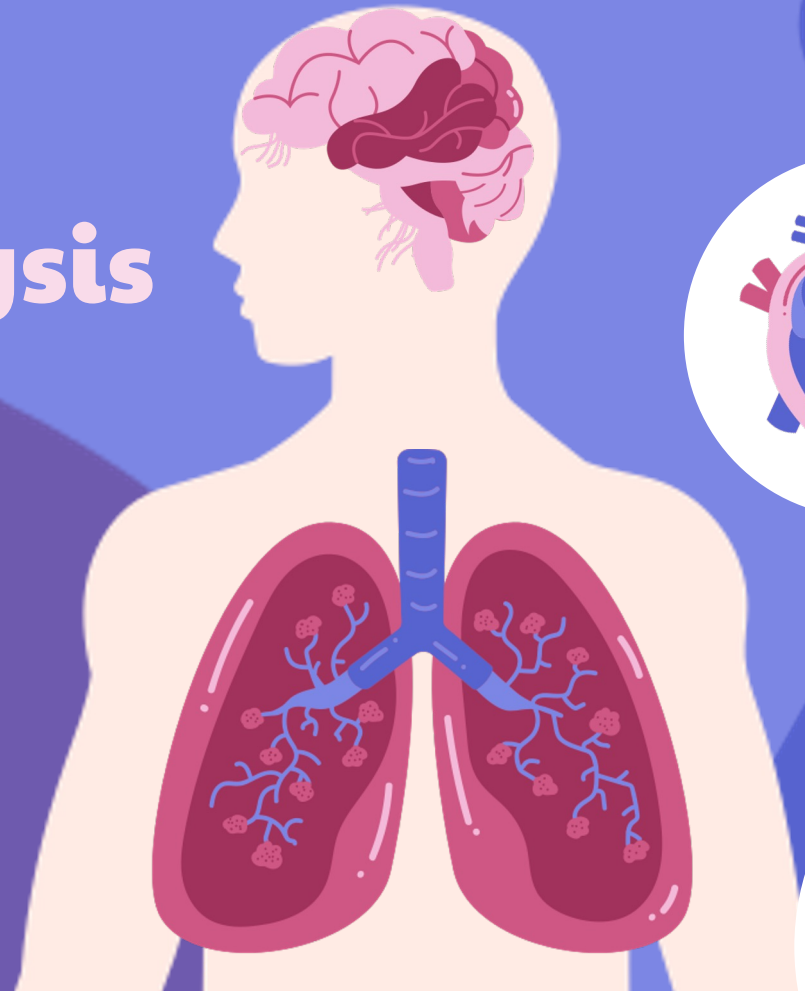
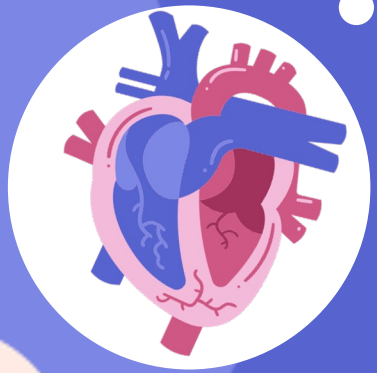
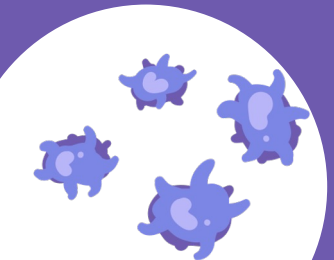
MSDS401: Data Analysis Project 2

Yuying Wang, Xinyi Wei, Yanqian Pan
CJ Tuan, Zeyang Yu

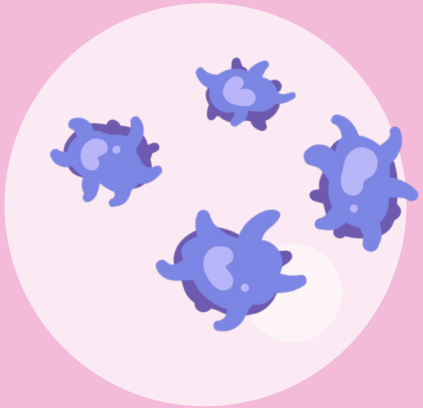




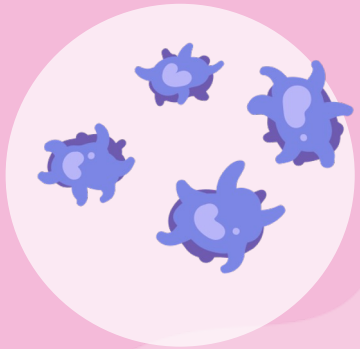


Agenda

Part 1 Descriptive Statistics
Part 2 Inferential Statistics
Part 3 Correlation
Part 4 Regression
Part 5 Conclusion



Part 1
Descriptive Statistics



Summary of Data

- The data we analyzed is about new daily covid-19 cases and deaths in European Union (EU) and European Economic Area (EEA) countries.
- Before cleaning, this dataset contains **28729 observations and 10 variables**
- Data source: (https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country)

Summary of Data

	dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoId	countryterritoryCode	popData2020
1	2022-10-23	23	10	2022	3557	0	Austria	AT	AUT	8901064
2	2022-10-22	22	10	2022	5494	4	Austria	AT	AUT	8901064
3	2022-10-21	21	10	2022	7776	4	Austria	AT	AUT	8901064
4	2022-10-20	20	10	2022	8221	6	Austria	AT	AUT	8901064
5	2022-10-19	19	10	2022	10007	8	Austria	AT	AUT	8901064
6	2022-10-18	18	10	2022	13204	7	Austria	AT	AUT	8901064
7	2022-10-17	17	10	2022	9964	8	Austria	AT	AUT	8901064
8	2022-10-16	16	10	2022	6606	12	Austria	AT	AUT	8901064
9	2022-10-15	15	10	2022	8818	6	Austria	AT	AUT	8901064
10	2022-10-14	14	10	2022	11751	10	Austria	AT	AUT	8901064
11	2022-10-13	13	10	2022	13068	14	Austria	AT	AUT	8901064
12	2022-10-12	12	10	2022	14305	13	Austria	AT	AUT	8901064
13	2022-10-11	11	10	2022	18498	11	Austria	AT	AUT	8901064
14	2022-10-10	10	10	2022	13369	10	Austria	AT	AUT	8901064
15	2022-10-09	9	10	2022	8689	10	Austria	AT	AUT	8901064
16	2022-10-08	8	10	2022	11230	12	Austria	AT	AUT	8901064
17	2022-10-07	7	10	2022	14391	18	Austria	AT	AUT	8901064
18	2022-10-06	6	10	2022	15560	10	Austria	AT	AUT	8901064
19	2022-10-05	5	10	2022	15291	13	Austria	AT	AUT	8901064
20	2022-10-04	4	10	2022	18531	16	Austria	AT	AUT	8901064
21	2022-10-03	3	10	2022	14200	16	Austria	AT	AUT	8901064

Data Cleaning

- Null Observations Identified: there are several missing data detected.
- Inconsistencies in Cases and Deaths Data: there are several unreasonable negative values in **“cases” and “deaths”** variables.
- Based on the requirement:
    - Develop a variable named “incidence_rate”, which are the daily new cases per 100,000 individuals for each country.
    - Construct a variable named “fatality_rate”, which are the new deaths per 100,000 individuals for each country.

After Cleaned

- After cleaning, our dataset now comprises **28313 observations and 12 variables.**

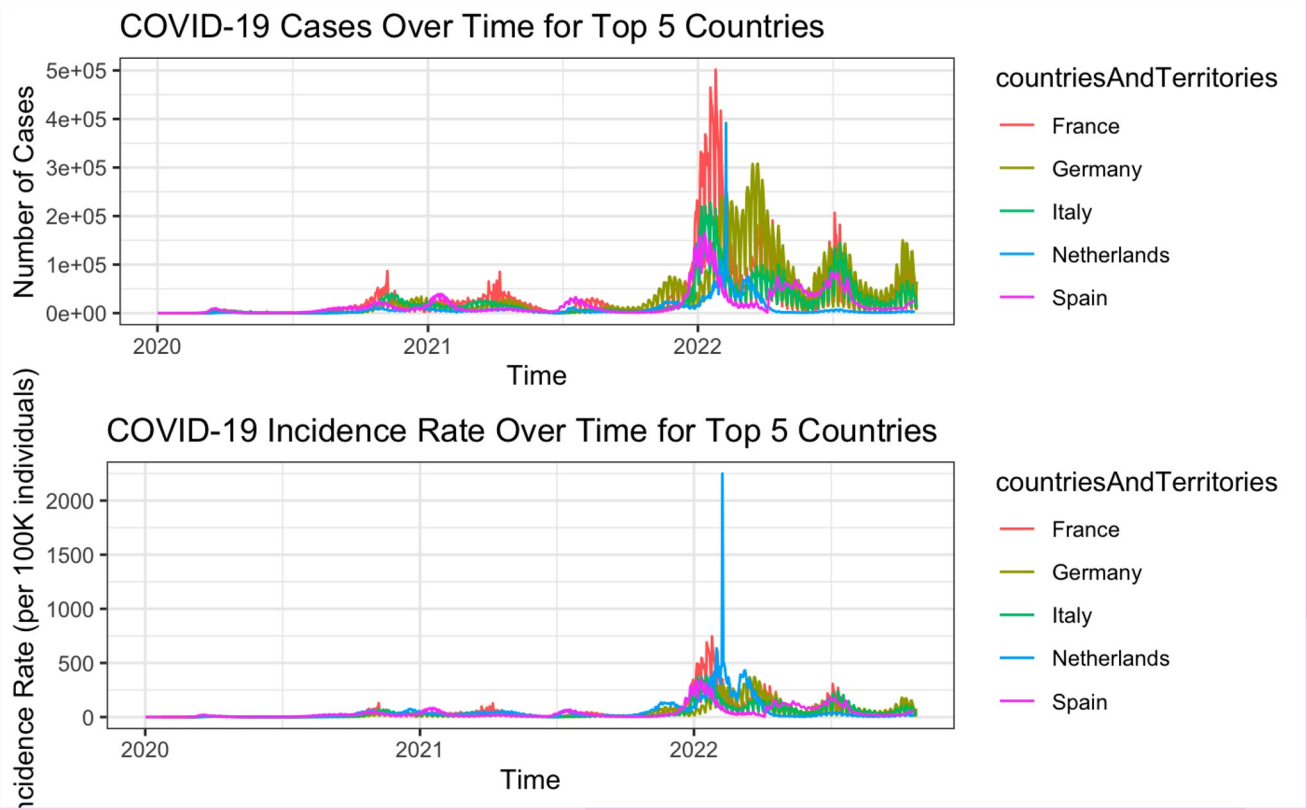
	dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoId	countryterritoryCode	popData2020	incidence_rate	fatality_rate
1	2022-10-23	23	10	2022	3557	0	Austria	AT	AUT	8901064	39.96151	0.00000000
2	2022-10-22	22	10	2022	5494	4	Austria	AT	AUT	8901064	61.72296	0.04493845
3	2022-10-21	21	10	2022	7776	4	Austria	AT	AUT	8901064	87.36034	0.04493845
4	2022-10-20	20	10	2022	8221	6	Austria	AT	AUT	8901064	92.35974	0.06740767
5	2022-10-19	19	10	2022	10007	8	Austria	AT	AUT	8901064	112.42476	0.08987690
6	2022-10-18	18	10	2022	13204	7	Austria	AT	AUT	8901064	148.34182	0.07864228
7	2022-10-17	17	10	2022	9964	8	Austria	AT	AUT	8901064	111.94167	0.08987690
8	2022-10-16	16	10	2022	6606	12	Austria	AT	AUT	8901064	74.21585	0.13481534
9	2022-10-15	15	10	2022	8818	6	Austria	AT	AUT	8901064	99.06681	0.06740767
10	2022-10-14	14	10	2022	11751	10	Austria	AT	AUT	8901064	132.01793	0.11234612
11	2022-10-13	13	10	2022	13068	14	Austria	AT	AUT	8901064	146.81391	0.15728457
12	2022-10-12	12	10	2022	14305	13	Austria	AT	AUT	8901064	160.71112	0.14604996
13	2022-10-11	11	10	2022	18498	11	Austria	AT	AUT	8901064	207.81785	0.12358073
14	2022-10-10	10	10	2022	13369	10	Austria	AT	AUT	8901064	150.19553	0.11234612
15	2022-10-09	9	10	2022	8689	10	Austria	AT	AUT	8901064	97.61754	0.11234612
16	2022-10-08	8	10	2022	11230	12	Austria	AT	AUT	8901064	126.16469	0.13481534
17	2022-10-07	7	10	2022	14391	18	Austria	AT	AUT	8901064	161.67730	0.20222302

Five Countries we chose

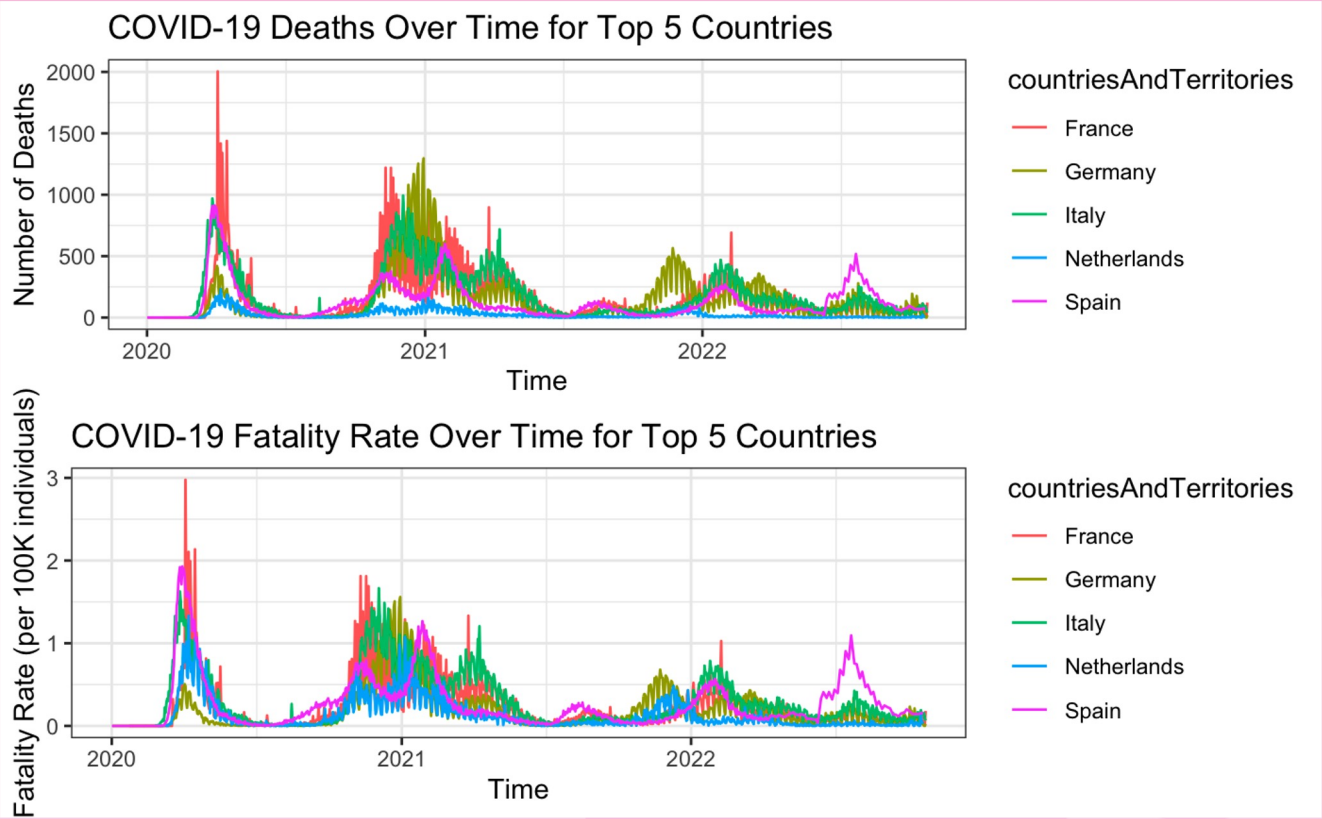
	countriesAndTerritories <fctr>	cases <int>
10	France	36952159
11	Germany	35287690
16	Italy	23359251
29	Spain	13564823
22	Netherlands	8494705

- Focused on countries with the highest COVID-19 case counts.
- Focused on countries that are among the most severely affected by the pandemic.
- Find out why COVID-19 spread quickly in those countries and the similarities between those countries

Exploring New cases & Incidence rates



Exploring New deaths & Fatality rates



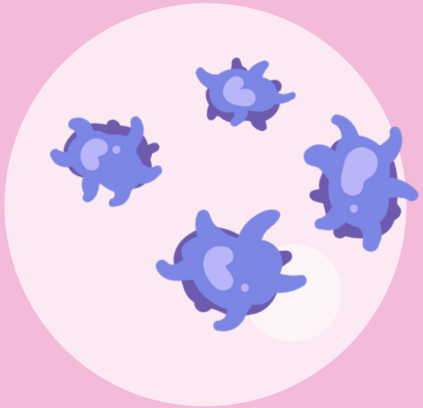
Explore case fatality rates per country

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1205  0.4259  0.6003  0.7968  0.8026  2.9715
```

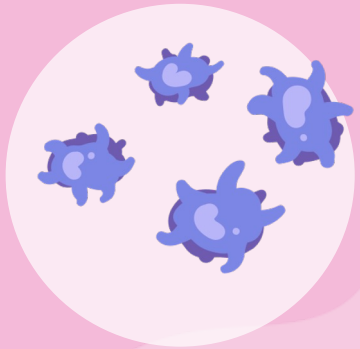
	countriesAndTerritories	cases	deaths	case_fatality_rate
14	Iceland	176725	213	0.1205262
5	Cyprus	599500	1188	0.1981651
7	Denmark	3221572	7116	0.2208860
22	Netherlands	8494705	22771	0.2680611
23	Norway	1454895	4190	0.2879933

	countriesAndTerritories	cases	deaths	case_fatality_rate
3	Bulgaria	1271735	37790	2.9715310
13	Hungary	2141513	47938	2.2385108
26	Romania	3246412	67179	2.0693307
24	Poland	6189562	118050	1.9072432
4	Croatia	1244692	17085	1.3726287

Formula:
(number of deaths/total number of cases)*100



Part 2
Inferential Statistics



Comparing the Incidence of Two Countries

- Prompt: Select two countries of your choosing and compare their incidence rates using hypothesis testing
  - Two Countries: France and Germany
  - $\mu 1$: the mean COVID-19 incidence rate per 100K people in France
  - $\mu 2$: the mean COVID-19 incidence rate per 100K people in Germany

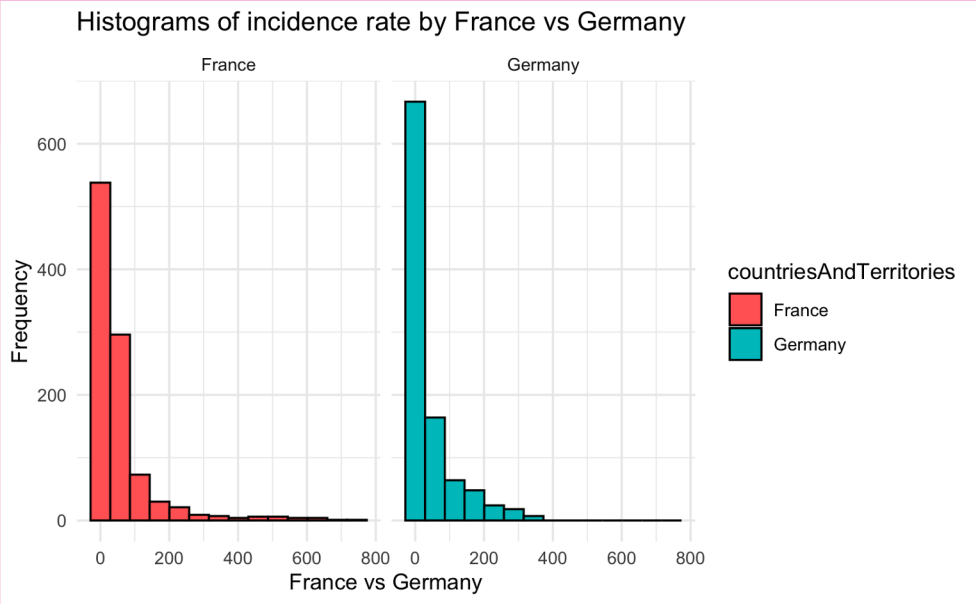
Null and Alternative Hypothesis

- Null & Alternative Hypothesis Statement

H0: $\mu 1 - \mu 2 = 0$
H0: There is no difference between the mean COVID-19 incidence rates per 100K people of France and Germany

Ha: $\mu 1 - \mu 2 \neq 0$
Ha: There is a difference between the mean COVID-19 incidence rates per 100K people of France and Germany

Histograms of Incidence Rate




```
Skewness for France incidence rate: 3.656043
Kurtosis for France incidence rate: 18.57803
Skewness for Germany incidence rate: 2.369289
Kurtosis for Germany incidence rate: 8.570674
```

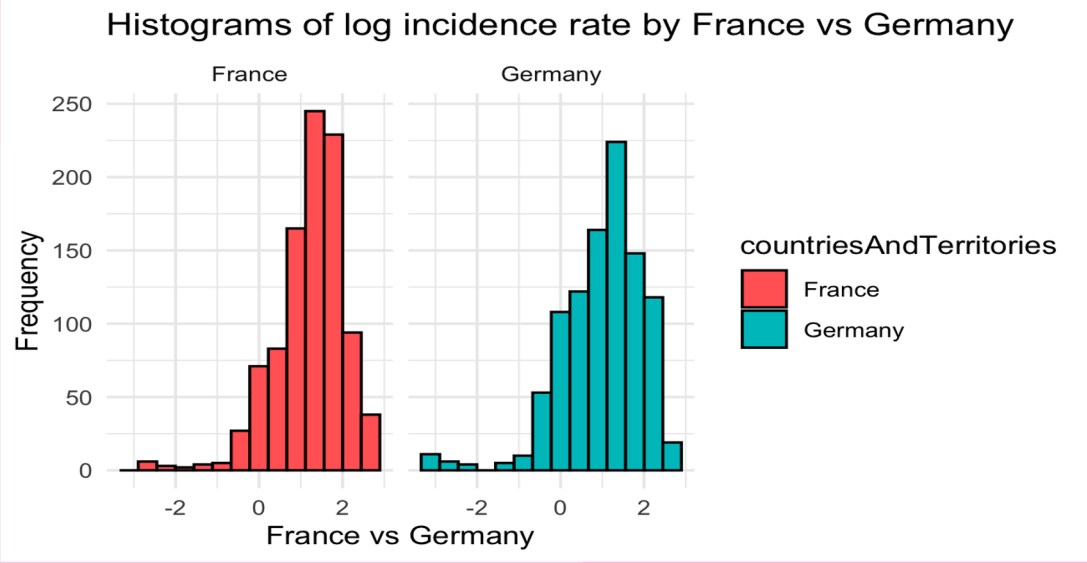
- The histograms are not normally distributed
  - Long, heavy tail
  - High skewness and kurtosis
- Normal distribution is a key assumption for parametric tests like the t-test, so conduct 3 inferential tests:
  - Independent two sample T-test assuming normal distribution
  - T-test after log transformation
  - Wilcoxon rank-sum test (a test used for non-normal data distribution)

(1)T-Test assuming normal distribution

```
##
##  Welch Two Sample t-test
##
## data:  France$incidence_rate and Germany$incidence_rate
## t = 3.2089, df = 1784.6, p-value = 0.001356
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.711288 19.524500
## sample estimates:
## mean of x mean of y
##  54.89014  42.77224
```

- The p-value is smaller than alpha 0.05 so we reject the null and accept the alternative

(2) T-Test after Log Transformation




```
Skewness for France incidence rate: -1.192788
Kurtosis for France incidence rate: 6.010383
Skewness for Germany incidence rate: -1.179382
Kurtosis for Germany incidence rate: 5.562636
```

- The histograms are more normally distributed than before log transformation
  - More normally distributed histograms
  - Lower skewness and kurtosis

(2) T-Test after Log Transformation

```
##
##  Welch Two Sample t-test
##
## data:  France$log10_incidence_rate and Germany$log10_incidence_rate
## t = 4.8681, df = 1950, p-value = 1.218e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1199236 0.2817398
## sample estimates:
## mean of x mean of y
## 1.1891016 0.9882699
```

- The p-value is smaller than alpha 0.05 so we reject the null and accept the alternative

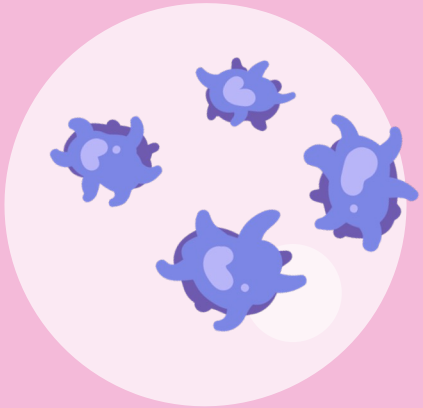
(3) Wilcoxon Rank-Sum Test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  incidence_rate by countriesAndTerritories
## W = 551524, p-value = 1.52e-05
## alternative hypothesis: true location shift is not equal to 0
```

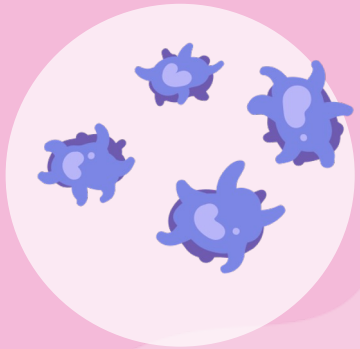
- The key assumption of parametric tests such as the t-test is the data must be normally distributed
- Wilcoxon rank-sum test is a non-parametric test that can be used when data is not normally distributed
- The p-value is smaller than alpha 0.05 so we reject the null and accept the alternative

Conclusion of Inferential Tests

- Our original data was not normally distributed so besides using the t-test, we also used log transformation and Wilcoxon rank-sum test to conduct inferential statistics
- All three hypothesis tests showed that the p-value is smaller than alpha 0.05, so we can confidently conclude to reject the null and accept the alternative
  - Ha: $\mu 1 - \mu 2 \neq 0$
  - Ha: There is a difference between the mean COVID-19 incidence rates per 100K people in France and Germany
- Because we use a confidence interval of 95%, we are 95% confident that there is a difference between the means of COVID-19 incidence rates between France and Germany



Part 3
Correlation



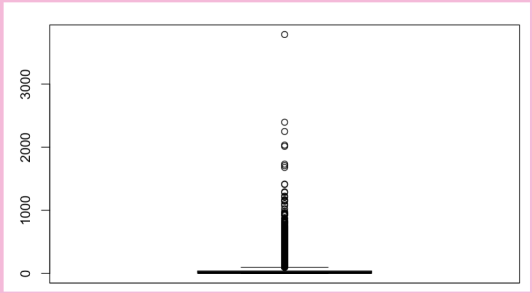
Correlation

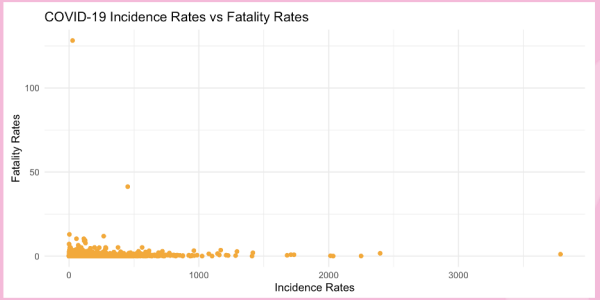
```
summary(data$incidence_rate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    2.581   13.219   42.208   41.011 3785.420
```

```
summary(data$fatality_rate)
```

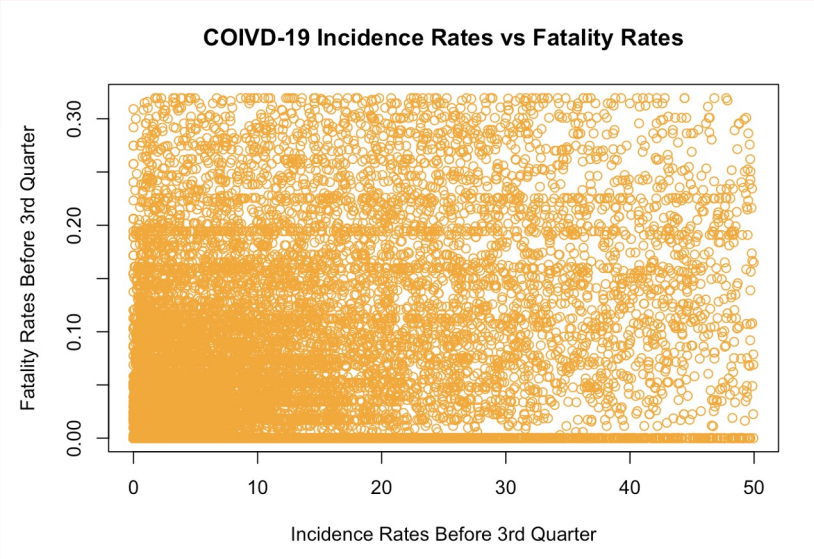
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00000   0.00000   0.08189   0.27030   0.32111 128.21679
```





- There is a huge gap **between 3rd quarter and the maximum** in both incidence rate variable and fatality rate variable
- When checking the number of outliers in both variables, we find that there are **too many outliers in both variables**
- Therefore, we **separate incidence rate and fatality rate into two parts**:
  - Before 3rd Quarter
  - After 3rd Quarter

Correlation

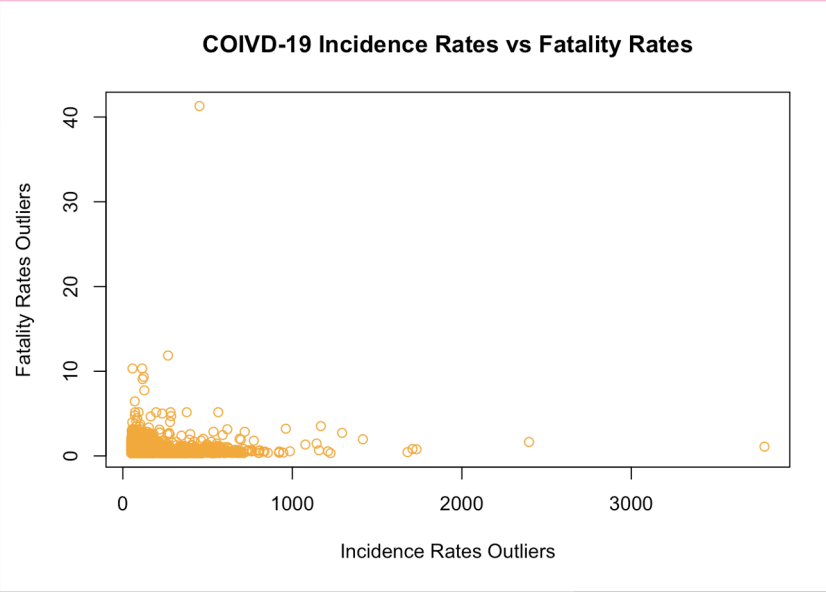



```
cor(tempi$incidence_rate, tempi$fatality_rate, use="complete.obs", method = "pearson")
```

```
## [1] 0.3962233
```

- The dense cluster suggest a wide range of data, but **do not indicate a clear correlation visually**
- Correlation coefficient of the incidence rate and fatality rate before 3rd quarter is approximately **0.396**, indicating a **moderate positive correlation** between incidence rates and fatality rates
- As incidence rates increase, fatality rates also increase

Correlation

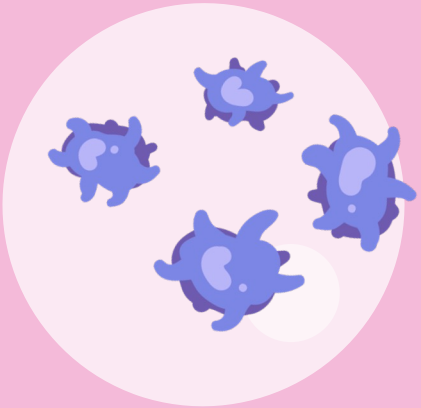



```
cor(tempb$incidence_rate, tempb$fatality_rate, use="complete.obs", method = "pearson")
```

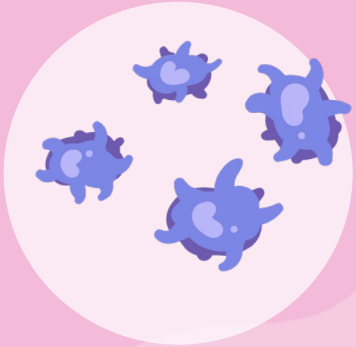
```
## [1] -0.04637824
```

- The data is **more spread out**, suggesting the relationship between incidence rates and fatality rates might differ from population with outlier consideration

- The result of -0.046 indicates a very **weak negative correlation**, demonstrating that when considering outliers, the **relationship between incidence rates and fatality rates is not significant**



Part 4
Regression



Understanding Data Frame

```
summary(model_df)
```

```
##      country        pop                sq_km          gdp_pps
##  Austria : 1   Min.   :  514564   Min.   :   316   Min.   : 51.0
##  Belgium : 1   1st Qu.: 5266795   1st Qu.: 60701   1st Qu.: 71.0
##  Bulgaria: 1   Median : 7926273   Median : 90723   Median : 95.5
##  Cyprus  : 1   Mean   :17305840   Mean   :192118   Mean   :100.2
##  Denmark : 1   3rd Qu.:13474040   3rd Qu.:342955   3rd Qu.:120.8
##  Finland : 1   Max.   :83166711   Max.   :551695   Max.   :190.0
##  (Other) :14
##     pop_dens         total_cases
##  Min.   :  13.94   Min.   :  115285
##  1st Qu.:  57.67   1st Qu.: 1319422
##  Median : 100.50   Median : 2570210
##  Mean   : 179.14   Mean   : 6496297
##  3rd Qu.: 121.55   3rd Qu.: 5430242
##  Max.   :1628.37   Max.   :36952159
##
```

The data comes from 20 countries, and there are 6 variables in total.
Except for the country, the other variables are numeric variables.

Initial Model Fitting

```
library(MASS)
rm<-lm(total_cases~pop+pop_dens+gdp_pps, data=model_df)
summary(rm)
```

```
##
## Call:
## lm(formula = total_cases ~ pop + pop_dens + gdp_pps, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8290586 -1092896   -81818  1657083  8955212
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.875e+06  2.787e+06  -1.390    0.183
## pop          4.285e-01  3.676e-02  11.658 3.13e-09 ***
## pop_dens     6.502e+02  2.432e+03   0.267    0.793
## gdp_pps      2.834e+04  2.559e+04   1.108    0.284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3706000 on 16 degrees of freedom
## Multiple R-squared:  0.8963, Adjusted R-squared:  0.8769
## F-statistic:  46.1 on 3 and 16 DF,  p-value: 4.251e-08
```

- Fit a linear model on the data with total cases as the response variable and population, population density, GDP as predictor variables.
- The adjusted r-squared of the initial model is about 0.88, suggesting that the model has a good fit.

Model Simplification

```
stepAIC(rm)
```

```
## Start:  AIC=608.56
## total_cases ~ pop + pop_dens + gdp_pps
##
##            Df  Sum of Sq        RSS    AIC
## - pop_dens  1 9.8181e+11 2.2077e+14 606.65
## - gdp_pps   1 1.6850e+13 2.3664e+14 608.04
## <none>                   2.1979e+14 608.56
## - pop       1 1.8668e+15 2.0866e+15 651.57
##
## Step:  AIC=606.65
## total_cases ~ pop + gdp_pps
##
##           Df  Sum of Sq        RSS    AIC
## - gdp_pps  1 1.7043e+13 2.3781e+14 606.14
## <none>                  2.2077e+14 606.65
## - pop      1 1.8723e+15 2.0931e+15 649.63
##
## Step:  AIC=606.14
## total_cases ~ pop
##
##        Df  Sum of Sq        RSS    AIC
## <none>                2.3781e+14 606.14
## - pop   1 1.8818e+15 2.1196e+15 647.89
```

```
##
## Call:
## lm(formula = total_cases ~ pop, data = model_df)
##
## Coefficients:
## (Intercept)          pop
##  -9.216e+05    4.286e-01
```

- Using the stepwise regression (stepAIC), we found that a model with only the population variable resulted in the lowest AIC, suggesting that it is the most parsimonious model.

Check with the smaller model

```
# New model without GDP and population density
rm2 <- lm(total_cases~pop, data=model_df)
rm2
```

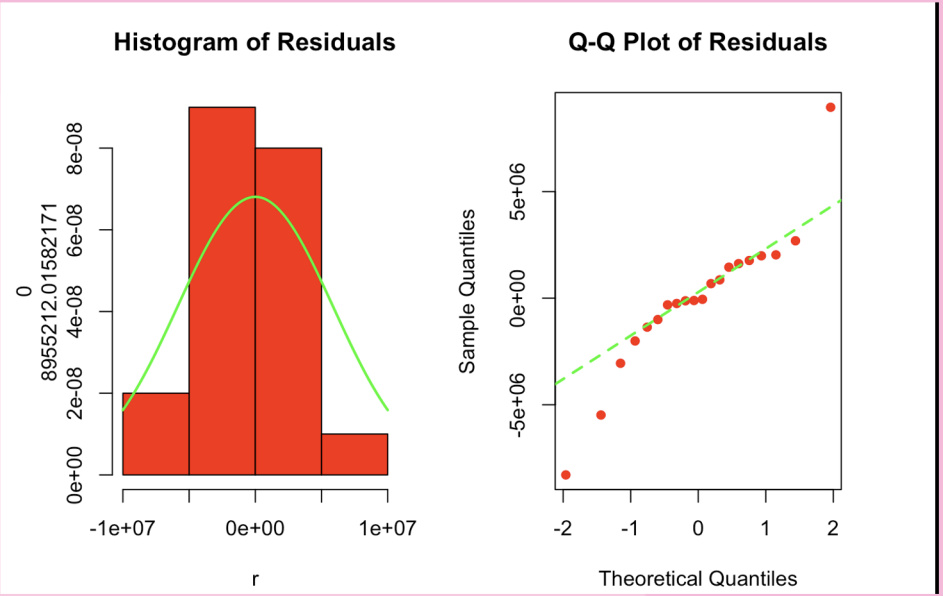
```
##
## Call:
## lm(formula = total_cases ~ pop, data = model_df)
##
## Coefficients:
## (Intercept)          pop
##  -9.216e+05    4.286e-01
```

```
summary(rm2)
```

```
##
## Call:
## lm(formula = total_cases ~ pop, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9159053  -814808   570800  1123412  9017900
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.216e+05  1.023e+06  -0.901     0.38
## pop          4.286e-01  3.592e-02  11.935 5.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3635000 on 18 degrees of freedom
## Multiple R-squared:  0.8878, Adjusted R-squared:  0.8816
## F-statistic: 142.4 on 1 and 18 DF,  p-value: 5.51e-10
```

- The adjusted R-squared of the smaller model is 0.8816, which performs a little bit better than the initial model.

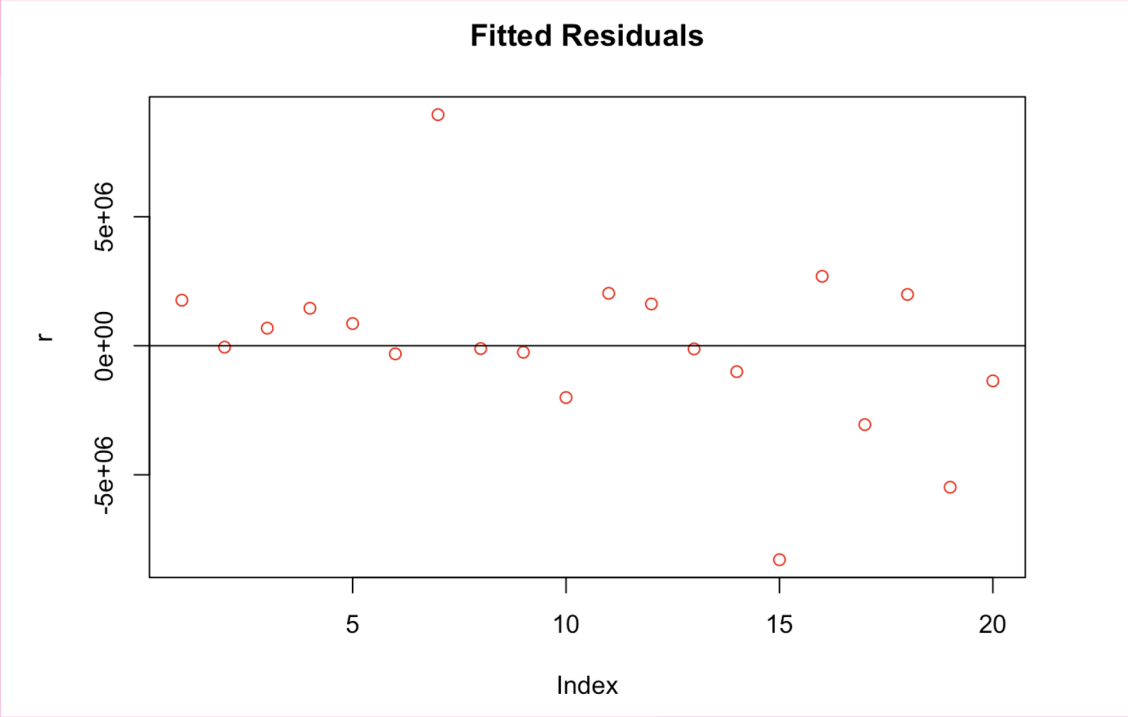
Reviewing Residuals




```
Skewness of Residuals = -0.009679622
Kurtosis of Residuals = 5.081594
```

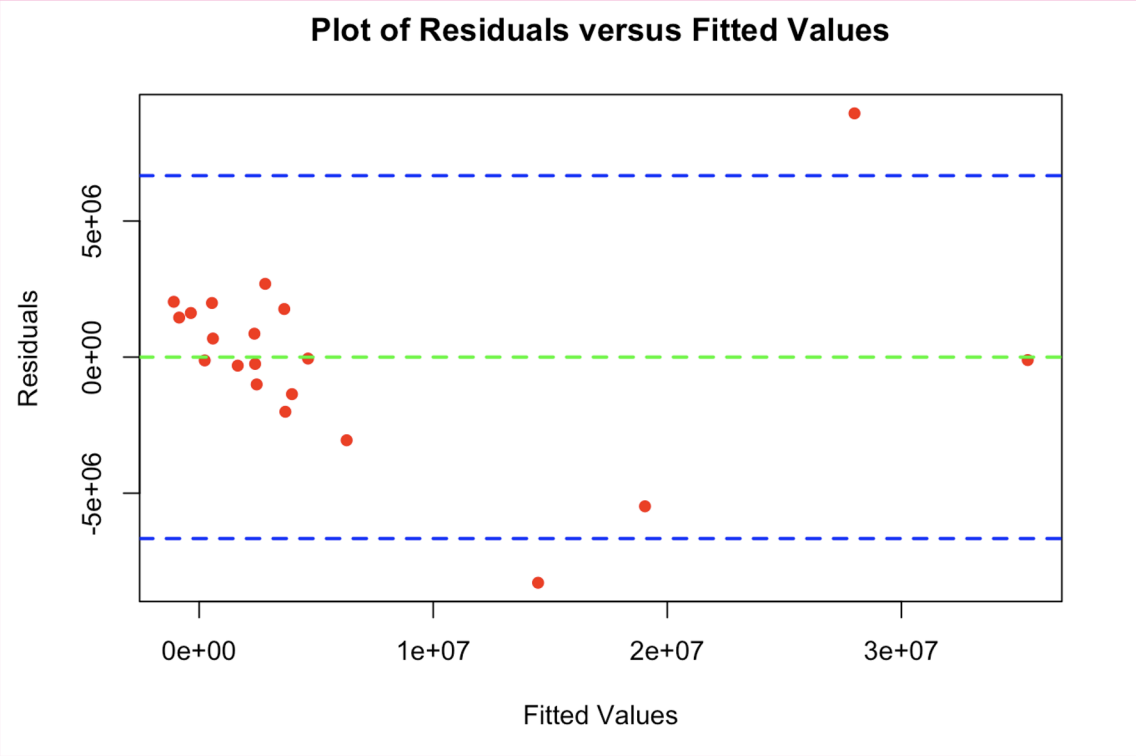
- The histogram shows a bell-shaped and symmetric distribution, indicating it closely approximates a normal distribution. The near-zero skewness and the mostly aligned Q-Q plot support this.
- However, the higher kurtosis value and tail deviations in the Q-Q plot indicate the existing of outliers.

Reviewing Residuals



The Fitted Residuals plot does not show any systematic pattern, implying that the model's predictive accuracy is consistent across the data range.

Reviewing Residuals



- The Plot of Residuals versus Fitted Values suggests that the model generally fits well, with most residuals clustering around the horizontal axis.
- However, a few points with high residuals, especially for higher fitted values, are evident. These points may indicate potential outliers that the model does not fully capture.

Model	Actual (L*)	Predict (L*)	Diff %***	Actual (N**)	Predict (N**)	Diff %***	MSE
Initial Model	301031	3947450	1211.3	8494705	7540820	11.2	7.1e+12
Smaller Model	301031	-653250	317.0	8494705	6539291	23.0	2.4e+12
Poisson Regression	301031	1757467	483.8	8494705	3593438	57.7	1.3e+13
Random Forest	301031	2169189	620.6	8494705	6380737	24.9	4.0e+12

*: Luxembourg
**: Netherlands
***: Diff %= abs(Predicted Value−Actual Value) / Actual Value * 100

- The varied performance of the models emphasize the importance of considering country-specific characteristics when modeling epidemiological data.
- Smaller countries like Luxembourg may have unique factors not adequately captured by general models.
- The significant deviations in predictions, especially for Luxembourg, highlight the need for more comprehensive models that can incorporate a wider range of variables.

Part 5: Conclusions

Discuss

- Data Limitations and External Factors:
  - Data is not only about numbers but is also influenced by a myriad of social, economic, and political factors that are difficult to quantify
  - External factors such as testing rates, public compliance with health measures, and government policies, are not always easily integrated into statistical models
- Future Directions and Improvements:
  - Incorporating additional data, such as public health interventions, and healthcare capacity, could enhance model accuracy
  - Collaboration across disciplines, including epidemiology, data science, and public health, is crucial for developing more robust models

Thank you!