SQL Data Cleaning and Analysis

Overview of Data Analysis Projects

This report presents two comprehensive projects that leverage advanced SQL techniques to clean and analyse large datasets, demonstrating the critical role of structured data in extracting meaningful insights.

Nashville Housing Data Cleaning

The Nashville Housing Data Cleaning project focused on addressing common data quality issues often encountered in real estate datasets. The dataset initially contained missing addresses, inconsistent date formats, and duplicate records. To resolve these issues, various SQL techniques were applied, including data conversion, joins, and window functions, to standardise and structure the dataset. By cleansing the data, the project ensured that the information was accurate and reliable for subsequent analysis. This work not only improved the quality of the dataset but also enhanced the ability to generate actionable insights on housing trends in the Nashville area, supporting decision-making processes for real estate stakeholders.

COVID-19 Data Analysis

The COVID-19 Data Analysis project involved extracting and analysing key global metrics, such as infection rates, death rates, and vaccination progress, from comprehensive pandemic datasets. Using SQL queries, the project aimed to identify patterns and trends across various countries, focusing on the impact of COVID-19 and vaccination efforts on public health. One of the most notable findings was the identification of countries with the highest COVID-19 infection rates. For example, the UK exhibited a significant infection rate of 12% of the total population, underscoring the severity of the pandemic in certain regions.

This analysis revealed important disparities between countries in terms of both infection and death rates, as well as vaccination coverage. Countries with higher vaccination rates generally experienced lower death rates, illustrating the critical role that vaccinations play in controlling the spread and impact of the virus. These insights can inform future public health strategies, emphasising the importance of timely interventions and equitable vaccine distribution across regions.

Conclusion

Both projects present the value of clean, structured data and the ability of SQL techniques to address data quality challenges while facilitating in-depth analysis. The Nashville Housing Data Cleaning project provided a reliable dataset for real estate insights, while the COVID-19 Data Analysis project offered valuable information on global pandemic trends. Together, these projects highlight the power of data-driven decision-making, emphasising the need for robust data cleaning methods and the application of advanced analytical techniques in solving real-world problems.

Django Project Overview: notapp

Overview of notapp Project

The notapp project is a simple Django application designed to demonstrate core functionalities such as database modeling, URL routing, dynamic templates, and admin panel integration. It features a Note model with fields for a title, content, and a creation timestamp, which are stored in a SQLite database. The app's home page dynamically displays all notes using Django's templating system, while the admin panel provides an interface for managing notes (add, edit, delete).

The project structure follows Django's modular design, with separate files for models, views, templates, and URLs. The application is easy to set up and run, requiring minimal configuration. It serves as a foundation for learning Django or building more complex applications, with potential for enhancements like user authentication, note editing via the web interface, and improved styling.

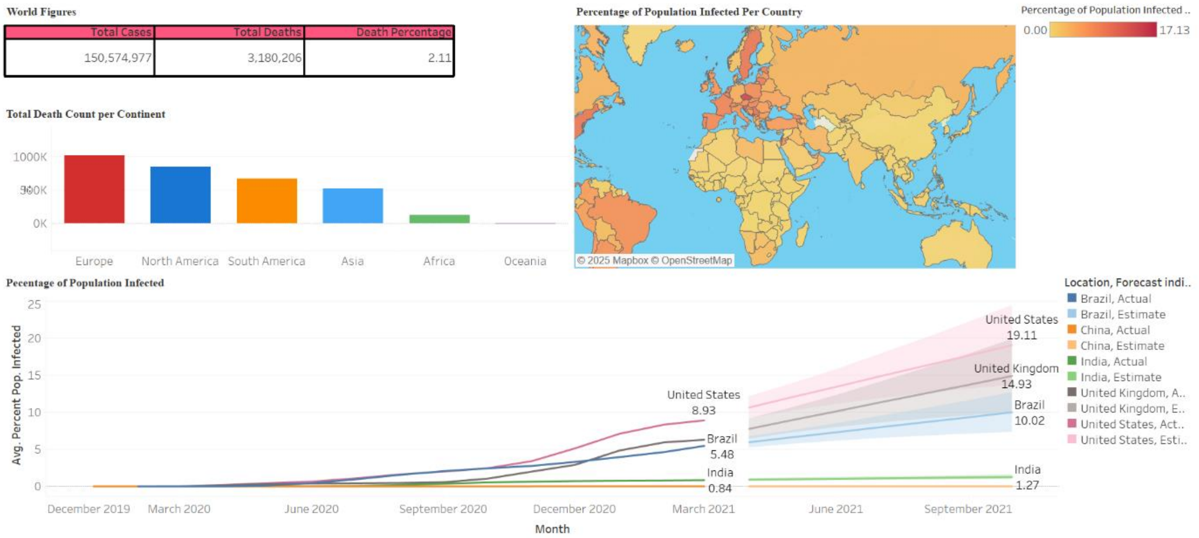
Visualising COVID-19 Trends with Tableau

Overview of Visualisation Project

This project demonstrates the use of Tableau Public to visualise global COVID-19 trends, showcasing insights derived from SQL-prepared data. As Tableau Public does not support direct SQL integration, the data was exported into Excel files before being visualised.

The project featured the following dashboards, designed to explore key pandemic metrics:

- Global COVID-19 Overview: A table showing the total number of COVID-19 cases, total deaths, and the overall global death rate, which was calculated at 2.11%.
- Total Death Count by Continent: A ranked bar chart of total death counts across continents, highlighting Europe as the hardest-hit region.
- Percentage of Population Infected (Choropleth Map): A colour-coded map displaying infection percentages by country, with the Czech Republic standing out at 15.33%— one of the highest recorded percentages globally.
- Predicted Infection Rates (Line Graph): A time-series graph projecting the percentage of the population infected. The model accurately forecasted a 19.11% infection rate for the United States in April 2021, reinforcing the reliability of trend-based predictions.



World Figures

Total Cases	Total Deaths	Death Percentage
150,574,977	3,180,206	2.11

Figure 1: Figure shows the tableau dashboard created presenting key figures such as death counts, the percentage of population infected and predicting future infection rates.

Conclusion

By analysing COVID-19 data, including peak infection periods, we generated accurate forecasts for countries such as Brazil, where the predicted infection rate of 10.02% aligned with reported figures. The project highlights the application of SQL hand-in hand with Tableau and how powerful the combination can be in data.