Report - City of Cape Town - Data Science Unit Code Challenge

Data Preparation

➢ Please ensure that the file sr_hex.csv is in the same directory as the script ETL_city.py

1. Data extraction

➢ Having issues with the S3 SELECT command – returns error ClientError: An error occurred (OverMaxRecordSize) when calling the SelectObjectContent operation: The character number in one record is more than our max threshold, maxCharsPerRecord: 1,048,576. Retried multiple times with no success. Resorted to using s3.get_object. Please see code used initially for S3 SELECT script used - commented out at the bottom of the script

2. Initial Data Transformation

➢ I chose the threshold of 212364 based on records were index == 0. This is essentially based on the fact that there were no values for Latitude and Longitude

3. Further Data Transformations (if applying for a Data Engineering Position)

➢ Centroid calculated by using geometry.centroid.values[0] – used by the shapely module (object.centroid which returns a representation of the object's geometric centroid (point).

➢ Please note that you need to install package odfpy (pip install odfpy) to open this file Wind_direction_and_speed_2020.ods. This is available natively in pandas 0.25

➢ Removed the following columns:

• reference_number

• directorate

• department

• branch

• section

• code_group

• code

• cause_code_group

• cause_code

• official_suburb

➢ Fields such as reference_number and official suburb may be linked to data that may contain personal information such as names, phone numbers, and addresses. Anonymizing this data helps protect the privacy of residents and reduces the risk of identity theft or other forms of privacy violation. Furthermore, service request data can reveal patterns of service requests across different

neighborhoods or demographic groups. Anonymizing this data can help prevent discrimination against certain groups or areas based on their service request history. Anonymizing service request data can help build trust between residents and government agencies by providing transparency in the handling of the data. Residents are more likely to trust that their data is being used in a responsible and ethical manner when they know it is being anonymized. Many countries have data protection regulations that require the anonymization of personal data. Anonymizing service request data helps ensure compliance with these regulations and avoids potential legal consequences. Anonymizing service request data can make it easier to share the data with other organizations for research or policy analysis purposes. Anonymized data is less sensitive and can be shared more widely without compromising individual privacy.