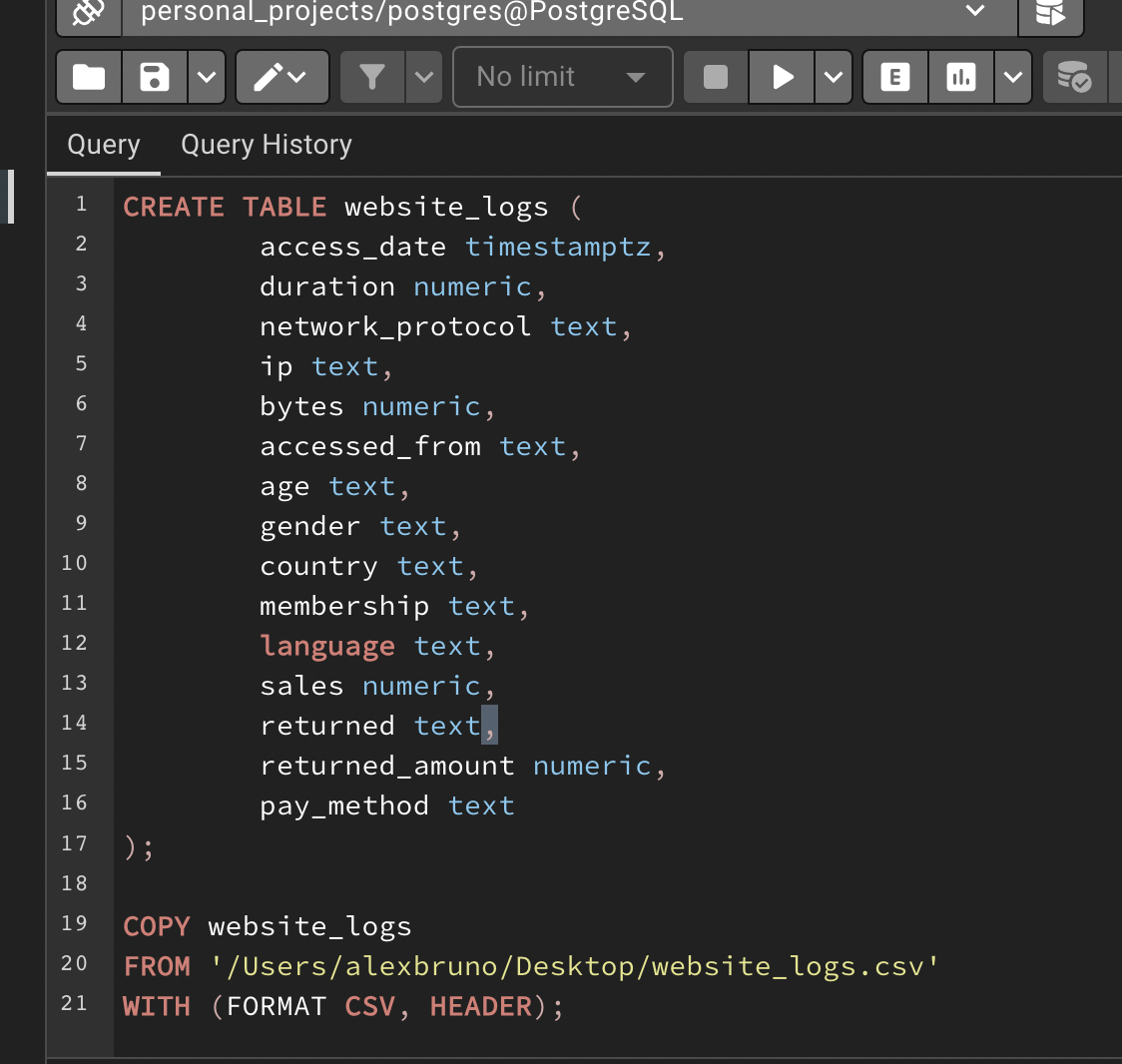
**SQL E-Commerce Data Project**

Source: found sample e-commerce website data from Kaggle.com

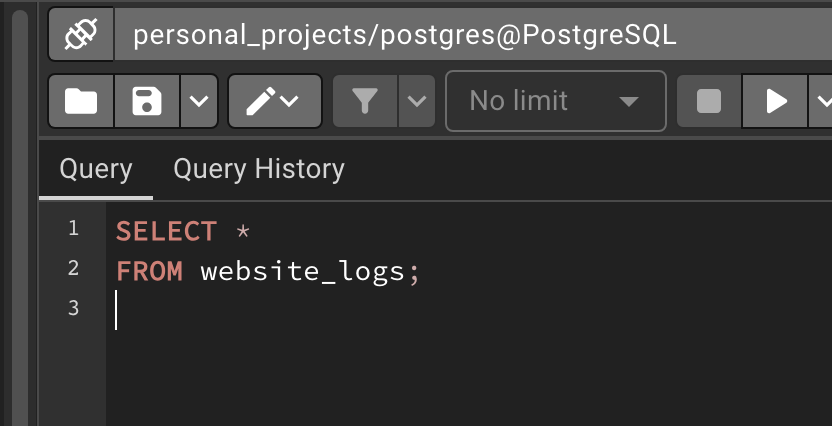
SQL type: PostgreSQL

Admin Tool: pgAdmin 4

Creating table within database and loading into pgAdmin 4:

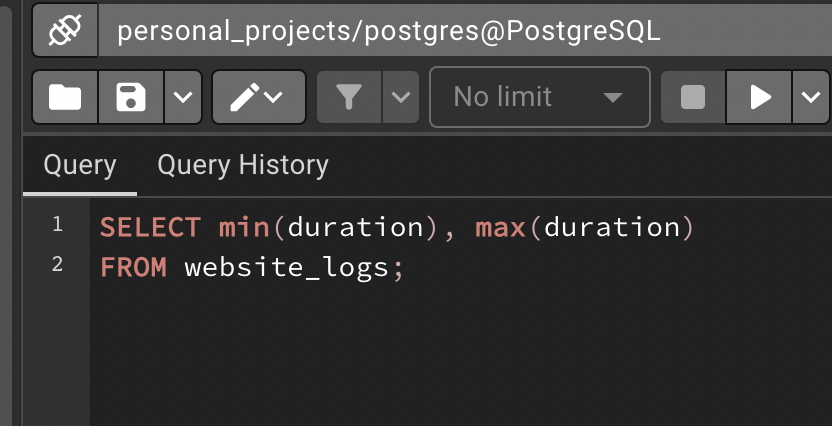


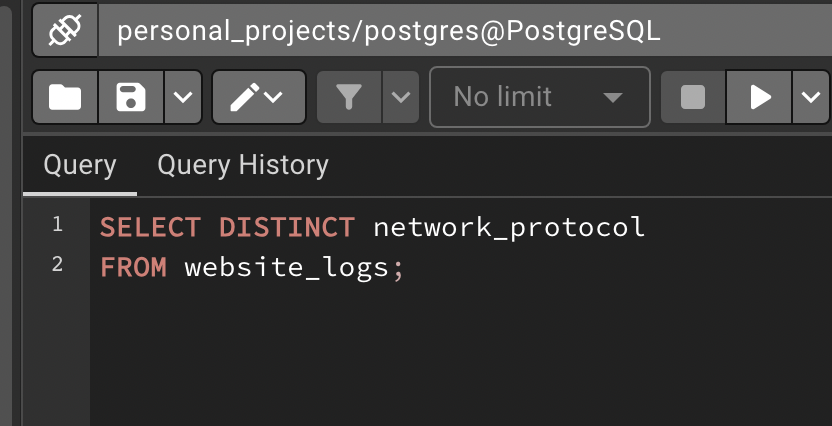
Initial data check:



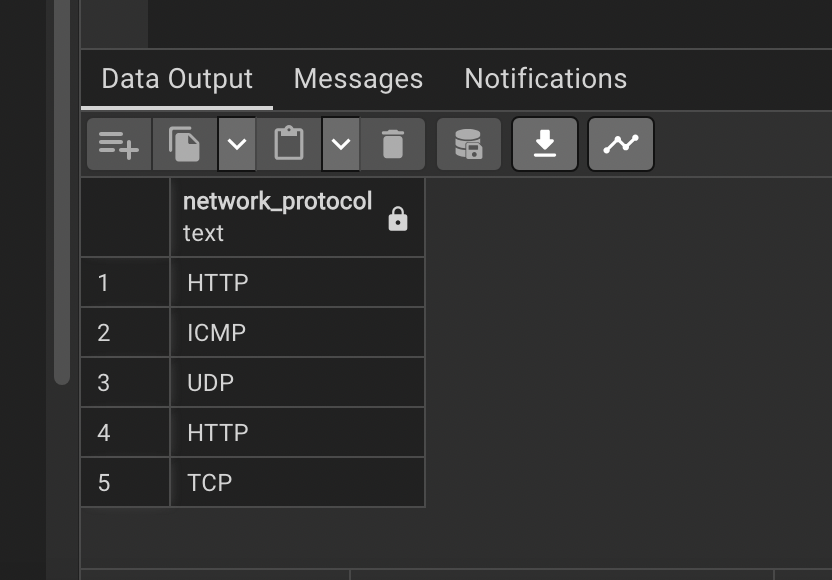
\*Loaded correctly with 172,838 rows.

Running DISTINCT clauses over any text columns, min/max over numeric columns, and WHERE column IS NULL for all columns. Checking for inconsistencies, errors, outliers, null values, etc. that could obscure any analyses:



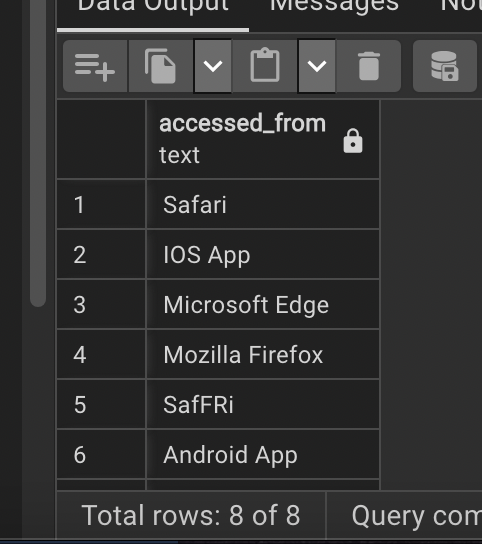


Error found during DISTINCT clause over network\_protocol column:



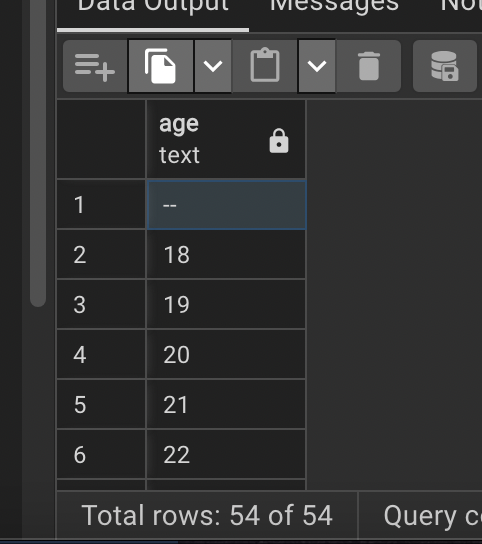
\*Upon further investigation, 37 rows that are a character length of 6 for HTTP. Added whitespace will need to be removed.

Error found during DISTINCT clause over accessed\_from column:



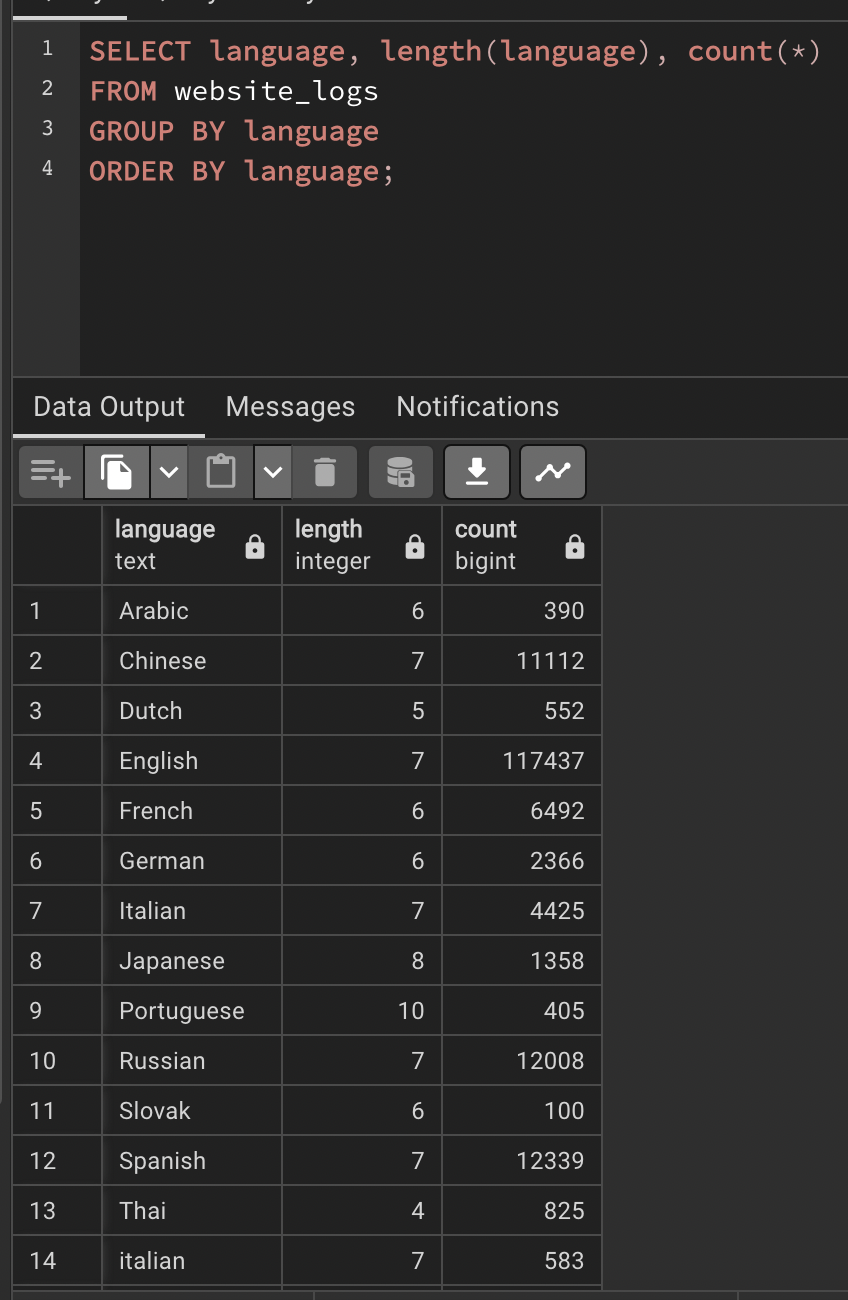
\*153 rows where “Safari” came in as “SafFRi” and will need to be corrected.

Error found during DISTINCT and ORDER BY age clause:



\*Around 88,000 rows came in as either NULL or ‘—’ and will need to change to something like ‘Unidentified.’

Error found during DISTINCT clause over language:



\*583 rows came in with Italian language as lowercase and will need to be merged with the other Italian rows that came uppercase.

Cleaning tasks found:

1. Strip whitespace from network protocol column for value ‘HTTP’ that is showing as 6 characters long.
2. Change rows in accessed\_from column that came in as “SafFRI” back to “Safari.”
3. Change rows in age column that were NULL or “—" to say “Unidentified.”
4. Turn rows with language – italian (lowercase) – to uppercase format to merge with the rest.

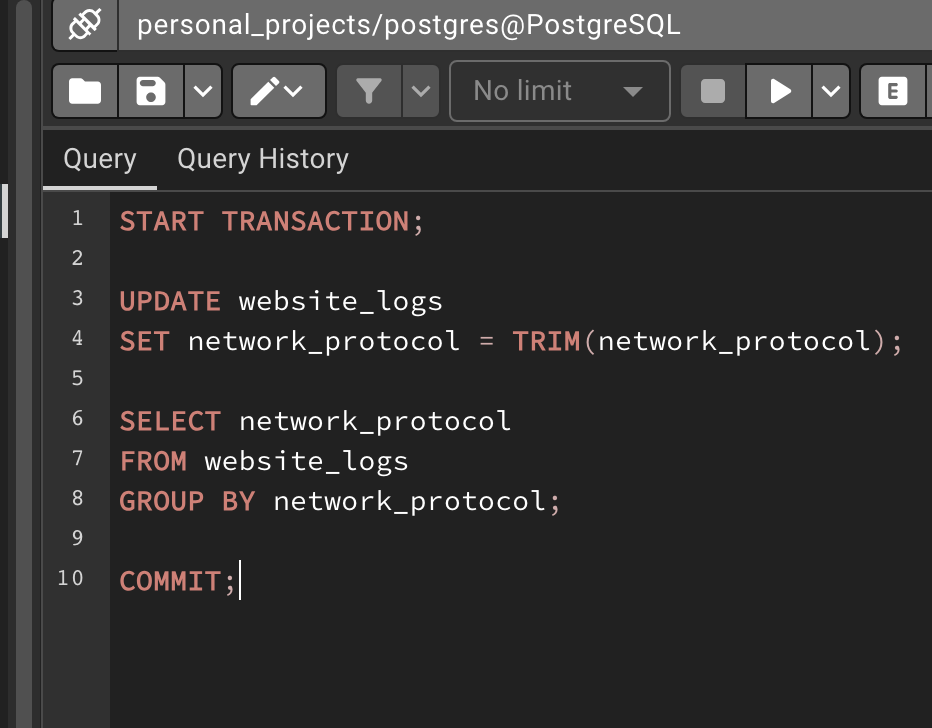
Other important findings to note when considering analysis:

1. 41,000 rows came with sales as $0. That will need to be factored in if performing any averages using that column.
2. Same thing as point 1 but with the returned\_amount column. 150,000 rows did not return anything.

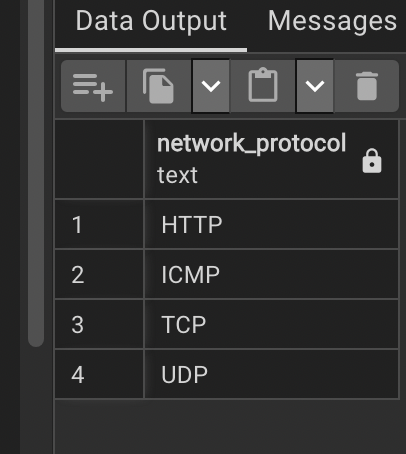
**Data Cleaning**

**Step one**: Strip whitespace from network protocol column. Specifically for value “HTTP” that is showing as six characters long, therefore not merging with the rest of the “HTTP” values.

Code for fixing that:

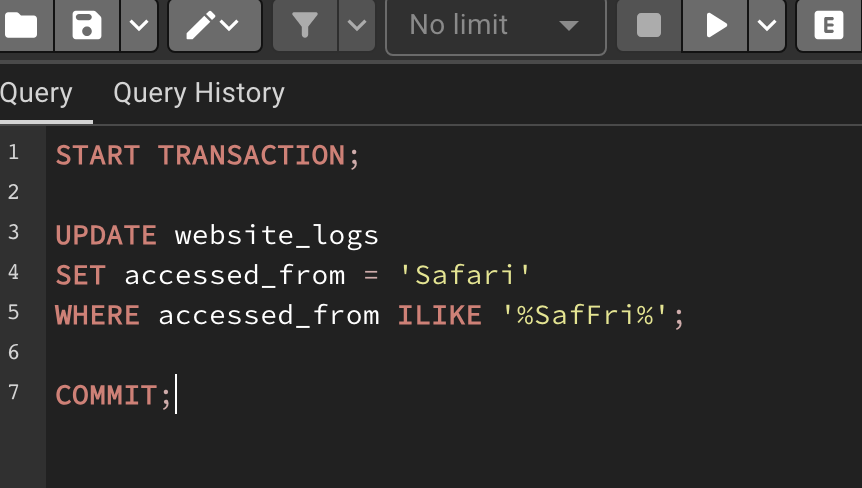


Corrected output now showing only 4 unique values for network\_protocol column:

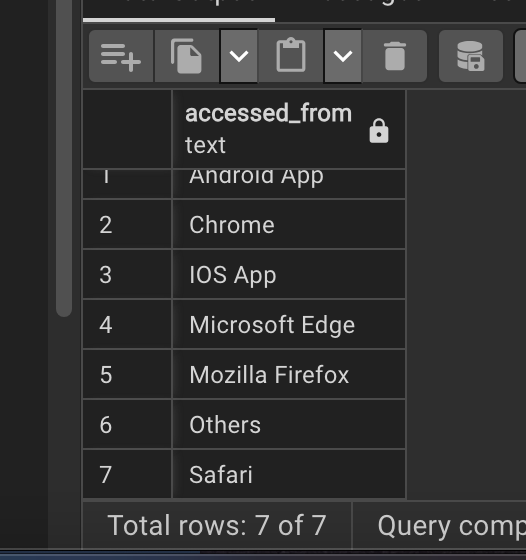


**Step two**: Change values in accessed\_from column that came in as ‘SafFri’ instead of ‘Safari.’

Code to fix that:

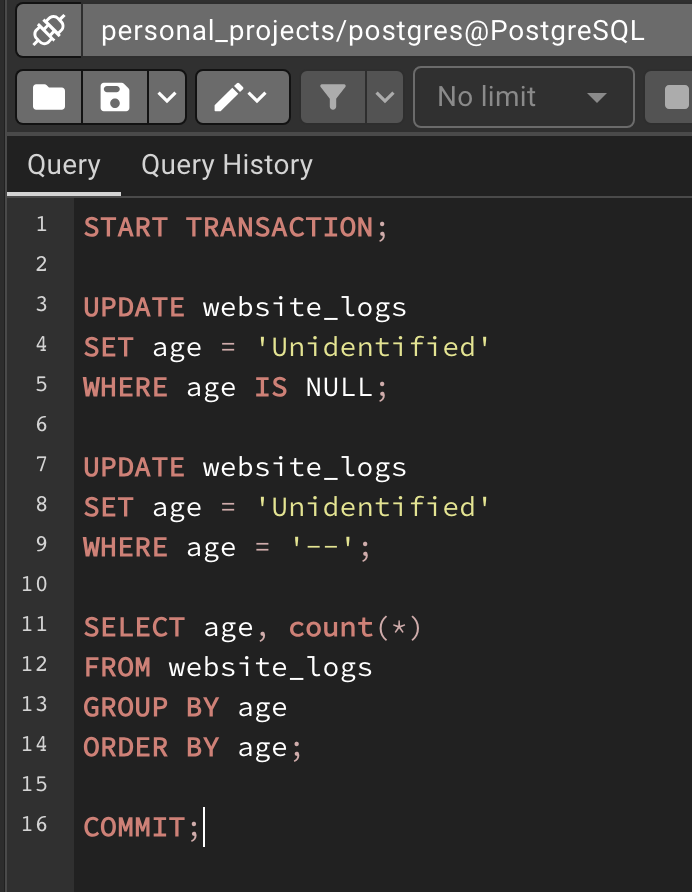


Output now standardized:

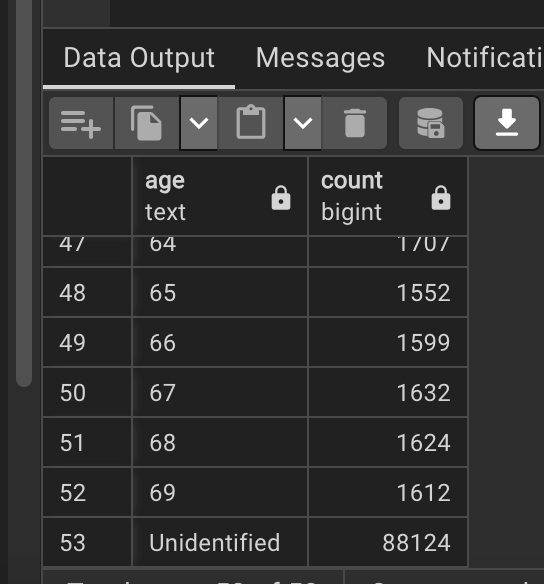


Step three: Change NULL and ‘—’ values in age column to say ‘Unidentified’ so that we can standardize format.

Code for that:

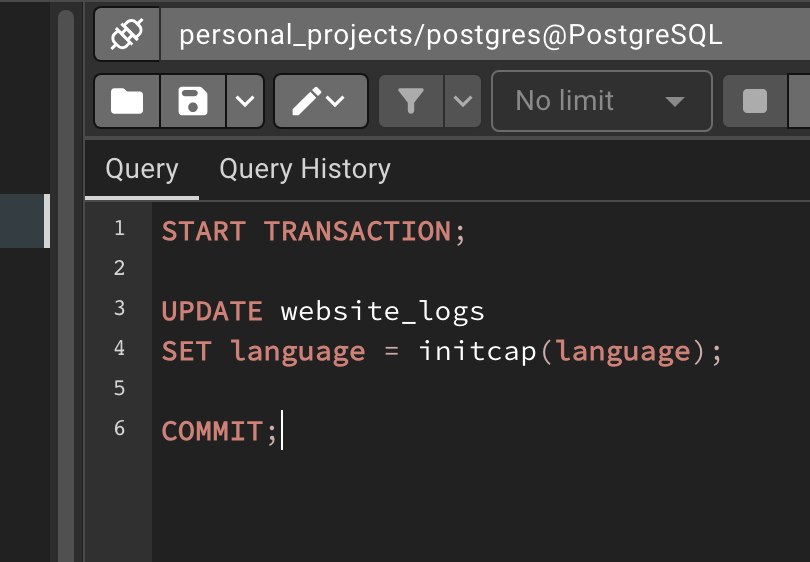


Output corrected and standardized with 88,124 values showing as ‘Unidentified’:

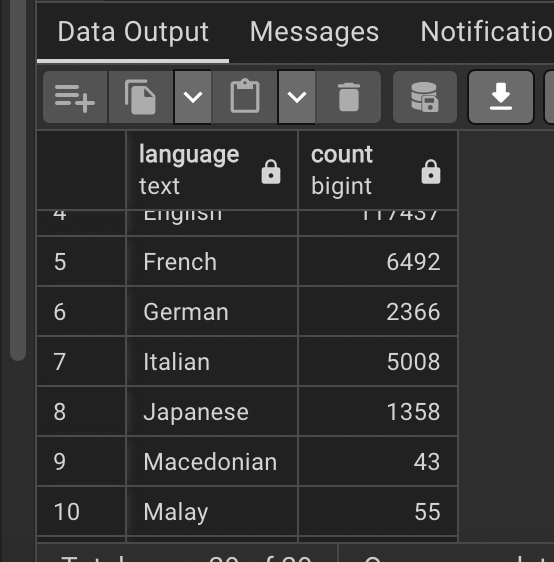


Step four: Turn rows with language in lowercase format to having the first letter uppercase. This will merge the double Italian values. Setting initcap to the entire column because there were other languages that came in as all lowercase. Those didn’t create duplicates but this way all the values will be standardized to one format.

Code for that:

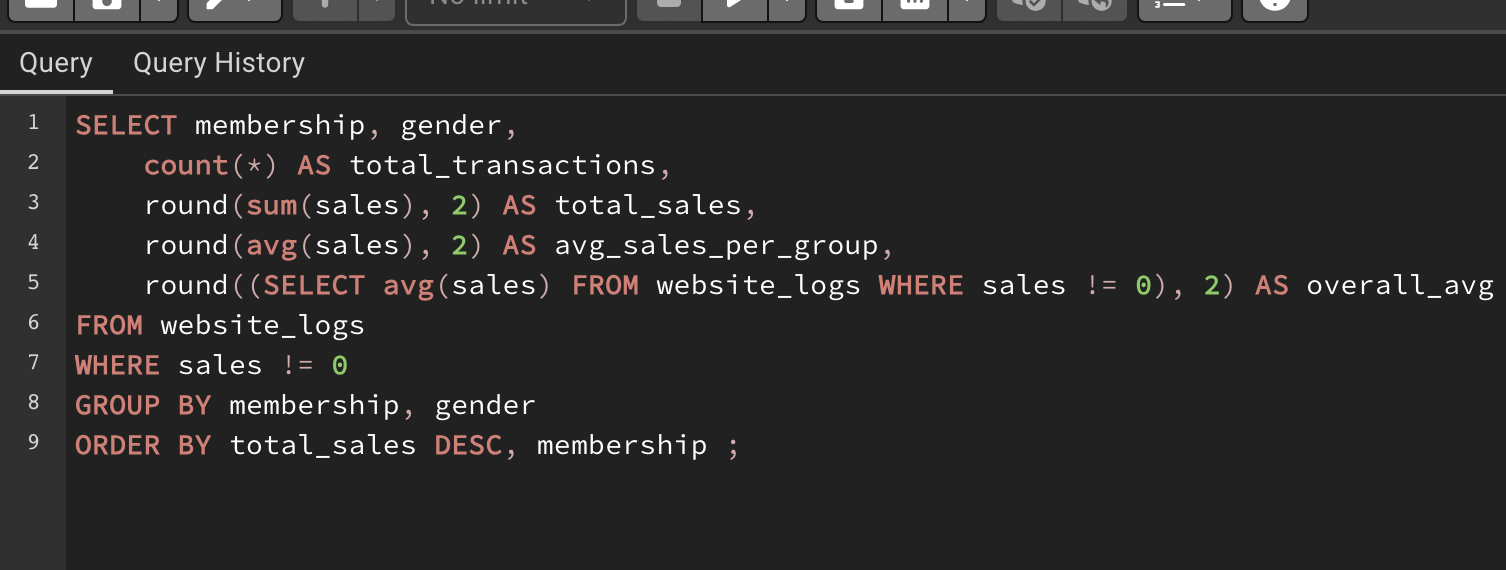


Output correctly merged and format standardized:

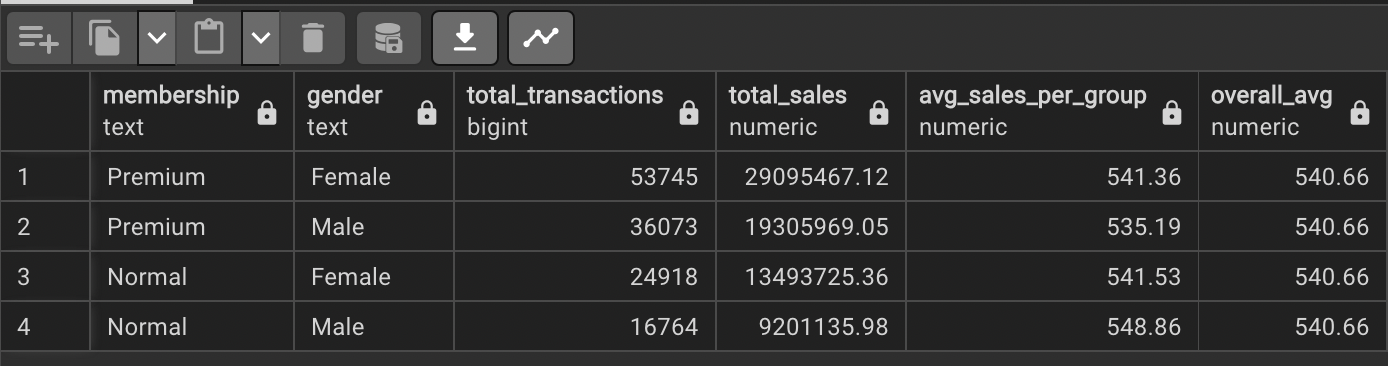


**EDA (Exploratory Data Analysis)**

1. Analysis of types of website membership (Premium and Normal) within each gender.

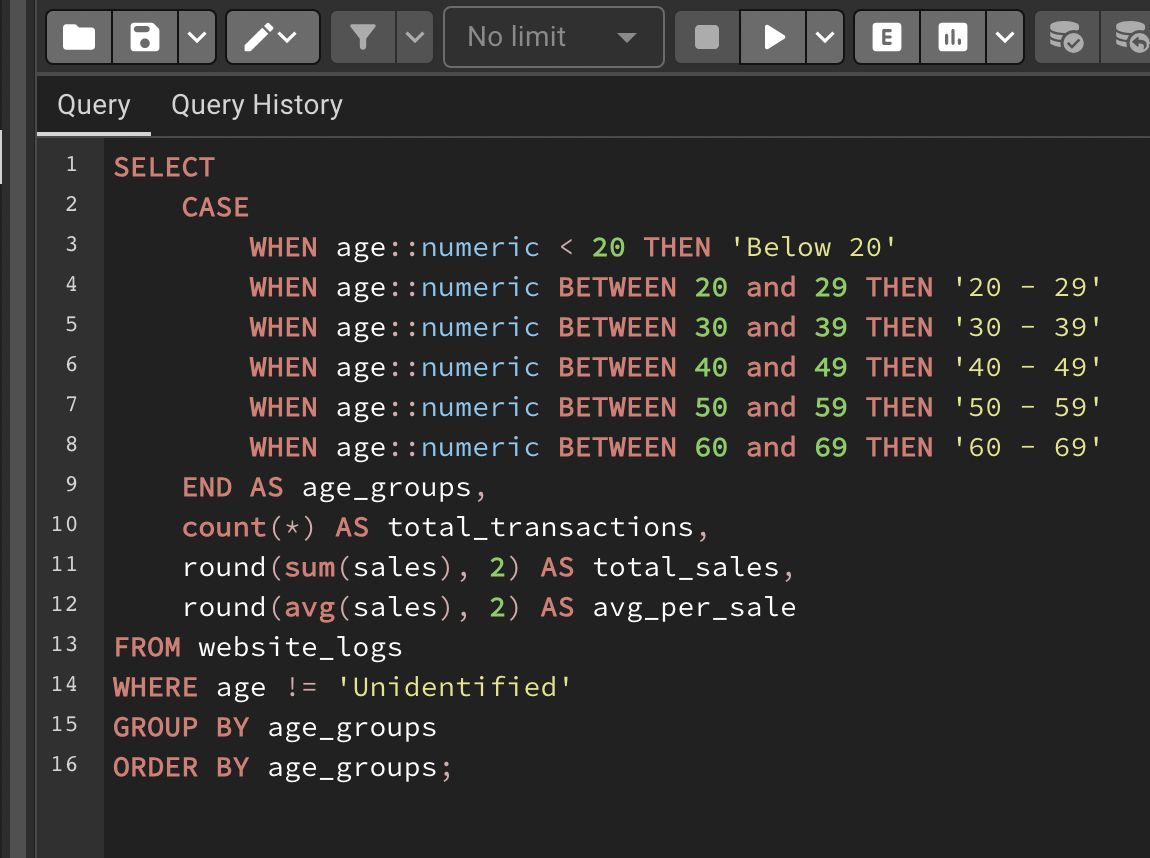


Results:

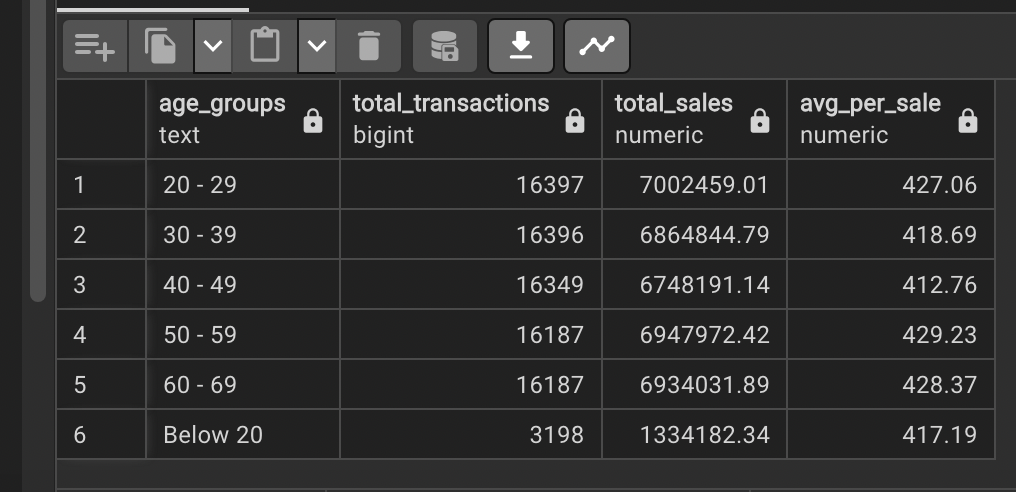


Takeaways:

1. Females with premium memberships had the highest amount of sales with a total of $29,095,467.12.
2. Males with a normal membership had the lowest in total sales ($9,201,135.98) yet they are leading in average amount spent per transaction. This group spent an average of $548.86 per sale which is the highest of all the groups and over $8 more than the overall average of $540.66 spent per transaction.
3. My recommendation for Sales/Marketing teams is that Female’s with premium memberships should be their target demographic. If they are looking to expand market share they should focus efforts on Males with normal memberships.
4. Exploring website activity by age groups.



Results:

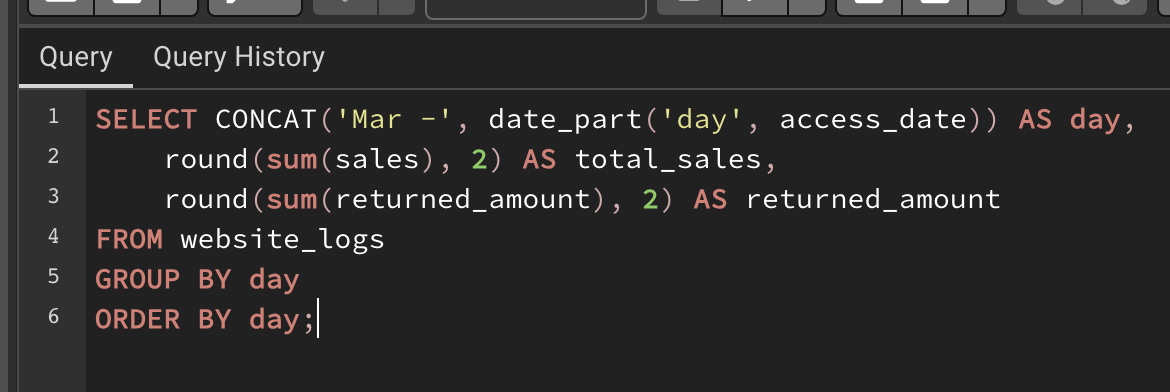


Takeaways:

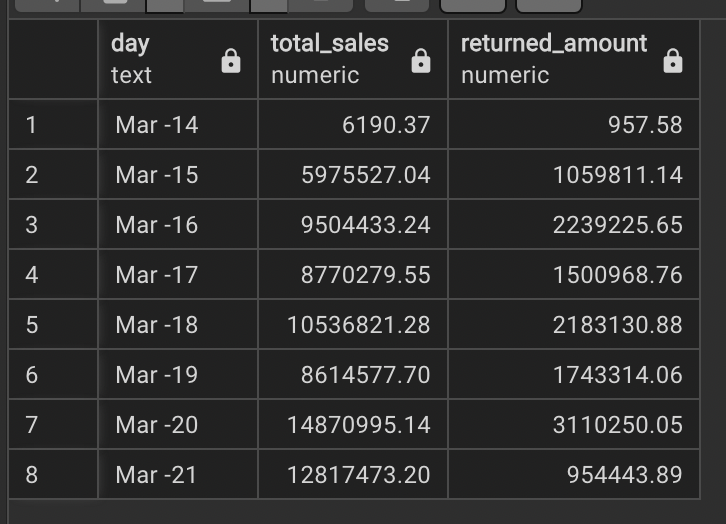
1. The total transactions per age group in this query are strikingly similar. I would imagine this is due to either A. this dataset was created by a Kaggle user who did not add diversity to the dataset in this respect or B. the product this website sells is a product sought after by all age groups.
2. People in their 30’s, 40’s, and below 20 are spending the lowest per transaction.
3. Exported to Excel to visualize sales per transaction among age group:

\*Clearly people in their 20’s, 50’s, and 60’s are the leaders in this category.

3. Exploring sales and returns activity by each day in this one week long dataset (3/14/2017 – 3/21/2017).



Results:



Visualized in Excel:

\*This would be an important visualization for Sales and Returns teams to have on hand. If the company had a visualization tool like Power Bi or Tableau ,it would be crucial to build a similar visualization within a dashboard to have this information to be constantly up to date for these teams.