PROJECT REPORT

PROJECT 2 - FOOD SERVICE DATA

Presented By

MOHD SAQUIB KHAN

Food Service Data Project 2

The spreadsheet has 3 tabs, that have the following:

- Total spend of his customers (restaurants) across various F&B categories, along with % share of purchases from Shaun's company
- Data about non-commercial establishments such as hospitals, schools across different geographic areas.
- Retail outlets spread across similar geographies.

Business Problem - "Opportunities Analysis" need to be determined. Sales Director has an annual conference coming up in less than a week to prepare for, but wanted to get this started without delay. In short, "it should the tell a story around where are our biggest opportunities"

What to show in the Analysis:

- Get a quick view of share of total spends and F&B spends for existing customers (restaurants)
- What are the adjacent opportunities from the non-commercial establishments?
- Get a view of retailer presence adjacent to his customers & prospects will help understand how much of these opportunities are being met / unmet.

Sheet 1 - Label Encoding used in the Dataset are as follows:

- **Units** = 501+ UNITS' = 3,
  'INDEPENDENT (1-9 UNITS)' = 5,
  '251-500 UNITS' = 2,
  '101-250 UNITS' = 1,
  '51-100 UNITS' = 4,
  '10-50 UNITS' = 0

- **No of emp** = 'OVER 50' = 4,
  '1 TO 4' = 0,
  '20 TO 49' = 2,
  '10 TO 19' = 1,
  '5 TO 9' = 3

- **Annual Sales** = $1,000,001 - $2,500,000' = 2,
  '<=$500,000' = 0,
  '$500,001 - $1,000,000' = 1
  ,'$2,500,001 - $5,000,000' = 3,
  '>$5,000,000' = 4

- **Year in Business =** (2 to 5 = 0),
  (5 yr. plus = 1)

Sheet 1 - Dataset after removing all Null-Values looked like this:

```
Int64Index: 888 entries, 0 to 999
Data columns (total 20 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   NAME                                   888 non-null    object
 1   CITY                                   888 non-null    object
 2   STATE                                  888 non-null    object
 3   ZIP                                    888 non-null    int64
 4   COUNTY_CODE                            888 non-null    float64
 5   COUNTY_NAME                            888 non-null    object
 6   SIMPLIFIED_MENU                        888 non-null    object
 7   UNITS                                  888 non-null    int32
 8   EMPLOYEES                              888 non-null    object
 9   YEARS_IN_BUSINESS                      888 non-null    object
 10  ANNUAL_SALES                           888 non-null    object
 11  AVERAGE_CHECK                          888 non-null    object
 12  FOOD_BEVERAGE_PURCHASE                 888 non-null    float64
 13  FOOD_PURCHASE                          888 non-null    float64
 14  BEVERAGE_PURCHASE                      888 non-null    float64
 15  TOTAL_CATEGORY_DRY                     888 non-null    float64
 16  TOTAL_CATEGORY_FRESH                   888 non-null    float64
 17  TOTAL_CATEGORY_FROZEN                  888 non-null    float64
 18  TOTAL_CATEGORY_PROTEIN                 888 non-null    float64
 19  % Share of F&B Purchases from Company  888 non-null    float64
dtypes: float64(9), int32(1), int64(1), object(9)
memory usage: 142.2+ KB
```

Exploratory Data Analysis using Python – Pandas Library:

1. We found that the Employee columns had a “Unknown” Value in unique counts. Which cannot be used for analysis.
2. The UNKNOWN employee rows have (5) Label Encoding in UNITS column, therefore output the ones whose units = 5
3. We gave a variable name (un5) to all the restaurants with Unit = 5 as per LE.
4. By using value_count () function we got to know that 5 TO 9 was the top count for UNITs = 5(1-9Units)
5. Therefore, replacing UNKNOWN Employee row value with (5 to 9).

6. Performed Label encoding to the no of Employee columns.

7. Label encoding the Year in business column for better statistical findings
   # 2 to 5 = 0
   # 5 yr. plus = 1

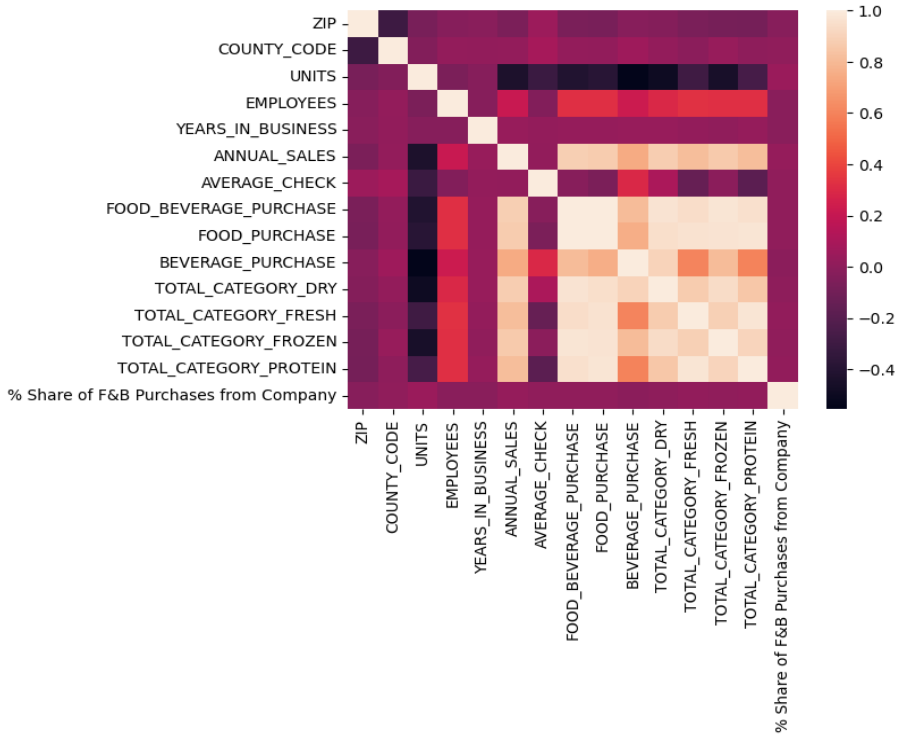
8. Replaced annual sales value to numeric type values rather than choosing Label Encoding, I replaced value by self for better understanding.

9. In the Average-Check column, found the UNCODED value, replaced UNCODED with null and deleting the null row.

10. AS all the Data Manipulation as well as Data cleaning operation were completed.

11. Those data which were converted to (numeric format) using Label Encoding can now be easily used for statistical analysis

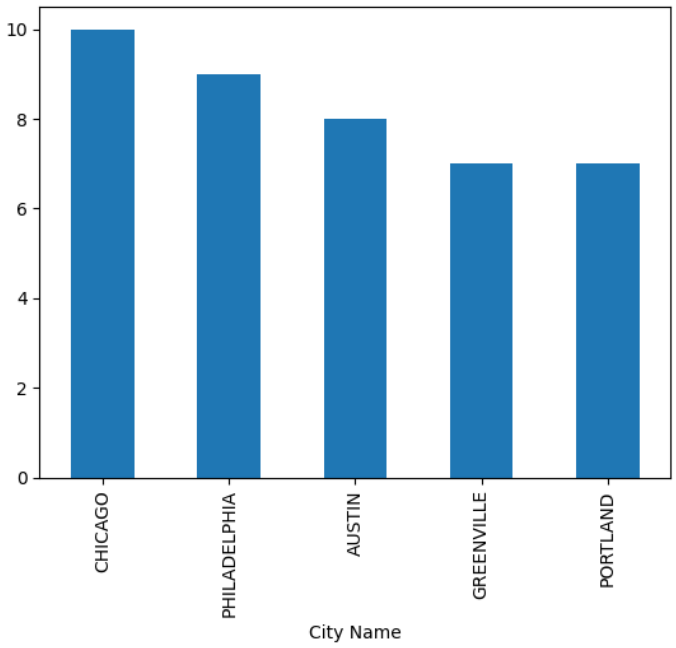
The Correlation Heatmap can be seen below:



Questions self-raised from the Data Base: Sheet 1

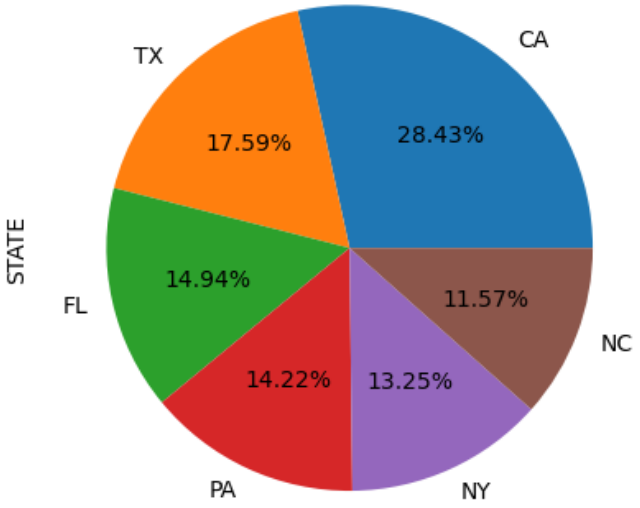
1 - Which City has most (restaurants) / Customers?

- The greatest number of customers are from Chicago city and then Philadelphia.



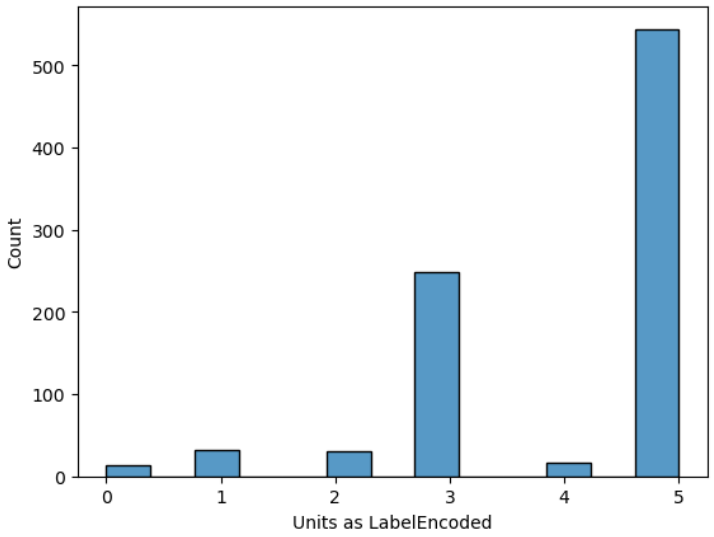
2 - Which State has most customers(restaurants)?

- 1st CA , 2nd Texas



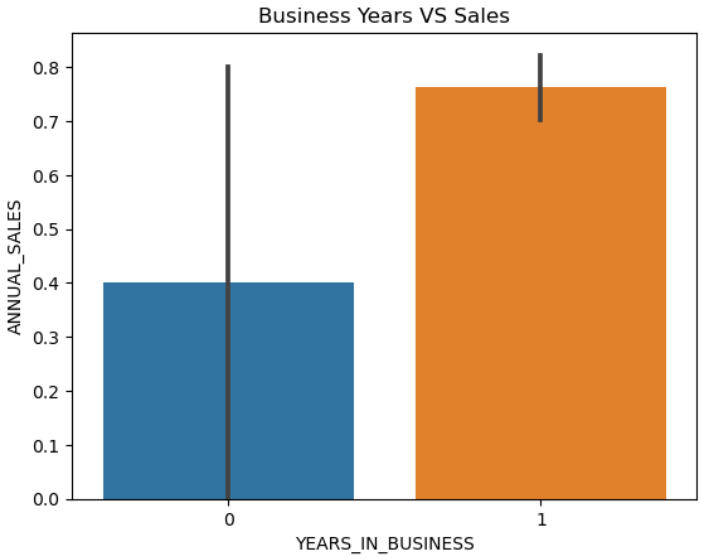
3 - What UNITS are most among the restaurants?

- We can see that most of the Restaurants have - INDEPENDENT Units (1-9) according to Label Encoding Value = (5)



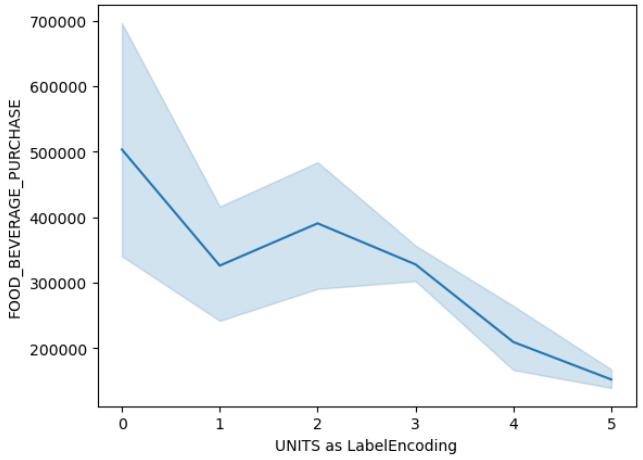
4 - Who performs better in Annual Sales with respect to (Years in business)?

- As shown in the above bar plot, the restaurants which are in business since (2-5) years are having more number of annual sales, than those in business for more than 5 years



5 - Which Units have most F&B Purchases?

- We can easily see the above line plot and determine, that those units - 0('10-50 UNITS') are having higher purchases
- There was an increase in purchase by Units - 2(251-500 UNITS), and then purchase gradually decreases.
- Its shocking to see the Restaurants having units - 3(500+ units) are having lower purchases than (10-50 units)



6 - What % share of F&B purchases is the max for the company's customer restaurants? And which restaurant has this max % share?

- Used Describe code - df ["% Share of F&B Purchases from Company"].describe()
- count 887.000000
- mean 0.486774
- std 0.288602
- min 0.001300
- 25% 0.231484
- 50% 0.489578
- 75% 0.736600
- max 0.999887

  Name: % Share of F&B Purchases from Company, dtype: float64

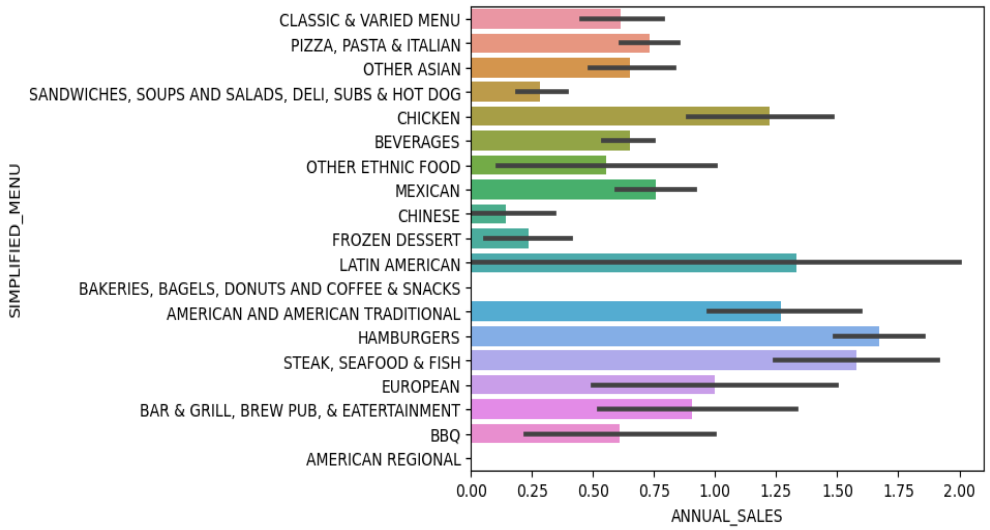
df["% Share of F&B Purchases from Company"].max()

0.9998868678449319 - This was the maximum %share of (CHINA BISTRO) restaurant in Danville City

- We can see that "China Bistro" has the most percent share of F&B purchases from Shaun's Company.

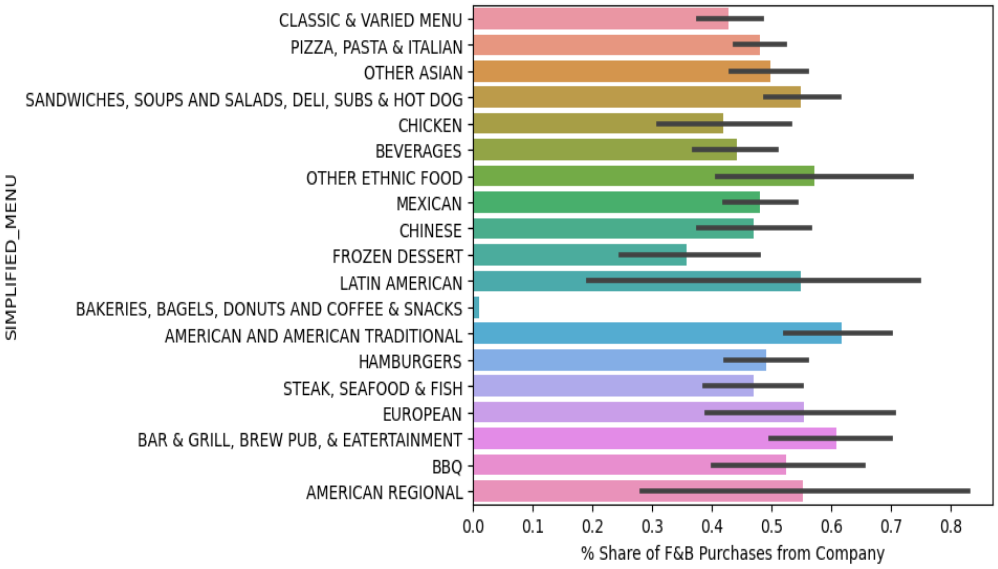
7 - Which Menu Style restaurant have more no of annual sales?

- As we can see from below chart: 1st Hamburgers
- After that 2nd Steak, Seafood & Fish category restaurants have most Annual sales



8 - Which Menu Style Restaurant has most % of F&B Share?

- 1st - (American and American Traditional),
- 2nd - (BAR & GRILL, BREW PUB, & EATERTAINMENT),
- 3rd (Other ethnic foods)
  having most %share of F&B purchases



9 - Total % share of F&B Purchases from Company:

By Menu-Style of the Restaurants:

- For AMERICAN AND AMERICAN TRADITIONAL = 22.857019189412007
- For BAR & GRILL, BREW PUB, & EATERTAINMENT = 12.763998314533591
- For ethnic foods = 5.144706661978111

CONCLUSION FROM CUSTOMER RESTAURANTS:

1 - AMERICAN AND AMERICAN TRADITIONAL

2 - BAR & GRILL, BREW PUB, & EATERTAINMENT

3 - OTHER ETHNIC FOOD

have the most % share of F&B Purchases from Shaun's company

- Hence, we get the insight that these category restaurants must be approached so as to get Maximum percent of share in their purchases.

Sheet 2 - Dataset after removing all Null-Values and deleting unnecessary columns looked like this:

```
del df2["AVERAGE_DAILY_CENSUS"]
del df2["PRIMARY_GPO"]
del df2["SECONDARY_GPO"]
del df2["TOTAL_PATIENT_DAYS"]
del df2["MSA"]
del df2["ZIP_PLUS4"]
del df2["CAPACITY_BEDS_RANGE"]
del df2["NUMBER_OF_SCHOOLS_RANGE"]
del df2["OPERATION_TYPE"]
del df2["OWNERSHIP_TYPE"]
```

```
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   NAME                            1000 non-null   object
 1   ADDRESS                         1000 non-null   object
 2   CITY                            1000 non-null   object
 3   STATE                           1000 non-null   object
 4   ZIP                             1000 non-null   int64
 5   COUNTY_CODE                     1000 non-null   int64
 6   COUNTY_NAME                     1000 non-null   object
 7   MARKET_SEGMENT                  1000 non-null   object
 8   EMPLOYEES_NON_COMMERCIAL        1000 non-null   object
 9   ENROLLMENT_RANGE                1000 non-null   object
 10  NUMBER_OF_MEALS_PER_DAY_RANGE   1000 non-null   object
```

Exploratory Data Analysis using Python – Pandas Library:

1. The dataset contains non-commercial data of MA (Massachusetts) State only.
2. The employee count column has Uncoded values too.
3. As most of the rows contain UNCODED values, hence no replacement can be done hence it is better to delete it and avoid analysis using this column.
4. ENROLLMENT_RANGE Column has 672 - Not Applicable values, hence deleting the column.
5. NUMBER_OF_MEALS_PER_DAY_RANGE column can help us identify the requirement of F&B Purchases from the company
6. NUMBER_OF_MEALS_PER_DAY_RANGE Column value are replaced using self-label encoding for better analysis:
   ('51 TO 100' = 50)
   ('101 TO 250' = 100)
   ('251 TO 500' = 250)
   ('501 TO 1000' = 500)
   ('1001 TO 2000' = 1000)
   ('>2000' = 2000)

Final Dataset for Sheet 2:

```
Data columns (total 9 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   NAME                            1000 non-null   object
 1   ADDRESS                         1000 non-null   object
 2   CITY                            1000 non-null   object
 3   STATE                           1000 non-null   object
 4   ZIP                             1000 non-null   int64
 5   COUNTY_CODE                     1000 non-null   int64
 6   COUNTY_NAME                     1000 non-null   object
 7   MARKET_SEGMENT                  1000 non-null   object
 8   NUMBER_OF_MEALS_PER_DAY_RANGE   1000 non-null   int64
```

Questions raised from sheet 2:

1 Which city consumes the maximum number of meals per day?

1. Top City with more than 2000 , 1000, 500, 250, 100, 50 meals per day
   1st - WORCESTER
   2nd - BOSTON

2 - Which Market Segment was among the top consumer of (Meals per day) in WORCESTE City?

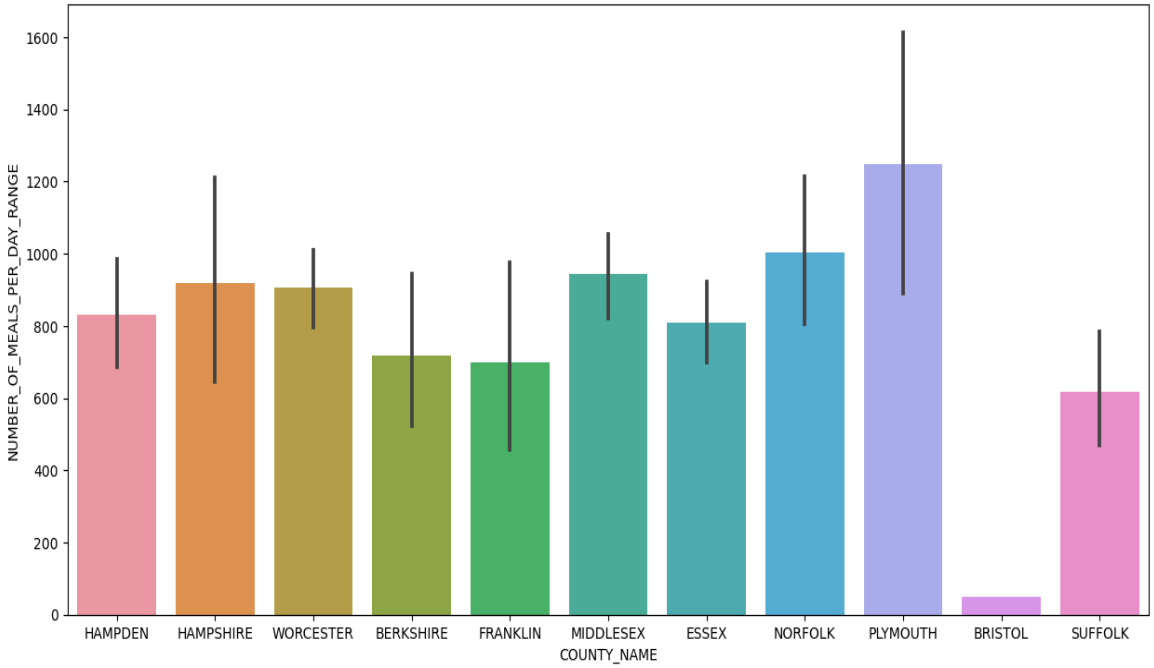
```
2. NURSING HOME                                                    15
3. OTHER HEALTHCARE ESTABLISHMENTS                                 12
4. df2.NUMBER_OF_MEALS_PER_DAY_RANGE[df2.MARKET_SEGMENT == "NURSING HOM
   E"].value_counts()
   2000    129
   1000     61
   500      31
```

5. This confirms that Nursing Homes consumes maximum amount of Meals per day.

3 - What is the total no of Meals per day in Worcester City consumed by Nursing homes?

6. Storing all Worcester City data in a Variable named (wr)
7. Storing all Nursing home in Worcester City in a variable named(wr_nh)
8. Showing the total no of Meals per day in Worcestor City consumed by Nursing homes are greater than 24500
9. Therefore, we have the adjacent opportunities from the non-commercial establishments to additionally sell into (Nursing Home) Market Segments in the city of WORCESTER and rest other cities too.

4 - Which County consumes the most no of meals per day?



- From the above chart, Plymouth County consumes most no of meals greater than 22450 per day.

CONCLUSION TO NON-COMMERCIAL SEGMENTS:

- We have the adjacent opportunities from the non-commercial establishments to additionally sell into (Nursing Home) Market Segments in the city of WORCESTER

- We also have opportunity to sell into the Plymouth County due to high consumption of meals.

Sheet 3 - Data Cleaning and EDA

Exploratory Data Analysis using Python – Pandas Library:

Deleting unnecessary columns from the database:

```
del df3["ZIP_PLUS4"]
del df3["MSA"]
del df3["years_in_business"]
```

```
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   NAME             1001 non-null   object
 1   CITY             1001 non-null   object
 2   STATE            1001 non-null   object
 3   ZIP              1001 non-null   int64
 4   COUNTY_CODE      974 non-null    float64
 5   COUNTY_NAME      974 non-null    object
 6   market_segment2  1001 non-null   object
 7   units            1001 non-null   object
 8   EMPLOYEES        1001 non-null   object
 9    annual_sales     1001 non-null   object
```

1. Replacing the UNITS values: with the minimum value from the range for analysis.

   ("INDEPENDENT (1-9 UNITS)", 1)
   ("51-100 UNITS", 51)
   ("101-250 UNITS", 101)
   ("251-500 UNITS", 251)
   ("501+ UNITS", 501)

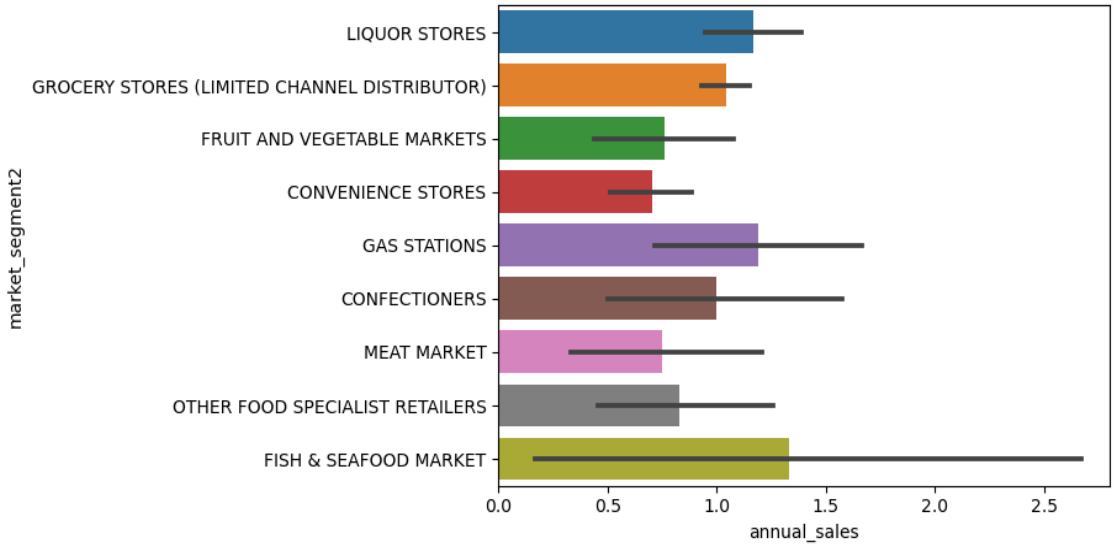
2. As the annual sales have “uncoded” value, hence deleting the rows is the best idea for analysis

3. Replacing annual sales values by self encoding.

   ('<=$500,000', 0)
   ('$500,001 - $1,000,000', 1)
   '$1,000,001 - $2,500,000', 2)
   ('$2,500,001 - $5,000,000', 3)
   ('>$5,000,000', 4)

Questions raised from sheet 2:

1 - Which market segment has the most number of annual sales?

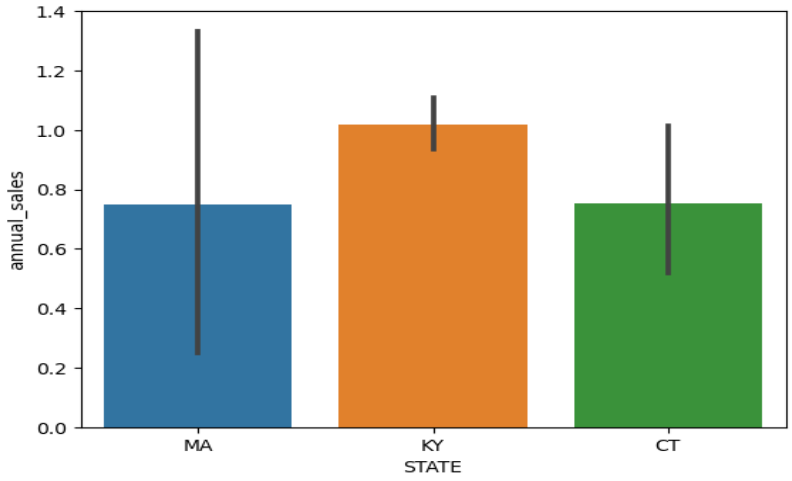


We can see from the chart that:

- 1 - Fish&Seafood Market has the maximum annual sales
- 2 - Gas Stations are on the 2nd for maximum sales

2 - Which states has most annual sales?

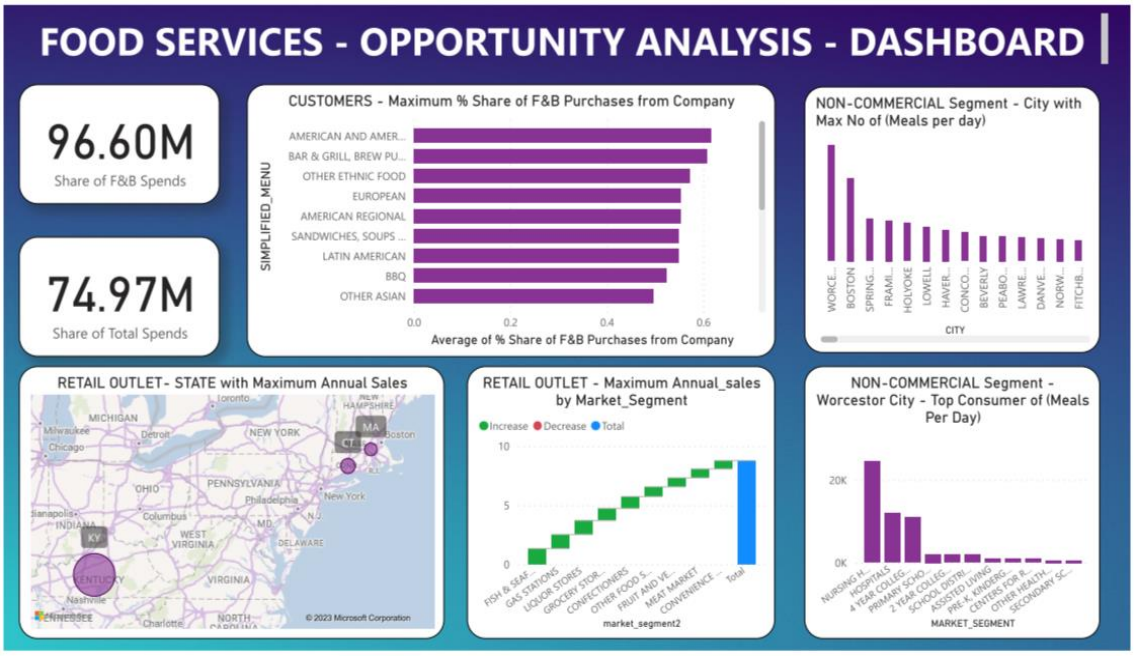
- We can see from the chart that Kentucky state, has most number of annual sales.



CONCLUSION TO RETAIL OUTLETS - SHEET 3 EDA :

- Biggest opportunities in Retail Outlets are in Kentucky State.
- Fish&Seafood Market Segment as well as different Gas stations must be our hit.

CONCLUSIONS BY EDA SHOWN IN DASHBOARD:



- Opportunities in various segments can easily be studied using the

  (OPPORTUNITY ANALYSIS DASHBOARD)