Introduction

The goal of our market research company is to help real estate agents and investors by providing data backed recommendation for identifying new investment opportunities by analyzing foot traffic data in respect to business density and demographic information to provide insights in different geographic locations. For the analysis here, the dataset is based on Safegraph, a company providing anonymized location data as a basis to track foot-traffic patterns through mobile phones. The dataset includes point-of-interest visits, their dwell time, the number of visits to specific POIs, and geographic regions in which these POIs are located. It incorporates external demographic information about median income and housing values to enrich the insights from the base data.

Since the firm is new, we mostly have clients from the restaurant industry. The first part of this research will focus on investment in new restaurant businesses (both Full service and Low Service Restaurants) and we will try to apply similar learning for grocery store category with an assumption that restaurants will source their raw material from the closest grocery stores.

Key Insights -

1. Which geographic areas exhibit a significant upward trend in foot traffic but have limited restaurant presence, as measured by NAICS industry classifications? (Refer Query 7 in Appendix)
2. Which geographic areas have a presence of restaurants but low count of grocery stores nearby?
3. How does foot traffic vary across different regions over time, and which areas have underutilized dwell times that could signal potential investment opportunities?
4. How do demographic characteristics (e.g., median income, housing value) correlate with traffic patterns, and which areas are underserved by restaurant businesses despite favorable demographic conditions?

Methodology

We have used techniques such as clustering and time-series analysis which will illustrate regions divided up based on growth rates of traffic, dwell time, and existence of businesses. The foot traffic patterns will be correlated with demographic trends, to identify the underserved regions that have the required amount of foot traffic and demographic trends to support new commercial investments. Additionally, we performed multiple exploratory data analysis and queries for the same can be found in the following section

Appendix

Query 1 – Query to fetch data for count of restaurants and places for a given CBG

```
SELECT v.poi_cbg,
COUNT(DISTINCT p.safegraph_place_id) AS num_places,
COUNT(DISTINCT p.safegraph_brand_ids) AS num_brands
FROM `mod-group-13-project.safegraph.visits` v
JOIN `mod-group-13-project.safegraph.places` p
ON v.safegraph_place_id = p.safegraph_place_id
GROUP BY v.poi_cbg
ORDER BY num_places DESC
```

Query 2 – Query to fetch data for foot traffic in different geographic areas (CBGs) over time

```
SELECT v.poi_cbg,
EXTRACT(YEAR FROM v.date_range_start) AS year,
EXTRACT(MONTH FROM v.date_range_start) AS month,
SUM(v.raw_visit_counts) AS total_visits
FROM `mod-group-13-project.safegraph.visits` v
GROUP BY v.poi_cbg, year, month
ORDER BY total_visits DESC
LIMIT 100;
```

Query 3 – Query to fetch the average dwell time in each CBG by examining the bucketed dwell time ranges.

```
WITH dwell_time_sums AS (
SELECT v.poi_cbg,
SUM(CAST(IFNULL(JSON_EXTRACT_SCALAR(REPLACE(v.bucketed_dwell_times, '"<5"',
'"less_than_5"'), '$.less_than_5'), '0') AS INT64)) AS dwell_time_under_5,
SUM(CAST(IFNULL(JSON_EXTRACT_SCALAR(REPLACE(v.bucketed_dwell_times, '"5-20"',
'"five_to_twenty"'), '$.five_to_twenty'), '0') AS INT64)) AS dwell_time_5_to_20,
SUM(CAST(IFNULL(JSON_EXTRACT_SCALAR(REPLACE(v.bucketed_dwell_times, '"21-60"',
'"twenty_one_to_sixty"'), '$.twenty_one_to_sixty'), '0') AS INT64)) AS
dwell_time_21_to_60,
SUM(CAST(IFNULL(JSON_EXTRACT_SCALAR(REPLACE(v.bucketed_dwell_times, '"61-240"',
'"sixty_to_twoforty"'), '$.sixty_to_twoforty'), '0') AS INT64)) AS
dwell_time_61_to_240,
SUM(CAST(IFNULL(JSON_EXTRACT_SCALAR(REPLACE(v.bucketed_dwell_times, '">240"',
'"greater_than_240"'), '$.greater_than_240'), '0') AS INT64)) AS dwell_time_above_240
FROM `mod-group-13-project.safegraph.visits` v
GROUP BY v.poi_cbg
)
```

```sql
SELECT poi_cbg,
(dwell_time_under_5 + dwell_time_5_to_20 + dwell_time_21_to_60 + dwell_time_61_to_240
+ dwell_time_above_240) AS total_dwell_time,
dwell_time_under_5, dwell_time_5_to_20, dwell_time_21_to_60, dwell_time_61_to_240,
dwell_time_above_240
FROM dwell_time_sums
ORDER BY total_dwell_time DESC;
```

Query 4 – This query fetches data for foot traffic variation throughout the day in different regions

```sql
SELECT v.poi_cbg,
EXTRACT(HOUR FROM v.date_range_start) AS hour,
ARRAY(SELECT value FROM UNNEST(SPLIT(REPLACE(JSON_EXTRACT(v.popularity_by_hour, '$'),
'[', ''), ']')) AS value) AS traffic_by_hour
FROM `mod-group-13-project.safegraph.visits` v
LIMIT 100;
```

Query 5 – This query fetched data to identify regions with high traffic but few businesses

```sql
WITH traffic_and_business AS (
SELECT v.poi_cbg,
SUM(v.raw_visit_counts) AS total_visits,
COUNT(DISTINCT p.safegraph_place_id) AS num_places
FROM `mod-group-13-project.safegraph.visits` v
JOIN `mod-group-13-project.safegraph.places` p
ON v.safegraph_place_id = p.safegraph_place_id
GROUP BY v.poi_cbg
)
SELECT poi_cbg, total_visits, num_places
FROM traffic_and_business
WHERE num_places < 5 -- Low business presence
ORDER BY total_visits DESC
LIMIT 100;
```

Query 6 – This query fetches data for conducting EDA for fetching restaurant datas and count of people (basis sex) visiting to these restaurants.

```sql
with b as (

select distinct safegraph_brand_id, brand_name, top_category, sub_category
FROM `mod-group-13-project.safegraph.brands` b
where sub_category in ('Full-Service Restaurants', 'Limited-Service Restaurants')
),
v as (
select safegraph_brand_ids, poi_cbg, sum(raw_visit_counts) as total_visits
```

```
from `mod-group-13-project.safegraph.visits` v
group by 1,2
),

d as (
select distinct cbg, sum(pop_m_total) as male, sum(pop_f_total) as female
from `mod-group-13-project.safegraph.cbg_demographics` d
group by 1
)

select d.cbg, b.brand_name, b.top_category, b.sub_category, v.total_visits, d.male,
d.female
from b join v on b.safegraph_brand_id = v.safegraph_brand_ids
join d on v.poi_cbg = d.cbg
```

Query 7 – Finalized Query which fetches the trend of traffic over different month and identifies areas with low business count but increase in traffic.

```
WITH traffic_trends AS (
-- Step 1: Calculate the trend in foot traffic over time for each CBG
SELECT p.safegraph_place_id,
v.poi_cbg,
v.date_range_start,
SUM(v.raw_visit_counts) AS total_visits,
EXTRACT(MONTH FROM v.date_range_start) AS month
FROM `mod-group-13-project.safegraph.visits` v
JOIN `mod-group-13-project.safegraph.places` p
ON v.safegraph_place_id = p.safegraph_place_id
GROUP BY p.safegraph_place_id, v.poi_cbg, month, v.date_range_start
),
business_density AS (
-- Step 2: Count the number of businesses in each geographic area (CBG) using the
NAICS code from brands
SELECT p.safegraph_place_id, v.poi_cbg,
COUNT(DISTINCT b.safegraph_brand_id) AS num_businesses
FROM `mod-group-13-project.safegraph.places` p
JOIN `mod-group-13-project.safegraph.brands` b
ON p.safegraph_brand_ids = b.safegraph_brand_id
JOIN `mod-group-13-project.safegraph.visits` v
ON p.safegraph_place_id = v.safegraph_place_id
GROUP BY p.safegraph_place_id, v.poi_cbg
),
recent_trends AS (
-- Step 3: Identify regions with increasing foot traffic
SELECT poi_cbg,
```

```
SUM(CASE WHEN month = 1 THEN total_visits ELSE 0 END) AS visits_Jan,
SUM(CASE WHEN month = 2 THEN total_visits ELSE 0 END) AS visits_Feb,
SUM(CASE WHEN month = 3 THEN total_visits ELSE 0 END) AS visits_Mar,
SUM(CASE WHEN month = 4 THEN total_visits ELSE 0 END) AS visits_Apr,
FROM traffic_trends
GROUP BY poi_cbg
HAVING visits_Apr > visits_Mar AND visits_Mar > visits_Feb and visits_Feb > visits_Jan
),
limited_business_areas AS (
-- Step 4: Filter regions with low business density (e.g., fewer than 5 businesses)
SELECT poi_cbg
FROM business_density
WHERE num_businesses < 5
)
-- Step 5: Combine results to get geographic areas with upward traffic trends but
limited business presence
SELECT distinct r.poi_cbg, r.visits_Jan, r.visits_Feb, r.visits_Mar, r.visits_Apr,
sum(bd.num_businesses)
FROM recent_trends r
JOIN limited_business_areas lba
ON r.poi_cbg = lba.poi_cbg
JOIN business_density bd
ON r.poi_cbg = bd.poi_cbg
group by 1,2,3,4,5
```