Final Report -
Optimal Locations and Strategy for a New Panda Express Competitor

Team 18

Team Members: Chi Wu, Hao-Chuan Lee

Intodcution:

As Chinese food lovers constantly in search of authentic Chinese fast food, we often find Panda Express underwhelming. Despite its lack of authenticity, the brand's widespread presence makes it difficult to overlook. This led us have the idea of creating a new Chinese fast-food franchise offering delicious and authentic Chinese cuisine, aimed at becoming a strong competitor to Panda Express. To ensure the success of our restaurant in the market, we would like to first conduct research on Panda Express to identify ideal locations for launching the business.

Research Questions:

1. **What are the top 10 counties with the highest "population-to-store" ratio?**

This ratio helps us determine which counties have a larger underserved customer base for each Panda Express location, indicating areas where there is a greater market potential for our restaurant in a less competitive environment.

- Method:

We first aggregate the total number of Panda Express locations and the total population within each county. We then calculate the population-to-restaurant ratio by dividing the total population by the number of Panda Express locations in each county. This ratio allows us to identify the top 10 counties with lower Panda Express density, indicating markets with less competition where new restaurants can potentially thrive.

- Result:

From the query result, the top 10 counties with the highest population-to-store ratio are Hamilton County, Middlesex County, Bronx County, Rogers County, Queens County, Westchester County, Kings County, Suffolk County, Lee County, and Bucks County, with four of these counties located in New York state. Detailed information can be found in the *Query Log A*.

2. **What are the top 10 counties with the highest customer growth?**

We analyzed Panda Express visit data from February to May (due to limited January data) to find the counties with the highest increase in visitors. These counties show significant growth in demand, suggesting strong market expansion opportunities

- Method:

We begin by examining the first and last months of recorded data for Panda Express in the visits table, which spans from January to May. We analyze the change in

visitor numbers between January-February and April-May to determine the growth trend of customers at Panda Express. Using these visitor counts, we identify the counties with the most significant positive growth. We chose to focus on raw visitor numbers instead of percentages because some counties had zero or one visitor during the January-February period, which could lead to misleading calculations and affect the accuracy of the results.

- Result:

From the query result in the query log result, the top 10 counties with the highest customer growth are Livingston Parish, Saline County, Butler County, Cullman County, Johnson County, Muskingum County, Onslow County, St. Joseph County, Washington County, Baldwin County. The detailed information could be find in the *Query Log B*.

3. **Find the difference between the top 10 ratio counties, the top 10 growth counties and the reference demography.**

After conducting the two analyses, we found that there were no overlapping counties in the results. To aid in our decision-making process, we selected income and age from the demographic table as additional comparison indices. This allows us to examine how the demographic profiles of these counties differ from those with high Panda Express density.

- Method:

We identified the average income and average age of counties with the lowest "population-to-store" ratio. These counties serve as a reference demographic since they have the highest density of Panda Express, suggesting that their population aligns well with the target market.(See *Query Log C*) Using this reference, we then calculated the income and age differences for the top 10 counties with the highest "population-to-store" ratio and the top 10 counties with the highest customer growth, allowing us to evaluate how these regions compare to the reference demographic profile.

- Result:

The top 10 counties experiencing an increase in customers share similar demographics with those that have a higher density of restaurants. These counties typically have younger populations and lower income levels. In contrast, counties with fewer Panda Express locations exhibit more distinct demographic differences compared to counties with a higher restaurant density, most of them are having elder and more wealthy population. The detailed result can be find in the *Query Log D*.

- Limitation:

After merging our trend table with the demographic information, we discovered discrepancies between these results and our previous query. Upon conducting a brief debugging session, we found that some counties lacked data in the demographic table (see *Query Log E*). This absence of data contributed to the discrepancies. Since demographic information is essential for our decision-making process, the latter part of this report will focus on the query that combines demographic data (*Query Log D*) to provide potential recommendations.

Conclusion:

Based on our research, we offer two strategic recommendations for our new restaurant. These recommendations present two distinct approaches for selecting new locations, each requiring unique brand strategies and market segmentation to ensure success in different market environments.

Recommendation A: Focus on Highest "Population-to-Store" Ratio

- Suggested Counties:
  Middlesex County, MA / Bronx County, NY / Rogers County, OK
- Advantage:
  The population-to-store rates indicated that there is a larger underserved market. These counties have the largest potential customer base with minimal competition, offering significant growth opportunities.
- Disadvantage:
  Demographics differ from successful Panda Express markets, suggesting lower familiarity with Chinese fast food. This might lead to the low acceptance of the new restaurants, making the potential market lower than expected and limiting the growth opportunities.
- Brand Strategy:
  We should focus on targeting higher-income, older customers by introducing premium ingredients and healthier menu options. This strategy would help differentiate the brand from Panda Express, appealing to a more health-conscious and wealthy customer segment.

Recommendation B: Focus on Highest Customer Growth

- Suggested Counties:
  Livingston Parish, LA / Butler County, MO / Johnson County, IA
- Advantage:
  These counties show strong existing demand for Chinese fast food. Additionally, these counties show positive visitors growth in Panda Express, which potentially implies a growing demand for Chinese fast food in the future.
- Disadvantage:
  The primary challenge is the high level of competition from Panda Expresses They already have brand recognition, customer loyalty, and market presence, making it harder to immediately break into the market.
- Brand Strategy:
  Since we are targeting a demographic similar to Panda Express' customer base, maintaining a comparable price point and style that appeals to a younger and relatively low income audience is essential. However, to capture more market share in these locations, we need to focus on building a strong brand image and creating a distinctive dine-in experience. By offering exceptional service, a unique atmosphere, and personalized touches, we can set ourselves apart from Panda Express in this competitive landscape. Additionally, engaging with the local community and leveraging technology for convenience—such as streamlined online ordering and loyalty programs—can further enhance customer loyalty and attract new patrons.

Query Log

A. Query for Question 1(Use Ratio As Base)

```
--Find County with highest population/restaurant ratio
WITH county AS (
  SELECT --Find How many restaurants in one county (Only based on data with visit
record)
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
    COUNT(DISTINCT street_address)--Avoid calculate the duplicated dataset
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
),

cbg_population AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,
    SUM(pop_total) AS total_population
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips
)
-- Select top ten ratio. These counties show that there is not many Panda Express
presence.
SELECT
  cf.state,
  cf.county,
  cp.total_population / c.numbers AS ratio
FROM county AS c
JOIN cbg_population AS cp ON c.full_fips = cp.full_fips
JOIN `team18-fa24-mgmt58200-final.team18data.cbg_fips` AS cf
  ON c.full_fips = CONCAT(cf.state_fips, cf.county_fips)
ORDER BY ratio DESC
LIMIT 10;
```

state	county	ratio
MA	Middlesex County	1,605,899
NY	Bronx County	1,427,056
OK	Rogers County	1,426,790
NY	Queens County	1,135,488
NY	Westchester County	968,738
NY	Kings County	858,923.67
OH	Hamilton County	815,790
MA	Suffolk County	801,162
FL	Lee County	756,570
PA	Bucks County	627,668

B. Query for Question 2(Use Visitor Growth As Base)

```
WITH county AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
     COUNT(DISTINCT street_address) AS numbers
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
```

```
),
--Find the first month of panda express' total visitors
-- There is no data before Jan ~ Feb
feb_visits AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
    SUM(raw_visitor_counts) AS feb_visitors
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE FORMAT_TIMESTAMP("%Y-%m", date_range_end) = '2020-02' --we use
date_range_end to group the total visitors
    AND safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
),
--Find the last month of panda express' total visitors
may_visits AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
    SUM(raw_visitor_counts) AS may_visitors
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE FORMAT_TIMESTAMP("%Y-%m", date_range_end) = '2020-05'
    AND safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
)
--Order by county where Panda Expresses are experiencing most increase in their
number of visitor
SELECT
  cf.state, cf.county,
  COALESCE(feb_visits.feb_visitors, 0) AS jan_to_feb_visitors,
  COALESCE(may_visits.may_visitors, 0) AS apr_to_may_visitors,
  COALESCE(may_visits.may_visitors, 0) - COALESCE(feb_visits.feb_visitors, 0) AS
visitor_difference,
FROM county AS c
JOIN `team18-fa24-mgmt58200-final.team18data.cbg_fips` AS cf
  ON c.full_fips = CONCAT(cf.state_fips, cf.county_fips)
LEFT JOIN feb_visits ON c.full_fips = feb_visits.full_fips
LEFT JOIN may_visits ON c.full_fips = may_visits.full_fips
GROUP BY cf.state, cf.county, c.numbers, feb_visits.feb_visitors,
may_visits.may_visitors
ORDER BY visitor_difference DESC
LIMIT 10;
```

state	county	jan_to_feb_visitors	apr_to_may_visitors	visitor_difference
LA	Livingston Parish	106	656	550
AR	Saline County	547	701	154
MO	Butler County	222	370	148
AL	Cullman County	632	778	146
IA	Johnson County	245	371	126
OH	Muskingum County	243	358	115
NC	Onslow County	547	648	101
IN	St. Joseph County	840	940	100
OK	Washington County	390	476	86
AL	Baldwin County	395	477	82

C. Query for Creating target demographic profile:

```sql
WITH county AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
     COUNT(DISTINCT street_address) AS numbers
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
),
--income index
income AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,
    SUM(
      (inc_lt10 * 7.5) +   -- Use midpoint to calculate the total income number
      (`inc_10-15` * 12.5) +
      (`inc_15-20` * 17.5) +
      (`inc_20-25` * 22.5) +
      (`inc_25-30` * 27.5) +
      (`inc_30-35` * 32.5) +
      (`inc_35-40` * 37.5) +
      (`inc_40-45` * 42.5) +
      (`inc_45-50` * 47.5) +
      (`inc_50-60` * 55) +
      (`inc_60-75` * 67.5) +
      (`inc_75-100` * 87.5) +
      (`inc_100-125` * 112.5) +
      (`inc_125-150` * 137.5) +
      (`inc_150-200` * 175) +
      (inc_gte200 * 200)  -- Assuming $200k for >200k range
    ) / NULLIF(inc_total, 0) AS average_income  -- Use `inc_total` as the
denominator
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips, inc_total
),
--age index
age AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,
    SUM(
      (`pop_m_10-14` * 12) +  --Use the midpoint to calculate the total age number
      (`pop_m_15-17` * 16) +
      (`pop_m_18-19` * 18.5) +
      (pop_m_20 * 20) +
      (pop_m_21 * 21) +
      (`pop_m_22-24` * 23) +
      (`pop_m_25-29` * 27) +
      (`pop_m_30-34` * 32) +
      (`pop_m_35-39` * 37) +
      (`pop_m_40-44` * 42) +
      (`pop_m_45-49` * 47) +
      (`pop_m_50-54` * 52) +
      (`pop_m_55-59` * 57) +
      (`pop_m_60-61` * 60.5) +
      (`pop_m_62-64` * 63) +
      (`pop_m_65-66` * 65.5) +
      (`pop_m_67-69` * 68) +
      (`pop_m_70-74` * 72) +
      (`pop_m_75-79` * 77) +
      (`pop_m_80-84` * 82) +
      (pop_m_gte85 * 90) +
      (`pop_f_lt5` * 2.5) +
      (`pop_f_5-9` * 7) +
      (`pop_f_10-14` * 12) +
      (`pop_f_15-17` * 16) +
      (`pop_f_18-19` * 18.5) +
      (pop_f_20 * 20) +
      (pop_f_21 * 21) +
      (`pop_f_22-24` * 23) +
```

```
      (`pop_f_25-29` * 27) +
      (`pop_f_30-34` * 32) +
      (`pop_f_35-39` * 37) +
      (`pop_f_40-44` * 42) +
      (`pop_f_45-49` * 47) +
      (`pop_f_50-54` * 52) +
      (`pop_f_55-59` * 57) +
      (`pop_f_60-61` * 60.5) +
      (`pop_f_62-64` * 63) +
      (`pop_f_65-66` * 65.5) +
      (`pop_f_67-69` * 68) +
      (`pop_f_70-74` * 72) +
      (`pop_f_75-79` * 77) +
      (`pop_f_80-84` * 82) +
      (pop_f_gte85 * 90)
    ) / NULLIF(pop_total, 0) AS average_age
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips, pop_total
),


cbg_population AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,
    SUM(pop_total) AS total_population
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips
),

--Use the average of the top ten counties with lowest population/restaurant ratio
(which mean the restaurant is more dense) as the current Panda Express customer
demographic profolio
dem_profile AS (SELECT
  AVG(inc.average_income) AS avg_income_per_county,
  AVG(a.average_age) AS avg_age_per_county,
  cf.state, cf.county,
  cp.total_population / c.numbers AS ratio
FROM county AS c
JOIN cbg_population AS cp ON c.full_fips = cp.full_fips
JOIN `team18-fa24-mgmt58200-final.team18data.cbg_fips` AS cf
  ON c.full_fips = CONCAT(cf.state_fips, cf.county_fips)
JOIN income AS inc
ON cp.full_fips = inc.full_fips
JOIN age AS a
ON cp.full_fips = a.full_fips
GROUP BY cf.state, cf.county, cp.total_population, c.numbers
ORDER BY ratio
LIMIT 10
)

--Find average number of these counties as the demographic profile
SELECT AVG(avg_income_per_county) AS income_index, AVG(avg_age_per_county) AS
age_index
FROM dem_profile
```

income_index	age_index
61.22945147	38.84621705

D. Combined Query for an overall review

```
WITH county AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
    COUNT(DISTINCT street_address) AS numbers
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
```

```sql
  GROUP BY full_fips
),
--income index
income AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,  -- Extract county FIPS code
    SUM(
      (inc_lt10 * 7.5) +   -- Use midpoint to calculate the total income number
      (`inc_10-15` * 12.5) +
      (`inc_15-20` * 17.5) +
      (`inc_20-25` * 22.5) +
      (`inc_25-30` * 27.5) +
      (`inc_30-35` * 32.5) +
      (`inc_35-40` * 37.5) +
      (`inc_40-45` * 42.5) +
      (`inc_45-50` * 47.5) +
      (`inc_50-60` * 55) +
      (`inc_60-75` * 67.5) +
      (`inc_75-100` * 87.5) +
      (`inc_100-125` * 112.5) +
      (`inc_125-150` * 137.5) +
      (`inc_150-200` * 175) +
      (inc_gte200 * 200)  -- Assuming $200k for >200k range
    ) / NULLIF(inc_total, 0) AS average_income
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips, inc_total
),
--age index
age AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,
    SUM(
      (`pop_m_10-14` * 12) +
      (`pop_m_15-17` * 16) +
      (`pop_m_18-19` * 18.5) +
      (pop_m_20 * 20) +
      (pop_m_21 * 21) +
      (`pop_m_22-24` * 23) +
      (`pop_m_25-29` * 27) +
      (`pop_m_30-34` * 32) +
      (`pop_m_35-39` * 37) +
      (`pop_m_40-44` * 42) +
      (`pop_m_45-49` * 47) +
      (`pop_m_50-54` * 52) +
      (`pop_m_55-59` * 57) +
      (`pop_m_60-61` * 60.5) +
      (`pop_m_62-64` * 63) +
      (`pop_m_65-66` * 65.5) +
      (`pop_m_67-69` * 68) +
      (`pop_m_70-74` * 72) +
      (`pop_m_75-79` * 77) +
      (`pop_m_80-84` * 82) +
      (pop_m_gte85 * 90) +
      (`pop_f_lt5` * 2.5) +
      (`pop_f_5-9` * 7) +
      (`pop_f_10-14` * 12) +
      (`pop_f_15-17` * 16) +
      (`pop_f_18-19` * 18.5) +
      (pop_f_20 * 20) +
      (pop_f_21 * 21) +
      (`pop_f_22-24` * 23) +
      (`pop_f_25-29` * 27) +
      (`pop_f_30-34` * 32) +
      (`pop_f_35-39` * 37) +
      (`pop_f_40-44` * 42) +
      (`pop_f_45-49` * 47) +
      (`pop_f_50-54` * 52) +
      (`pop_f_55-59` * 57) +
```

```
      (`pop_f_60-61` * 60.5) +
      (`pop_f_62-64` * 63) +
      (`pop_f_65-66` * 65.5) +
      (`pop_f_67-69` * 68) +
      (`pop_f_70-74` * 72) +
      (`pop_f_75-79` * 77) +
      (`pop_f_80-84` * 82) +
      (pop_f_gte85 * 90)
    ) / NULLIF(pop_total, 0) AS average_age
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips, pop_total
),
cbg_population AS (
  SELECT
    SUBSTRING(cbg, 1, 5) AS full_fips,
    SUM(pop_total) AS total_population
  FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
  GROUP BY full_fips
),
-- there is no visit data in jan
feb_visits AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
    SUM(raw_visitor_counts) AS feb_visitors
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE FORMAT_TIMESTAMP("%Y-%m", date_range_end) = '2020-02'
    AND safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
),
may_visits AS (
  SELECT
    SUBSTRING(poi_cbg, 1, 5) AS full_fips,
    SUM(raw_visitor_counts) AS may_visitors
  FROM `team18-fa24-mgmt58200-final.team18data.visits`
  WHERE FORMAT_TIMESTAMP("%Y-%m", date_range_end) = '2020-05'
    AND safegraph_brand_ids = "SG_BRAND_4a453a402f10c481d2fe4cefaa1d2b09"
  GROUP BY full_fips
)

SELECT
  cf.state, cf.county,
  COALESCE(feb_visits.feb_visitors, 0) AS feb_visitors,
  COALESCE(may_visits.may_visitors, 0) AS may_visitors,
  COALESCE(may_visits.may_visitors, 0) - COALESCE(feb_visits.feb_visitors, 0) AS
visitor_difference,
  cp.total_population / c.numbers AS ratio,
  AVG(inc.average_income) - 61.2294514666073 AS income_difference,-- the number of
61.2294514666073 is from our demographic profile query
  AVG(a.average_age)-38.846217052179739 AS age_difference,-- the number of
38.846217052179739 is from our demographic profile query
  c.numbers
FROM county AS c
JOIN cbg_population AS cp ON c.full_fips = cp.full_fips
JOIN `team18-fa24-mgmt58200-final.team18data.cbg_fips` AS cf
  ON c.full_fips = CONCAT(cf.state_fips, cf.county_fips)
JOIN income AS inc ON cp.full_fips = inc.full_fips
JOIN age AS a
ON cp.full_fips = a.full_fips
LEFT JOIN feb_visits ON c.full_fips = feb_visits.full_fips
LEFT JOIN may_visits ON c.full_fips = may_visits.full_fips
GROUP BY cf.state, cf.county, cp.total_population, c.numbers,
feb_visits.feb_visitors, may_visits.may_visitors
ORDER BY visitor_difference DESC -- by changing the ordering method, we will get
two different overall result for review
LIMIT 10;
```

The difference of demographic index (Use Ratio as Base)

state	county	ratio	income_difference	age_difference
MA	Middlesex County	1,605,899	154.8462118	13.74625275
NY	Bronx County	1,427,056	54.63513168	10.18690685
OK	Rogers County	1,426,790	63.77799574	4.503239081
NY	Queens County	1,135,488	155.9728818	20.59773946
NY	Westchester County	968,738	119.0595791	10.72303199
NY	Kings County	858,923.67	171.3543648	23.66771472
OH	Hamilton County	815,790	52.32297642	6.766679724
MA	Suffolk County	801,162	71.6133783	6.644517535
FL	Lee County	756,570	41.20566194	16.50089136
PA	Bucks County	627,668	68.21762264	7.066572011

The difference of demographic index (Use Trend as Base)

state	county	feb_visitors	may_visitors	visitor_difference	income_difference	age_difference
LA	Livingston Parish	106	656	550	20.60624119	-0.748756726
MO	Butler County	222	370	148	-7.731486731	2.418108135
IA	Johnson County	245	371	126	19.78021479	-2.448621818
OH	Muskingum County	243	358	115	-0.1254211443	2.89570086
NC	Onslow County	547	648	101	1.459430837	-5.841813128
IN	St. Joseph County	840	940	100	13.86560061	1.81810254
OK	Washington County	390	476	86	11.62755686	0.8907667729
MO	Cass County	361	441	80	28.42808225	2.761800131
IL	Winnebago County	630	694	64	16.18753068	4.238103853
TN	Loudon County	464	528	64	14.11241949	5.917028539

E. Checking the reason for discrepancy:

```
--Find fibs of one of the missing county in the combined query result
SELECT *
FROM `team18-fa24-mgmt58200-final.team18data.cbg_fips`
WHERE county LIKE "Baldwin County"
AND state LIKE "AL"
--Check if there is any data in the cbg table
SELECT *
FROM `team18-fa24-mgmt58200-final.team18data.cbg_demographics`
WHERE cbg LIKE '01003%';
```

There is no data to display.