Postal Codes of All Airports in the World

I. Introduction

a. Purpose

This project intends to solve one of Anywhr's data engineering challenge problems, which is to determine the postal codes of airports in the whole world.

b. Duration

The duration of the project is from 26 February to 28 February, within a 48-hour constraint.

II. Methodology

The problem is to be solved through Python programming, specifically through a Python web crawler program and API calls.

a. Crawler program

HTTP requests will be made to the website https://www.world-airport-codes.com/, where the airport names, cities, countries, and geolocations of all airports in the world are stored. The crawler program is to be coded using the BeautifulSoup Python library, which helps to parse and scrape an HTML webpage. This crawler program will retrieve the above mentioned four types of information about airports in the world.

b. Google Geocoding API

Geolocations are to be passed to the Google Geocoding API, which helps to determine postal codes based on latitudes and longitudes. If this API cannot determine the postal code, another API will be called.

c. Nominatim API

If Google's API cannot determine a postal code, then Nominatim API will be called via Python's GeoPy library to determine the postal code. If the postal code still cannot be identified from this API, the program should give up and place a 'None' in the postal code field.

III. Possible Reasons for Failures to Find Postal Codes

It may be possible that for some airports, postal codes cannot be determined eventually. Three reasons may account for this failure. First, the country that the airport belongs to does not use postal codes at all. Second, the postal code of the airport is not stored in the database of either Google Maps or Nominatim. Third, the geolocation obtained for the airport lacks accuracy.

IV. Expected Outcome

The final result is to be saved in a CSV file, containing the following fields: airport name, city, country, and postal code.