Data Analysis Project
P3: Data Analysis and Visualization
Impact of Covid-19 on air traffic in the U.S.
(Group 05)

Yannick Martin and Sascha Tran

University of Basel
Databases (CS244) course
Autumn Semester 2022

Table of Contents

Data Analysis Project P3: Data Analysis and Visualization Impact of Covid-19 on air traffic in the U.S. (Group 05) 1
Yannick Martin and Sascha Tran

1 Introduction 3
2 Goals 3
3 Datasets 3
3.1 Airline Employment Data 3
3.2 Coronavirus (Covid-19) Case Surveillance Data in the United States 4
3.3 COVID-19 Vaccination Trends in the United States 5
3.4 Carrier On-Time Performance Dataset 6
3.5 Tools 8
4 Integrated Database 8
4.1 Proceeding 8
4.2 Entity Relationship Model 9
4.3 Logical Schema 12
5 Methods 13
5.1 Data Integration 13
5.2 Data Access 14
5.3 Analysis Queries 14
6 Results 14
6.1 Overview Number of New Corona Cases 15
6.2 Overview Number of Performed Flights 16
6.3 Cancelled Flights 17
6.4 New Corona Cases 18
6.5 Correlation between new Corona Cases and Number of Flights . . 19
6.6 Delayed Flights 21
6.7 Airline Employees 22
6.8 Covid Vaccination 23
6.9 Correlation between new Corona Cases and Vaccination 24
7 Analysis 25
7.1 New Corona cases over Time: 25
7.2 Performed Flights overview: 26
7.3 Correlation between the cancelled flights and the number of Corona cases 27
7.4 Non-weather related delays 27
7.5 Influence of Corona vaccination 28
8 Summary 29
9 References 30
10 Attachment 30
10.1 States in the heatmaps 30

1 Introduction

The COVID-19 pandemic has had a profound impact on the global economy and has disrupted nearly every sector, including the air travel industry. In the United States, the pandemic has led to a significant reduction in the number of flights being operated, as well as changes to the way flights are operated and the services they offer.

This project aims to examine the relationship between COVID-19 and flight operations in the United States by analyzing data on cancelled flights and COVID-19 cases in the different US states, as well as data on non-weather related delays and vaccination rates. Statistical techniques and visualization tools will be used to assess the correlation between these variables, and the implications of the findings will be considered for the future of the air travel industry.

The scope of this project is limited to the impact of COVID-19 on flight operations in the United States of America and does not cover the broader impacts of the pandemic on the global air travel industry.

2 Goals

As already mentioned in the introduction, the analysis of this project is based on the following questions:

- Is there a correlation between the cancelled flights and the number of Corona cases in the different US states?
- Have there been more non-weather related delays since the start of the pandemic, for example due to vaccination certificate checks or staff shortages at airports?
- Is there a relationship between the number of flights and the number of COVID-19 vaccinations in the different states in the United States?

3 Datasets

For the database project the following datasets and their corresponding Entity Relationship (ER) diagrams were required.

3.1 Airline Employment Data

Information:

- **Source:** https://www.transtats.bts.gov/Employment/

- **Size:** 34 KB

- **Format:** .xls

- **Description:** This dataset consists of the total number of employees of all airlines in the U.S. in every month since 1990. Besides the total number, the number of parttime and fulltime employees is also available.

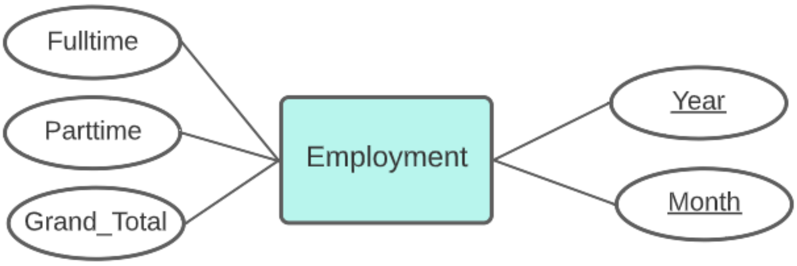


Fig. 1: ER diagram of the Airline Employment dataset

3.2 Coronavirus (Covid-19) Case Surveillance Data in the United States

- **Source:** https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

- **Size:** 4'851 KB

- **Format:** .csv

- **Description:** Several attributes regarding the number of Corona cases and Corona deaths per state and date are present in this dataset. The first data start in the year 2020.

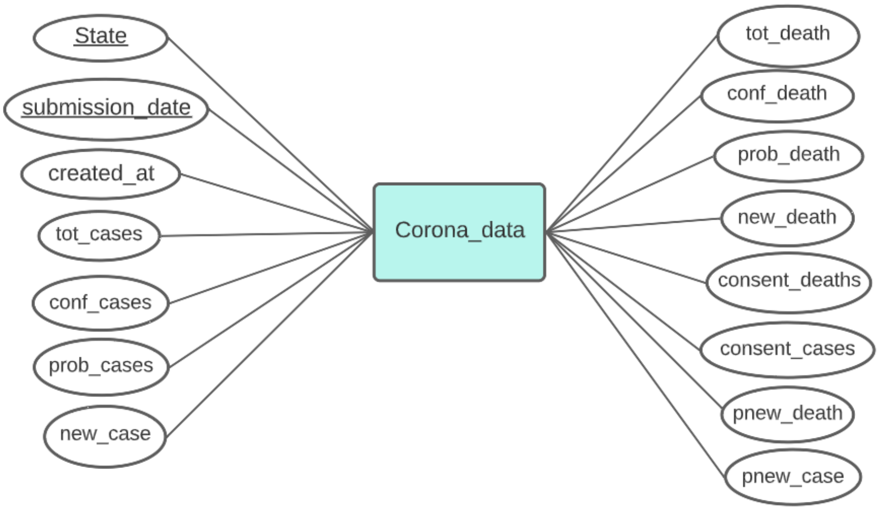


Fig. 2: ER diagram of the COVID-19 cases dataset

3.3 COVID-19 Vaccination Trends in the United States

- **Source:** https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Age-and-Sex-Trends-in-the-Uni/5i5k-6cmh

- **Size:** 130'543 KB

- **Format:** .csv

- **Description:** Among many attributes, this dataset consists of the number of vaccinated people per state and date starting from 2020.

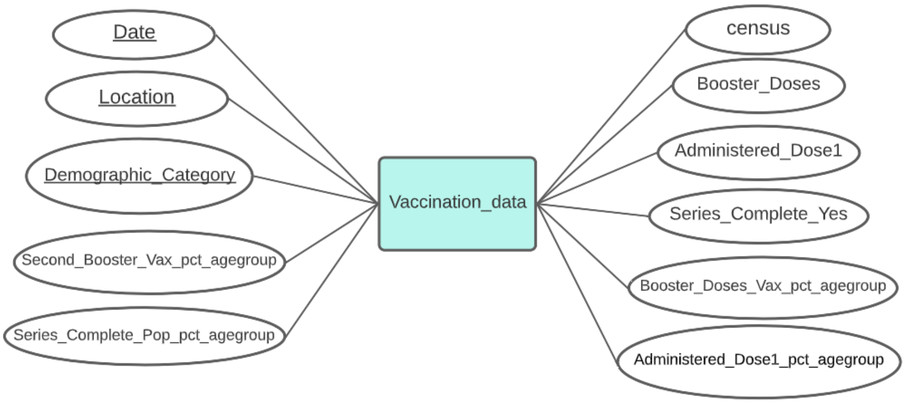

Date
Location
Demographic_Category
Second_Booster_Vax_pct_agegroup
Series_Complete_Pop_pct_agegroup
Vaccination_data
census
Booster_Doses
Administered_Dose1
Series_Complete_Yes
Booster_Doses_Vax_pct_agegroup
Administered_Dose1_pct_agegroup

Fig. 3: ER diagram of the COVID-19 Vaccination dataset

3.4 Carrier On-Time Performance Dataset

- **Source:** https://transtats.bts.gov/DL_SelectFields.aspxgnoyr_VQ=FGJQO_fu146_anzr=b0-gvzr

- **Size:** total approx. 13 GB

- **Format:** .csv

- **Spatial Resolution**: for each month one file

- **Description:** This dataset consists of a time range of July 2017 to July 2022. Its size is only approximate since it was necessary to download one dataset (200 MB) for each month and year resulting in 60 downloads. The dataset consists of several attributes regarding the flight performance. For our analysis, we are especially interested in the following attributes:
  - OriginCityName
  - Cancelled
  - Flightdate
  - all attributes of Cause of Delay

In the following ER diagram, not all attributes are present. To be precise, all diverted related attributes are represented as "all diverted attributes". This is because the same attributes would have been listed again for a case where a flight was diverted only with a prefix div. Among other things, the attributes of diverted flights are not of the project's interest.

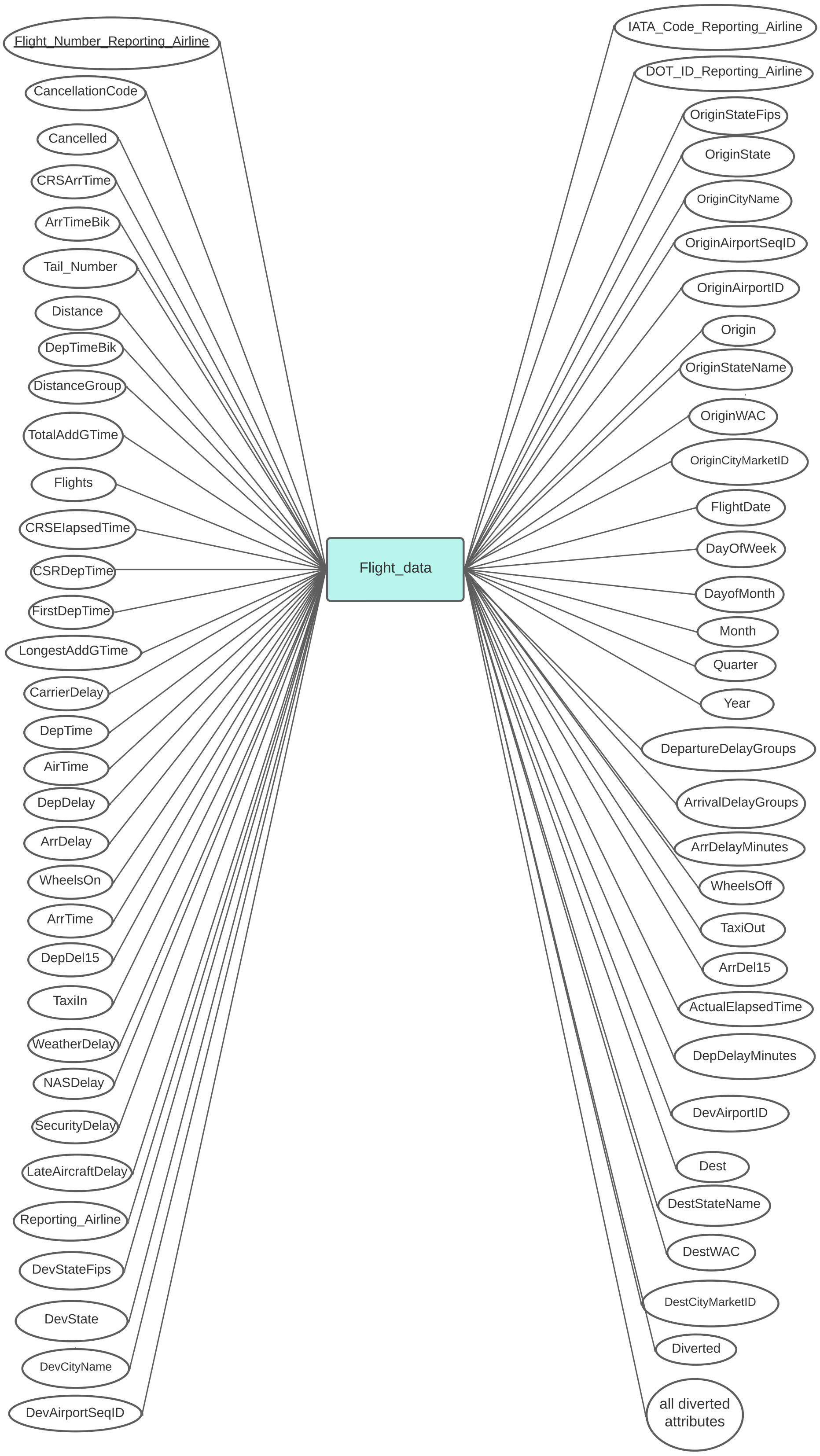


Fig. 4: ER diagram of the Flight dataset

3.5 Tools

During our project we used the following tools:

- Python
- Overleaf
- PostgreSQL
- pgAdmin4
- LucidChart

We used Python to preprocess our data sets before loading them into the database. In addition, we also used Python to create our plots. For this we mainly used the Python modules Pandas, Psycopg2, Seaborn, and Bokeh. We used Overleaf to define our PDF with Latex. PostgreSQL is our DBMS and we used pgAdmin4 as graphical interface for our database. Last but not least, we used LucidChart to create our entity relationship models.

4 Integrated Database

This section describes the changes from the individual datasets to the integrated schema. It especially outlines the major changes. Additionally, the ER diagram of the integrated schema as well as the relational schema can be found in this section.

4.1 Proceeding

Linking the individual datasets

There are two attributes which all datasets have in common, namely the state and date attributes. Therefore, our goal was to connect all our datasets using State and Date. They might be named differently in each dataset. Therefore, an unified format for both attributes is required for the integration. The state attribute in each dataset already has the same format. As for the date-format, we have looked into each dataset about their format:

- Flight_Data: 1999-01-31
- Corona_data: 31/01/1999
- Vaccination_data: 01/31/1999
- Employment: 1999 , 1 (as two attributes Year, Month)

We have decided on the date-format "yyyy-mm-dd". Since the Employment dataset only consists of the attributes Year and Month, the first day of every month was set.

Changing the datasets

There are many changes that occurred for the integrated database and therefore the interested reader is referred to the ER diagram of the integrated schema (Figure 5).

The main changes and remarks to Figure 5 are the following:

- The integrated ER diagram doesn't consist of all attributes, that were represented in the first chapter. For example, in the original datasets (Flight dataset) there were the attributes DestAirportID and OriginAirportID which are now represented as AirportID in the integrated ER model. Further differences can be seen in the attachment.

- There are many attributes in the "Flight" entity. In order to keep a good overview of the diagram, the attributes are attached to one line that leads to the correct entity.

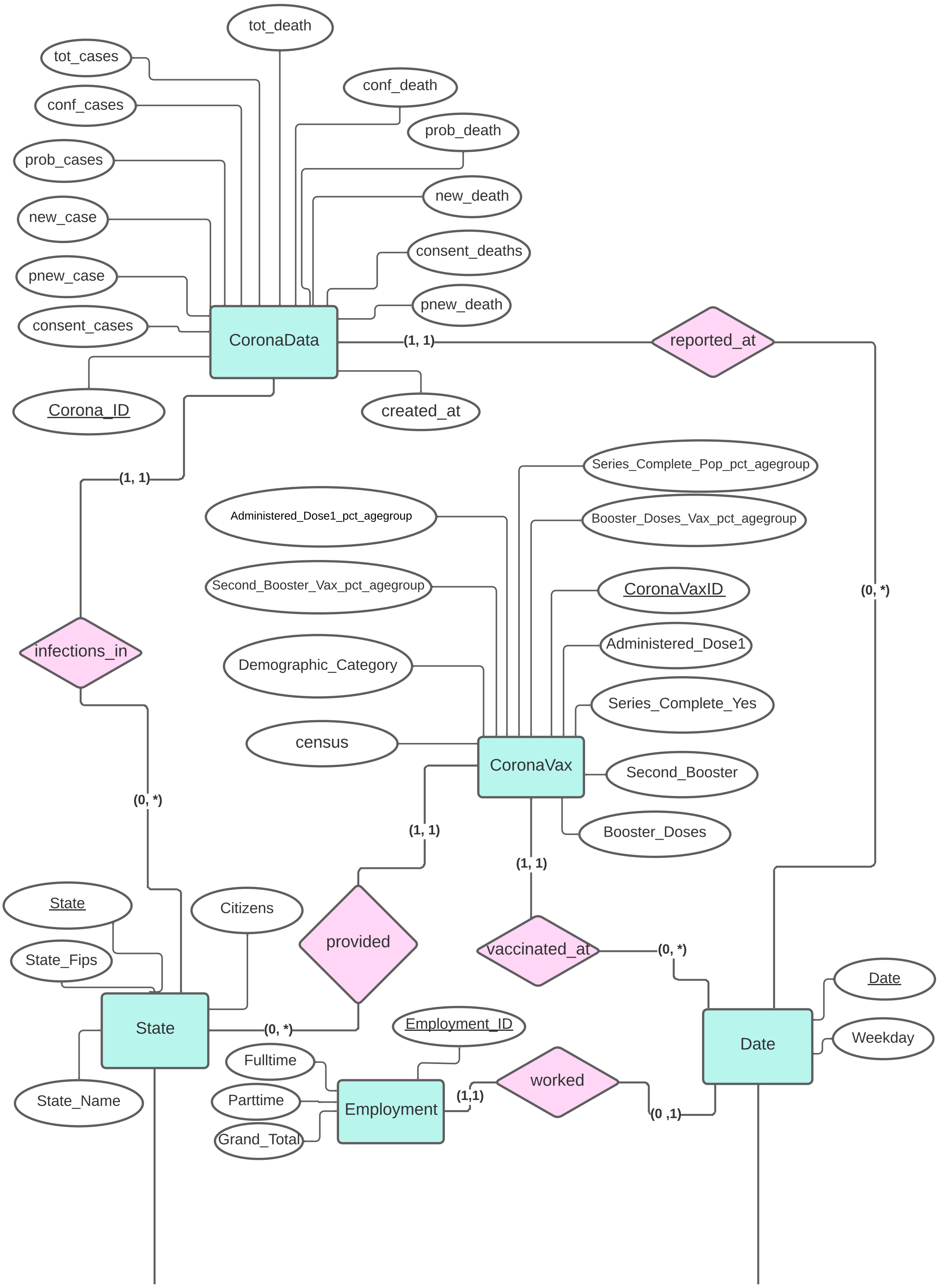
- Artificial IDs were added to many some entities (which are also marked in the table). Namely: Corona_ID, Employment_ID, Delayed, Vaccination_ID

- There is an entity "Performed_Flight" which inherits all the attributes of the entity "Flight". This entity was created because whenever a flight is cancelled, no flight can be performed, delayed or diverted. Therefore, many attributes would be redundant.
  The same goes for the subentities of "Performed_Flight". Not all performed flights must necessarily be delayed or diverted.

- There is no entity "Cancelled_Flight" because only one attribute (CancellationCode) would belong to this entity.
  Therefore, this attribute simply belongs to the entity "Flight".

- The attribute "created_at" of the entity "Corona_data" is actually a date and would fit into the entity "Date". However, there exists another date-attribute from this dataset, namely "submission_date", which will be of much more use in our analysis.
  For completeness reasons, the attribute "created_at" stays in the entity "CoronaData".

4.2 Entity Relationship Model


tot_death
tot_cases
conf_death
conf_cases
prob_death
prob_cases
new_death
new_case
consent_deaths
pnew_case
pnew_death
consent_cases
CoronaData
(1, 1)
reported_at
Corona_ID
created_at
(1, 1)
Series_Complete_Pop_pct_agegroup
Administered_Dose1_pct_agegroup
Booster_Doses_Vax_pct_agegroup
Second_Booster_Vax_pct_agegroup
CoronaVaxID
(0, *)
infections_in
Administered_Dose1
Demographic_Category
Series_Complete_Yes
census
CoronaVax
Second_Booster
(0, *)
(1, 1)
Booster_Doses
(1, 1)
State
Citizens
provided
State_Fips
vaccinated_at
(0, *)
Date
State
Employment_ID
(0, *)
Weekday
Date
Fulltime
worked
State_Name
Parttime
(1,1)
Employment
(0 ,1)
Grand_Total

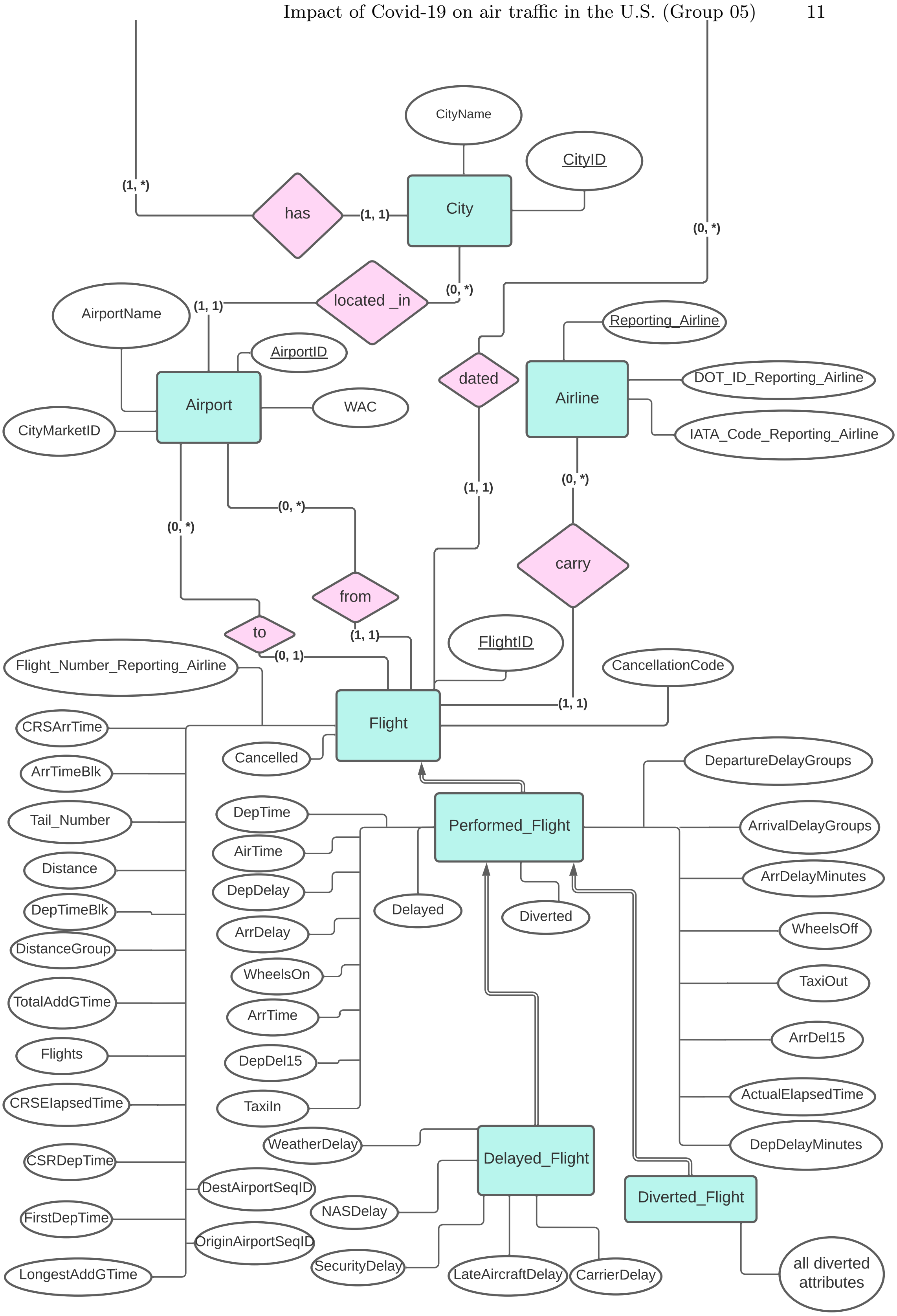

CityName
CityID
(1, *)
has
(1, 1)
City
(0, *)
located _in
(0, *)
(1, 1)
AirportName
AirportID
Reporting_Airline
DOT_ID_Reporting_Airline
dated
Airline
Airport
WAC
CityMarketID
IATA_Code_Reporting_Airline
(1, 1)
(0, *)
(0, *)
(0, *)
carry
from
to
(1, 1)
FlightID
(0, 1)
Flight_Number_Reporting_Airline
CancellationCode
(1, 1)
Flight
CRSArrTime
ArrTimeBlk
Cancelled
DepartureDelayGroups
Tail_Number
DepTime
Performed_Flight
ArrivalDelayGroups
AirTime
Distance
DepDelay
ArrDelayMinutes
DepTimeBlk
Delayed
Diverted
ArrDelay
DistanceGroup
WheelsOff
WheelsOn
TaxiOut
TotalAddGTime
ArrTime
ArrDel15
Flights
DepDel15
ActualElapsedTime
CRSElapsedTime
TaxiIn
DepDelayMinutes
WeatherDelay
CSRDepTime
Delayed_Flight
DestAirportSeqID
Diverted_Flight
NASDelay
FirstDepTime
OriginAirportSeqID
all diverted attributes
LongestAddGTime
SecurityDelay
LateAircraftDelay
CarrierDelay

Fig. 5: ER diagram of the integrated schema

4.3 Logical Schema

CoronaData(Corona_ID, Date, State, tot_cases, conf_cases, prob_cases, new_case, pnew_case, tot_death, conf_death, prob_death, new_death, pnew_death, created_at, consent_cases, consent_deaths)

CoronaVax(CoronaVaxID, Date, Location, Demographic_Category, census, Administered_Dose1, Series_Complete_Yes, Booster_Doses, Second_Booster, Administered_Dose1_pct_agegroup, Series_Complete_Yes_pct_agegroup, Booster_Doses_pct_agegroup, Second_Boster_pct_agegroup)

State(State, State_Name, State_Fips, Citizens)

Employment(Employment_ID, Fulltime, Parttime, Grand_Total, Date)

Date(Date, Weekday)

City(CityID, CityName, State)

Airline(Reporting_Airline, DOT_ID_Reporting_Airline, IATA_Code_Reporting_Airline)

Airport(AirportID, AirportName, WAC, CityMarketID, CityID)

Flight(Flight_ID,Flight_Number_Reporting_Airline, CRSArrTime, ArrTimeBlk, Tail_Number, Distance, DepTimeBlk, DistanceGroup, TotalAddGTime, Flights, CRSElapsedTime, CSRDepTime, FirstDepTime, LongestAddGTime, Cancelled, CancellationCode, OriginAirportSeqID, DestAirportSeqID, FlightDate, Reporting_Airline, DestAirportID, OrigAirportID)

Performed_Flight(Flight_ID, DepTime, AirTime, DepDelay, ArrDelay, WheelsOn, ArrTime, DepDel15, TaxiIn, Delayed, Diverted, DepartureDelayGroups, ArrivalDelayGroups, ArrDelayMinutes, WheelsOff, TaxiOut, ArrDel15, ActualElapsedTime, DepDelayMinutes)

Delayed_Flight(Flight_ID,CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay)

Diverted_Flight(Flight_ID, all diverted attributes)

5 Methods

This section describes how the data was integrated.

5.1 Data Integration

- Python

  - **flightDataMerge.ipynb**
    Reads the Flight datasets (each flight dataset from 07-2017 up to 07-2022) and writes them into one big file. Note that the paths are in the format '1_2022.csv'. Hence, save the files accordingly or change the path. This file is a Jupyter-Notebook because it was more convenient to check whether the merge worked or not without rereading the datasets every time.

  - **adjustDataset.py**
    This script does the following:
    - Converts the Employment dataset (format: .xlsx) to .csv.
    - Creates a column 'Date' in the Employment dataset out of the column 'Month' and 'Year'. Since there is no day given, the first day of each month was put (e.g. 2022-02-01).
    - Adjusts the date format of the Corona-Cases and Corona-Vaccination dataset.
    - Add a column consisting of an artificial ID for the datasets Corona-Cases, Corona-Vaccination and Employment.
- SQL

  - **1_creates_Table_for_Datasheets.sql**
    Creates empty tables for the four "original" datasets. (They slightly might be already adjusted from the py-scripts.)

  - **2_load_Data_into_Tables.sql**
    Loads the data of the four "original" datasets into the tables.

  - **3_creates_Tables_with_Data.sql**
    Here the integrated schema is implemented. Every entity that was planned, if not revised, is created in this file (without foreign key constraints).

- **4_create_foreignkey_constraint.sql**
  Foreign key constraints are set in this file.

- **5_add_population_for_state.sql**
  Adds the current number of people in each state.

A guide to reproduce our database can be seen in the chapter Guide of our Readme.md file on Gitlab[1].

5.2 Data Access

You can download our *SQL* Dump from this link: `https://drive.switch.ch/index.php/s/1QmWWZlMbc71FLt`

5.3 Analysis Queries

For our analysis, we have used pretty standard queries that were introduced in the lecture. We have very often used aggregate functions to sum up values such as all Corona cases from the different states in every month. We also often used the count aggregate function to count how many flights there were in a certain period.
We have all our queries with comments in the file queries.sql in our repository to see what each query looks like exactly.

To visualize our results from the queries, we use several methods. We use heatmaps, lineplots and scatterplots. We created the heatmaps with seaborn, the scatterplots with matplotlib and all other plots with bokeh. Our code to generate the plots is also uploaded on our Gitlab in the Analysis folder.

6 Results

Disclaimer: All plots are visible on the file "AllPlots.html" which can be found in the repository in "Analysis/html" The line plots are interactive (follow the description on the website).

The United States of America consists of many states. Therefore, it was decided to focus mainly on two chosen states, California and New York, for answering the questions in Section 2.

Besides the plots for the analysis goals, there is also an overview regarding the number of Corona cases and performed flights given.

Remark: The number of new Corona cases and the number of performed/cancelled flights are the sum of each month.

[1] `https://git.scicore.unibas.ch/databases-hs22/group-5`

6.1 Overview Number of New Corona Cases

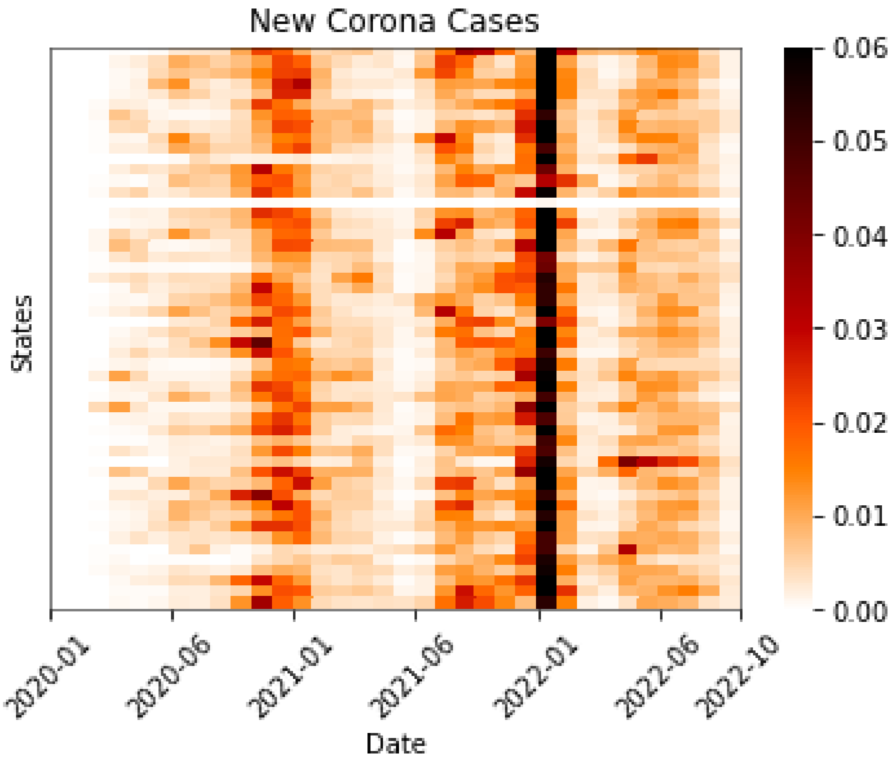


Fig. 6: Heatmap of the number of new Corona cases in each state.

Figure 6 is a heatmap of the number of new Corona cases per month in each state over the time of January 2020 to October 2022.

The y-axis of this map consists of all the states. Since there are too many states to fit into this figure without overlap, the y-axis label can be seen in the attachment (Table 3). The order of labels in the attachment is also the order of the figure from the top to the bottom.

The number of people in each state is different. Therefore, the number of new Corona cases was divided by the total number of people in each state. The numbers can be seen on the colour scale next to the plot.

The darker the colour, the more new Corona cases are there relative to the size of the population in each state.

Observation: In general, we can observe that the darker areas are in winter times, especially in January.
Also, we can see, that the relative number of new Corona cases is overall higher in the year 2021 than in 2020 and even higher in 2022. The differences are especially visible in January 2021 to January 2022 as well as summer 2021 to summer 2022.
These observations can be made over all states.

6.2 Overview Number of Performed Flights

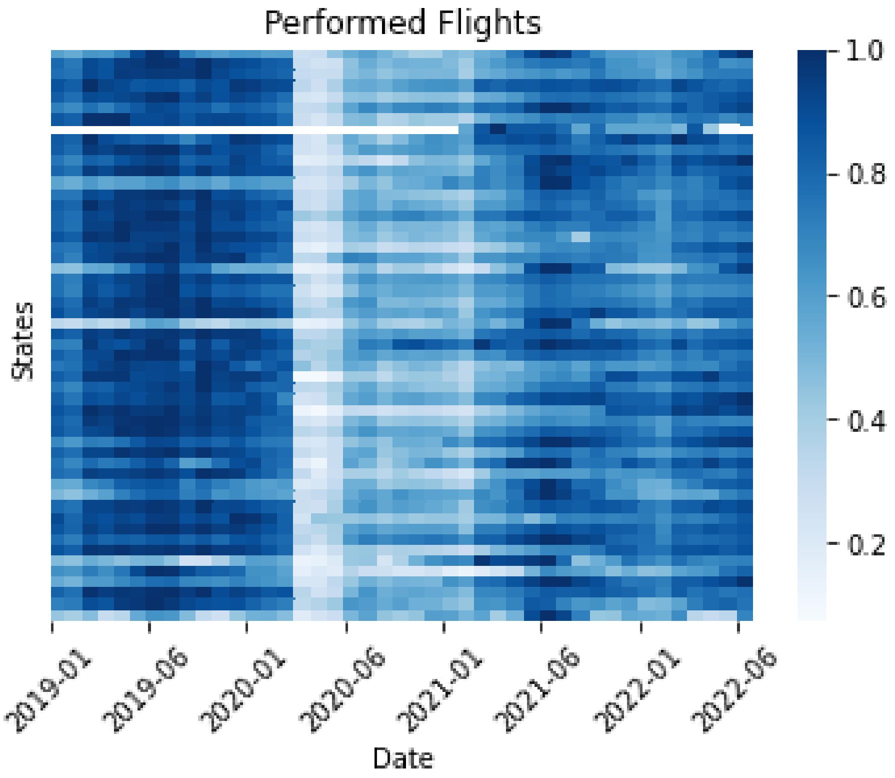


Fig. 7: Heatmap of the number of performed flights in each state.

Figure 7 is a heatmap of the number of performed flights per month in times between January 2019 and July 2022.

For the same reasons as in Section 6.1 the y-axis' labels are available in the attachment.

Also, because there is a different number of capacities regarding the airports in each state, the number of performed flights has to be scaled. In this figure, the number of performed flights is divided by the biggest number of performed flights in each state.

The darker the colour, the more flights were performed in each state.

Observation: There is an overall trend of all states visible.
In general, we can see that there was a collapse in the relative number of performed flights from March 2020 to June 2020. In all states, we can see a "recovery" process from that one collapse.

6.3 Cancelled Flights

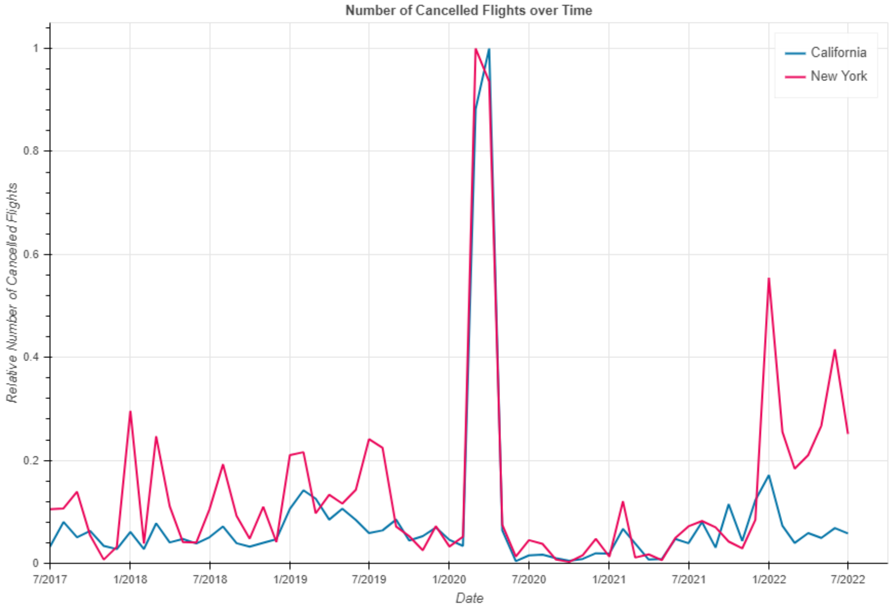


Fig. 8: Relative number of cancelled flights over time for California and New York states.

This figure represents the number of cancelled flights over the time span of July 2017 up to July 2022.

There is a curve for the state of California and one for the state of New York.

The number of cancelled flights is scaled equally as in Figure 7 also for the same reason.

Observation: We can see a peak for each curve. The peaks are around March and April in the year 2020.

6.4 New Corona Cases

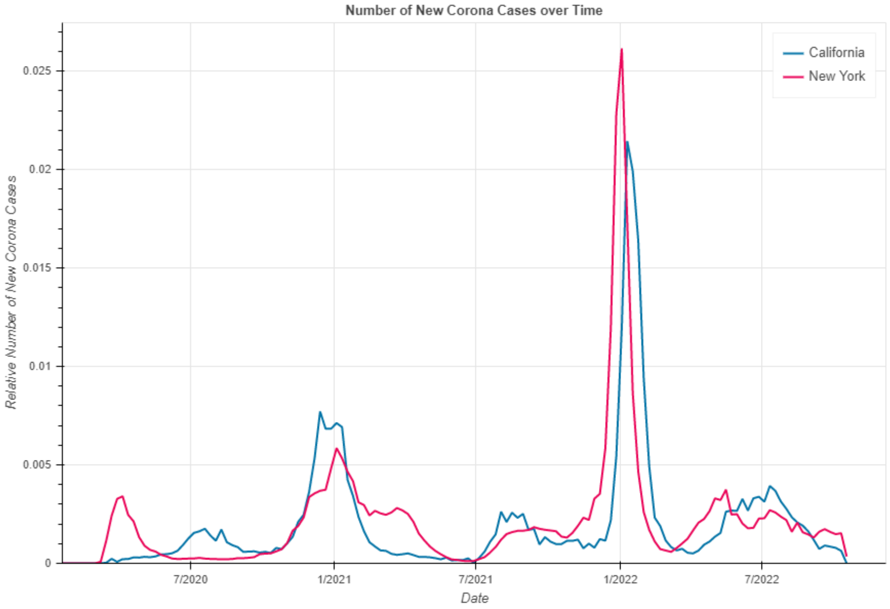


Fig. 9: Relative number of new Corona cases over time for the states California and New York.

This figure represents the number of new Corona cases per month over a time span of January 2020 up to October 2022.

There is a curve for the state of California and one for the state of New York.

The number of new corona cases is scaled equally as in Figure 6 also for the same reason.

Observation: Similarly to Figure 6 there are two main peaks visible for each curve. The peak in 1/2022 is the highest for both curves. Interestingly, there is

also a third "peak" of the curve of New York at the beginning.
In 1/2021, the peak of California is higher. On the other hand, the peak of New York is higher in 1/2022.

6.5 Correlation between new Corona Cases and Number of Flights

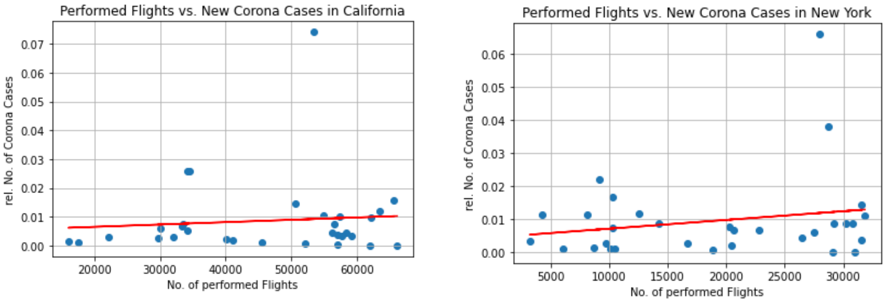


Fig. 10: Relative number of new Corona cases in connection with the relative number of performed flights for the states California (left) and New York (right).

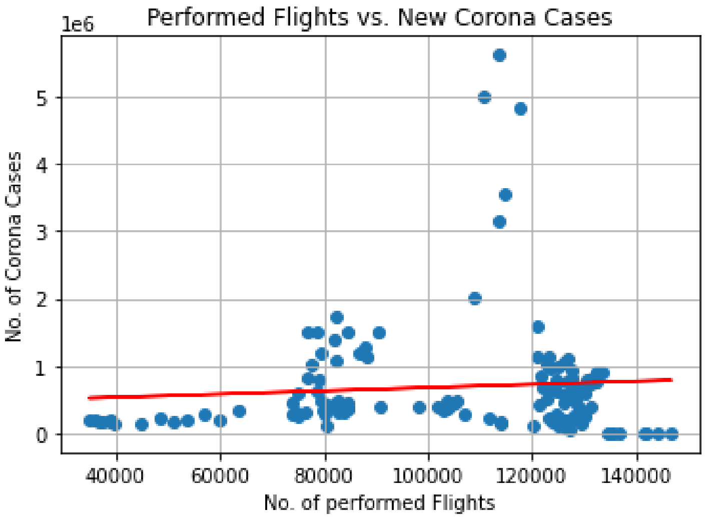


Fig. 11: Number of new Corona cases in connection with the number of flights for the USA.

Both figures (fig. 10 and fig. 11) are scatterplots representing the number of performed flights with the relative number of new Corona cases.

Every datapoint represents the number of new Corona cases and the number of performed flights in a month spanning from January 2020 to July 2022.

The red line is the least squared polynomial fit of the first degree.

Figure 10 visualizes the plots of the states of California (left) and New York (right). The number of new Corona cases in this figure is constructed the same as in Figure 6.

As for Figure 10, the axes consist of the absolute values and the plot summarizes the relation of the whole country.

The following table[2] displays the Pearson correlation coefficient and p-value of each plot.

state	**correlation coefficient**	**p-value**	**figure**
California	0.087	0.64	fig. 10 (left)
New York	0.20	0.30	fig. 10 (right)
United States of America	0.079	0.37	fig. 11

Table 1: List of the Pearson correlation coefficient and the p-value of each plot in fig. 10 and fig. 11. The "Figure" column links to the corresponding figure.

Observation: There is a slight trend observable in both figures. However, many datapoints do not align with the trend line at all.

[2] `https://www.tablesgenerator.com/latex_tables`, (January 2023)

6.6 Delayed Flights

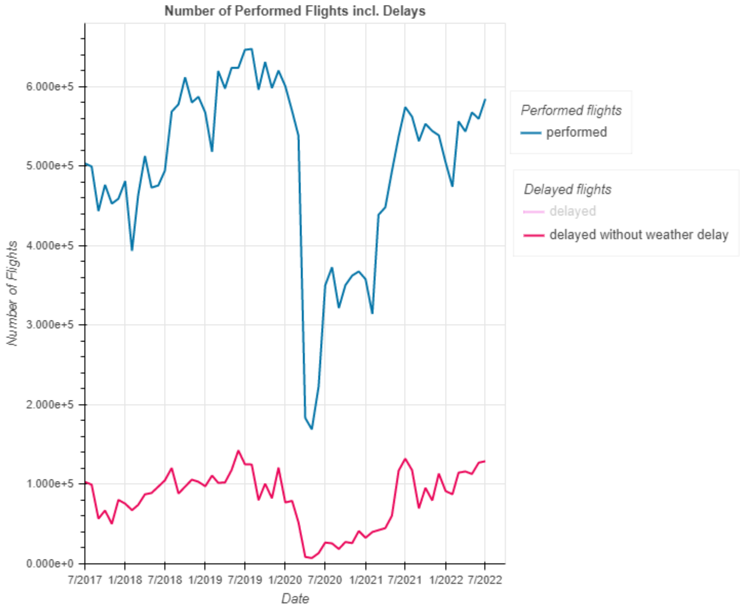


Fig. 12: Number of performed flights and number of delayed flights without weather delay over time in the USA.

This figure represents the number of performed flights and the number of delayed flights without weather delay per month over a time span of July 2017 up to July 2022.

Observation: We can see that the curve of performed flights and the curve of delayed flights (without weather delay) are more or less parallel. The only difference is that there was a much bigger decrease around March 2020 for of the performed flights' curve than the delayed curve.

6.7 Airline Employees

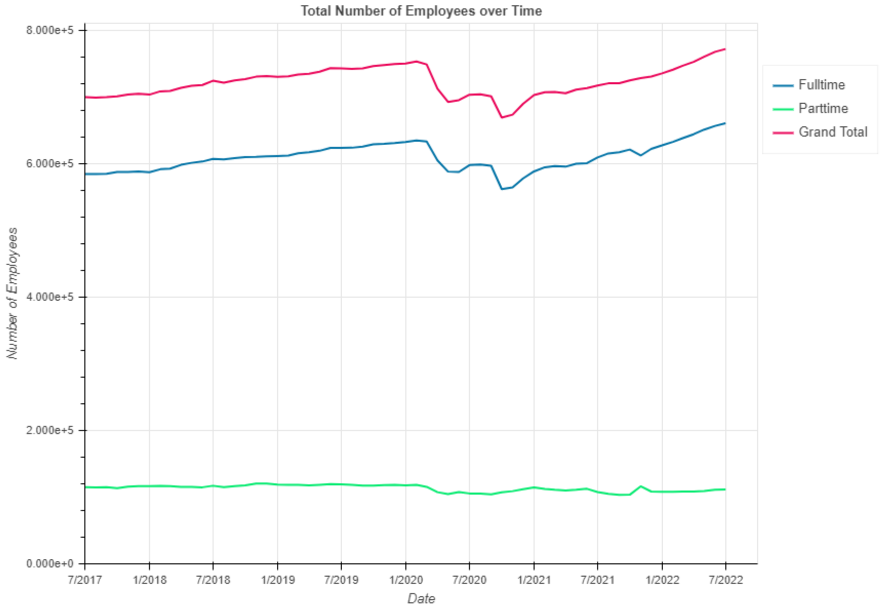


Fig. 13: Number of full-time, part-time and total numbers of employees over time in the USA.

This figure represents the number of airline employees per month over a time span of July 2017 up to July 2022.

There is a curve for full-time, part-time and the total numbers of employees.

Observation: The curves Fulltime and Grand Total are quite parallel. Interestingly, Parttime curve has fewer (and especially less strong) fluctuations than the other curves.

6.8 Covid Vaccination

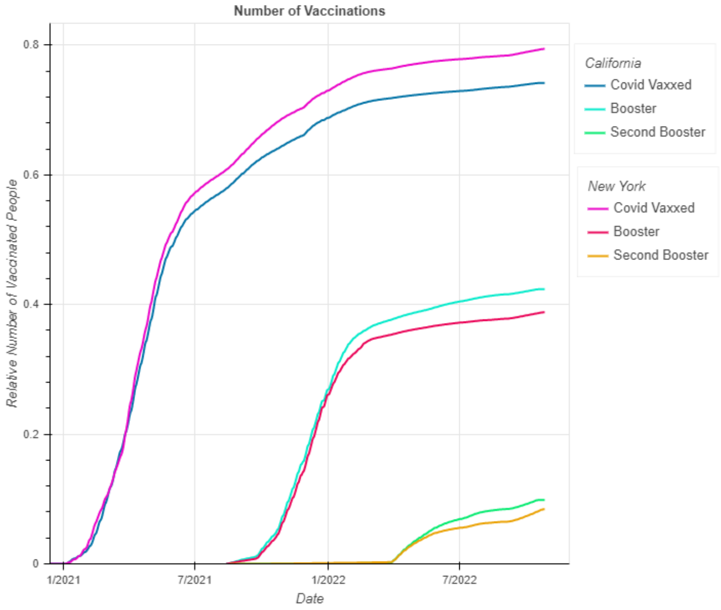


Fig. 14: Number of the Covid-19 Vaccination over time for the states California and New York.

This figure represents the number of vaccinated people per month over a time span of January 2021 up to October 2022.

There is a curve for the first vaccination, booster and second booster each for California and New York.

Observation: In general, fewer people got a first or even second booster than they got themselves fully vaccinated (dose 1 and dose 2). In comparison to California, fewer people got a first or second booster in New York.

6.9 Correlation between new Corona Cases and Vaccination

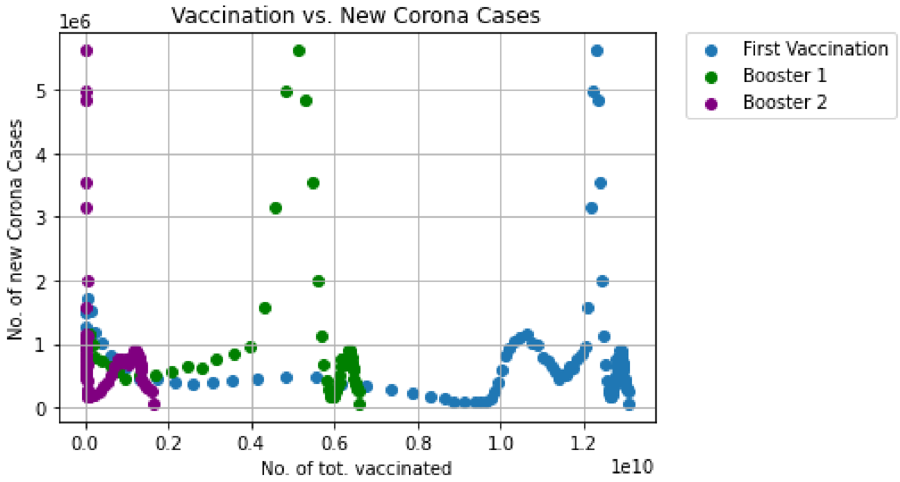


Fig. 15: Number of new Corona cases in connection with the number of each kind of vaccination.

Figure 15 a scatterplot representing the absolute number of new Corona cases with the number of each kind of vaccination, First Vaccination (dose 1 and dose 2), Booster 1 and Booster 2.

Every datapoint represents the total number of new Corona cases and each vaccination type of all states in a month spanning from January 2021 to October 2022.

The following table displays the Pearson correlation coefficient and p-value of each plot.

State	**Correlation Coefficient**	**p-Value**
First Vaccination	0.011	0.91
Booster 1	-0.040	0.76
Booster 2	-0.29	0.020

Table 2: List of the Pearson correlation coefficient and the p-value for each kind of vaccination [2].

Observation: There is no trend line observable in all three kinds of vaccination. All correlation coefficients are very small.

7 Analysis

In this section, we want to analyze our results based on our main questions (section 2). Furthermore, we try to explain the observations made in section 6.

Remark: We can talk about strong correlation when the correlation coefficient is between 0.5 and 1 respectively -0.5 to -1. If the p-value is 0.05 or less, the correlation between two instances most likely occurred by chance.

Disclaimer: A statistical correlation simply means that two variables are correlated.
However, correlation does not imply causation which means there is no guarantee that one factor causes the other.
It is important to keep this in mind when interpreting statistical results and applying them to the real world. It is possible that other factors influence the association, that the association only exists under certain conditions or even that this correlation occurred by chance.
Therefore, it is important to interpret results with caution and always consider the limited implications of statistics, especially since we have not taken into account several factors such as the actions of the U.S. government, availability of Corona tests, incubation period and much more.
If we had tried to consider all these aspects, we would have had to consult even more data and collaborate with virologists, which would have stretched the scope of this work.
This is why the following analysis is also mainly based on our experience and knowledge throughout the pandemic in Europe rather than on further sources.

7.1 New Corona cases over Time:

Figure 6 and Figure 9 are being analyzed in this section.
We can summarize our observations as follows:

- There are two main peaks visible in both figures in January 2021 and January 2022:
  Based on our knowledge and experience, the disease seems to be more contagious at low temperatures which are mostly in winter.

- The number of new Corona cases is higher in 2022 than in 2021 (especially visible in Figure 6):
  We have observed that, based on our experience, the Corona variant was less contagious in 2021 in comparison to 2022. Another plausible reason is that people became tired of the pandemic, and became more careless such as ignoring the distance between people and therefore the number of new Corona cases is higher in 2022.
  Additionally, there were fewer restrictions on public life from the U.S. government in 2022. As a result, fewer people stayed at home which might have

led to more contagions[3].

- The number of new Corona cases is higher in 2021 than in 2020:
  In 2020, the disease was quite new to the country. Therefore, we assume that the Corona disease was not recognized as such in the beginning and also there might not be many, if at all, testing stations.

- There is already a peak (though small) in March 2020 in New York but in California, a small peak first appears in July 2020. (see fig. 9):
  This means that the Coronavirus first spread within New York before spreading within California. You can see that the first peak in New York is higher than in California. A reason could be that California was maybe able to learn from other states and implement measures before the virus spread.

- The peak in January 2021 is higher in California than in New York but the other way around in January 2022. (fig. 9):
  The Corona fall curves look pretty much the same for the states of New York and California. Although New York is on the east coast and California is on the west coast. There are small differences in the peaks as described earlier. The reason for this could be different measures taken by local governments, how stringent the measures were or even the timing of the implementation of the measures. We would have expected that there would be larger differences between the two states since California and New York are 3918 kilometres apart.[4] However, the graph shows us that it is not the case.

7.2 Performed Flights overview:

We analyze Figure 7 in this section.

We can see that the number of flights plummeted in March 2020 in every state of the USA and remained so low up until June 2020. This is due to the curfew[5] imposed by the government in America. Until spring 2021, the number of flights remained low. From then on, the number of flights slowly increased and consequently, so more flights were performed. In February 2022, we again see a small drop in all states.
If we compare this graph with the Corona heat map (fig. 6) we can see that logically the big drop in flights is when the lockdown was introduced. But more interesting is that for example in June 2021 we can see that we had few new Corona cases, but again an increased number of flights. You can also see very clearly that the number of flights slightly decreased again at the highest peak of the number of Corona cases in January 2022.

[3] https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_States, (January 2023)

[4] https://www.luftlinie.org/Kalifornien/New-York, (January 2023)

[5] https://en.wikipedia.org/wiki/COVID-19_lockdowns, (January 2023)

7.3 Correlation between the cancelled flights and the number of Corona cases

In this section, we will use the figures 8, 9, 10 and 11 for analysis.

In section 2, we wanted to know whether or not there is a correlation between the number of Corona cases and the number of cancelled flights.

In Figure 8, we have observed that there is a peak for both curves in March 2020 and April 2020.
This peak most likely occurred when the lockdown was called out in both states. (Starting date: California, 19.03.2020 and New York, 22.03.2020 [5]. Therefore, we can not see a real correlation between the number of cancelled flights to the number of new Corona cases. There was a peak at that time but it is smaller than at other times. (see Section 7.1)

Since the number of cancelled flights only shows where flights had to be cancelled at short notice and not how many flights really did not fly compared to a normal month, we have decided to compare the number of performed flights and not the number of cancelled flights (fig. 10 and fig. 11) to measure the correlation.
The Pearson correlation coefficient of all plots did not exceed 0.20 (and are of positive value) with the state of New York having the highest correlation coefficient (0.20). Also, all p-values exceed 0.30 which means that the correlations might have occurred by chance.
These numbers mean that in general, we have a slight tendency that the more flights were performed, the more new Corona cases we have. However, we can not guarantee that the correlation did not occur by chance resp. that there is a correlation at all.

This tendency can be explained that there are more people close to each other in a plane which increases the risk of getting sick. So if the number of performed flights increases, more people most likely take a flight and therefore more people are prone to get sick.

Interestingly, the correlation coefficient of the plot with the state New York is by far larger than the average and California. We can explain this by looking at the Figure 8 where we can see a slight peak in 1/2022 for New York. This plot shows the number of cancelled flights instead of the performed flights, but since the number of cancelled flights directly correlates with the number of performed flights, this plot can be used for the conclusion.

7.4 Non-weather related delays

We would have expected to see clearly that there were more delays due to Covid-19 certification checks or lack of staff, as we saw in the media in Europe.
However, on the plot Figure 12 we can see that the curve with the number of

delayed flights is parallel to the curve of performed flights. So there is no other effect after the occurrence of Corona.

We can see a peak in July 2021 in the number of delayed flights which is the same as the peak in July 2019, although there were more flights in 2019 than in 2021.
However, there is not really a trend in the following months, and we would have expected a bigger effect. Therefore, it could be that there was an impact, but it was relatively small, so it could also be a normal fluctuation.

The plot (Figure 13) shows the airlines' employees. There you can see a drop in the full-time curve at the beginning of the pandemic by 10,000 employees. The number of part-time employees remained constant. In July 2021, the number of employees was already back at the same level as in January 2019.
We would have expected a greater effect as well.
However, this also supports our previous plot and conclusion(Figure 12) that there were no more delays because there were enough employees.

7.5 Influence of Corona vaccination

Our observation in Figure 14 was that in general, fewer people got a first or even second booster than they got themselves fully vaccinated (dose 1 and dose 2). In comparison to California, fewer people got a first or second booster in New York.
One possible reason for this could be that there were stricter vaccination rules at the workplaces in New York. However, this was relaxed in the course of time and thus fewer boosters were vaccinated[6]. However, it is very good that in both states almost 80 % have gotten vaccinated.

In Figure 15, we wanted to examine the correlation between new Corona cases and the number of vaccinations. Since the number of vaccinated people always increased steadily but the new Corona case numbers still varied widely, it is actually impossible to show a correlation.

What we would have hoped for is to show that fewer people would have been infected by the Corona vaccination.
However, since other factors, mainly the measures, play a big role in how many people get infected, we can not show the expected or desired correlation. Here, it might have been better to compare the number of new Corona deaths relative to the number of cases with the vaccination numbers, as this would have been more meaningful. But even such a graph would have been "distorted" because it was more difficult to detect Corona at the beginning of the pandemic than towards the end of the pandemic.

[6] `https://www.jacksonlewis.com/publication/no-more-covid-19-vaccine-mandate-new-york-city-s-private-sector`, (January 2023)

8 Summary

In summary, we can say that the different U.S. states have behaved very similarly in terms of Corona case numbers and the number of flights. To see a correlation between cancelled flights and the number of Corona cases is very difficult because the number of cancelled flights only includes the very late cancelled flights. But one can see a weak correlation between the number of cases and the number of flights performed. However, the correlation is statistically not significant. This means that it is not guaranteed that when more flights are operated, the number of Corona cases increases.

Corona had no significant impact on flight delays in the United States. And therefore does not correspond to our expectations, which were based on newspaper articles from Europe.
The number of employees in the U.S. did not drop as rapidly as it did in Europe (based on what we have generally grasped from the news). This aligns with the number of delayed flights which did not drop a lot during the pandemic either.

For us, it was difficult to see a correlation between the number of people vaccinated to the Corona cases or the number of flights. Because the number of vaccinated persons increased steadily and the other data fluctuatedTherefore, we can not make a concrete statement.
In the end, we would like to emphasize that even if we did not find a correlation between the number of cases and the number of infections, there may still be one. Since we have left out some concepts like the measures.
It is also possible that we have found a correlation between data that does not exist in reality.

9 References

[1]: https://git.scicore.unibas.ch/databases-hs22/group-5

[2]: https://www.tablesgenerator.com/latex_tables

[3]: https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_States

[4]: https://www.luftlinie.org/Kalifornien/New-York

[5]: https://en.wikipedia.org/wiki/COVID-19_lockdowns

[6]: https://www.jacksonlewis.com/publication/no-more-covid-19-vaccine-mandate-new-york-city-s-private-sector

10 Attachment

10.1 States in the heatmaps

state	**state_name**
AK	Alaska
AL	Alabama
AR	Arkansas
AZ	Arizona
CA	California
CO	Colorado
CT	Connecticut
DE	Delaware
FL	Florida
GA	Georgia
HI	Hawaii
IA	Iowa
ID	Idaho
IL	Illinois
IN	Indiana
KS	Kansas
KY	Kentucky
LA	Louisiana
MA	Massachusetts
MD	Maryland
ME	Maine
MI	Michigan
MN	Minnesota
MO	Missouri
MS	Mississippi
MT	Montana
NC	North Carolina
ND	North Dakota
NE	Nebraska
NH	New Hampshire
NJ	New Jersey
NM	New Mexico
NV	Nevada
NY	New York
OH	Ohio
OK	Oklahoma
OR	Oregon
PA	Pennsylvania
PR	Puerto Rico
RI	Rhode Island
SC	South Carolina
SD	South Dakota
TN	Tennessee
TX	Texas
UT	Utah
VA	Virginia
VI	U.S. Virgin Islands
VT	Vermont
WA	Washington
WI	Wisconsin
WV	West Virginia
WY	Wyoming

Table 3: List of all the states in the heatmaps (fig. 6 and fig. 7).