# MSDS 420 Final Project Phase 2 – Design & Development

**Requirement 1:** (10 Points) **Create** a **table** that has the list of **queries** for the group that you selected from the groups (**Group A, Group B, Group C**) listed in the **Final Project - Requirements Specification.pdf**, and add to it the **3 queries** that your created in **Phase 1** that belong to the group you selected.

Query
How many reviews for asin ID B000FPDHKA received an overall rating of 5?
How many reviews were submitted in January of 2014?
How many reviews received a helpfulness score?
How many products are there in every product subcategory?
How many product subcategories are in the provided datasets?
What is the average review rating of all reviews?
How many unique reviewers?
How many reviews for every product subcategory?
What is the average size (characters/words) of the review text/body?
How many reviews submitted every January for years: 2011, 2012, 2013, 2014?
How many reviews submitted for every product subcategory every May for years: 2011, 2012, 2013, 2014?
What is the average price of every product subcategory that has an average overall rating that is greater than 3?
What are the top 3 reviewed products names in every product subcategory?

**Requirement 2:** (10 Points) Which database that you have selected to implement the queries above? What was your **benchmark-metrics** for the **database selected**?

I have selected PostgreSQL, specifically PGAdmin to conduct the queries in the chart. The benchmark performance metrics used in my selection of PGAdmin were design, response time and query performance. PGAdmin allowed for easy importing of the cleaned data. After cleaning the data using Python, I was able to easily import the data in the necessary tables. When I came across issues, PGAdmin was clear in defining what needed to be addressed to complete upload. Previous assignments solidified my preference for PGAdmin's interface compared to others. PGAdmin is very user friendly, specifically the ability to expand tables to see columns to ensure data is coded correctly. In addition, the ERD Diagram option that provides a visual of the schema. Lastly, PGAdmin's performance is fast and efficient.

**Requirement 3:** (10 Points) Use the database you selected above, **design** and **develop** a **database application** (insert the data into the database) to host Amazon datasets (Reviews and Metadata).

To design and develop a data application within PGAdmin I first had to clean and pre-process the data for import. I used python/jupyter notebook and the pandas library to clean the data.

#### Cleaning Metadata:

I used pandas to convert the previous csv to an excel and remove the lines with null spaces. I used Microsoft Excel to remove additional commas and quotations that would prevent PGAdmin from importing.

```
□ ↑ ↓ 古 〒 盲
   [59]: import pandas as pd
              df = pd.read_excel('data.xlsx')
              # Display the first few rows of the dataframe
              print(df.head())
             asin price
0 0970408285 40.75 http://ecx.images-amazon.com/images/I/41kQjPIF...
1 730113188 NaN http://ecx.images-amazon.com/images/I/41EfFOPw...
2 800002N7HY 2.29 http://ecx.images-amazon.com/images/I/41EfFOPw...
3 800002N7HY 9.13 http://ecx.images-amazon.com/images/I/41EfFOPw...
                   B00004SQH9 16.29 http://ecx.images-amazon.com/images/I/41eg0Hvy...
              0 High-pressure hose 90 end & hose washer instal.
                   NaN
Built of durable thermoplastic, Leviton Power ...
Built of durable thermoplastic, Leviton Power ...
Coleman Cable 3536 14/3 General-Use Appliance ...
                                                                                      categories
                 categories

[['Appliances', 'Parts & Accessories', 'Washer...

[['Appliances', 'Freezers', 'Chest Freezers']]

[['Appliances', 'Parts & Accessories', 'Range ...

[['Appliances', 'Parts & Accessories', 'Range ...

[['Appliances', 'Parts & Accessories', 'Refrig...
                  Leviton 5050 50 Amp, 125/250 Volt, NEMA 10-50R... Leviton Leviton 7313 30 Amp, 125 Volt, NEMA Tt-30R, 2P... Leviton Coleman Cable 3536 14/3 General-Use Appliance ... Coleman Cable
                                                                                                  NaN
              2 {'also_bought': ['B000BPFZHM', 'B000VU1KC6', '...
3 {'also_bought': ['B002OUMU66', 'B007HFT034', '...
4 {'also_bought': ['B000XSNDS0', 'B0022NH420', '...
[61]: print(df1.head())
                  B00004TBK4
                                        8.54 http://ecx.images-amazon.com/images/I/51qKi62b...
           10 B00004TBJT 15.61 http://ecx.images-amazon.com/images/I/51oEVW5w...
           13 B00004UE6V 12.74 http://ecx.images-amazon.com/images/I/31K6MMRC...
14 B00004W4UJ 13.95 http://ecx.images-amazon.com/images/I/31Cf6Wj%...
                  B00004W4UK 13.95 http://ecx.images-amazon.com/images/I/31Cf6Wj%...
                  Springfield has been producing thermometers an...
                  The 90078 Springfield 13.25 inch Thermometer w...
                 Here's a nifty little item for consumers of ca...
Heat tamers are made of heavyweight steel to p...
                  Heat tamers are made of heavyweight steel to p..
                 [['Appliances', 'Parts & Accessories', 'Humidi...
[['Appliances', 'Parts & Accessories', 'Humidi...
                 [['Appliances', 'Trash Compactors']]
[['Appliances', 'Trash Compactors']]
[['Appliances', 'Parts & Accessories', 'Range ...
[['Appliances', 'Parts & Accessories', 'Range ...
            13
                  Springfield 90113-1 Color Track Humidity Meter... Springfield
                 Springfield 90078 Round 13.25inch Outdoor Ther... Springfield
                        Norpro 1305 Can Crusher Norpro
Nordic Ware 8 Inch Heat Tamer and Burner Plate Nordic Ware
            13
            15
                       Nordic Ware 8 Inch Heat Tamer and Burner Plate Nordic Ware
                  {'also_bought': ['B0030SZ4P0', 'B006Z70C0K', '...
                 { also_bought': ['B00305Z4P0', 'B006Z70C0K', '...
{'also_bought': ['B000AS22P8', 'B00004TBJI', '...
{'also_viewed': ['B00006ST35', 'B002718TE0', '...
{'also_bought': ['B000W24RW8', 'B000I1W08C', '...
{'also_bought': ['B000W24RW8', 'B000I1W08C', '...
            13
                      {'Home & Kitchen': 17812}
                            {'Home & Kitchen': 5100}
           13 {'Industrial & Scientific': 6625}
                   {'Kitchen & Dining': 3723}
{'Kitchen & Dining': 3723}
[63]: df1.to_excel('findata.xlsx', index=False)
```

## Cleaning Appliance Review Data:

I attempted this first; it took a bit of trial and error. I encountered encoding issues that were preventing pandas from reading the provided file. I had pandas ignore and replace the values that were causing that error to occur. Then, I went on to clean the data by dropping the null values. Again I used Microsoft excel to remove the unwanted values to make upload to PGAdmin easier.

```
[34]: import pandas as pd
[13]: with open('Amazon_Appliances_Reviews.xls') as f:
[14]: df=pd.read_csv('Amazon_Appliances_ReviewsCSV.csv', encoding="cp1252")
         # Specify the file path
file_path = 'C:/Users/jamia/MSDS 420/Final Project/Phase 1/Final_Project_Phase_1/Amazon/Amazon_Appliances_ReviewsCSV.csv'
         # Read the CSV file line by line to handle errors
lines = []
with open(file_path, 'r', encoding='cp1252', errors='replace') as file:
    for line in file:
        lines.append(line)
         # Convert Lines to a single string
data = ''.join(lines)
          # Read the CSV data into a dataframe
         from io import StringIO
df1 = pd.read_csv(StringIO(data))
         # Display the first few rows of the dataframe
print(df.head())

        reviewer_id
        asin_id
        reviewer_name

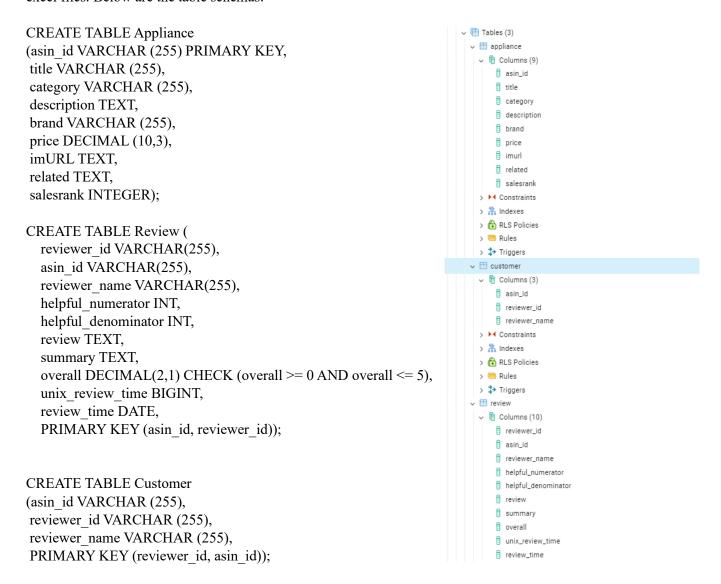
        0 A27H0WZBUB38FF
        970408285
        Steve

        1 A24HQ894NFSTF5
        7301113188
        Maha Saqfalhait shopaholic!;)

         2 AXE83MK90ZEVZ B00002N7HY
3 A2J7X7ZIH2EWB1 B00002NATH
4 AJQFNOFTZ7GOX B00002NATH
                                                                             Barthbill
             helpful_numerator helpful_denominator \
        0 could have been longer though, well made and e...
1 I like these containers so much i have ordered...
2 I tworks, no fires, etc. Why not 5 stars? Ho...
3 Fast shipping, works great
What can I say? It is the usual Leution high q...
           [18]: import pandas as pd
            Read the file with potential encoding issues
         file_path = 'C:/Users/jamia/MSDS 420/Final Project/Phase 1/Final_Project_Phase_1/Amazon/Amazon_Appliances_ReviewsCSV.csv'
          # Open the file with cp1252 encoding and read into panda
         with open(file_path, 'r', encoding='cp1252', errors='replace') as file:
df2 = pd.read_csv(file)
         # Display the first few rows of the dataframe
print(df2.head())
                                      asin_id reviewer_name
         reviewer_id asin_id
0 A27HOWZBUBJ8FF 970408285
         Barthbill
             helpful_numerator helpful_denominator \
                                                                       review overall \
         0 could have been longer though. well made and e...
1 I like these containers so much i have ordered...
2 It works, no fires, etc. Why not 5 stars? Ho...
3 Fast shipping. Works great
         4 What can I say? It is the usual Leviton high q...
         [19]: # Specify the file path to save the CSV file
  output_file_path = 'Amazon_App_ReviewsNEW.csv'
         df2.to_csv(output_file_path, index=False)
[20]: with open('Amazon_App_ReviewsNEW.csv') as f:
         <_io.TextIOWrapper name='Amazon_App_ReviewsNEW.csv' mode='r' encoding='cp1252'>
```

## **Creating Tables**

I then created the database within PGAdmin and populated the tables with the necessary information found in the excel files. Below are the table schemas:



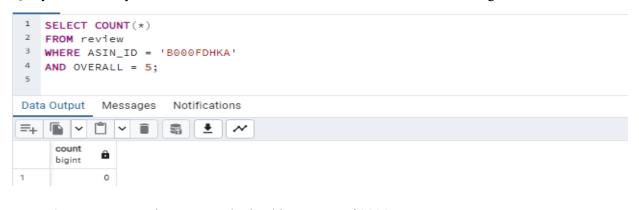
Lastly, I uploaded the clean data into the appropriate tables in PGAdmin. See below. I pulled the appropriate columns from the review data to create a new dataset to import into the customers table. Unfortunately, I deleted the record below, but they are present in the table as you will see in the queries.

	PID	Туре	Server	Object	Start Time V	Status	Time Taken (sec)
8	89344	Import Data	PostgreSQL 14 (localhost:5432)	Amazon Reviews and Metadata/public.appli	5/27/2024, 1:41:09 PM	Finished	0.22
8	19676	Import Data	PostgreSQL 14 (localhost:5432)	Amazon Reviews and Metadata/public.custo	5/27/2024, 11:30:37 AM	Finished	2.14

**Requirement 4:** (20 Points) **Create code** for the **queries** listed in Requirement\_1 using the database you selected in Requirement\_2. You are allowed to use gpt-3.5-trubo to assist you in the code generation.

Requirement 5: (25 Points) Execute the code for every query you created in Requirement 4.

Query 1: How many reviews for asin ID B000FPDHKA received an overall rating of 5?



Query 2: How many reviews were submitted in January of 2014?

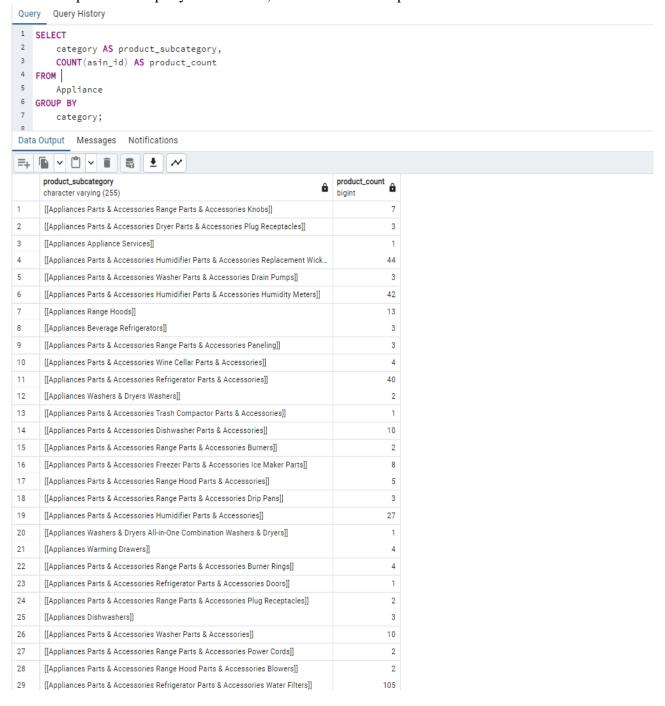


#### Query 3: How many reviews received a helpfulness score?

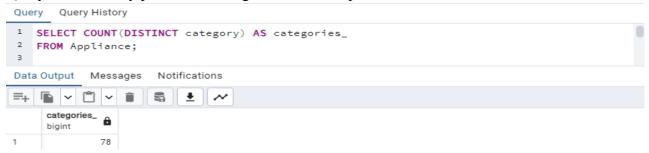


Query 4: How many products are there in every product subcategory?

\*\*The output for this query has 78 rows, it is shown in Panopto video\*\*



Query 5: How many product subcategories are in the provided datasets?



Query 6: What is the average review rating of all reviews?

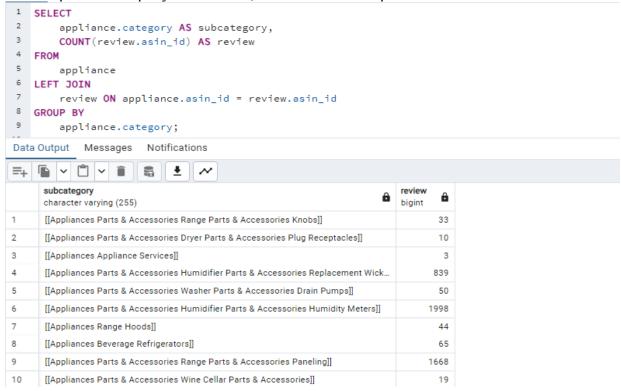


Query 7: How many unique reviewers?



Query 8: How many reviews for every product subcategory?

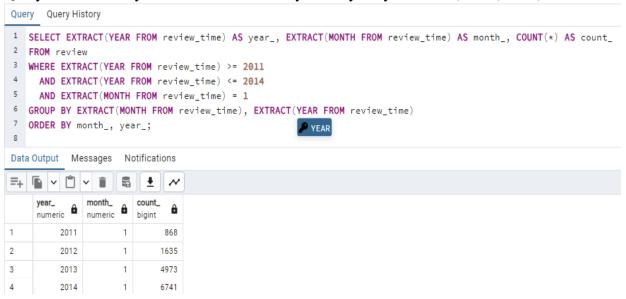
\*The output for this query has 78 rows, it is shown in Panopto video\*



Query 9: What is the average size (characters/words) of the review text/body?

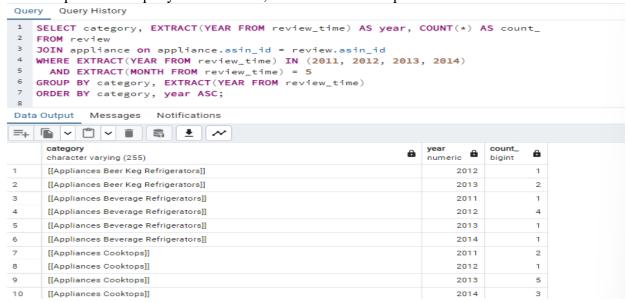


Query 10: How many reviews submitted every January for years: 2011, 2012, 2013, 2014?



Query 11: How many reviews submitted for every product subcategory every May for years: 2011, 2012, 2013, 2014?

\*The output for this query has 135 rows, it is shown in Panopto video\*



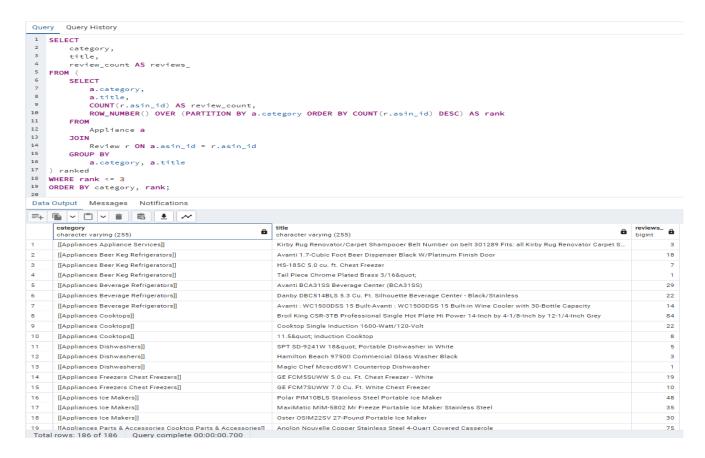
Query 12: What is the average price of every product subcategory that has an average overall rating that is greater than 3?

\*The output for this query has 75 rows, it is shown in Panopto video\*

```
Query Query History
 1 SELECT
 2
          appliance.category AS subcategory,
 3
          AVG(appliance.price::numeric) AS average_price
 4
    FROM
 5
          Appliance
 6
    JOIN
 7
          Review ON appliance.asin_id = review.asin_id
 8
    WHERE
9
          review.overall > 3
10
   GROUP BY
11
          appliance.category
12 ORDER BY
13
          appliance.category;
14
                             Notifications
Data Output Messages
     subcategory
                                                                                    average_price
      character varying (255)
                                                                                    numeric
1
       [[Appliances Appliance Services]]
                                                                                        6.18000000000000
2
       [[Appliances Beer Keg Refrigerators]]
                                                                                     210.5235714285714
3
       [[Appliances Beverage Refrigerators]]
                                                                                     442.4281818181818
4
       [[Appliances Cooktops]]
                                                                                     196.4168085106382
5
       [[Appliances Dishwashers]]
                                                                                     408.01000000000000
6
       [[Appliances Freezers Chest Freezers]]
                                                                                     322.7172727272727
       [[Appliances Ice Makers]]
7
                                                                                     167.8959016393442
       [[Appliances Parts & Accessories Cooktop Parts & Accessories]]
                                                                                      44.8139705882352
8
9
       [[Appliances Parts & Accessories Dishwasher Parts & Accessories Baskets]]
                                                                                      37.86500000000000
       [[Appliances Parts & Accessories Dishwasher Parts & Accessories Door Springs]]
                                                                                      19.95000000000000
10
11
       [[Appliances Parts & Accessories Dishwasher Parts & Accessories Panels]]
                                                                                       54.77000000000000
       [[Appliances Parts & Accessories Dishwasher Parts & Accessories Racks]]
                                                                                       48.3072591006423
12
       [[Appliances Parts & Accessories Dishwasher Parts & Accessories Trim Kits]]
                                                                                      17.6161904761904
13
14
       [[Appliances Parts & Accessories Dishwasher Parts & Accessories]]
                                                                                        6.6182234957020
       [[Appliances Parts & Accessories Dryer Parts & Accessories Lint Screens]]
                                                                                      17.3765060240963
15
16
       [[Appliances Parts & Accessories Dryer Parts & Accessories Plug Receptacles]]
                                                                                       15.712222222222
       [[Appliances Parts & Accessories Dryer Parts & Accessories Power Cords]]
                                                                                      21.26000000000000
17
18
       [[Appliances Parts & Accessories Dryer Parts & Accessories Vents]]
                                                                                       16.5035820895522
19
       [[Appliances Parts & Accessories Dryer Parts & Accessories]]
                                                                                      43.7279565217391
       [[Appliances Parts & Accessories Fasteners]]
                                                                                        4.40000000000000
20
21
       [[Appliances Parts & Accessories Food Waste Disposer Parts & Accessories]]
                                                                                      79.9900000000000
                                                                                      61.7478260869565
22
       [[Appliances Parts & Accessories Freezer Parts & Accessories Ice Maker Parts]]
```

## Query 13: What are the top 3 reviewed products names in every product subcategory?

\*The output for this query has 186 rows, it is shown in Panopto video\*



**Requirement 6:** (25 Points) Provide and **record** a **live run** with a walkthrough that demonstrates the **correctness** of the **output** produced for every **query** that you executed in Requirement 5.

Link to video:

https://northwestern.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ea880d88-a555-4763-b9e0-b17e0123c21a