ORGANIZING AND ANALYZING ONLINE ORDERS

GOALS

- Create a database which stores the information of the online orders
- Streamline the storage and minimize storage space
    - Modify the data within the SQL server instead of downloading and doing it in VSCode
- Add more data to create more insightful analysis, but also challenge myself more
- Find potential groups of consumers to expand into

DATA

- Data sourced from Brazilian E-Retailer olist (Details are anonymized, otherwise real)
    - Wanted to get real data so I could actually analyze and get real information for potential future uses unlike the Northwind Traders dataset
- Downloaded from Kaggle
- Over 100k orders

TOOLS

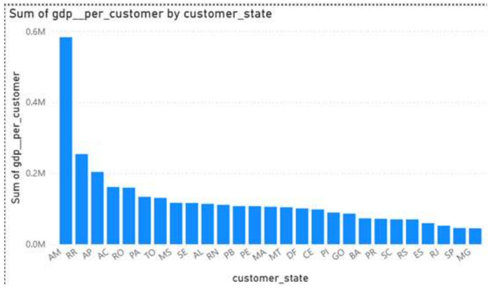
- PowerBI to analyze the data without having to manually write all the code, not to mention it gets all the relations right off the bat.
- DBSchema to manipulate database
- PostgresSQL because it's easiest DB system I had to deal with, especially concerning PowerBI connections
- Python for some data conversions, like converting table to linear variance

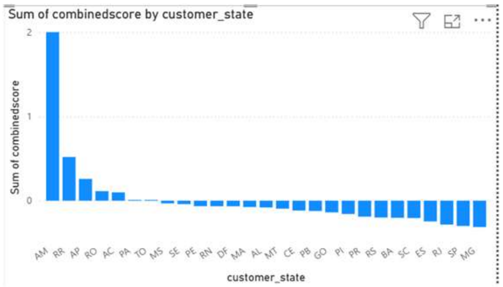
ANALYSIS: POTENTIAL EXPANSIONS

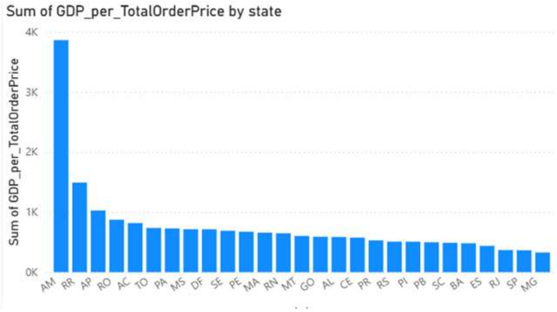
- One metric I'm using for potential expansions is using GDP per Customer(GDPCu), which indicates the amount of the economy is left on the table. Long story short, the higher the , the more potential reach there is into the market.
  - However the biggest shortfall is it doesn't account for population size, so a micro market with massive upside potential is not going to be as attractive
  - Doesn't account for factors as to *why* there even is such a high GDPCu
- Another metric I'm using is GDP per TotalOrderPrice(We'll refer to it as amount paid)(GDPAP), or the amount spent per GDP of a province. This one focuses more on the money itself than the customer base.
  - The shortfalls are also the same as GDPCu
- What I'm planning to do to narrow down which potential market is which is calculate those 2 values for each province, get their linear variance score, and add them up together for a final undersaturation score.

POTENTIAL EXPANSIONS (CONT)

- As we can see the most undersaturated markets based on the GDPCu and GDPAP metrics are provinces in the Amazon Region, which isn't exactly known for ease of transport. However, what's particularly interesting is the Amazonas(AM) Province is extremely higher than the rest.
- Provinces I think we should target; AM, DF, PE, and RN
- Why I didn't immediately choose the other top scorers in undersaturation is because small population size and Amazon factor

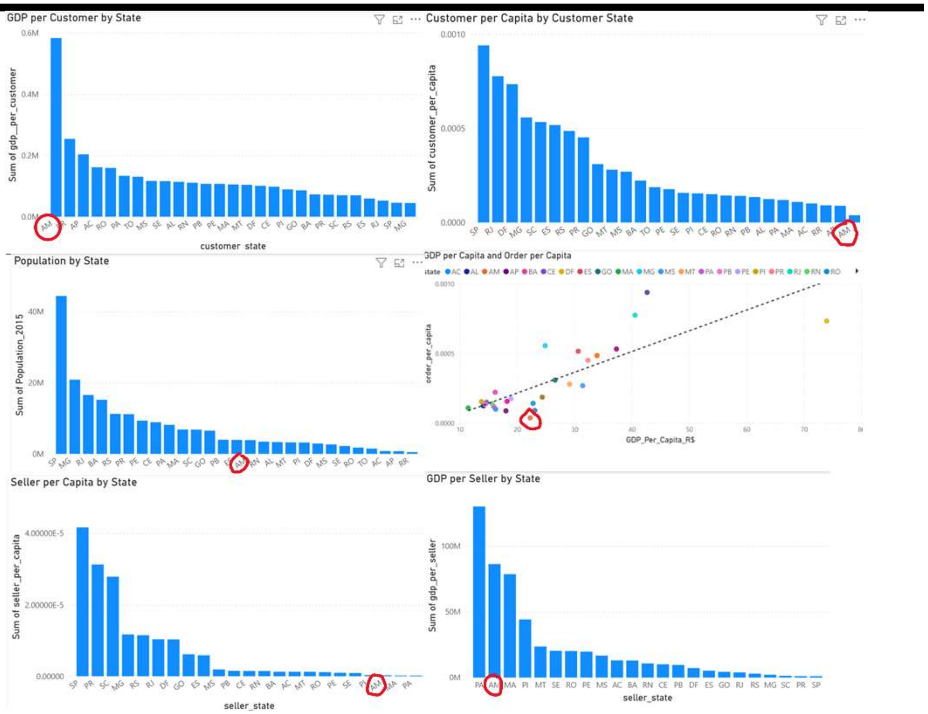


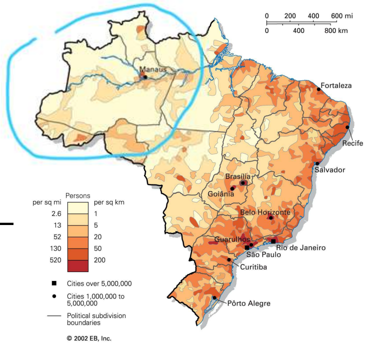




AMAZONAS (AM)

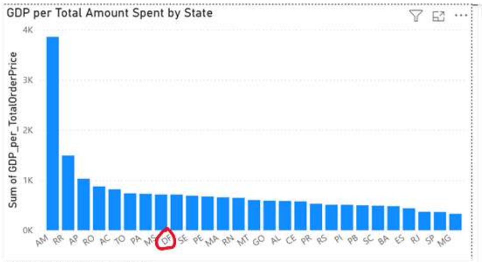
- Reasons for expansion:
  - By far largest GDPCu and GDPAP
  - By far lowest Customer per Capita
  - Half of population concentrated in city of Manaus
  - Population not tiny
  - Orders per Capita in relation to the GDP per capita is especially low
- Challenges
  - Because it’s an Amazon province, it’s infrastructure will be quite weak, which is especially a challenge for building distribution centers/
  - The concentration of sellers in Amazonas is abysmal; It’s bottom 3 in sellers per capita and top 3 in GDP per seller (GDPS)
- Moves:
  - Dump all focus on Manaus
  - Somehow incentivize more sellers in that region
  - Market more to attract customers

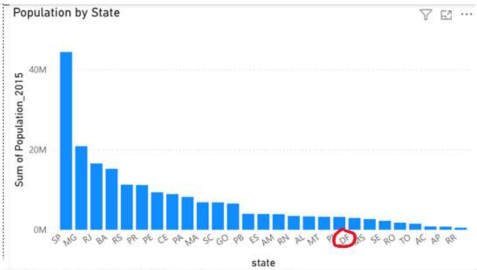


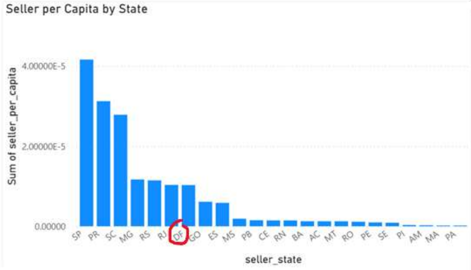


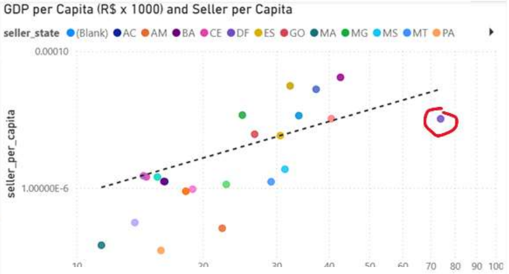
DISTRITO FEDERAL (DF)

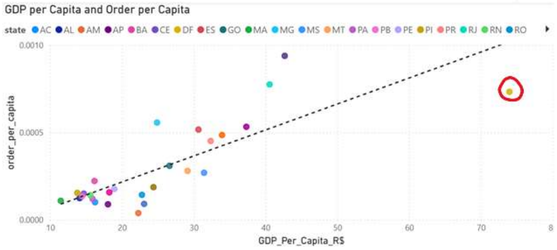
- Reasons for expansion:
  - Somewhat undersaturated when using GDPAP metric
  - It's a city, which means better infrastructure already provided
  - Moderate population size
  - Closer to main population centers of Brazil, so orders are easier to reach
  - Seller Per Capita relatively low to GDP per capita trend.
  - Orders per Capita in relation to GDP per capita is quite below the trend
- Challenges
  - Not much apparent
  - It's GDP per capita is an outlier, so it might end up bucking stat trends
- Moves:
  - Customer concentration is good, so awareness isn't necessarily the problem, perhaps we need to incentivize them to order more
  - Incentivize more sellers

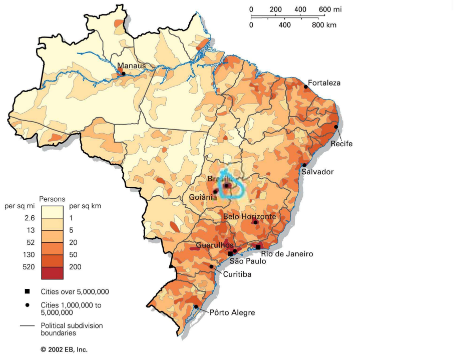

GDP per Total Amount Spent by State


Population by State


Seller per Capita by State

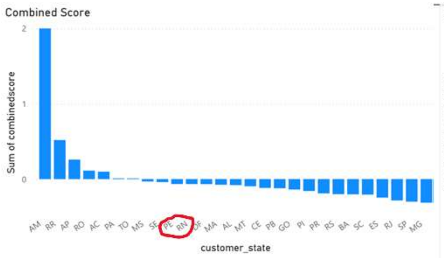

GDP per Capita (R$ x 1000) and Seller per Capita

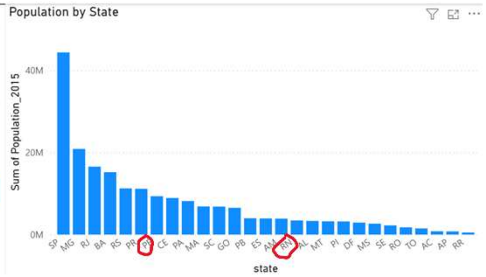

GDP per Capita and Order per Capita

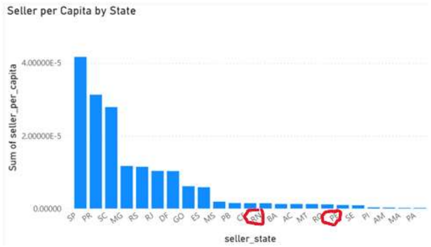


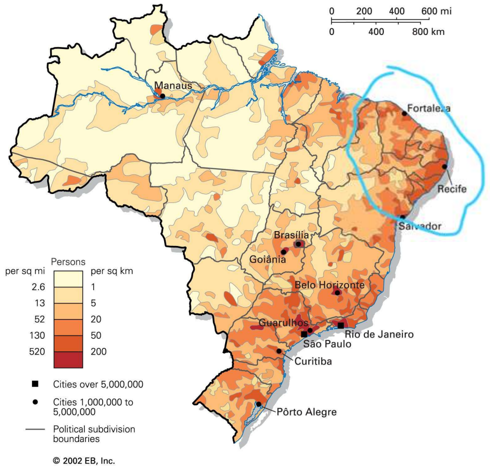
PERNAMBUCO (PM) AND RIO GRANDE DO NORTE (RN)

- Reasons for expansion:
  - Somewhat high undersaturation score
  - Large population
  - In populated region, so reach to everyone else is quite good
- Challenges:
  - Known to be poor region, so infrastructure likely to have shortfalls
  - Seller per capita in region pretty low
- Moves:
  - Incentivize more sellers in region


Combined Score

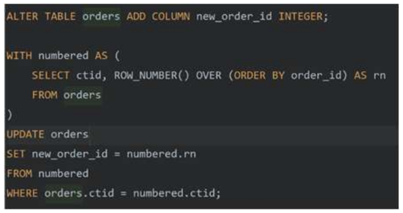

Population by State


Seller per Capita by State



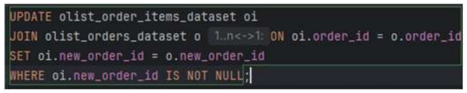
© 2002 EB, Inc.

HOW I MODIFIED THE DATA



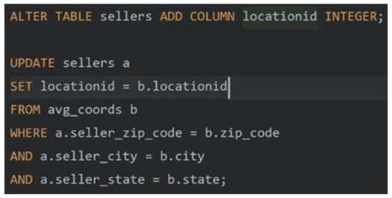
```sql
ALTER TABLE orders ADD COLUMN new_order_id INTEGER;

WITH numbered AS (
    SELECT ctid, ROW_NUMBER() OVER (ORDER BY order_id) AS rn
    FROM orders
)
UPDATE orders
SET new_order_id = numbered.rn
FROM numbered
WHERE orders.ctid = numbered.ctid;
```



```sql
UPDATE olist_order_items_dataset oi
JOIN olist_orders_dataset o 1..n<->1: ON oi.order_id = o.order_id
SET oi.new_order_id = o.new_order_id
WHERE oi.new_order_id IS NOT NULL;
```

- Simplified the IDs from a 20 something character long text to an integer for data compression
- Reduced the zip codes to integers for data compression
- Gave location ID for combinations of zip code, city, and state because I found some zip codes to be in more than one city. Took the location information from a geolocation table and put it in an avg_coords table to be a reference. As of now there isn't real use, I just wanted to give myself a challenge to play with data.
- Added state population and GDP tables for deeper analysis because it wasn't part of the olist dataset and those 2 data values are the most basic fundamentals for analyzing states and populations

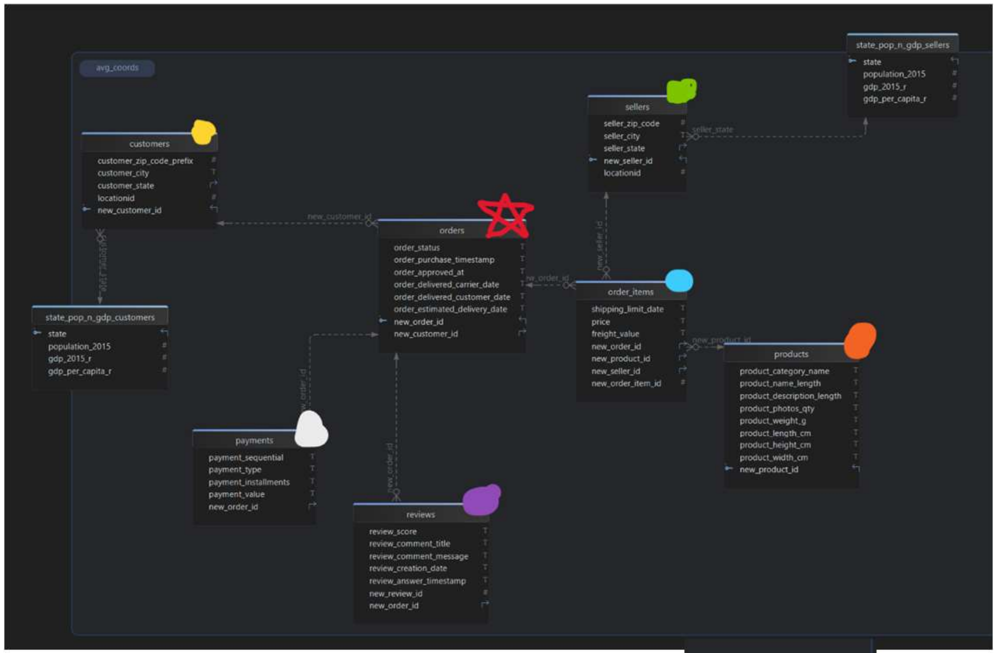


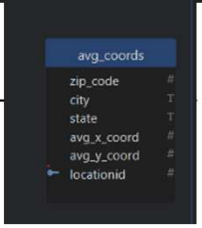
```sql
ALTER TABLE sellers ADD COLUMN locationid INTEGER;

UPDATE sellers a
SET locationid = b.locationid
FROM avg_coords b
WHERE a.seller_zip_code = b.zip_code
AND a.seller_city = b.city
AND a.seller_state = b.state;
```

DATA ORGANIZATION

- Orders are the core of this data (Red Star)
- Customers and Sellers have their own State GDP and Population tables to prevent data loop
- Sellers are connected to order items instead of orders because that is how the data started out as
- There is the Avg Coords table as reference for the locationIDs. No real use for it.





CHALLENGES AND SOLUTIONS TO THIS PROJECT

- PowerBI is a great tool to relate data and also create new tables on the fly, but it’s terrible at keeping things organized. Basically I ended up creating a lot of new tables, but the data pile was growing too large to be able to effectively keep track of.
    - I broke the project down into multiple PowerBI files as a workaround. Not great, but better than having this hard to organize data dump.
- Connecting to DB isn’t always reliable. I started off with using MySQL but long story short, I couldn’t connect it to PowerBI because the connector wouldn’t work, so I ended up switching to PostGres and didn’t have any major issues.