I535 Spring 2023

Project Link: https://data.world/adityamhaske/projectmgmt

Aditya Sanjay Mhaske (admhaske)

4. Practice the following steps:

1. Create a new project (1Pt).

Project Link: https://data.world/adityamhaske/projectmgmt

2. Find a dataset on the data.world platform, add it to your project (connect to it) and describe the data (1 Pt)
    - ○ COVID-19 Open Research Dataset (CORD-19): https://data.world/kgarrett/covid-19-open-research-dataset

The COVID-19 Open Research Dataset (CORD-19)

The COVID-19 Open Research Dataset (CORD-19) is a collection of scientific papers, preprints, and other academic publications about the SARS-CoV-2 virus and the COVID-19 sickness. The Allen Institute for AI generated the dataset in collaboration with prominent academic institutes such as the National Institutes of Health and the World Health Organization.

The CORD-19 dataset is intended to aid academics, healthcare professionals, and policymakers in understanding the SARS-CoV-2 virus and the COVID-19 sickness. It comprises almost 100,000 research papers on themes including viral transmission, vaccine development, and the social and economic consequences of the epidemic.

3. Find some additional relevant data on the web (could be an image, a document, more data, etc.), add it to your dataset (1 Pt).
    - ○ liz-friedman/us-covid-19-data-from-nytimes (Additional Data set used): https://data.world/liz-friedman/us-covid-19-data-from-nytimes

2.liz-friedman/us-covid-19-data-from-nytimes

The "liz-friedman/us-covid-19-data-from-nytimes" dataset is a collection of COVID-19 data for the United States, sourced from the New York Times. The data includes information such as the number of confirmed cases, deaths, and new cases by state and by date. The data is updated regularly and provides a snapshot of the COVID-19 pandemic in the United States.

In terms of data type, the "liz-friedman/us-covid-19-data-from-nytimes" dataset is primarily structured data, with the data organized in tabular format, with each row representing a record and each column representing a feature.

In terms of use case, this dataset is valuable for researchers, analysts, and policymakers who are interested in understanding the impact of the COVID-19 pandemic in the United States. The data can be used to track the spread of the disease over time, compare the impact of the pandemic across different states, and inform decisions about public health measures.

4. Create at least two SQL queries to get filtered meaningful information from your data, and save it to your project (1 Pt).
    - Query 1:

```
SELECT us_state_or_territory, COUNT(*) FROM us_states

GROUP BY us_state_or_territory;
```

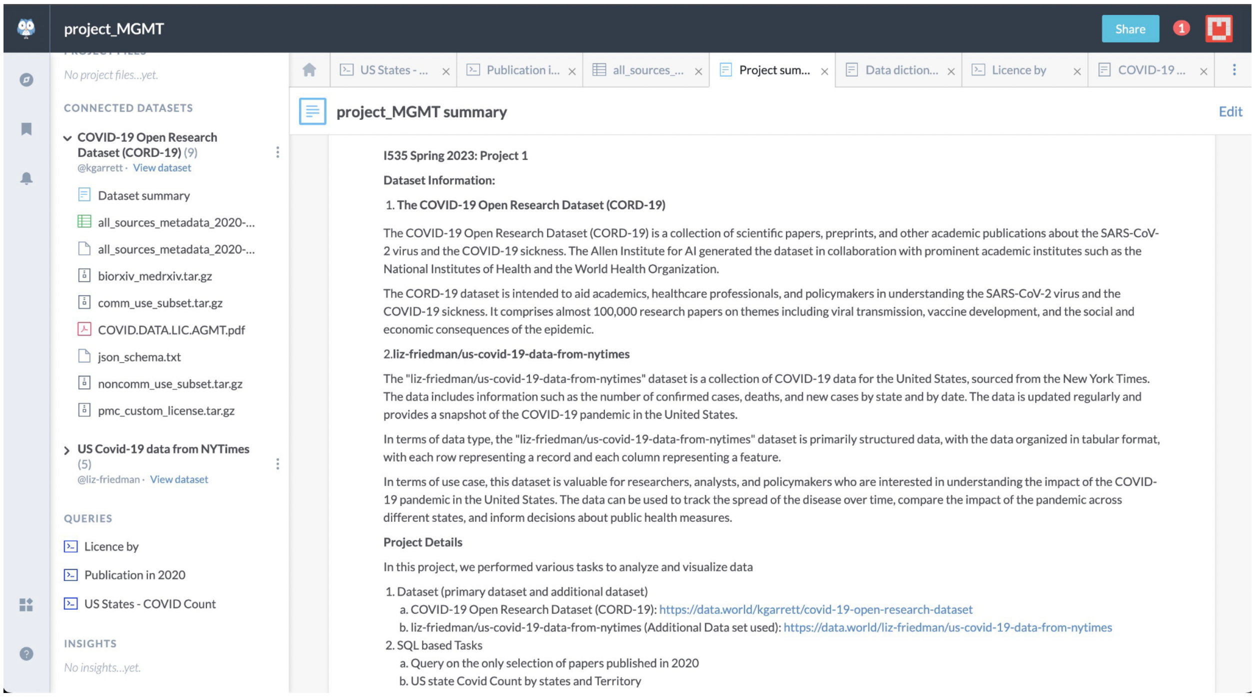
    - Query 2:

```
SELECT title, abstract FROM all_sources_metadata_2020_03_13

WHERE publish_time = 2020;
```

5. Build at least two charts to visualize your data - one from the full dataset and one from a query, and save your charts to the project (1 Pt).

5. Take the following screenshots (3 Pts):

1. **Your project summary (has to have the words "I535 Spring 2023" in it and a brief description of your data).**



I535 Spring 2023: Project 1

Dataset Information:

1. **The COVID-19 Open Research Dataset (CORD-19)**

The COVID-19 Open Research Dataset (CORD-19) is a collection of scientific papers, preprints, and other academic publications about the SARS-CoV-2 virus and the COVID-19 sickness. The Allen Institute for AI generated the dataset in collaboration with prominent academic institutes such as the National Institutes of Health and the World Health Organization.

The CORD-19 dataset is intended to aid academics, healthcare professionals, and policymakers in understanding the SARS-CoV-2 virus and the COVID-19 sickness. It comprises almost 100,000 research papers on themes including viral transmission, vaccine development, and the social and economic consequences of the epidemic.

2.**liz-friedman/us-covid-19-data-from-nytimes**

The "liz-friedman/us-covid-19-data-from-nytimes" dataset is a collection of COVID-19 data for the United States, sourced from the New York Times. The data includes information such as the number of confirmed cases, deaths, and new cases by state and by date. The data is updated regularly and provides a snapshot of the COVID-19 pandemic in the United States.

In terms of data type, the "liz-friedman/us-covid-19-data-from-nytimes" dataset is primarily structured data, with the data organized in tabular format, with each row representing a record and each column representing a feature.

In terms of use case, this dataset is valuable for researchers, analysts, and policymakers who are interested in understanding the impact of the COVID-19 pandemic in the United States. The data can be used to track the spread of the disease over time, compare the impact of the pandemic across different states, and inform decisions about public health measures.

Project Details

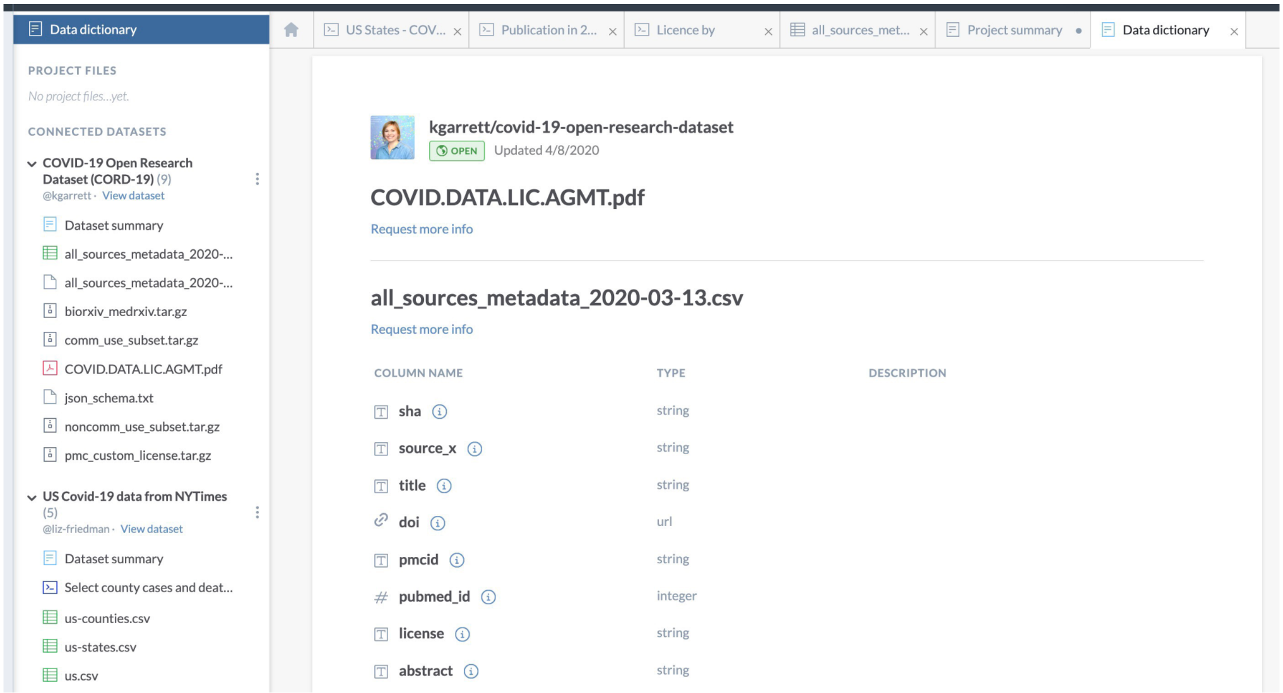
In this project, we performed various tasks to analyze and visualize data

1. Dataset (primary dataset and additional dataset)
   a. COVID-19 Open Research Dataset (CORD-19): https://data.world/kgarrett/covid-19-open-research-dataset
   b. liz-friedman/us-covid-19-data-from-nytimes (Additional Data set used): https://data.world/liz-friedman/us-covid-19-data-from-nytimes
2. SQL based Tasks
   a. Query on the only selection of papers published in 2020
   b. US state Covid Count by states and Territory

2. Your data dictionary.

Dataset Used for this project:

1. COVID-19 Open Research Dataset (CORD-19): https://data.world/kgarrett/covid-19-open-research-dataset
2. liz-friedman/us-covid-19-data-from-nytimes (Additional Data set used): https://data.world/liz-friedman/us-covid-19-data-from-nytimes



kgarrett/covid-19-open-research-dataset
OPEN Updated 4/8/2020

COVID.DATA.LIC.AGMT.pdf

Request more info

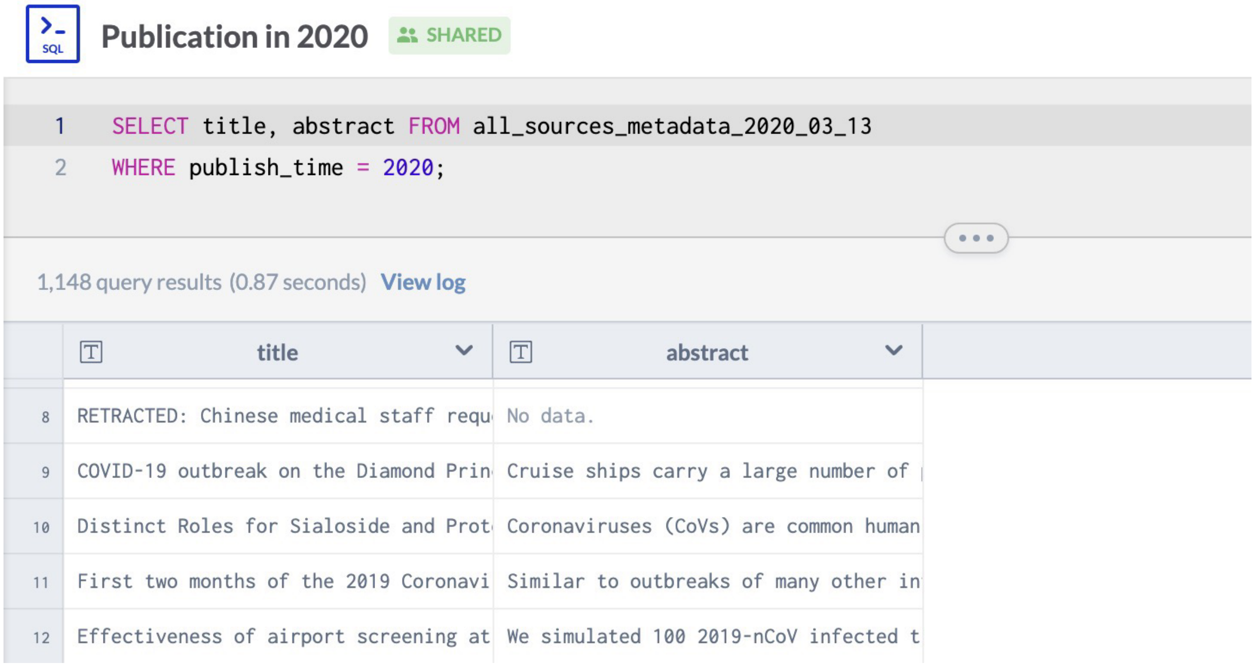
all_sources_metadata_2020-03-13.csv

Request more info

COLUMN NAME	TYPE	DESCRIPTION
sha	string	
source_x	string	
title	string	
doi	url	
pmcid	string	
pubmed_id	integer	
license	string	
abstract	string	

3. One of your queries.

1. Selected title and abstract from dataset published in 2020

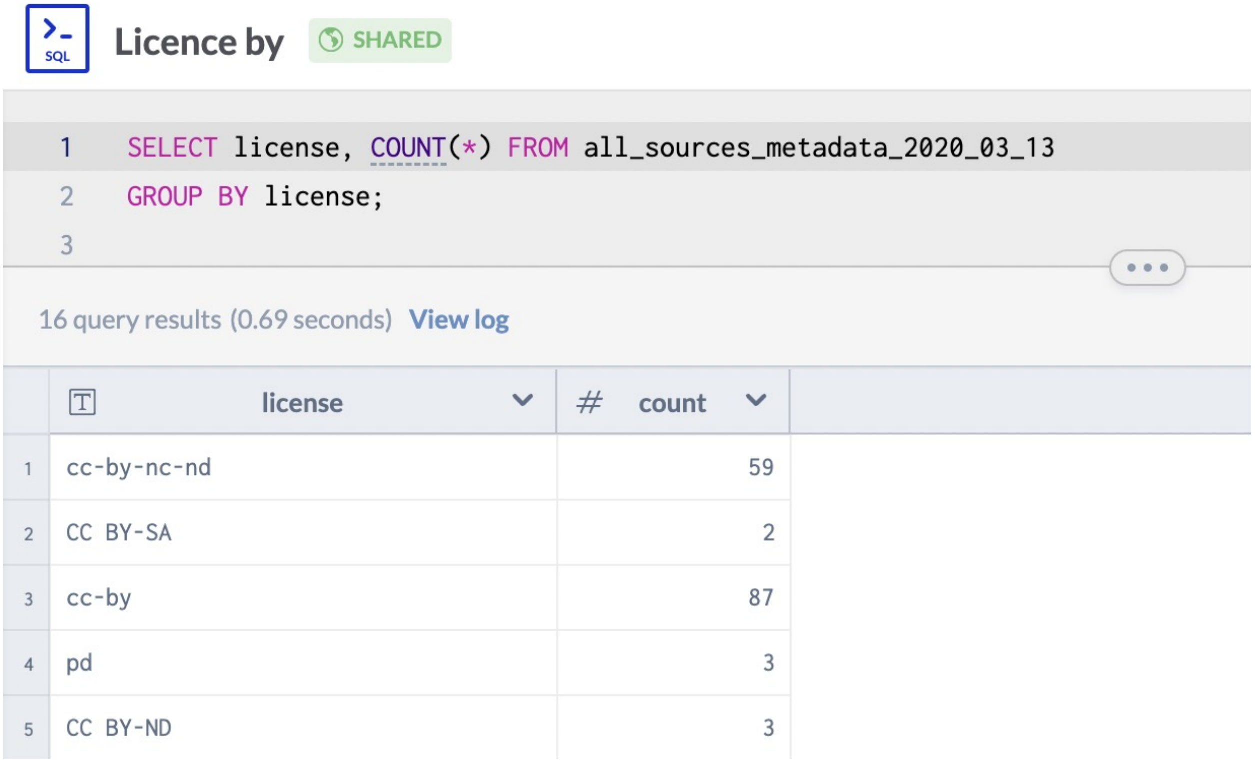


```sql
SELECT title, abstract FROM all_sources_metadata_2020_03_13
WHERE publish_time = 2020;
```

1,148 query results (0.87 seconds) View log

	title	abstract
8	RETRACTED: Chinese medical staff requ	No data.
9	COVID-19 outbreak on the Diamond Prin	Cruise ships carry a large number of
10	Distinct Roles for Sialoside and Prot	Coronaviruses (CoVs) are common human
11	First two months of the 2019 Coronavi	Similar to outbreaks of many other in
12	Effectiveness of airport screening at	We simulated 100 2019-nCoV infected t

2. Count licensed by



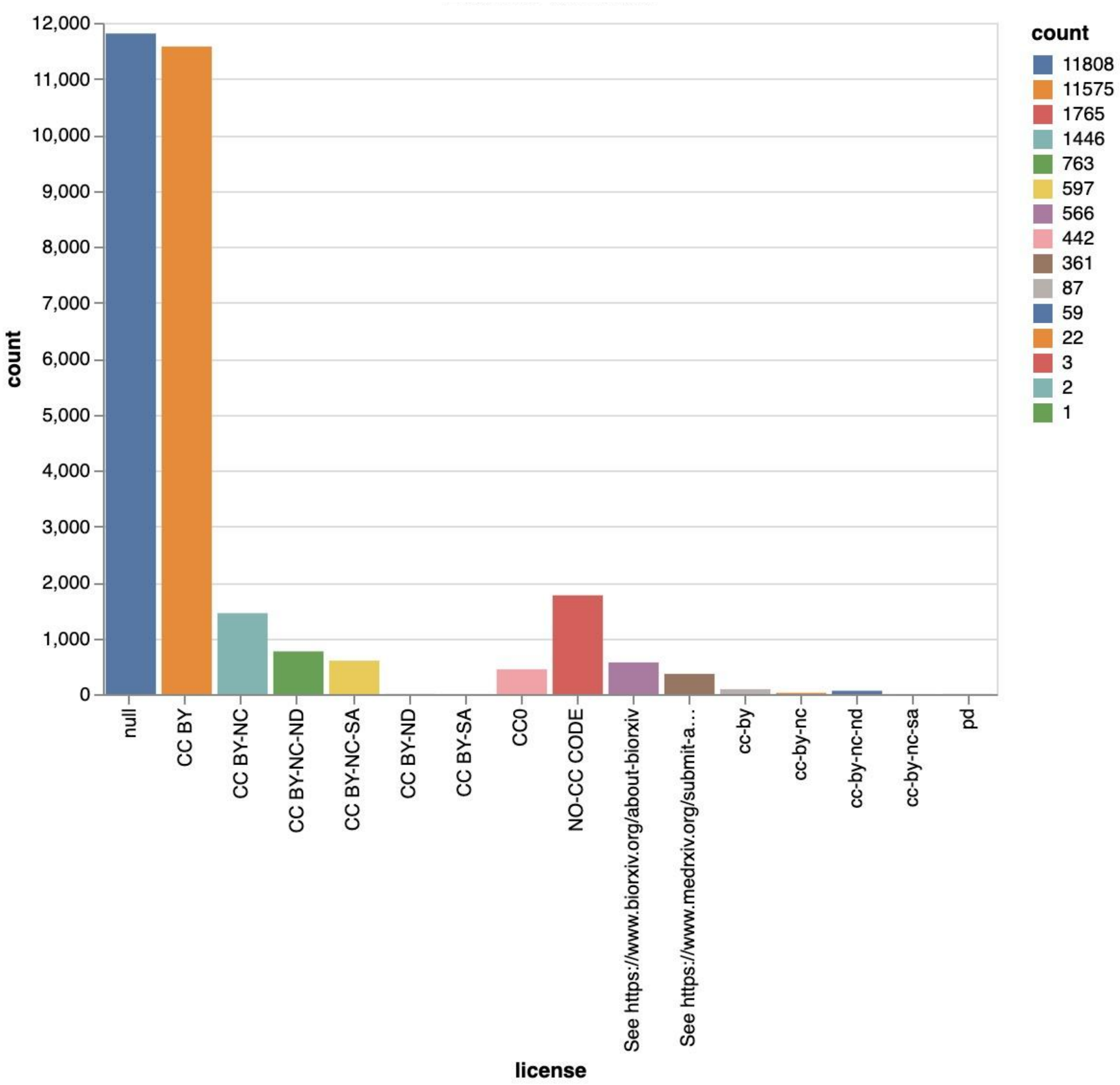
```sql
SELECT license, COUNT(*) FROM all_sources_metadata_2020_03_13
GROUP BY license;

```

16 query results (0.69 seconds) View log

	license	count
1	cc-by-nc-nd	59
2	CC BY-SA	2
3	cc-by	87
4	pd	3
5	CC BY-ND	3

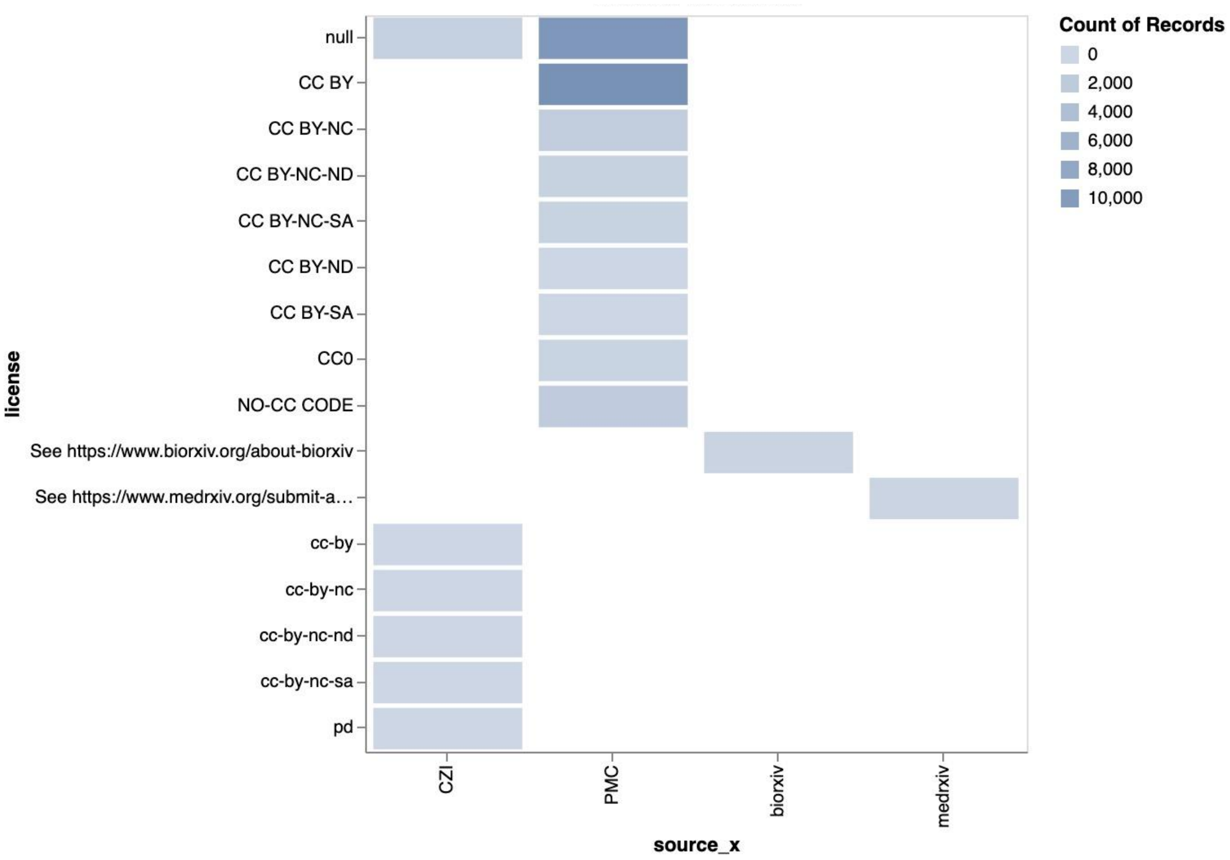
4. **Your visualization.**

1. Visualization 1: Count Vs License
   Visualization based on SQL query:



2. Visualization 2: Source Vs License

Visualization based on Full Dataset:



6. Incorporate your screenshots into a brief (2-3 paragraphs) report and submit it (3 Pts). Your report should answer the following questions:

1. **What is your impression of this platform? What do you see as its advantages and disadvantages?**

According to my perspective, the data.world is a powerful and versatile platform for data analysis and management, but it may not be suitable for all users, especially those who are just starting with data analysis. Data.world platform is a better version of kaggle.com, because of the integration of various tools for visualization and other data science-related platforms. It is really useful to get insights into data.

Advantages of data.world is a platform that allows multiple users to collaborate on a project, for teamwork and sharing results. This platform integrates with various tools and platforms, for data visualization, database management, and programming. Platform has a large repository of public datasets, allowing users to access a wide range of data for their analysis and projects.

Disadvantages of data.world are platform has a wide range of features and integrations, which can make it overwhelming or confusing for some users. As well as with any public repository of data, the quality of the data in data.world can vary.

2. **What dataset have you used and how would you describe it using the terms you learned in the lecture and readings?**

I used COVID-19 Open Research Dataset (CORD-19). CORD-19 for this project. Multi-disciplinary dataset of scientific articles and preprints related to the COVID-19 pandemic.
In terms of the concepts learned in lectures and readings, CORD-19 can be described as follows:

- Big Data: With over 200,000 articles, CORD-19 is a large and complex dataset
- Structured Data: The data in CORD-19 is structured in a tabular format, with each article represented as a row and the attributes of the article (such as title, abstract, and publication date) represented as columns.
- Relational Data: The data in CORD-19 can be related to other datasets, such as datasets of COVID-19 cases and deaths, by linking the data on common attributes such as the location and date of the event.
- Unstructured Data: Some of the data in CORD-19, such as the full text of articles, is unstructured and requires preprocessing and analysis to extract meaningful information.

In summary, CORD-19 is a complex and multi-faceted dataset that provides valuable information for COVID-19 research and can be analyzed using a variety of techniques and tools.