

# Progress Report 1

## Automated Product Inquiry Response System for Light Recycling Program



**Student Name:** Namesh Mathara Arachchi Vidanalage

**Student ID:** 300359798

**Course:** CSIS 4495 – Applied Research Project

**Section:** 002

## Contents

<b>Work Logs – Summarized.....</b>	<b>3</b>
<b>Work Logs – Detailed .....</b>	<b>3</b>
<b>28/01/2025 .....</b>	<b>3</b>
<b>Researched methods to extract texts and images from a PDF (product guide) .....</b>	<b>3</b>
<b>31/01/2025 – 05/02/2025 .....</b>	<b>4</b>
<b>Exported the Lights Product Guides to Word and created the Products Table in Excel.....</b>	<b>4</b>
<b>Cleaned data to remove inconsistencies .....</b>	<b>7</b>
<b>06/02/2025 .....</b>	<b>8</b>
<b>Wrote a script to upload product data from offline excel file to SQL Server DB table (ONLY textual data) .....</b>	<b>8</b>
<b>07/02/2025 – 09/02/2025 .....</b>	<b>9</b>
<b>Implemented an ID to number Product Categories correctly .....</b>	<b>9</b>
<b>Changed the structure of the Products Table to suit the user requirement .....</b>	<b>9</b>
<b>Drafted and finalized the progress report 1 .....</b>	<b>9</b>
<b>Repository Check-in Details.....</b>	<b>10</b>

# Work Logs – Summarized

Work Logs					
Phase	Date	#	Number of Hours	Tr	Description of Work Done
Project Proposal	13/01/2025	2			Conducted initial research on PCA's operations and identified key inefficiencies in product inquiry management.
Project Proposal	14/01/2025	1			Gathered and analyzed the existing product catalog and historical inquiry data to assess feasibility for ML integration.
Project Proposal	15/01/2025	2			Met with the staff handling Lights Product Inquiries to review: - The product catalog and past inquiries for data understanding. - The entire process from receiving to closing an inquiry.
Project Proposal	16/01/2025	2			Researched suitable ML techniques for text and image-based product classification.
Project Proposal	20/01/2025	2			Outlined project scope, methodology, and initial timeline.
Project Proposal	21/01/2025	1			Discussed project expectations and obtained consent from PCA with my manager
Project Proposal	25/01/2025	3			Drafted the initial project proposal and refined research objectives.
Project Proposal	26/01/2025	3			Finalized the initial project proposal and refined research objectives.
Progress Report 1	28/01/2025	2			Researched methods to extract texts and images from a PDF (product guide). One method is using python - PyMuPDF, pdfplumber for text and table data extraction and PyMuPDF, Pillow (PIL) for image extraction. The other method is to get a paid subscription of Adobe Acrobat and export it to Word/Excel.
Progress Report 1	31/01/2025	2			Export the Lights Product Guide to Word and create the Products Table in Excel. The guide's structure is unsuitable for an SQL Server database, as products are categorized rather than listed individually. Some images are merged and must be separated so each row represents a single product.
Progress Report 1	02/02/2025	3			Add 150 'included' lights products from the product guide to the Products Table in Excel
Progress Report 1	03/02/2025	3			Add 184 'included' lights products from the product guide to the Products Table in Excel (finished all the included products in product guide)
Progress Report 1	04/02/2025	3			Fixed the inconsistencies in the Products Table - e.g. same category in different names, mismatching characters (En Dash (-) and Hyphen (-)), etc.
Progress Report 1	05/02/2025	2			Add 56 'excluded' lights products from the product guide to the Products Table in Excel (finished all the excluded products in product guide)
Progress Report 1	06/02/2025	2			Wrote the script to upload product data from offline excel file to SQL Server DB table (ONLY textual data)
Progress Report 1	07/02/2025	2			Implemented an ID to number Product Categories correctly with provisions for new product categories in between current ones
Progress Report 1	08/02/2025	3			Changed the structure (pivoted by Province) of the Products Table to suit the user requirement
Progress Report 1	09/02/2025	2			Drafted and finalized the progress report 1

[Link to Work Logs Sheet](#)

## Work Logs – Detailed

**28/01/2025**



### Researched methods to extract texts and images from a PDF (product guide)

I explored various methods to extract text and images from a product guide in PDF format. The primary approaches considered were:

1. **Python-Based Approach** – Utilizing libraries such as PyMuPDF and pdfplumber for text and table data extraction, and PyMuPDF along with Pillow (PIL) for image extraction.
2. **Adobe Acrobat Pro** – Exporting data, including tables, directly to Word or Excel.

The product guide does not follow a simple tabular format; instead, it presents a complex table structure with multiple images categorized within a single row, merged cells, and inconsistent labeling. Product categories and descriptions often lack uniformity, making automated extraction more challenging.

Screenshot of Product Guide

CATEGORY	DESCRIPTION	EHF
1. Fluorescent tubes measuring less than or equal to 2 ft	<p>Includes all diameters and light outputs, shaped fluorescent tubes, and UV-A and UV-B tubes.</p> 	\$0.20
2. Fluorescent tubes measuring greater than 2 ft and up to or equal to 4 ft		\$0.40
3. Fluorescent tubes measuring greater than 4 ft		\$0.80
4. Compact Fluorescent Lights (CFL)/ Screw-In Induction Lamps	<p>Fluorescent bulbs that are typically similar in size and intended to replace an incandescent (traditional) light bulb, including pin-type sockets, covered CFLs and various output wattages. Includes screw-in induction lamps.</p> 	\$0.15

Python-based extraction is most effective when the table structure remains consistent across all pages, and the data follows a predictable layout, such as fixed column positions. However, given that the dataset contains only 400 records, fully automating the extraction process in Python is not a viable option. Manual verification would still be necessary, and developing custom handling for each inconsistency would require more effort than manually reviewing and formatting the data in Excel.

Considering these challenges and time constraints, I opted to export the data into Excel using Adobe Acrobat Pro and then manually format it to create a final dataset suitable for export to SQL Server.

## 31/01/2025 – 05/02/2025





### Exported the Lights Product Guides to Word and created the Products Table in Excel

There are two product guides:

1. Product Guide for *British Columbia, Manitoba, Prince Edward Island, and Québec*
2. Product Guide for *Ontario*

The Ontario guide is separate because it contains certain confidential information specific to the province. Each product guide primarily details the included products under the Lights Recycling Program for the respective province.

*Included Products in Product Guide*

<div>Light Recycling Product Guide British Columbia</div>		
CATEGORY	DESCRIPTION	EHF
10. Fixture Category A – Emergency / Egress Lights	<p>Emergency/Egress Lights. Does not include exit signs without attached light heads (refer to excluded list for further details.)</p> 	\$0.15
10. Fixture Category A – Small Outdoor Fixtures	<p>Bollard</p> 	\$0.15
	<p>Post Lighting (consumer applications only)</p> 	
	<p>Path, Walkway, Garden, In-Grade, Border, Step Lights (non-solar powered only). Solar powered equivalents included in Category 9</p> 	

Additionally, it provides examples of excluded products applicable to that province.

*Excluded Products in Product Guide*

**Light Recycling Product Guide  
British Columbia**

### Excluded Items for BC

The following products are exempt from EHF's and are not collected as part of BC's light recycling program.

Replacement lamps used in excluded fixtures and products should be charged EHF's and included in the program if they are sold separately and can be disposed of separately from the fixture or product.

**"Light Containing" Products:**

Products containing lights with a primary purpose that is not to illuminate or assist in the illumination of space are outside the scope of this program, including, but not limited to:

- Products covered by other schedules of the BC Recycling Regulation and for management in other product stewardship programs in BC. Examples include large appliances, small appliances, medical equipment and electronic products.
- Products containing lights with a primary purpose of signaling or displaying information. See below for detailed examples.

<ol style="list-style-type: none"> <li>1. Alarms, phones and devices for the visually impaired</li> <li>2. Aquarium equipment</li> <li>3. Auto fixtures</li> <li>4. Back lit signs</li> <li>5. Ventilation fans</li> <li>6. Black light equipment</li> <li>7. Bug zappers</li> <li>8. Camera and video accessories</li> </ol>	<ol style="list-style-type: none"> <li>18. Lava lamps</li> <li>19. Light up shoes, hats, collars, and clothes</li> <li>20. Marine and aeronautical fixtures</li> <li>21. Mirror ball lights</li> <li>22. Neon signs</li> <li>23. Plasma balls</li> <li>24. Propane and gas powered lights</li> </ol>	<ol style="list-style-type: none"> <li>30. Decorative sculptures and statues with one or more integrated lights where the primary purpose of the product is decorative and the contained lights are designed to light the decoration itself and not to illuminate surrounding space</li> <li>31. Tanning beds</li> <li>32. Umbrellas with integrated lights</li> </ol>
---	--	--

After exporting the data from PDF to Word, a total of 337 included products (across 20 product categories) and 56 excluded products were compiled into an Excel file, resulting in a master Products Table containing 393 products.

**Note:** As mentioned in the Project Proposal, there are two sources of product data: the product guide, which serves as the base document, and a spreadsheet containing past product inquiries. However, the spreadsheet has data integrity issues that require time to resolve and clean.

Maintaining data integrity is crucial when training a Machine Learning (ML) model, as the quality, accuracy, and consistency of the dataset directly impact the model's performance, reliability, and generalization.

For this reason, I have started the project by using the product guide to create the products table. Once the issues in the spreadsheet have been addressed, I can merge its data into the products table.

## Cleaned data to remove inconsistencies

After creating the master table in Excel, the next step was data cleaning to eliminate inconsistencies. Some of the key inconsistencies identified and removed include:

- **Inconsistent Product Category Labeling**

Includes all HID technologies that contain mercury, such as High-Pressure Sodium (HPS), Mercury Vapor and Metal Halide, as well as UV-C / Germicidal lamps and tubes, Tubular Induction lamps (circular, square, U, etc.), UHP replacement lamps (projector etc.), neon replacement lamps, etc.
Includes all HID technologies, such as High Pressure Sodium, Low Pressure Sodium (HPS), Mercury Vapour and Metal Halide, as well as UV-C / Germicidal lamps and tubes, Tubular Induction lamps (circular, square, U etc.), UHP replacement lamps (projector etc.), Neon replacement lamps, etc.
Includes all HID technologies, such as High-Pressure Sodium (HPS), Low Pressure Sodium (LPS), Mercury Vapor and Metal Halide, as well as UV-C / Germicidal lamps and tubes, Tubular Induction lamps (circular, square, U etc.), UHP replacement lamps (projector etc.), Neon replacement lamps, etc.
Includes all HID technologies, such as High-Pressure Sodium (HPS), Low Pressure Sodium, Mercury Vapor and Metal Halide, Tubular Induction lamps (circular, square, U, etc.), UHP replacement lamps (projector etc.), neon replacement lamps, etc.

High Intensity Discharge (HID), <u>Germicidal</u> , Special Purpose and Other
High Intensity Discharge (HID), Special Purpose and Other

Includes all diameters and light outputs, shaped fluorescent tubes, and UV-A and UV-B tubes
Includes all diameters and light outputs, shaped fluorescent tubes, and UV-A and UV-B tubes.

Fluorescent tubes measuring greater than 2 ft and up to or equal to 4 ft
Fluorescent tubes measuring greater than 2 ft and up to or equal to 4 ft

Compact Fluorescent Lights (CFL)/ <u>Screw-</u>
In Induction Lamps
Compact Fluorescent Lights (CFL)/ Screw-In Induction Lamps



- **Inconsistent Use of Characters like en dash, hyphen, space, forward slash**

Rope/Strip/Ribbon/Tape Lights
-------------------------------

Products are reported and applied fees in increments of 10 meters. Products of 10 meters or less are applied one recycling fee. Products greater than 10 meters are charged one recycling fee per 10-meter increment (i.e., 38 meters of rope lights would be reported as 4 units and assessed four fees). Members may choose to calculate a fee rate/unit sold to apply at point of sale and then bundle this into increments of 10 meters/\$0.15 for reporting purposes.
--

Path, Walkway, Garden, In-Grade, Border, Step Lights (non-solar powered only)
---

Path/Walkway/Garden/In-Grade/Border/Step Lights (solar powered only)
--

Fixture Category A - Emergency / Egress Lights
--

Designated Small Fixtures/ Decorative Light Strings
---

Fixture Category A - Linear Fixtures (including linear shop lights and linear pool/fountain fixtures)
---

Wall Mount/Small Flood - including commercial "wall packs" and flood lights less than 250 W
---

Security Lighting (with or without integrated cameras)- Including residential-type security floodlights
---

Lamp-holders (stand-alone and for more than one lamp)
---

Lamp-holders (stand-alone and single lamp only)
---

Recessed/Pot - Fee is only applied to the housing if housing and trims are sold separately.
---

Light Emitting Diodes (LED) - Bulbs
-------------------------------------

Light Emitting Diodes (LED) - Tubes and Other
---

## 06/02/2025

### Wrote a script to upload product data from offline excel file to SQL Server DB table (ONLY textual data)

Currently, the script reads data from the offline Excel file (**Products Table**) and connects to the local instance of SQL Server. It first checks whether the **lights\_product\_inquiry** table in the **product\_inquiry** database contains any existing data. If data is present, the script clears the table and resets the ID before inserting new records.

Once the company's SQL Server and Azure Blob Storage environments are set up and I receive the necessary access (expected next week), the script will be modified accordingly. The updated version will:

1. Insert product details into the SQL Server table.
2. Upload product images to Azure Blob Storage.
3. Store the image URLs in the SQL Server table.



**07/02/2025 – 09/02/2025**

## Implemented an ID to number Product Categories correctly

Implemented a consistent ID system to number product categories accurately, ensuring space for new categories between existing ones. The previous system lacked consistency, assigning different category numbers to the same product category across provinces. The new system provides adequate spacing, accommodates both main and subcategories, and ensures uniqueness across all provinces.

## Changed the structure of the Products Table to suit the user requirement

The user's requirement is that when a product inquiry is submitted by a member, the user reviews the product details, and the image provided. They then identify the relevant product category from the product guide and inform the member whether the product is included in the recycling program, along with its EHF value.

The user provides this information for all provinces because an inquiry may not specify a particular province. In such cases, pivoting by province is beneficial, as it allows the ML model to find the best-matching record from the dataset. Once matched, all relevant information for all provinces is available within the corresponding row.

*Current SQL Server Table*

Results		Messages											
	id	image_url	source	product_category_id	product_category	product_description	british_columbia	manitoba	ontario	prince_edward_island	quebec	created_at	
1	1	NULL	Product Guide	1101	Fluorescent tubes measuring less than or equal t...	Includes all diameters and light outputs...	0.2	1	C	0.4	0.35	2025-02-09 03:26:01.827	
2	2	NULL	Product Guide	1101	Fluorescent tubes measuring less than or equal t...	Includes all diameters and light outputs...	0.2	1	C	0.4	0.35	2025-02-09 03:26:01.827	
3	3	NULL	Product Guide	1101	Fluorescent tubes measuring less than or equal t...	Includes all diameters and light outputs...	0.2	1	C	0.4	0.35	2025-02-09 03:26:01.827	
4	4	NULL	Product Guide	1101	Fluorescent tubes measuring less than or equal t...	Includes all diameters and light outputs...	0.2	1	C	0.4	0.35	2025-02-09 03:26:01.827	
5	5	NULL	Product Guide	1601	Fluorescent tubes measuring greater than 2 ft an...	Includes all diameters and light outputs...	0.4	1.8	C	0.8	0.65	2025-02-09 03:26:01.827	
6	6	NULL	Product Guide	1601	Fluorescent tubes measuring greater than 2 ft an...	Includes all diameters and light outputs...	0.4	1.8	C	0.8	0.65	2025-02-09 03:26:01.827	
7	7	NULL	Product Guide	1601	Fluorescent tubes measuring greater than 2 ft an...	Includes all diameters and light outputs...	0.4	1.8	C	0.8	0.65	2025-02-09 03:26:01.827	
8	8	NULL	Product Guide	1601	Fluorescent tubes measuring greater than 2 ft an...	Includes all diameters and light outputs...	0.4	1.8	C	0.8	0.65	2025-02-09 03:26:01.827	
9	9	NULL	Product Guide	2101	Fluorescent tubes measuring greater than 4 ft	Includes all diameters and light outputs...	0.8	2.5	C	1.1	1.05	2025-02-09 03:26:01.827	
10	10	NULL	Product Guide	2101	Fluorescent tubes measuring greater than 4 ft	Includes all diameters and light outputs...	0.8	2.5	C	1.1	1.05	2025-02-09 03:26:01.827	
11	11	NULL	Product Guide	2101	Fluorescent tubes measuring greater than 4 ft	Includes all diameters and light outputs...	0.8	2.5	C	1.1	1.05	2025-02-09 03:26:01.827	
12	12	NULL	Product Guide	2101	Fluorescent tubes measuring greater than 4 ft	Includes all diameters and light outputs...	0.8	2.5	C	1.1	1.05	2025-02-09 03:26:01.827	
13	13	NULL	Product Guide	2601	Compact Fluorescent Lights (CFL) / Screw-In Indu...	Fluorescent bulbs that are typically simil...	0.15	1	C	0.3	0.25	2025-02-09 03:26:01.827	
14	14	NULL	Product Guide	2601	Compact Fluorescent Lights (CFL) / Screw-In Indu...	Fluorescent bulbs that are typically simil...	0.15	1	C	0.3	0.25	2025-02-09 03:26:01.827	
15	15	NULL	Product Guide	2601	Compact Fluorescent Lights (CFL) / Screw-In Indu...	Fluorescent bulbs that are typically simil...	0.15	1	C	0.3	0.25	2025-02-09 03:26:01.827	
16	16	NULL	Product Guide	2601	Compact Fluorescent Lights (CFL) / Screw-In Indu...	Fluorescent bulbs that are typically simil...	0.15	1	C	0.3	0.25	2025-02-09 03:26:01.827	

After pivoting by province, the total number of included products is reduced from 337 to 229 unique products.

## Drafted and finalized the progress report 1

Prepared the Progress Report 1 as per the guidelines mentioned in the Progress Report Template.

# Repository Check-in Details

The files and folders I have checked into the GitHub repository since the project proposal are as follows:

- [Latest Products Table \(Excel\)](#)
- [Lights Product Guide \(PDF\)](#)
- [Python Script to upload product data from excel file to SQL Server DB table](#)
- [Lights Product Categories Summary](#)
- [Lights Product Categories and EHF's Summary](#)
- [SQL script to CREATE TABLE lights\\_product\\_inquiry in SQL Server](#)

Regular check-ins were maintained to ensure all progress was tracked in the repository.