**National Pollutant Inventory**

A factory with smoke coming out of it

Description automatically generated with low confidenceMap

Description automatically generated

**Technical Report**

|  |
| --- |
| **1.) NPI capstone data analysis using SQL and Python** |

**Excel**

* Creating and renaming the columns to Data Dictionary

**Python**

* Importing all required libraries
* Make connection with PostgreSQL through jupyter notebook



* Uploading all the Nine CSV files to PostgreSQL

1. anzsic\_2006.csv
2. emission\_reduction\_techniques.csv
3. emissions.csv
4. facility\_secondary\_anzsics.csv
5. facilities.csv
6. reports.csv
7. substances.csv
8. transfers.csv
9. transfer\_destinations.csv

|  |
| --- |
| * 1. **Data Cleaning** |

**1.1a) Cleaning anzsic\_2006 data**

* Reading the data from the anzsic\_2006 table
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls and removing columns – refer to Data Dictionary; and Juypter Notebook for code
* Writing SQL quires to get the necessary information about the data and plotting­­ using pandas

Text

Description automatically generated

* Saving the cleaned file as CSV (cleaned\_anzsic.csv) **– refer to the Cleaned data folder**

**1.1b) Cleaning emission-reduction-techniques data**

* Reading the data from the emission-reduction-techniquestable
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Juypter Notebook for code**
* Using python codes to get more insights from data **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_reduction\_tech.csv) **– refer to the Cleaned data folder**

**1.1c) Cleaning emissions data**

* Reading the data from the emissionstable
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Data Dictionary; and Juypter Notebook for code**
* Using python codes to get more insights from data **– refer to Juypter Notebook for code**
* Impute the null values with a mean of those columns **(example:** air\_total\_emission\_kg, air\_fugitive\_emission\_kg based upon the relative year**)**
* Saving the cleaned file as CSV (Cleaned\_emissions.csv) **– refer to the Cleaned data folder**

**1.1d) Cleaning facilities data**

* Reading the data from the facilitiestable
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Juypter Notebook for code**
* Using python codes to get more insights from data **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_facilities.csv) **– refer to the Cleaned data folder**

**1.1e) Cleaning facility\_secondary\_anzsics data**

* Reading the data from the facility\_secondary\_anzsics table
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_facility\_secondary.csv) **– refer to the Cleaned data folder**

**1.1f) Cleaning reports data**

* Reading the data from the reports table
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_reports.csv) **– refer to the Cleaned data folder**

**1.1g) Cleaning substances data**

* Reading the data from the substances table
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Juypter Notebook for code**
* Using python codes to get more insights from data **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_substances.csv) **– refer to the Cleaned data folder**

**1.1h) Cleaning transfer\_destinations data**

* Reading the data from the transfer\_destinations table
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_transfers.csv) **– refer to the Cleaned data folder**

**1.1i) Cleaning transfers data**

* Reading the data from the transfers table
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes, and removing columns **– refer to Data Dictionary; and Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_transfer\_destination.csv) **– refer to the Cleaned data folder**

|  |
| --- |
| * 1. **Data Joining using cleaned files** |

### 1.2a) Joining cleaned anzsic and reports tables

* Joining the anzsic and reports tables
* Using python codes to do Exploratory data analysis **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_join\_rep\_anz\_data.csv) **– refer to the Cleaned\_Join\_data folder**

### 1.2b) Joining cleaned substances and emissions tables

* Joining the substances and emissions tables
* Using python codes to do Exploratory data analysis **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_join\_em\_sub\_data.csv) **– refer to the Cleaned\_Join\_data folder**

### 1.2c) Joining cleaned facilities and Facility Secondary ANZSICs tables

* Joining the facilities and Facility Secondary ANZSICs tables
* Using python codes to do Exploratory data analysis **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_join\_facilities\_SecFacilityANZ\_data.csv) **– refer to the Cleaned\_Join\_data folder**

### 1.2d) Joining cleaned emission reduction technique and emission tables

* Joining the emission reduction technique and emission tables
* Using python codes to do Exploratory data analysis **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_join\_emission\_emissionReductionTech\_data.csv) **– refer to the Cleaned\_Join\_data folder**

### 1.2e) Joining cleaned transfers and transfer destinations table

* Joining the transfers and transfer destinations tables
* Using python codes to do Exploratory data analysis **– refer to Juypter Notebook for code**
* Saving the cleaned file as CSV (Cleaned\_join\_transfers\_TransDestination\_data.csv) **– refer to the Cleaned\_Join\_data folder**

### 1.2f) Joining All cleaned tables

* Joining all tables to create a master table and saving the file as CSV (master\_data.csv) **– refer to the Cleaned\_Join\_data folder**

**Tableau**

* Cleaned\_anzsic.csv
* Join Cleaned\_emission.csv with Cleaned\_substances using substance\_id -> substance\_id
* Join Cleaned\_emission.csv with Cleaned\_reduction\_tech.csv using report\_id -> report\_id
* Join Cleaned\_facilities.csv with Cleaned\_facility\_secondary.csv using facility\_id -> facility\_id
* Join Cleaned\_transfers.csv with Cleaned\_transfer\_destination.csvusing transfer\_destination\_id -> transfer\_destination\_id

|  |
| --- |
| **2.) EDA with Machine learning model** |

**Python**

* Importing all the required libraries
* Reading the data from the Cleaned\_join\_em\_sub\_data.csv file
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes and removing columns, label encoding, and one hot encoding, and doing Exploratory data analysis **– refer to Juypter Notebook (EDA with Machine learning models) for code**
* Using Linear regression Machine learning algorithm to train the model
* Getting predictions from the model

|  |
| --- |
| **3.) Predictive analysis (Time Series Analysis) of the total emission of substances** |

**Python**

* Importing all the required libraries
* Reading the data from the Cleaned\_join\_em\_sub\_data.csv file
* Checking the datatypes, null values, shape of data, and description (mean, max, min, percentage of distribution, and count)
* Handling nulls, changing datatypes and removing columns, and doing Exploratory data analysis to get the mean of “**air\_point\_emission\_kg**” by report year **– refer to Juypter Notebook (EDA with Machine learning models) for code**
* Saving the file as CSV (time\_series\_air\_total\_emission.csv) for time series analysis**– refer to the Cleaned\_data folder**

**Tableau**

* Use “time\_series\_air\_total\_emission.csv” to forecast the total substance emission in the air for the next five years using air\_total\_emission -> Report Year

|  |
| --- |
| **Sample Analysis of EDA with Machine learning models using Python** |

**Size of the data frame**

Graphical user interface, text, application, email

Description automatically generated

**Trained Linear Regression Model**

Graphical user interface, text, application, email

Description automatically generated

**Time Series Analysis**

**Impact of total emission in the air by year**

