Proportion of Deaths") +
  theme_minimal()
```

```
         date proportion_deaths   clinic
1  2021-01-01       0.092742484 Clinic A
2  2021-01-02       0.040630688 Clinic A
3  2021-01-03       0.032035033 Clinic A
4  2021-01-04       0.009366700 Clinic A
5  2021-01-05       0.058568899 Clinic A
6  2021-01-06       0.056779536 Clinic A
7  2021-01-07       0.092499234 Clinic B
8  2021-01-08       0.027992151 Clinic B
9  2021-01-09       0.057771039 Clinic B
10 2021-01-10       0.090639006 Clinic B
11 2021-01-11       0.099757720 Clinic B
12 2021-01-12       0.007267178 Clinic B
```



6. The effect of handwashing highlighted

Starting from the summer of 1847 the proportion of deaths is drastically reduced and, yes, this was when Semmelweis made handwashing obligatory.

The effect of handwashing is made even more clear if we highlight this in the graph.

```
# From this date handwashing was made mandatory
handwashing_start = as.Date('1847-06-01')

# Add a TRUE/FALSE column to monthly called handwashing_started
# .... YOUR CODE FOR TASK 6 ....
monthly <- monthly %>%
  mutate(handwashing_started = date >= handwashing_start)
# Plot monthly proportion of deaths before and after handwashing
# .... YOUR CODE FOR TASK 6 ....
ggplot(monthly, aes(x = date, y = proportion_deaths, color = handwashing_started, group = clinic)) +
  geom_line() +
  labs(title = "Proportion of Deaths by Date per Clinic",
       x = "Date",
       y = "Proportion of Deaths",
       color = "Handwashing Started") +
  theme_minimal()
```

```
ERROR while rich displaying an object: Error in `geom_line()`:
! Problem while computing aesthetics.
i Error occurred in the 1st layer.
Caused by error in `FUN()`:
! object 'clinic' not found

Traceback:
1. tryCatch(withCallingHandlers({
 .     if (!mime %in% names(repr::mime2repr))
 .         stop("No repr_* for mimetype ", mime, " in repr::mime2repr")
 .     rpr <- repr::mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler), error = outer_handler)
2. tryCatchList(expr, classes, parentenv, handlers)
3. tryCatchOne(expr, names, parentenv, handlers[[1L]])
4. doTryCatch(return(expr), name, parentenv, handler)
5. withCallingHandlers({
 .     if (!mime %in% names(repr::mime2repr))
 .         stop("No repr_* for mimetype ", mime, " in repr::mime2repr")
 .     rpr <- repr::mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler)
6. repr::mime2repr[[mime]](obj)
```

7. More handwashing, fewer deaths?

Again, the graph shows that handwashing had a huge effect. How much did it reduce the monthly proportion of deaths on average?

```
# Calculating the mean proportion of deaths
# before and after handwashing.

# .... YOUR CODE FOR TASK 7 HERE ....
monthly_summary <- monthly %>%
  group_by(handwashing_started) %>%
  summarise(mean_proportion_deaths = mean(proportion_deaths))
# Printing out the summary.
monthly_summary
```

	handwashing_st...	mean_proportion_dea...
1	False	0.105
2	True	0.0211

Rows: 2 Expand

8. A statistical analysis of Semmelweis handwashing data

It reduced the proportion of deaths by around 8 percentage points! From 10% on average before handwashing to just 2% when handwashing was enforced (which is still a high number by modern standards). To get a feeling for the uncertainty around how much handwashing reduces mortalities we could look at a confidence interval (here calculated using a t-test).

```r
# Calculating a 95% Confidence intrerval using t.test
test_result <- t.test( proportion_deaths ~ handwashing_started, data = monthly)
test_result
```

```
    Welch Two Sample t-test

data:  proportion_deaths by handwashing_started
t = 9.6101, df = 92.435, p-value = 1.445e-15
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 0.06660662 0.10130659
sample estimates:
mean in group FALSE  mean in group TRUE
         0.10504998          0.02109338
```

9. The fate of Dr. Semmelweis

That the doctors didn't wash their hands increased the proportion of deaths by between 6.7 and 10 percentage points, according to a 95% confidence interval. All in all, it would seem that Semmelweis had solid evidence that handwashing was a simple but highly effective procedure that could save many lives.

The tragedy is that, despite the evidence, Semmelweis' theory — that childbed fever was caused by some "substance" (what we today know as *bacteria*) from autopsy room corpses — was ridiculed by contemporary scientists. The medical community largely rejected his discovery and in 1849 he was forced to leave the Vienna General Hospital for good.

One reason for this was that statistics and statistical arguments were uncommon in medical science in the 1800s. Semmelweis only published his data as long tables of raw data, but he didn't show any graphs nor confidence intervals. If he would have had access to the analysis we've just put together he might have been more successful in getting the Viennese doctors to wash their hands.

```r
# The data Semmelweis collected points to that:
doctors_should_wash_their_hands <- TRUE
```

Hidden output