Antibiotic resistance patterns in Staphylococcus isolates in clinical settings

Data Analytics - July 2024

Presented by
Verónica Larroy

Content

1

Project overview

Some context about me!

1 Project overview

Why did I choose this project?

Staphylococcus

120K

diagnosed in the US

20K

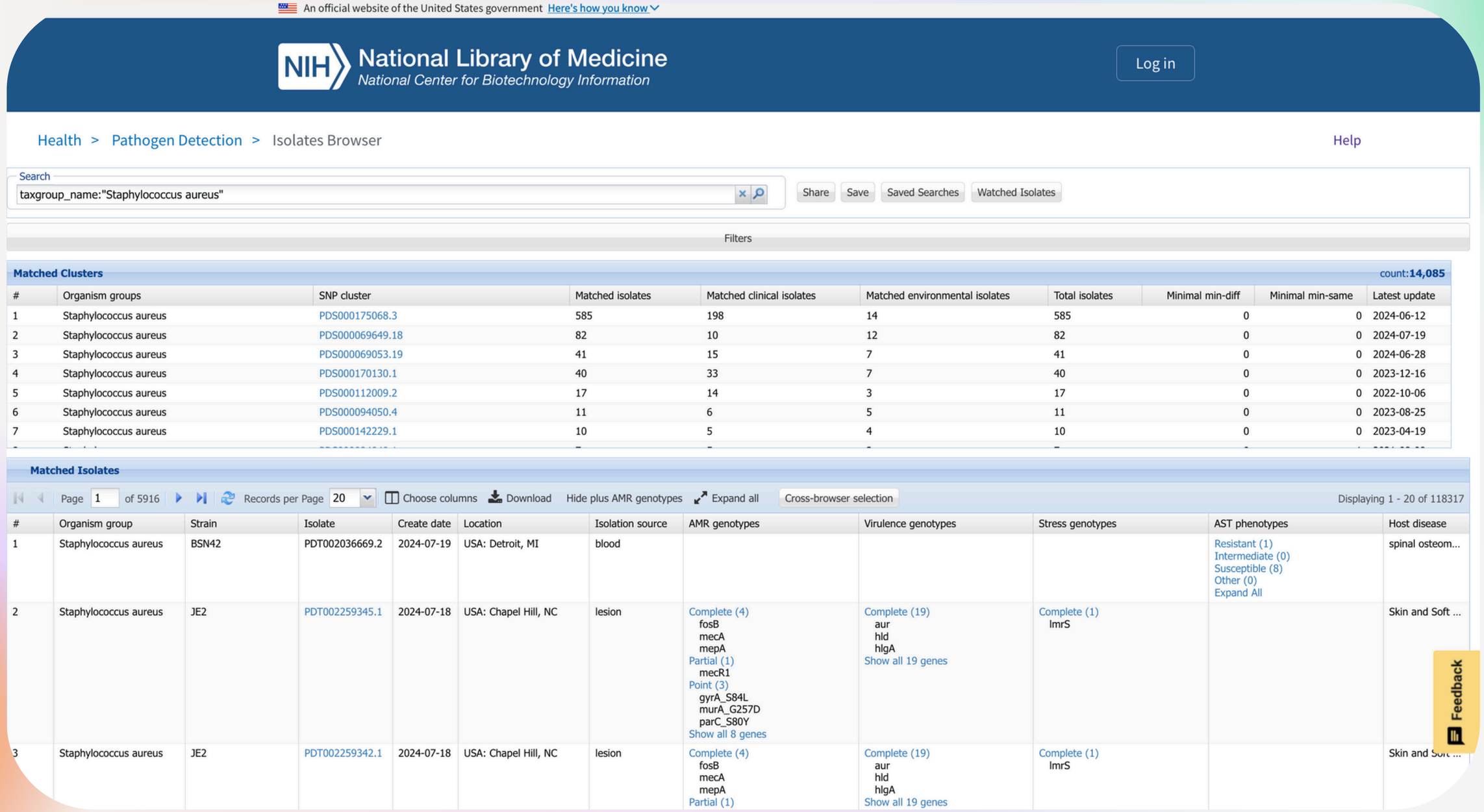
annual deaths in the US due to antibiotic resistance

2 Research questions & datasets

1. What is the diversity of Staphylococcus species in different clinical sources

2. What are the antibiotic resistance patterns of Staphylococcus isolates from various clinical sources?

2 Research questions & datasets



An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Log in

Health > Pathogen Detection > Isolates Browser

Help

Search

taxgroup_name:"Staphylococcus aureus"

Share Save Saved Searches Watched Isolates

Filters

Matched Clusters count:**14,085**

#	Organism groups	SNP cluster	Matched isolates	Matched clinical isolates	Matched environmental isolates	Total isolates	Minimal min-diff	Minimal min-same	Latest update
1	Staphylococcus aureus	PDS000175068.3	585	198	14	585	0	0	2024-06-12
2	Staphylococcus aureus	PDS000069649.18	82	10	12	82	0	0	2024-07-19
3	Staphylococcus aureus	PDS000069053.19	41	15	7	41	0	0	2024-06-28
4	Staphylococcus aureus	PDS000170130.1	40	33	7	40	0	0	2023-12-16
5	Staphylococcus aureus	PDS000112009.2	17	14	3	17	0	0	2022-10-06
6	Staphylococcus aureus	PDS000094050.4	11	6	5	11	0	0	2023-08-25
7	Staphylococcus aureus	PDS000142229.1	10	5	4	10	0	0	2023-04-19

Matched Isolates

Page 1 of 5916 Records per Page 20 Choose columns Download Hide plus AMR genotypes Expand all Cross-browser selection Displaying 1 - 20 of 118317

#	Organism group	Strain	Isolate	Create date	Location	Isolation source	AMR genotypes	Virulence genotypes	Stress genotypes	AST phenotypes	Host disease
1	Staphylococcus aureus	BSN42	PDT002036669.2	2024-07-19	USA: Detroit, MI	blood				Resistant (1) Intermediate (0) Susceptible (8) Other (0) Expand All	spinal osteom...
2	Staphylococcus aureus	JE2	PDT002259345.1	2024-07-18	USA: Chapel Hill, NC	lesion	Complete (4) fosB mecA mepA Partial (1) mecR1 Point (3) gyrA_S84L murA_G257D parC_S80Y Show all 8 genes	Complete (19) aur hld hlgA Show all 19 genes	Complete (1) lmrS		Skin and Soft ...
3	Staphylococcus aureus	JE2	PDT002259342.1	2024-07-18	USA: Chapel Hill, NC	lesion	Complete (4) fosB mecA mepA Partial (1)	Complete (19) aur hld hlgA Show all 19 genes	Complete (1) lmrS		Skin and Soft ...

Feedback

3 Data wrangling and cleaning

```python
#Renaming the relevant columns:
df.rename(columns={"#Organism group": "organism_group",
           "Strain": "strain",
           "Isolate": "isolate",
           "Create date": "create_date",
           "Location": "location",
           "Isolation source": "isolation_source",
           "AMR genotypes": "amr_genotypes",
           "Virulence genotypes": "virulence_genotypes",
           "Stress genotypes": "stress_genotypes",
           "AST phenotypes": "ast_phenotypes",
           "Host disease": "host_disease"
           },
           inplace=True
           )
```

✓ 0.0s

```python
#deleting non-relevant rows:
#I realised that there are rows that should not be included in the analysis. I will delete those rows since my analysis is strictly focused on clinical settings:
df = df[df["isolation_source"].str.contains("Nursing home")==False]
df = df[df["isolation_source"].str.contains("household surface")==False]
df = df[df["isolation_source"].str.contains("Facility")==False]
df = df[df["isolation_source"].str.contains("laboratory mutant")==False]
df = df[df["isolation_source"].str.contains("host")==False]
```

✓ 0.0s

```python
#Lets clean the columns:
#DATE: To clean it, we convert it to datetime and then, create 2 new columns, one for the year and another for the month.
df["create_date"] = pd.to_datetime(df["create_date"])
df["year"] = df["create_date"].dt.year
df["month"] = df["create_date"].dt.month
df.drop(columns="create_date", axis=1, inplace=True)
```

✓ 0.0s

```python
#casting my relevant columns so that I can later work with them:
df["isolation_source"] = df["isolation_source"].astype(str)
df["virulence_genotypes"] = df["virulence_genotypes"].astype(str)
df["stress_genotypes"] = df["stress_genotypes"].astype(str)
df["host_disease"] = df["host_disease"].astype(str)
```

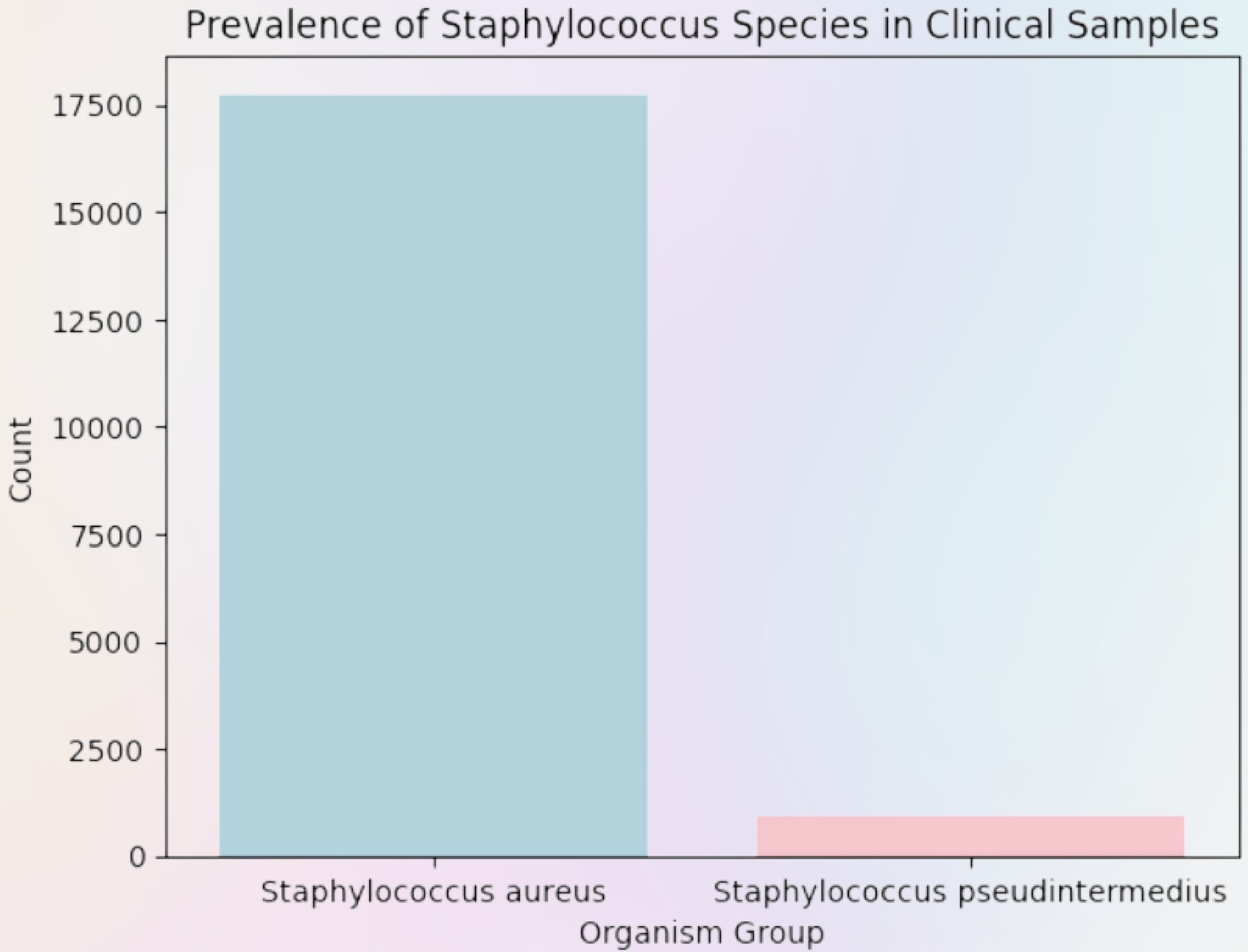
✓ 0.0s

```python
#Cleaning functions:
def get_state(location: str) -> str:
    state_eq = {
        "Alabama": ["Alabama","AL"],
        "Alaska": ["Alaska","AK"],
        "Arizona": ["Arizona","AZ"],
        "Arkansas": ["Arkansas","AR"],
        "California": ["California","CA"],
        "Colorado": ["Colorado","CO"],
        "Connecticut": ["Connecticut","CT"],
        "Delaware": ["Delaware","DE"],
        "Florida": ["Florida","FL"],
```

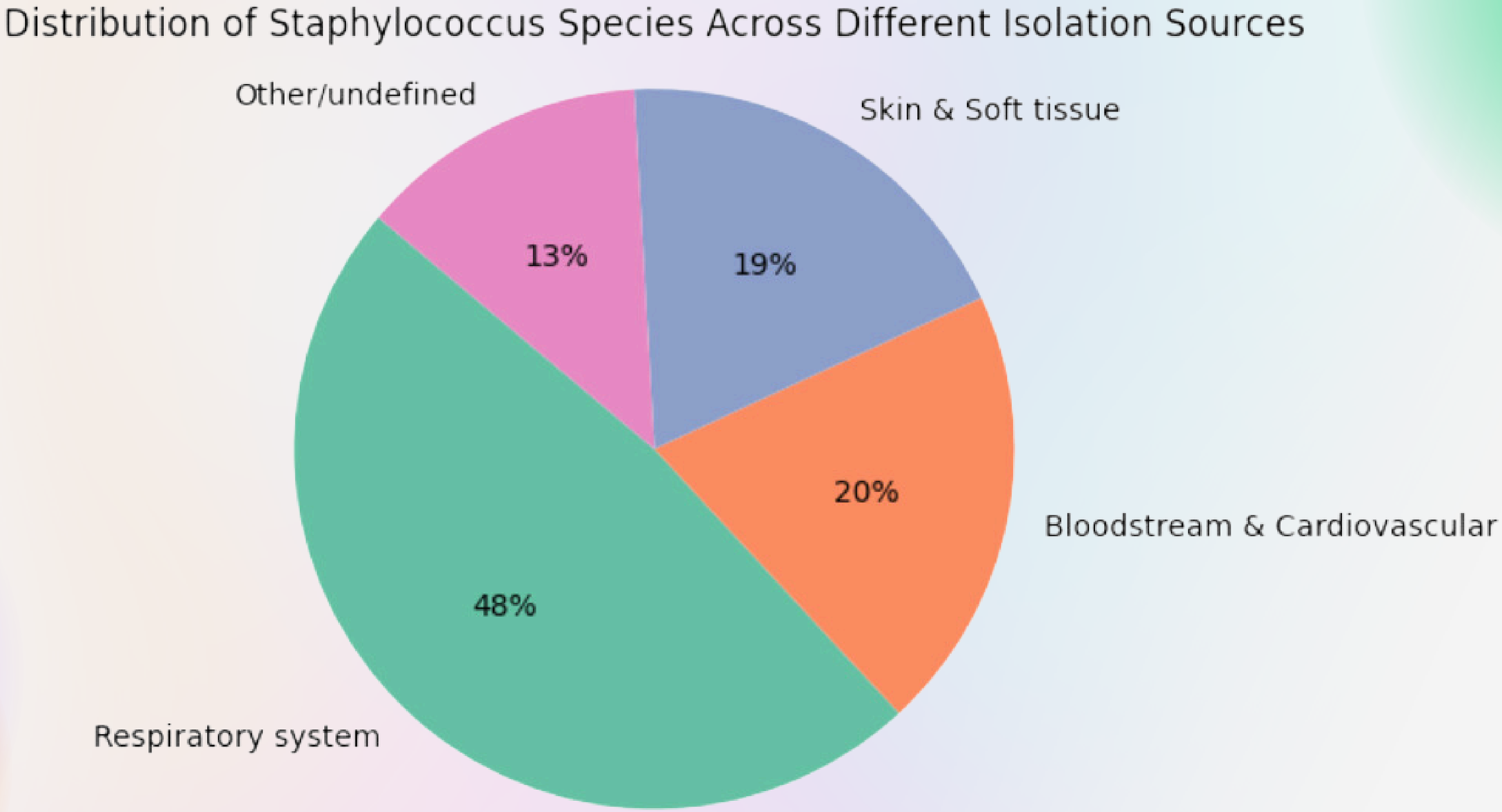
Some challenges

- Duplicates
- Formatting issues with genotype strings

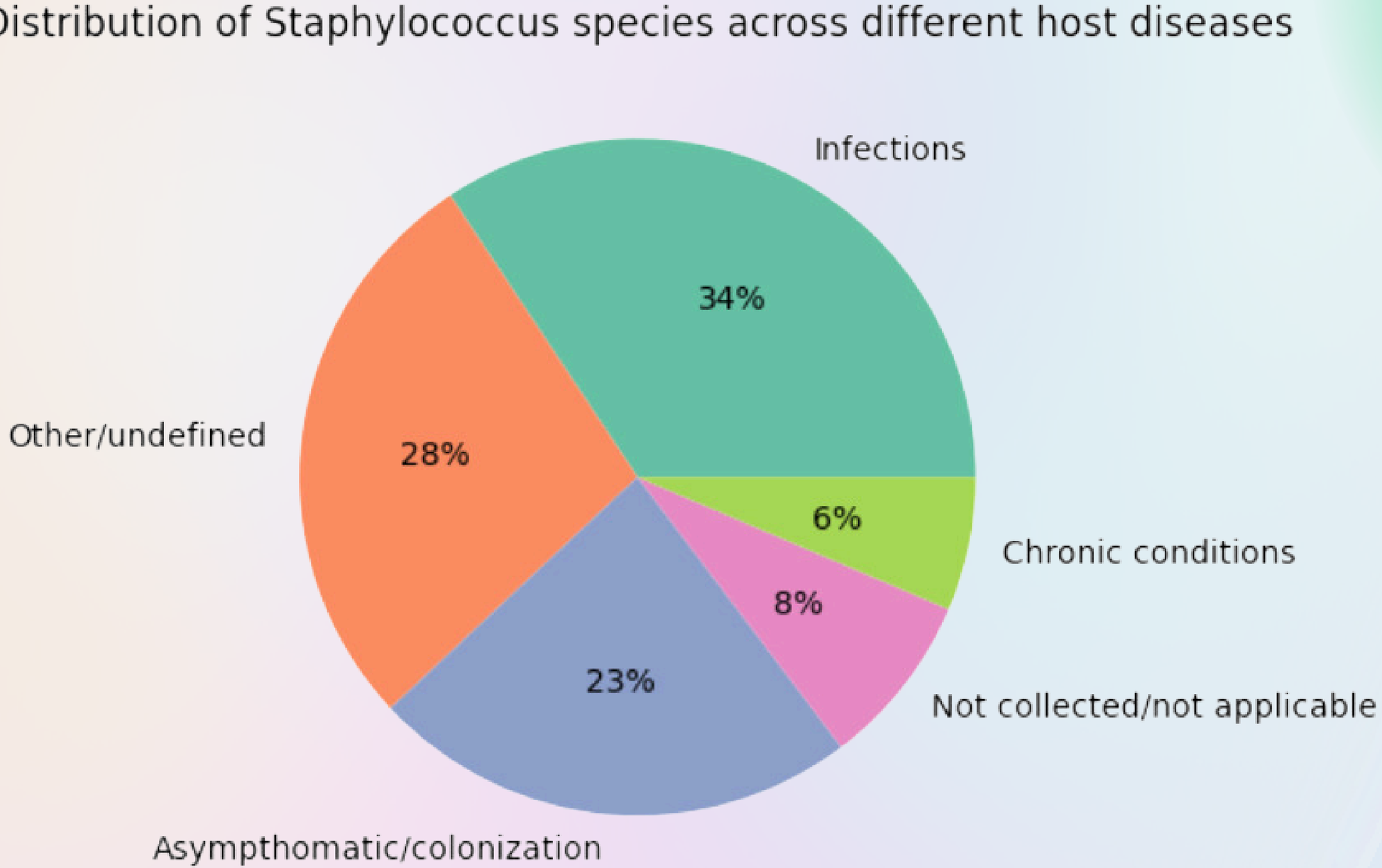
4 EDA & visualizations



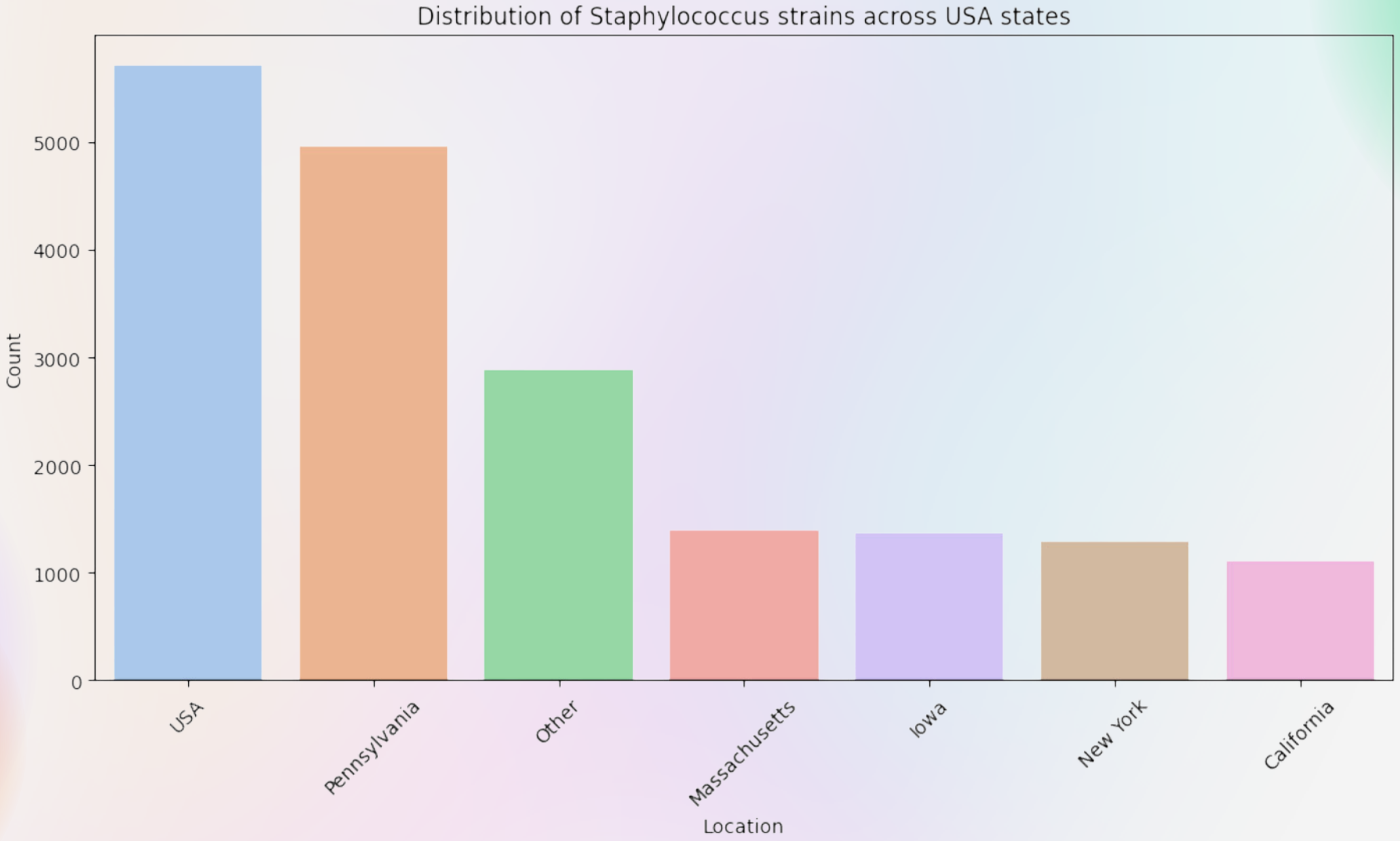
4 EDA & visualizations



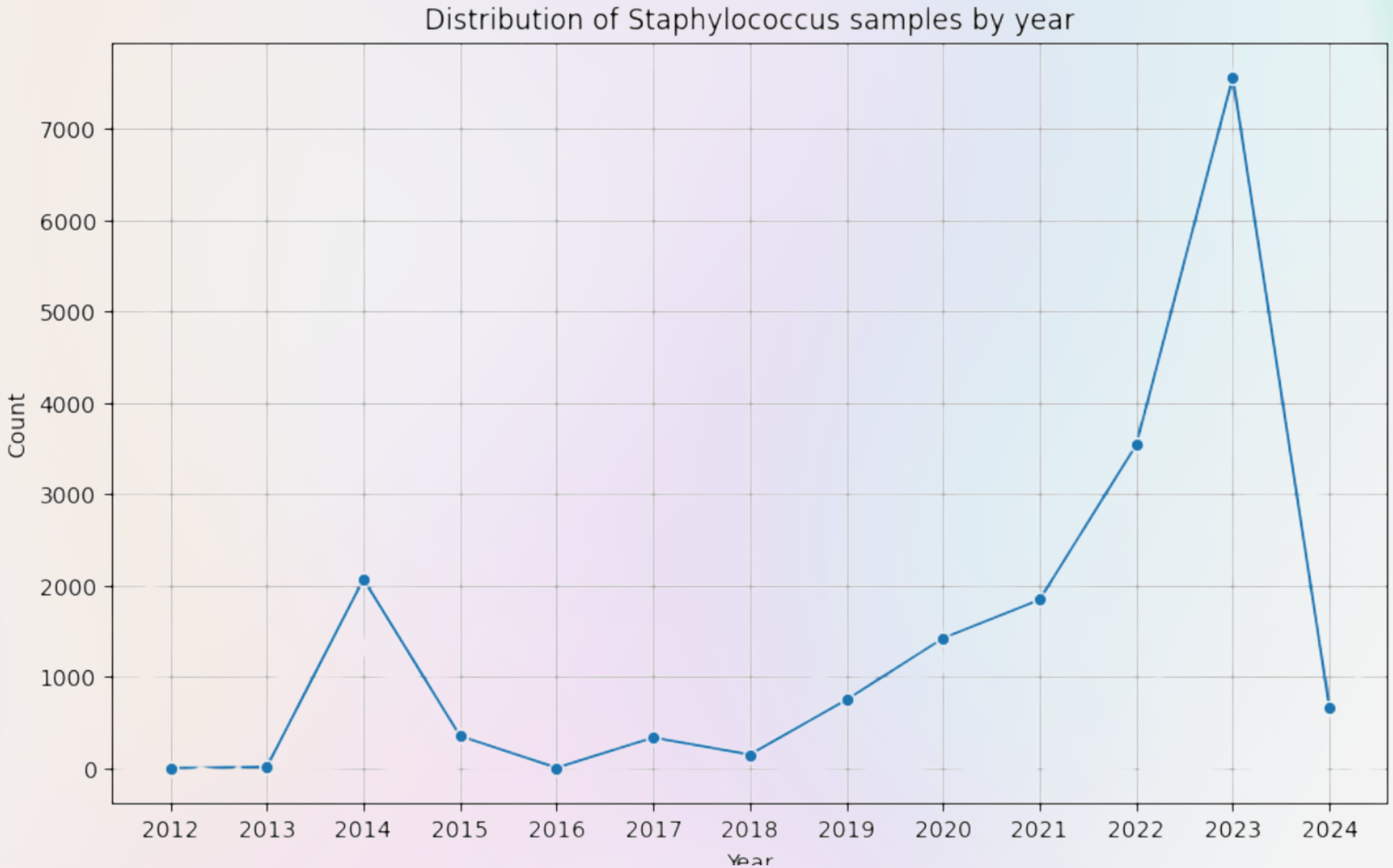
4 EDA & visualizations



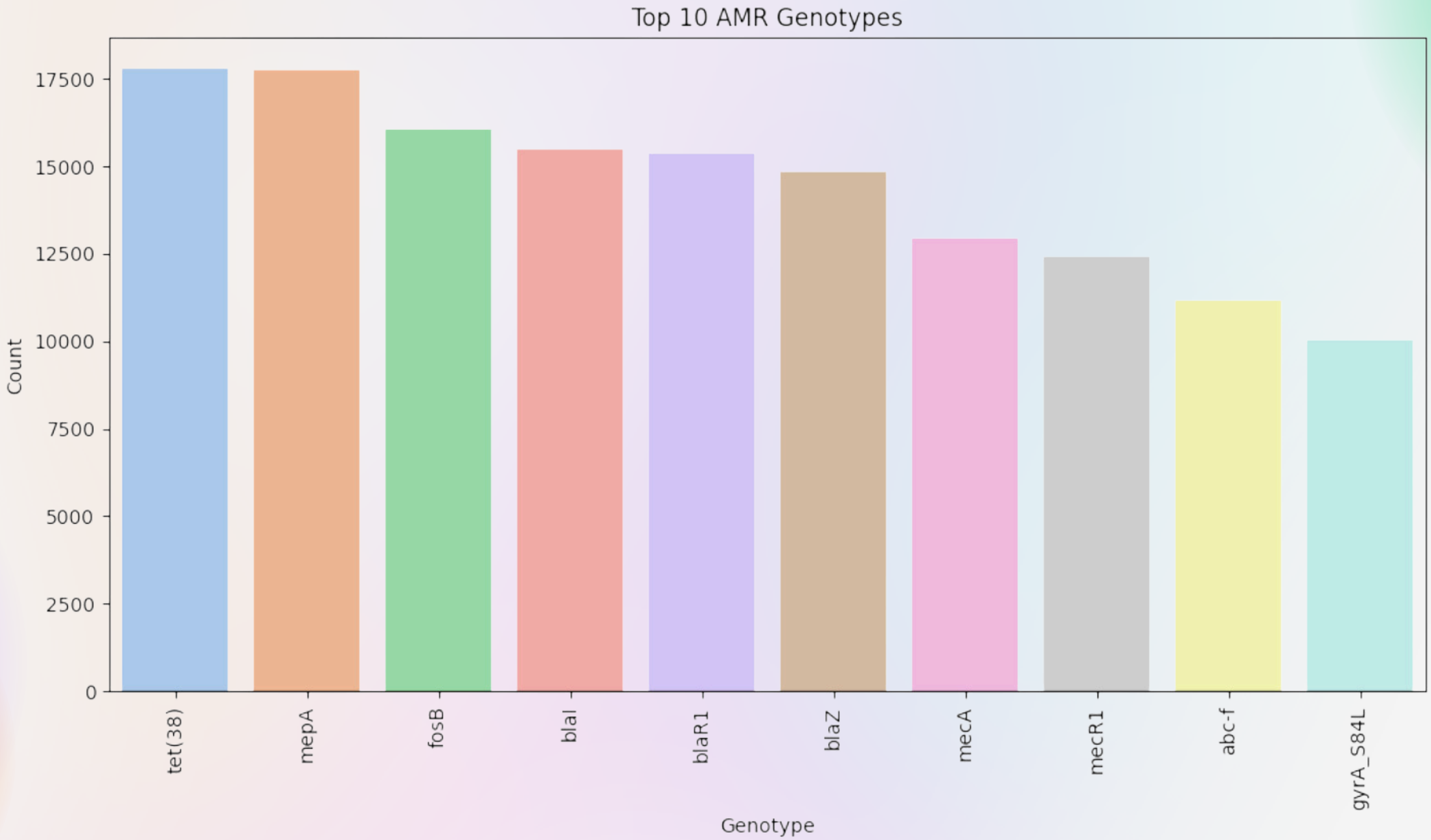
4 EDA & visualizations



4 EDA & visualizations



4 EDA & visualizations



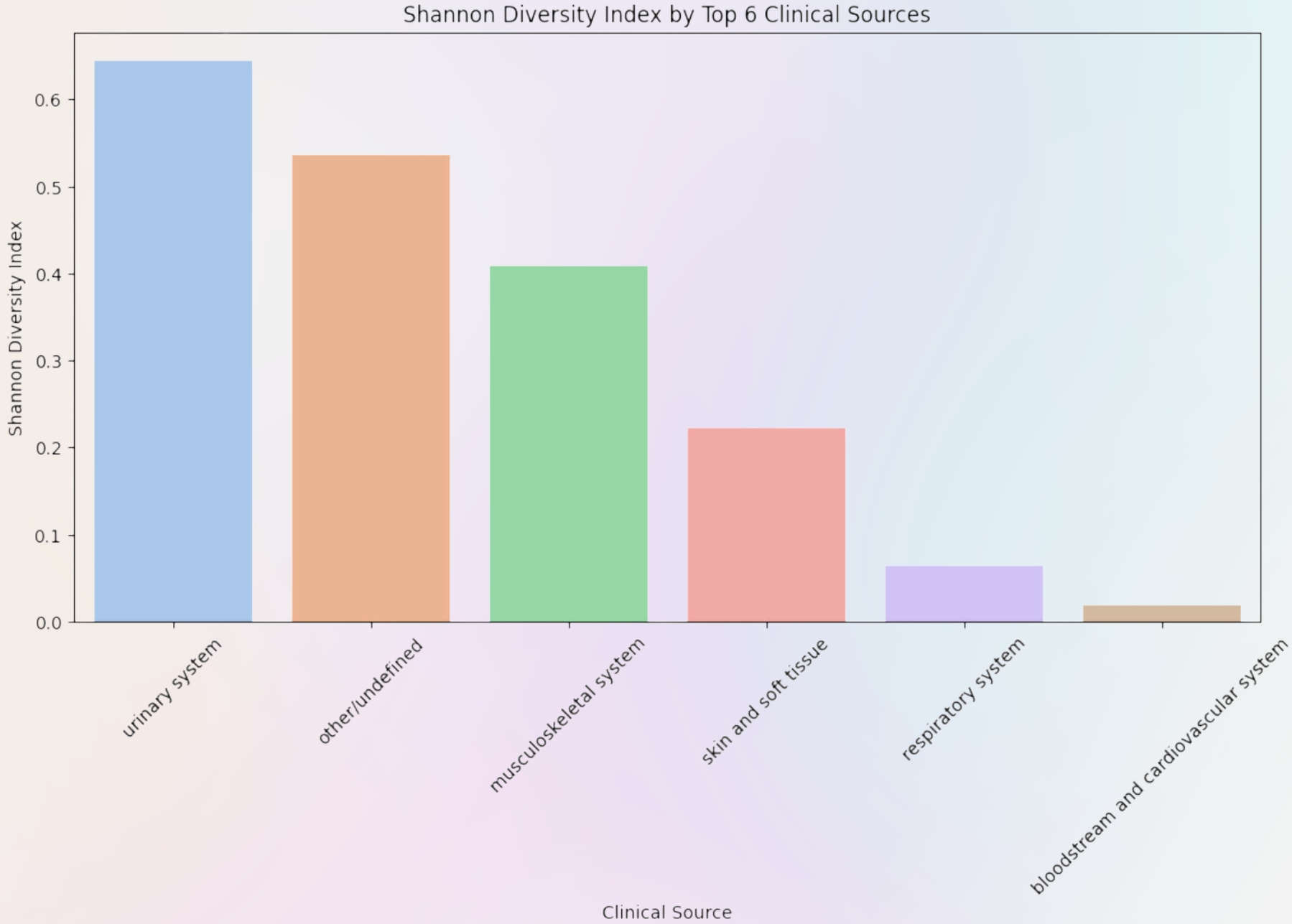
4 EDA & visualizations

1. What is the diversity of Staphylococcus species in different clinical sources

Shannon Diversity index

$$H = -\sum[(pi) \times \log(pi)]$$

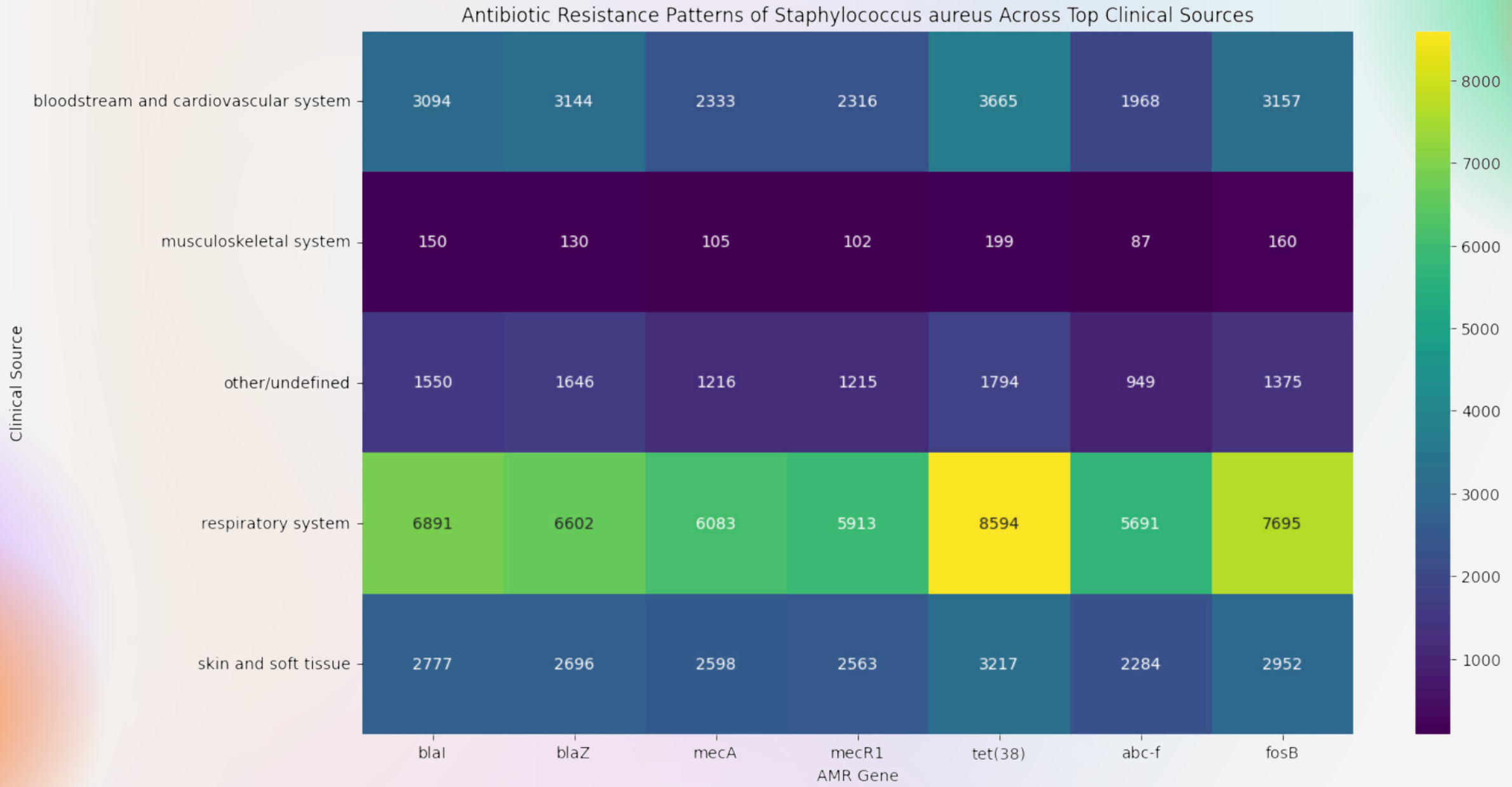
4 EDA & visualizations



4 EDA & visualizations

2. What are the antibiotic resistance patterns of Staphylococcus isolates from various clinical sources?

4 EDA & visualizations

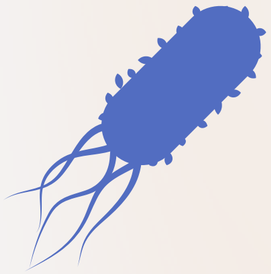


5 Major obstacles

- Doing the math
- Finding the right libraries to do the math

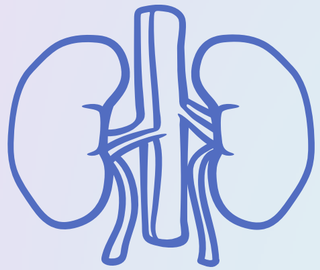
6 Conclusion & insights

Key takeaways



High prevalence of S. aureus

Respiratory system is the primary isolation source.



The urinary system accounts for major Staphylococcus diversity

AMR genes: tet(38), mepA, fosB are most common.

Thank you!!

Data Analytics - July 2024

Presented by
Verónica Larroy