Local Dynamics of COVID-19 Transmission in Queensland

Prepared by: Gabriel Wu
Purpose: Portfolio Submission for Entry-Level Data Analyst Role
Date: 19th August 2025

Table of Contents

1. Executive Summary....1
2. Introduction....1
3. Objectives....1
4. Methodology / Data Overview....1
5. Data Cleaning and Preparation....2
6. Exploratory Data Analysis (EDA)....3
    a) Age Group Analysis....3
    b) Temporal Trends....4
    c) Geographic Distribution....5
    d) Source Attribution....7
7. Key Insights and Interpretation....8
8. Recommendations....9
9. Conclusion....9
10. References....10
11. Appendices....10

1. Executive Summary

This report explores COVID-19 trends in Queensland from July 2024 to July 2025, using open government data. Key findings highlight a shift in vulnerability toward young children and older adults, seasonal case peaks aligned with flu periods, and high case concentrations in Metro North and Metro South. Data limitations, especially the high number of cases marked “Under Investigation,” affect transmission insights. Recommendations include improving data structure, ensuring complete reporting, and targeting public health measures around seasonal risks and vulnerable groups.

2. Introduction

COVID-19 has significantly impacted public health systems worldwide. In Queensland, understanding the localised dynamics of transmission is critical for guiding policy, improving preparedness, and supporting vulnerable populations. This report explores patterns in the spread of COVID-19 from July 2024 to July 2025, focusing on key variables such as age group, geography, and source of infection.

The data-driven approach taken here supports evidence-based decision-making by public health bodies and aims to identify both trends and anomalies to inform future strategies.

3. Objectives

This report aims to:

- Identify **vulnerable age groups** and shifts in case patterns over time
- Analyse the **temporal and demographic spread** of COVID-19 in Queensland
- Assess **regional case distribution** across LGAs and HHS
- Investigate **transmission sources** and data completeness

4. Methodology / Data Overview

Tool Used

This analysis was conducted using Python in Jupyter Notebook for its efficiency and flexibility. Python’s libraries, including **Pandas** for data cleaning and transformation, **NumPy** for numerical operations, and **Matplotlib** and **Seaborn** for data visualisation, were used to explore trends across demographics, time, and regions in detail.

Throughout the project, I demonstrated a range of Python skills, from foundational concepts to more advanced techniques. Additionally, AI-assisted coding tools were used to validate and optimise code before execution, ensuring both accuracy and efficiency.

Datasets Used

The dataset used in this analysis was sourced from the **Queensland Government Open Data Portal** on 10 August 2025 and is licensed under the **Creative Commons Attribution 4.0** licence. It consists of two interrelated tables, each containing 1,854,305 records and nine attributes, offering a comprehensive view of COVID-19 case data across Queensland.

The first table, titled **Queensland COVID-19 Case Line List – Age Groups**, focuses on the demographic breakdown of cases. Key fields include NOTIFICATION_DATE, which enables identification of case peaks over time, and AGE_GROUP_5Y, which supports the analysis of age-specific trends and vulnerabilities in five-year groupings. Available here

The second table, **Queensland COVID-19 Case Line List – Location & Source of Infection**, provides a geographic and epidemiological lens. Key columns include NOTIFICATION_DATE for tracking temporal trends, HHS to explore the performance and coverage of Hospital and Health Services, LGA_NAME to assess case density and responses at the local government level, and SOURCE_INFECTION to distinguish between community transmission and cases acquired overseas. Available here

Limitations and Data Discrepancies

A notable limitation of the dataset is the significant proportion of cases (98.1%) classified as “Under Investigation” in the SOURCE_INFECTION field, which restricts the reliability of any analysis focused on transmission sources.

Additionally, inconsistencies between the HHS and LGA_NAME fields may impact the accuracy of regional insights and complicate alignment across administrative boundaries. For example, cases where Brisbane City appears under non-Brisbane HHS regions represent unusual combinations. Each unique HHS and LGA_NAME pairing was counted to detect such inconsistencies and note potential outliers. However, this was not explored further in the EDA, as it was outside the scope of this project's objectives.

These anomalies were flagged and filtered during analysis where needed, but no corrections were applied to preserve data fidelity and accurately reflect the original source.

5. Data Cleaning and Preparation

To ensure the dataset was analysis-ready, a series of data cleaning and preparation steps were undertaken across both tables.

A missing value analysis revealed that 1.46% of LGA_NAME entries were missing. These null values were retained to preserve the completeness of the dataset without introducing imputation bias. According to the data source, “Location variables are masked as null in instances when a case does not usually reside in Queensland” (Queensland Government, 2025) which provides context for some of these missing values. While some inconsistencies were also identified between the HHS and LGA_NAME fields, no corrections were made in order to maintain data fidelity and accurately reflect the original reporting.

Non-essential columns were removed to streamline the dataset for the intended analytical objectives. In the Age Groups table, the _id column was excluded. In the Location & Source of Infection table, the _id, POSTCODE, SA2_CODE, and SA2_REGION columns were removed, as they were not relevant to the scope of this analysis.

Although each table contained unique row identifiers, a decision was made not to join them directly, as mismatches in granularity and row alignment posed a risk of introducing inaccuracies or duplicating data.

Several formatting adjustments were applied to support data visualisation and time-series analysis. The NOTIFICATION_DATE field was converted to datetime format, enabling accurate chronological sorting and aggregation. In the Age Groups table, the AGE_GROUP_5Y field was cleaned to remove extraneous text such as the word “years,” facilitating cleaner category labels in visual outputs.

No additional deduplication was required. According to the data source, the dataset had already been internally reviewed through the Notifiable Conditions System (NoCS), ensuring that duplicate or repeat entries within the same reinfection period were appropriately integrated (Queensland Government, 2025). This level of preprocessing reduced the risk of inflating case counts due to duplication.

6. Exploratory Data Analysis (EDA)

The findings from the charts highlight several key trends in the demographic, temporal, and geographic distribution of COVID-19 cases from the start of the pandemic to the current period. These trends show how the impact has shifted across population groups and locations over time.

a) Age Group Analysis

Figure 1 shows that the 25-29 age group recorded the highest overall case count, with over 164,000 cases throughout the pandemic. However, during the more recent period (July 2024 to July 2025), the 0-4 age group (4,838 cases) and the 75-79 age group (4,994 cases) emerged as the most affected, as illustrated in Figure 2.

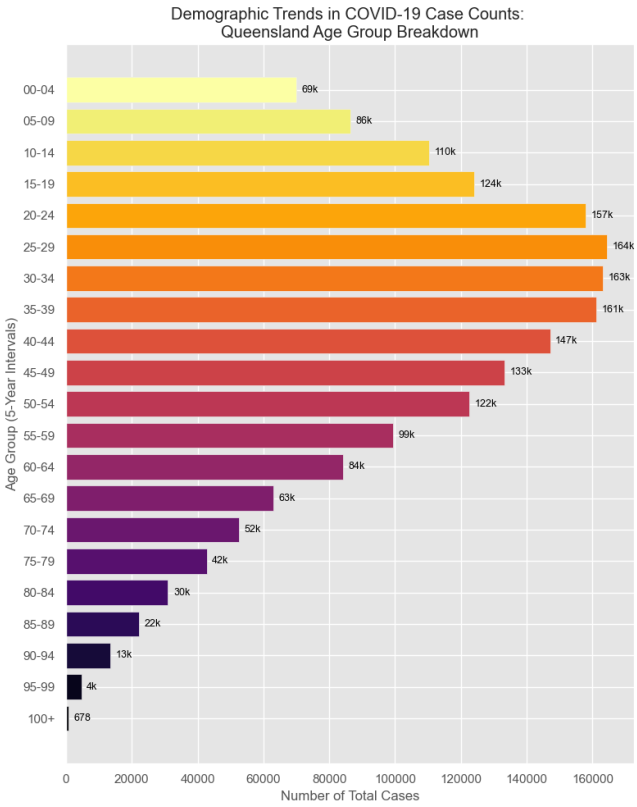

Figure 2: COVID-19 case distribution by age group in Queensland, highlighting the 25-29 cohort as the most affected over the full period.

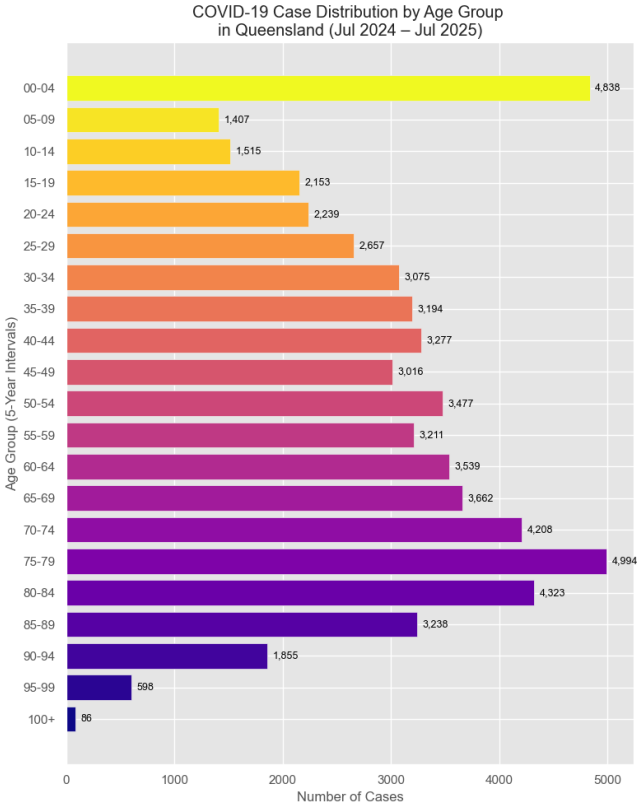

Figure 1: COVID-19 case counts by age group in Queensland from July 2024 to July 2025, with the 0-4 and 75-79 age groups showing the highest incidence during this period.

b) Temporal Trends

In Figure 3, case numbers peaked in October 2021 before declining and stabilising after January 2023. Figure 4 shows that peak activity across all HHS regions occurred in both July and December 2024, with the lowest activity observed in September 2024 and April 2025.

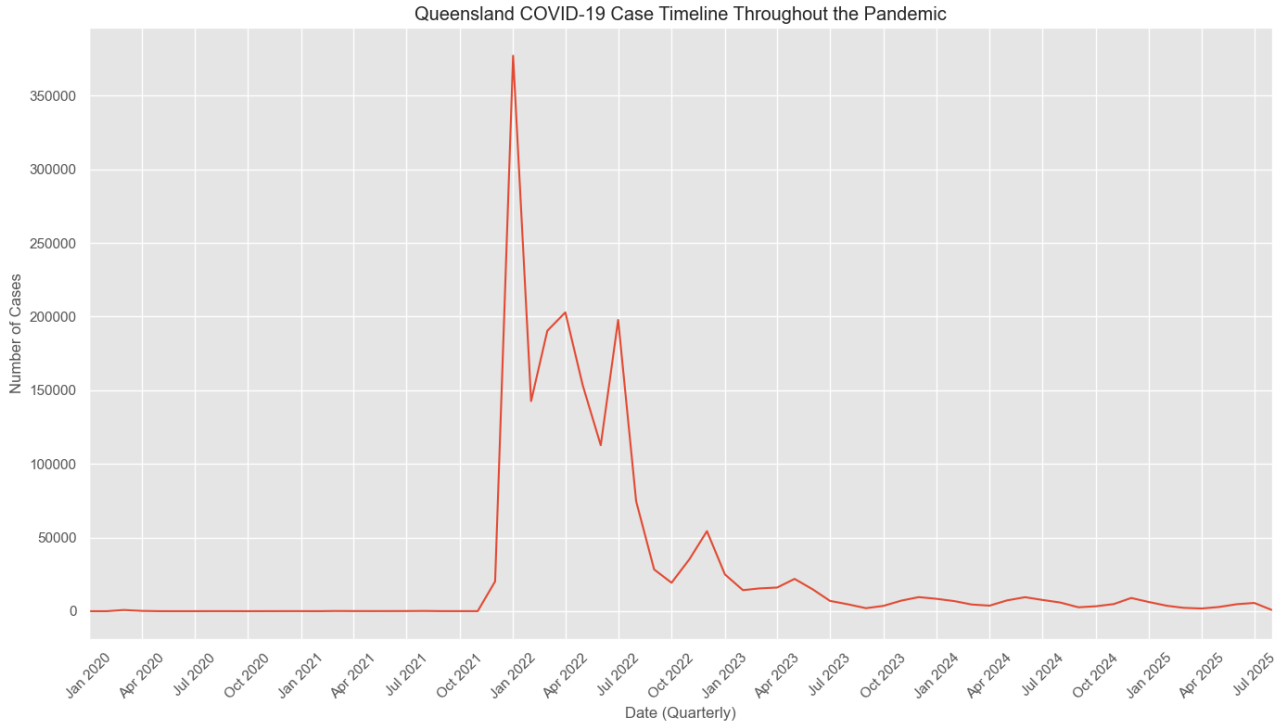

Figure 3: COVID-19 case trends in Queensland throughout the pandemic, peaking in October 2021 before declining and stabilising after January 2023.

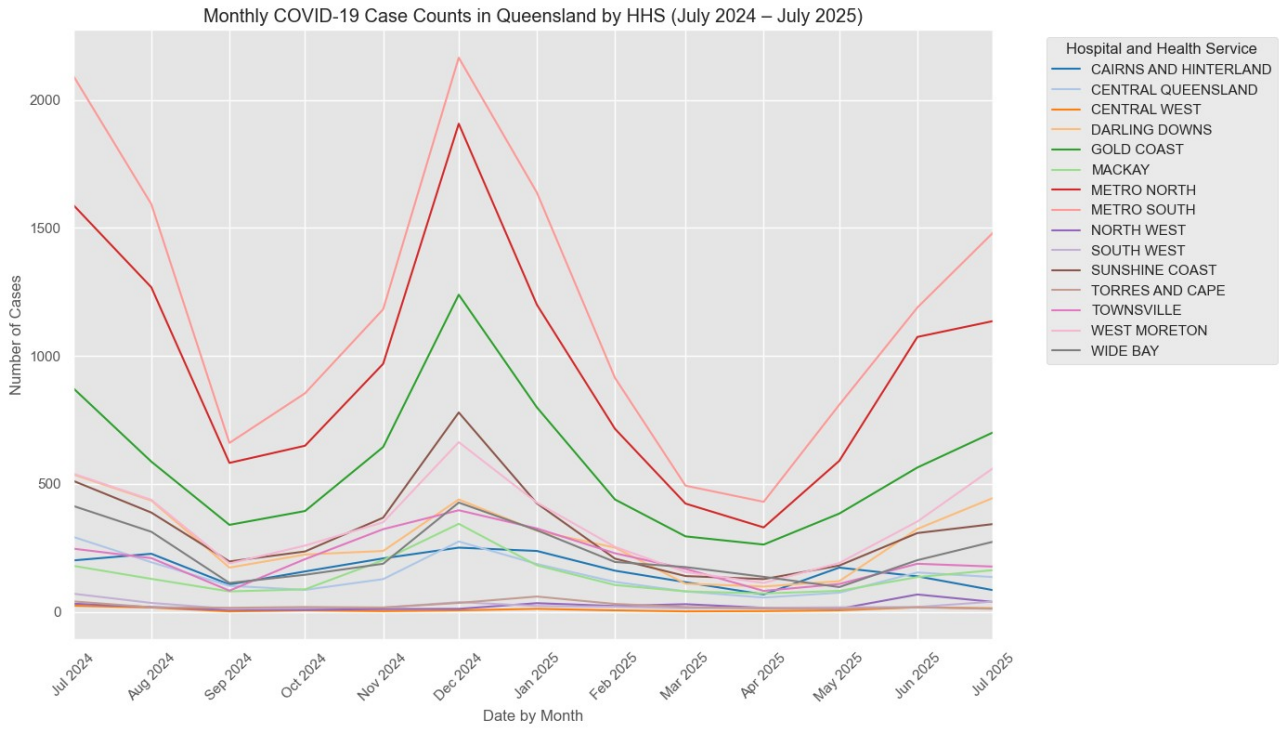

Figure 4: Monthly COVID-19 case counts across Queensland's HHS regions from July 2024 to July 2025, showing peaks in July and December 2024, with lows in September 2024 and April 2025.

c) Geographic Distribution

Figure 5 shows that Brisbane City had the highest case count, with over 475,000 cases, while the Gold Coast recorded just under half that amount, at 216,000. The remaining LGAs exhibited a gradual decline in case numbers, with percentage decreases between intervals ranging from approximately 2.4% to 29.3%, indicating a consistent downward trend across regions.

In Figure 6, during the 12-month period from July 2024 to July 2025, Metro North and Metro South together accounted for approximately 46.1% of total cases, while the remaining Hospital and Health Services (HHS) regions comprised the other 53.9%.

Figure 7 presents a heatmap illustrating the relationship between HHS and Local Government Areas (LGAs). To improve readability and focus on high-impact areas, the chart was filtered to include only the top ten LGAs. Most HHS regions align closely with specific LGAs, for example, Cairns and Hinterland corresponds to Cairns Regional. However, some HHS regions, such as Wide Bay, do not strongly correlate with a single dominant LGA, possibly indicating service to smaller or more dispersed communities.

Another key observation is that Metro North serves Brisbane City and Moreton Bay Regional, accounting for a combined 27.6% of cases. Metro South serves Brisbane City, Logan City, and Redland City, with a combined total of 29.8% of cases.

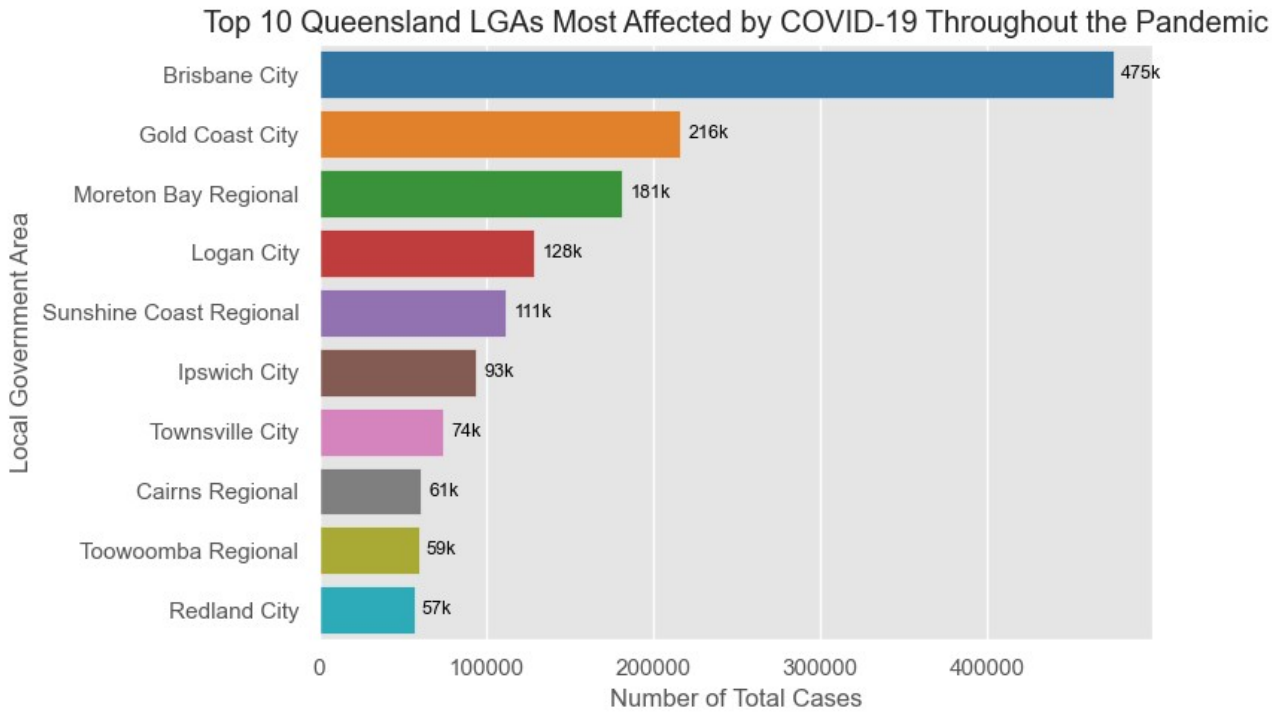

Figure 5: Total COVID-19 cases across the ten most affected LGAs in Queensland, with Brisbane City recording the highest count, followed by Gold Coast. Case counts decline steadily across remaining LGAs.

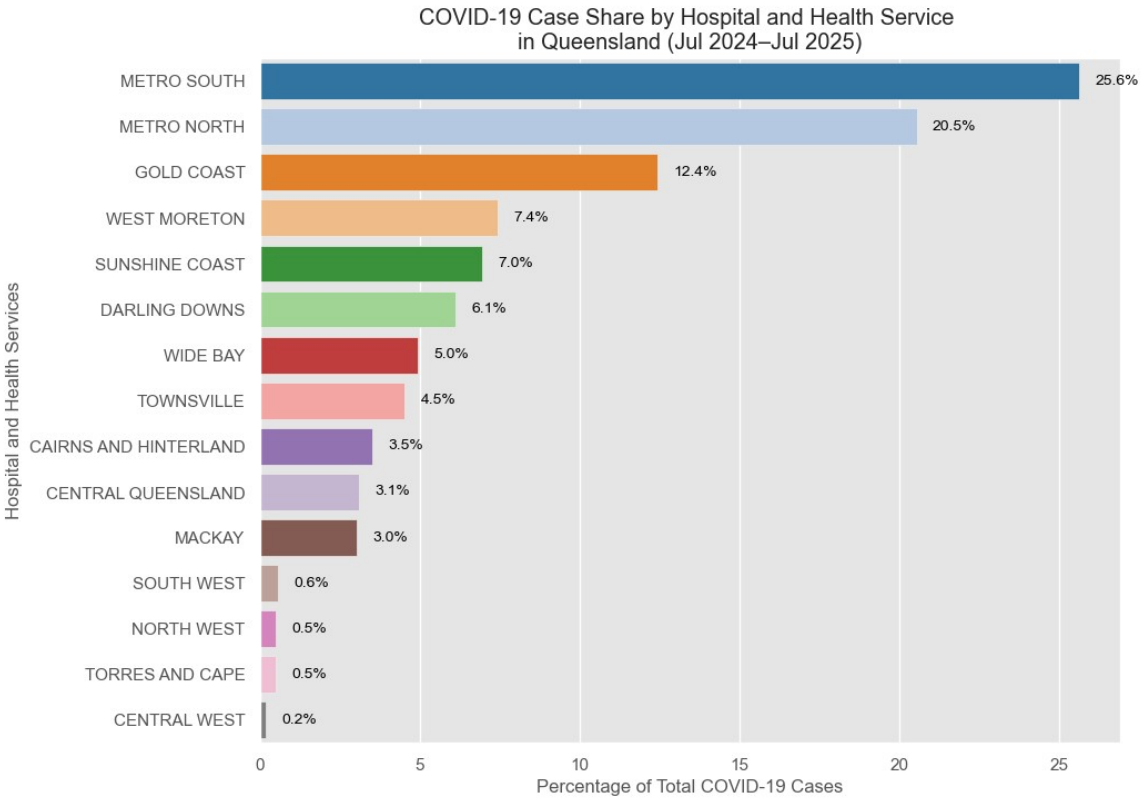

Figure 6: COVID-19 case share by Hospital and Health Service in Queensland between July 2024 and July 2025, showing Metro North and Metro South collectively accounting for 46.1% of all reported cases.

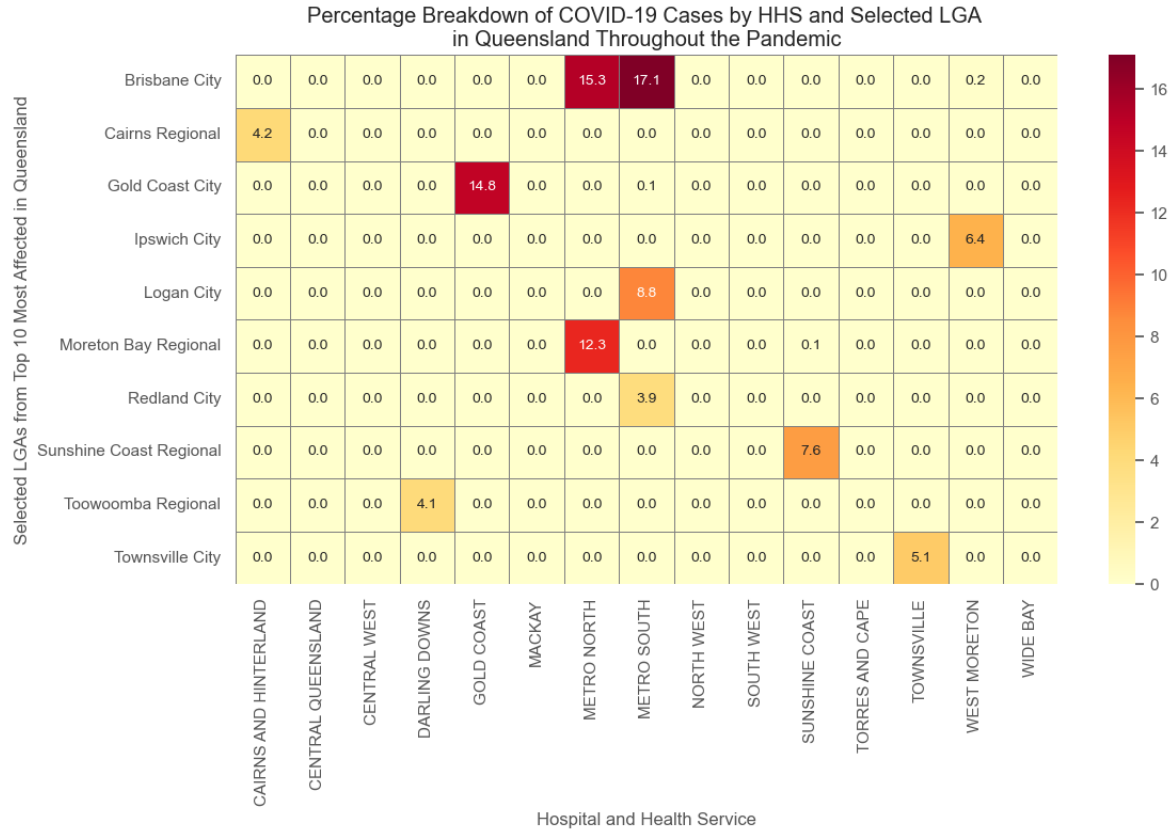

Figure 7: COVID-19 case percentages by HHS and top ten LGAs, highlighting alignment patterns and regional variation in service coverage.

d) Source Attribution

A high number of cases classified as “Under Investigation” was removed, as it significantly skewed the data. In addition, the “Not Applicable” category, which contained only 39 entries, was also excluded due to its negligible impact on the analysis.

Figure 8 presents a heatmap of HHS × SOURCE_INFECTION, revealing limitations in interpreting local versus overseas transmission trends. A significant number of locally acquired cases through contact with confirmed cases and/or known clusters were recorded in Cairns and Hinterland, totalling 12,236 cases. Thousands of confirmed cases were also identified in Torres and Cape, Townsville, and Wide Bay.

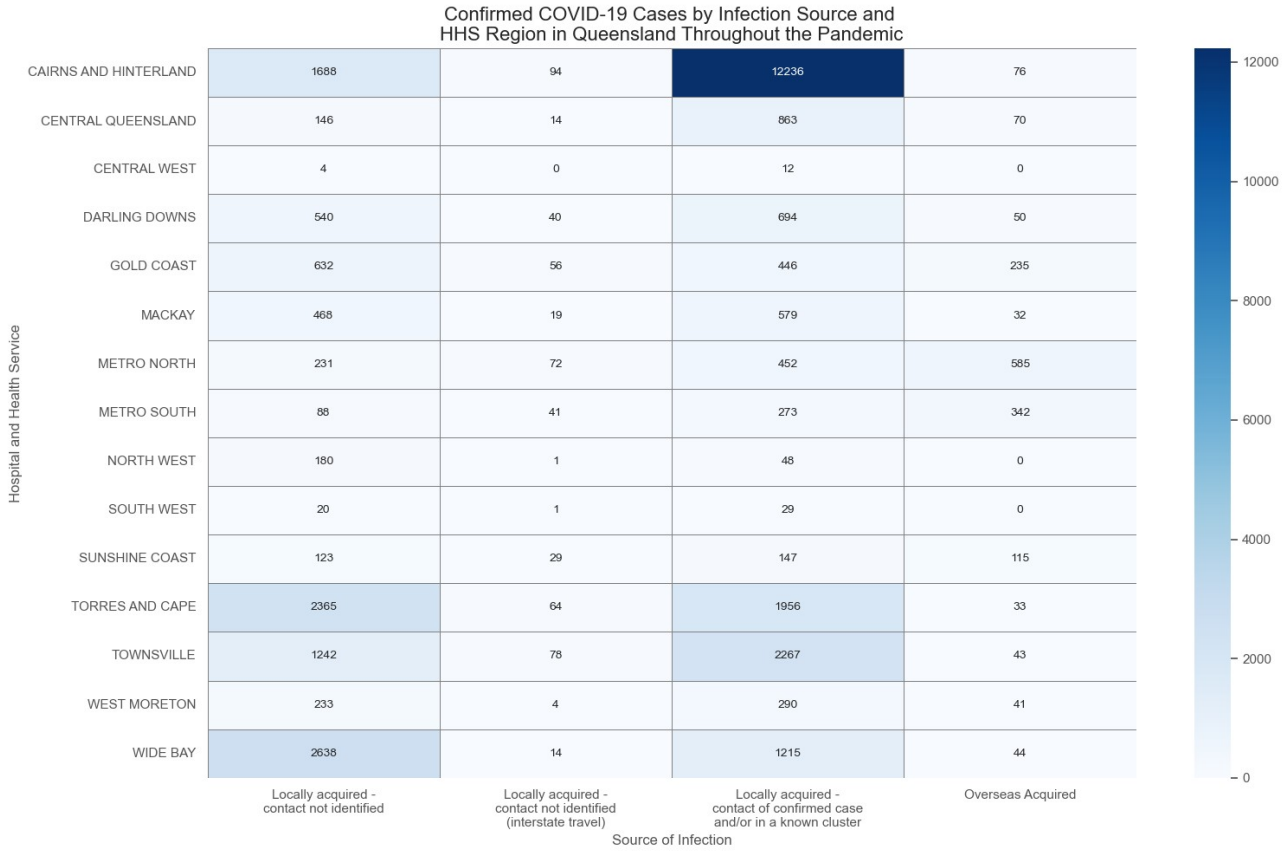

Figure 8: Heatmap of confirmed COVID-19 cases by infection source and HHS region in Queensland, excluding “Under Investigation” and “Not Applicable” cases. Locally acquired cases were highest in Cairns and Hinterland, with notable counts in Torres and Cape, Townsville, and Wide Bay.

7. Key Insights and Interpretation

This section identifies key insights from the charts and interprets potential causes or correlations behind observed impacts.

Demographic Shift:

While the 25-29 age group recorded the highest total case count throughout the pandemic (over 164,000 cases), more recent data indicates a shift in vulnerability. Young adults may have developed immunity through vaccination or prior infection, while children aged 0-4 (4,838 cases) and older adults aged 75-79 (4,994 cases) are now experiencing higher rates. This change may be due to localised clusters in settings such as kindergartens and aged care facilities, highlighting evolving risk profiles.

Seasonal Trends:

Following the decline and stabilisation of case numbers after January 2023, data from July 2024 to July 2025 shows a clear seasonal pattern. Peaks in mid-winter and early summer align with traditional flu seasons. This trend underscores the importance of seasonal public health interventions, including vaccination and public awareness campaigns.

Geographic Anomalies:

The mapping between Hospital and Health Services (HHS) and Local Government Areas (LGAs) is complicated by cross-boundary healthcare. Based on the data, the most widely utilised services throughout the pandemic were located in Brisbane. Metro North accounted for 27.6% of cases, and Metro South accounted for 29.8%, likely due to residents seeking care outside their residential HHS regions.

This trend remains consistent in the 12-month period from July 2024 to July 2025, with Metro North and Metro South together accounting for approximately 46.1% of total cases. Brisbane City alone reported over 475,000 cases across the pandemic, further highlighting the significance of these health regions.

Data Quality Issues in Source Attribution:

A large portion of cases (98.1%) were classified as “Under Investigation,” severely limiting the ability to draw conclusions about transmission pathways. However, among the subset of confirmed data, significant local transmission was observed, particularly in Cairns and Hinterland, likely linked to social gatherings or clusters in confined settings.

There is some indication that these confirmed clusters, combined with demographic vulnerability in the 0-4 and 75-79 age groups, may be correlated. However, due to the small sample size and limited source attribution, further investigation with more comprehensive data is necessary to validate this pattern.

8. Recommendations

Based on the analysis, the following recommendations may improve data quality and support better public health outcomes.

Data Improvement:

- Introduce a regional classification column to enhance spatial analysis and support clearer segmentation by area type.
- Include a shared unique identifier across both the Age groups and Location & Source of Infection tables to enable accurate joins without risk of misalignment.
- Encourage more complete case investigations to reduce the volume of “Under Investigation” records.

Public Health Action:

- Increase public health messaging and hygiene campaigns during seasonal high-risk periods (April-June and October-December).
- Launch targeted vaccination efforts ahead of peak seasons, with special focus on vulnerable groups such as newborns and the elderly.
- Promote workplace health accommodations in densely populated areas, particularly in Metro North and Metro South.

Further Analysis:

- Explore cross-HHS healthcare usage to better understand patient movement and service delivery beyond residential boundaries.
- Investigate potential correlations between age groups, locations, and source of infection to uncover deeper demographic insights.

9. Conclusion

This report highlights the evolving dynamics of COVID-19 in Queensland, including notable shifts in demographic vulnerability, particularly among younger children and older adults, and consistent seasonal patterns that mirror traditional flu activity.

Regional analysis reveals significant variation across Hospital and Health Services, influenced in part by cross-boundary healthcare usage and data inconsistencies in LGA attribution. While the dataset provides meaningful insights, limitations such as the high volume of “Under Investigation” cases hinder a complete understanding of transmission pathways.

To address these challenges, improvements in data completeness and structure such as urban/rural classification and clearer source attribution are essential. Targeted public health interventions during seasonal peaks, along with strategic resource planning in high-density areas like Metro North and Metro South, will support a more effective response. Continued analysis, including predictive modelling, will be critical to anticipating and managing future outbreaks with greater precision.

10. References

- Queensland Government Open Data (2025, August 6). *Queensland COVID-19 Case Line List – Age Groups*. Retrieved from: https://www.data.qld.gov.au/dataset/queensland-covid-19-case-line-list-age-groups
- Queensland Government Open Data (2025, August 6). *Queensland COVID-19 Case Line List – Location & Source of Infection*. Retrieved from: https://www.data.qld.gov.au/dataset/queensland-covid-19-case-line-list-location-source-of-infection

11. Appendices

All code, charts, and detailed analysis steps are documented in the accompanying Jupyter Notebook. You can view or download it here:

Jupyter Notebook: https://github.com/gabywu/Queensland-Covid-Case/blob/main/qld-covid19-2025.ipynb

```
#Import Libraries
import pandas as pd
import numpy as np

import seaborn as sns
sns.set(style="whitegrid")

import matplotlib
import matplotlib.dates as mdates
import matplotlib.pyplot as plt
plt.style.use('ggplot')
from matplotlib.pyplot import figure
from matplotlib.colors import LinearSegmentedColormap

%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8)

# Read in the Data (Downloaded from 10/08/2025)
# File path omitted for privacy/security
df_age = pd.read_csv(r'...\opendata_qld_covidcase_agegrp.csv')
# SA2_CODE contains both string and numeral, which was set to float64
df_loc = pd.read_csv(r'...\opendata_qld_covidcase_loc.csv',
    dtype={'SA2_CODE': 'float64'}, low_memory=False
)
```

Analyse Table Contents

```
df_age.head()

df_loc.head()

df_age.dtypes

df_loc.dtypes
```

```
# Define and call function to check missing values in age groups table
missing_df_age = df_age.isnull().sum()
null_percent_df_age = df_age.isnull().mean() * 100
def print_missing_info(title, df):
    print(title)
    print(df)
    print()
print_missing_info("Number of Missing Nulls in Age Groups Table", missing_df_age)
print_missing_info("Number of Missing Nulls in Age Group Table",
                   null_percent_df_age.apply(lambda x: f"{x:.6f}%"))
```

```
# Define and call function to check missing values in location table
missing_df_loc = df_loc.isnull().sum()
null_percent_df_loc = (df_loc.isnull().mean() * 100).round(2)
```

```
def print_missing_info(title, df):
    print(title)
    print(df)
    print()

print_missing_info("Number of Missing Nulls in Location & Source of Infection Table",
missing_df_loc)
print_missing_info("Percentage of Missing Nulls in Location & Source of Infection Table",
                   null_percent_df_loc.apply(lambda x: f"{x}%"))


# list all unique names while excluding missing values using dropna()
def print_distinct_values(df, columns, label):
    for col in columns:
        distinct_values = df[col].dropna().unique()
        print(f"{label} Column: {col}")
        print(f"Distinct names ({len(distinct_values)}): {distinct_values}\n")

# For df_age (all columns)
print_distinct_values(df_age, df_age.columns, "Age Groups")

# For df_loc (specific columns)
columns_to_count = ['HHS', 'LGA_NAME', 'SOURCE_INFECTION']
print_distinct_values(df_loc, columns_to_count, "Location Infection")


# Count the occurrences of each unique name and calculate its percentage of the total
def print_column_value_percents(df, columns, label, decimals=1, include_na=False):
    total = len(df)
    print(f"\nCounts and Percentages for {label} Table (base = {total:,} rows)")
    for col in columns:
        print(f"\nColumn: {col}")
        # Include or drop NaN as a category
        counts = df[col].value_counts(dropna=not include_na)
        # header row
        print(f"{'Value':<70}{'Count':>12}{'Percent':>12}")
        print("-" * 94)
        for value, count in counts.items():
            pct = (count / total) * 100 if total else 0.0
            name = "NA" if pd.isna(value) else str(value)
            # Print row with fixed-width columns for clear visual alignment
            print(f"{name:<70}{count:>12,}{pct:>11.{decimals}f}%")
        # If excluding NaNs, still show how many are missing
        if not include_na:
            missing = df[col].isna().sum()
            if missing:
                mp = (missing / total) * 100
                print(f"{'Missing':<70}{missing:>12,}{mp:>11.{decimals}f}%")
# Age Groups
print_column_value_percents(df_age, ['AGE_GROUP_5Y'], 'Age Groups')

# Location & Source of Infection
columns_to_count = ['HHS', 'LGA_NAME', 'SOURCE_INFECTION']
print_column_value_percents(df_loc, columns_to_count, 'Location & Source of Infection')


# Check for inconsistencies between HHS and LGA
df_loc[['HHS', 'LGA_NAME']].value_counts().reset_index(name='count').sort_values(by='count',
ascending=False)


# Search for unusual combinations
# Example: cases where Gold Coast LGA appears with non-Gold Coast HHS
df_loc[(df_loc['LGA_NAME'] == 'Brisbane City') &
    (~df_loc['HHS'].isin(['METRO NORTH', 'METRO SOUTH']))
]


# Count distinct combinations between HHS and LGA using a function
def group_and_count_unique(df, group_col, count_col, label=None):
    result = df.groupby(group_col)[count_col].nunique().sort_values(ascending=False)
    if label:
        print(f"\n{label}")
    print(result)

# Count distinct HHSs per LGA
group_and_count_unique(df_loc, 'LGA_NAME', 'HHS', label="Unique HHSs per LGA")
```

```
# Count distinct LGAs per HHS
group_and_count_unique(df_loc, 'HHS', 'LGA_NAME', label="Unique LGAs per HHS")
```

Cleaning Data

```
# Drop columns that aren't needed for this analysis
df_loc = df_loc.drop(columns=['_id', 'POSTCODE', 'SA2_CODE', 'SA2_REGION'])
df_age = df_age.drop(columns=['_id'])


# Remove "Years" in the age group by applying regex
df_age['AGE_GROUP_5Y'] = df_age['AGE_GROUP_5Y'].str.replace(r'\s*years$', '', regex=True)


# Convert NOTIFICATION_DATE to datetime and print dtype with table name
for table_name, df in [('df_age', df_age), ('df_loc', df_loc)]:
    df['NOTIFICATION_DATE'] = pd.to_datetime(df['NOTIFICATION_DATE'], dayfirst=True)
    print(f"[{table_name}] dtype: {df['NOTIFICATION_DATE'].dtype}")


# Review the table after it has been cleaned
df_loc
df_age
```

Exploratory Data Analysis

```
# Displaying age group distribution

# Predefined age order
age_order = ['00-04', '05-09', '10-14', '15-19', '20-24', '25-29', '30-34', '35-39',
    '40-44', '45-49', '50-54', '55-59', '60-64', '65-69', '70-74', '75-79',
    '80-84', '85-89', '90-94', '95-99', '100+']
age_order.reverse()

# Set age group as a categorical with custom age order
df_age['AGE_GROUP_5Y'] = pd.Categorical(df_age['AGE_GROUP_5Y'], categories=age_order, ordered=True)

# Count cases for each age group
counts = df_age['AGE_GROUP_5Y'].value_counts().reindex(age_order)

# Create gradient colors
cmap = plt.get_cmap("inferno")
colours = cmap(np.linspace(0, 1, len(counts)))

# Function to format numbers
def format_number(val):
    if val >= 1000:
        return f"{int(val // 1000)}k"
    else:
        return str(val)

plt.figure(figsize=(8, 10))
bars = plt.barh(age_order, counts.values, color=colours)

# Add value labels at the end of each bar
for bar, value in zip(bars, counts.values):
    plt.text(
        bar.get_width() + max(counts.values)*0.01,
        bar.get_y() + bar.get_height()/2,
        format_number(value), va='center',
        ha='left', fontsize=9, color='black'
    )

plt.title("Demographic Trends in COVID-19 Case Counts: \nQueensland Age Group Breakdown")
plt.xlabel('Number of Total Cases')
plt.ylabel('Age Group (5-Year Intervals)')
plt.tight_layout()
plt.show()


# Displaying age group distribution after filtering by date
```

```
# Filter data between July 2024 and July 2025
start_date = pd.Timestamp('2024-07-01')
end_date = pd.Timestamp('2025-07-31')
df_filtered_date = df_age[(df_age['NOTIFICATION_DATE'] >= start_date) &
                          (df_age['NOTIFICATION_DATE'] <= end_date)].copy()

# Define and reverse age order
age_order = ['00-04', '05-09', '10-14', '15-19', '20-24', '25-29', '30-34', '35-39',
             '40-44', '45-49', '50-54', '55-59', '60-64', '65-69', '70-74', '75-79',
             '80-84', '85-89', '90-94', '95-99', '100+']
age_order.reverse()

# Set age group as categorical with custom order
df_filtered_date['AGE_GROUP_5Y'] = pd.Categorical(df_filtered_date['AGE_GROUP_5Y'],
                                                  categories=age_order, ordered=True)

# Count cases for each age group
counts = df_filtered_date['AGE_GROUP_5Y'].value_counts().reindex(age_order)

# Create gradient colors
cmap = plt.get_cmap("plasma")
colours = cmap(np.linspace(0, 1, len(counts)))

# Function to format with comma as thousands separator
def format_number(val):
    return f"{val:,}"

# Plot horizontal bar chart
plt.figure(figsize=(8, 10))
bars = plt.barh(age_order, counts.values, color=colours)

# Add value labels at the end of each bar
for bar, value in zip(bars, counts.values):
    plt.text(
        bar.get_width() + max(counts.values)*0.01,
        bar.get_y() + bar.get_height()/2,
        format_number(value), va='center',
        ha='left', fontsize=9, color='black'
    )

plt.title("COVID-19 Case Distribution by Age Group \nin Queensland (Jul 2024 – Jul 2025)")
plt.xlabel("Number of Cases")
plt.ylabel("Age Group (5-Year Intervals)")
plt.tight_layout()
plt.show()
```

 # Generate case timeline from the start of the pandemic

```
# Extract only the date and group by month for smoother trends
df_loc['MONTHLY_DATES'] = df_loc['NOTIFICATION_DATE'].dt.to_period('M').dt.to_timestamp()

# Group by MONTHLY_DATES
timeline_data = df_loc.groupby('MONTHLY_DATES').size().reset_index(name='Cases')

plt.figure(figsize=(14, 8))
sns.lineplot(data=timeline_data, x='MONTHLY_DATES', y='Cases')

# Format x-axis to show selected months
plt.gca().xaxis.set_major_locator(mdates.MonthLocator(bymonth=[1, 4, 7, 10]))  # Jan, Apr, Jul, Oct
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%b %Y'))  # e.g., Jan 2021

plt.title('Queensland COVID-19 Case Timeline Throughout the Pandemic')
plt.xlabel('Date (Quarterly)')
plt.ylabel('Number of Cases')
plt.xticks(rotation=45)
# Removing the empty dates that was created from mdates
plt.xlim(timeline_data['MONTHLY_DATES'].min(), timeline_data['MONTHLY_DATES'].max())
plt.tight_layout()
plt.show()
```

 # Generate monthly cases from the start the last 12 month

```
# Convert NOTIFICATION_DATE to monthly period, then back to timestamp (first day of month)
```

```
df_loc['MONTHLY_DATES'] = df_loc['NOTIFICATION_DATE'].dt.to_period('M').dt.to_timestamp()

# Filter data between July 2024 and July 2025
start_date = pd.Timestamp('2024-07-01')
end_date = pd.Timestamp('2025-07-01')
df_filtered_date = df_loc[(df_loc['MONTHLY_DATES'] >= start_date) & (df_loc['MONTHLY_DATES'] <=
end_date)]

# Group by DATE and HHS
timeline_data = df_filtered_date.groupby(['MONTHLY_DATES', 'HHS']).size().reset_index(name='Cases')

# Plotting
plt.figure(figsize=(14, 8))
sns.lineplot(data=timeline_data, x='MONTHLY_DATES', y='Cases', hue='HHS', palette='tab20')

# Set monthly ticks (show every month)
plt.gca().xaxis.set_major_locator(mdates.MonthLocator(interval=1))
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%b %Y'))

plt.title('Monthly COVID-19 Case Counts in Queensland by HHS (July 2024 – July 2025)')
plt.xlabel('Date by Month')
plt.ylabel('Number of Cases')
plt.legend(title='Hospital and Health Service', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=45)
plt.xlim(start_date, end_date)  # Ensures only the selected timeline is shown
plt.tight_layout()
plt.show()
```

 # Displaying COVID-19 case counts broken down by LGA

```
top_lga = df_loc['LGA_NAME'].value_counts().head(10)

# Function to format numbers
def format_number(val):
    if val < 1000:
        return str(val)
    elif val < 100000:
        return f"{int(val // 1000)}k"
    else:
        return f"{int(val // 1000):,}k"

fig, ax = plt.subplots(figsize=(8, 5))
ax = sns.barplot(x=top_lga.values, y=top_lga.index, hue=top_lga.index,
    palette='tab10', legend=False, ax=ax)

ax.set_title("Top 10 Queensland LGAs Most Affected by COVID-19 Throughout the Pandemic")
ax.set_xlabel("Number of Total Cases")
ax.set_ylabel("Local Government Area")

# Add value labels at the end of each bar
for patch, value in zip(ax.patches, top_lga.values):
    x = patch.get_width()
    y = patch.get_y() + patch.get_height() / 2
    ax.text(x + max(top_lga.values)*0.01,  # small padding to the right
        y, format_number(value), va='center',
        ha='left', fontsize=9, color='black')
plt.tight_layout()
plt.show()
```

 # Showing case percentages per HHS after applying date filter

```
# Filter data between July 2024 and July 2025
start_date = pd.Timestamp('2024-07-01')
end_date = pd.Timestamp('2025-07-31')
df_filtered_date = df_loc[(df_loc['MONTHLY_DATES'] >= start_date) &
                          (df_loc['MONTHLY_DATES'] <= end_date)]

# Count occurrences of each HHS
hhs_counts = df_filtered_date['HHS'].value_counts()

# Convert counts to percentages
hhs_percentages = (hhs_counts / hhs_counts.sum()) * 100
```

```
# Sort for better visualisation
hhs_percentages = hhs_percentages.sort_values(ascending=False)

# Plot the percentages
plt.figure(figsize=(10, 8))
ax = sns.barplot(x=hhs_percentages.values, y=hhs_percentages.index,
                 hue=hhs_percentages.index, palette='tab20',
                 dodge=False, legend=False)
# Add percentage labels to the end of each bar
for i, (value, name) in enumerate(zip(hhs_percentages.values,
                                      hhs_percentages.index)):
    ax.text(value + 0.5, i, f"{value:.1f}%",
            va='center', ha='left',
            fontsize=10, color='black')
plt.title("COVID-19 Case Share by Hospital and Health Service \nin Queensland (Jul 2024–Jul 2025)")
plt.xlabel("Percentage of Total COVID-19 Cases")
plt.ylabel("Hospital and Health Services")
plt.show()
```

 # Visualising case concentration using a heatmap of HHS vs selected LGAs

```
# Define the LGAs you want to include
selected_lgas = ['Brisbane City', 'Gold Coast City', 'Moreton Bay Regional', 'Logan City',
    'Sunshine Coast Regional', 'Ipswich City', 'Townsville City',
    'Cairns Regional', 'Toowoomba Regional', 'Redland City']

# Create the counted DataFrame
df_counts = (df_loc[['HHS', 'LGA_NAME']].value_counts()
             .reset_index(name='count').sort_values(by='count', ascending=False))

# Filter to only selected LGAs
df_filtered_lgas = df_counts[df_counts['LGA_NAME'].isin(selected_lgas)]

# Pivot the data for heatmap
heatmap_data = df_filtered_lgas.pivot(index='LGA_NAME', columns='HHS', values='count').fillna(0)

# Convert counts to percentage of total
heatmap_percent = (heatmap_data / heatmap_data.values.sum()) * 100

# Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(heatmap_percent, annot=True, fmt='.1f',
            cmap='YlOrRd', linewidths=0.5, linecolor='gray')
plt.title('Percentage Breakdown of COVID-19 Cases by HHS and Selected LGA \nin Queensland Throughout
the Pandemic')
plt.xlabel('Hospital and Health Service')
plt.ylabel('Selected LGAs from Top 10 Most Affected in Queensland')
plt.tight_layout()
plt.show()
```

 # Visualising COVID-19 case origins by HHS and source of infection

```
# Create a new column with line-broken versions of long SOURCE_INFECTION values
df_loc['SOURCE_INFECTION_BREAK_LINE'] = df_loc['SOURCE_INFECTION'].replace({
    'Locally acquired - contact of confirmed case and/or in a known cluster':
    'Locally acquired -\ncontact of confirmed case\nand/or in a known cluster',
    'Locally acquired - contact not identified':
    'Locally acquired -\ncontact not identified',
    'Locally acquired - contact not identified interstate travel':
    'Locally acquired -\ncontact not identified\n(interstate travel)'
})

# Removing "Under Investigation" and "Not Applicable" data
filtered_df = df_loc[~df_loc['SOURCE_INFECTION_BREAK_LINE'].isin(["Under Investigation", "Not
Applicable"])]

# Grouping the filtered data
heatmap_data = filtered_df.groupby(['HHS',
'SOURCE_INFECTION_BREAK_LINE']).size().reset_index(name='Cases')

# Pivot the data: rows = HHS, columns = SOURCE_INFECTION
heatmap_matrix = heatmap_data.pivot(index='HHS', columns='SOURCE_INFECTION_BREAK_LINE',
values='Cases').fillna(0)
```

```
plt.figure(figsize=(16, 10))
sns.heatmap(heatmap_matrix, cmap='Blues', annot=True,
            fmt='g', linewidths=0.5, linecolor='gray')

plt.title('Confirmed COVID-19 Cases by Infection Source and \nHHS Region in Queensland Throughout
the Pandemic',
          fontsize=16)
plt.xlabel('Source of Infection')
plt.ylabel('Hospital and Health Service')
plt.xticks(rotation=0, ha='center')
plt.tight_layout()
plt.show()
```