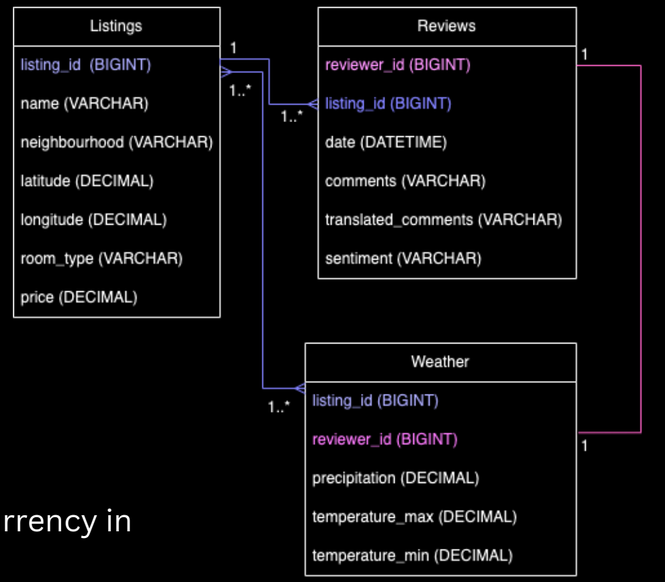
Deloitte. Data Engineer Case Study

Objective

The objective of this case study is to collect data and come up with insights regarding Airbnb listings in Cape Town, and identify if there is any correlation between ratings and the weather.

Solution Architecture

Data Model



Steps followed

1. Download Cape Town, December 2023 Reviews and listings data from Inside Airbnb website
2. Import data in Python notebook
3. Clean data
4. Transform data
5. Analyse reviews and determine sentiment (good/neutral/bad)
6. Extract Weather data from Open-Meteo API
7. Design data model
8. Load data into MySQL
9. Visualise insights using Power BI

Challenges

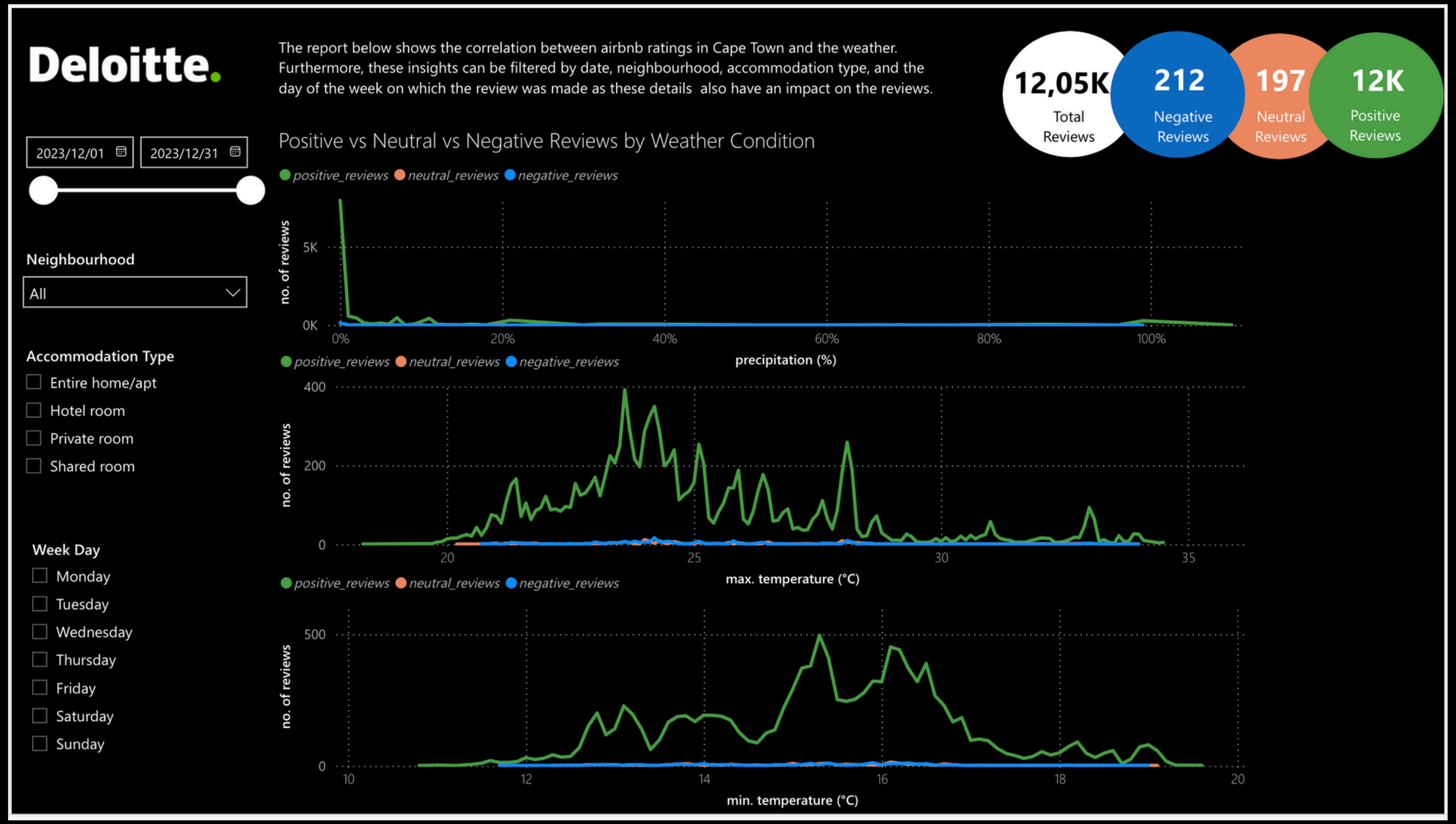
- Large dataset caused efficiency issues during processing
- Proposed weather API (Open Weather) has limits on no. of daily API calls on the free tier
- Review data was not numeric or in a standardized format
- Some reviews were not written in English

Solutions

- Implemented concurrency in Python
- Researched alternatives and used Open-Meteo API
- Used VADER Sentiment Analyzer library to analyse and group reviews into categorical data (good/neutral/bad)
- Used Google translate API to translate non-English comments before analysis

Dashboard

Link to Power BI Dashboard for dynamic interaction with filters: [Power BI Dashboard]



Future improvements

To measure the extent of the correlation statistically, we can calculate the correlation co-efficient between the number of total reviews, number of positive/neutral/negative reviews and the weather data, and other categorical features of the listings data which can be encoded. These numbers can be visualised using a correlation matrix.

Conclusion

The general pattern between Airbnb reviews in Cape Town and the weather in December 2023. is that the no. of negative and neutral reviews are very low irrespective of weathe. There is a high number of positive reviews and there is a positive spike in the no. of positive reviews when precipitation is close to 0%.

The no. of positive reviews tends to increase as the max. temperature of the day increases to warmer temperatures (between 20-25 degrees Celcius), but then decreases significantly as temperature gets extremely hot.

It is also interesting to note how the neighbourhood, accommodation type and the day of the week on which the review was made impact the no. of reviews. Reviews made on Mondays and Tuesdays tend to be more positive while reviews made during the week are more neutral.

Furthermore, reviews for hotel rooms are mostly positive irrespective of weather, and some neighbourhoods receive more positive reviews than others irrespective of weather.

See link to live Power BI Dashboard for dynamic filtering of the visuals based on date, day of the week, neighbourhood and accommodation type.