ETL Project

**Team**: Hephzibah and Shibani

**Topic**: ETL queries to understand the infrastructure development and greenhouse gases in Ontario

**Data Sources**:

1. <https://data.ontario.ca/dataset/ontario-builds-key-infrastructure-projects/resource/35dc5416-2b86-4a79-b3e6-acbfe004c81a>
2. <http://data.ec.gc.ca/data/substances/monitor/greenhouse-gas-reporting-program-ghgrp-facility-greenhouse-gas-ghg-data/>

**Data Formats for Extraction**:

* Infrastructure Development in **csv**
* Greenhouse Emissions in **csv**

**Data Transformation**:

Transformed the raw data in the following ways:

* Filtered by province
* Renamed the columns
* Added columns with default values
* Dropped the null values
* Used count function to determine how much of data was lost
* Removed the duplicate values
* Rounded the latitude, longitude and total emissions to 2-digits after the decimals
* Joined the tables in SQL to display relevant information using inner join
* Set the indices

**Data Loading**:

* PostgreSQL database

**Flask:**

* Flask Application to demonstrate the final view

**Challenges Faced:**

* Syntax limitation in the SQL view join
* Displaying the transformed data in Flask in a table format. The challenge was to loop over the results in the render\_template