Database, Data Warehouse Technology for U.S. Chain-Restaurant Menu Item Nutrition Data Analysis

Foroozan Akhavan, Minh-Tam Pham, Zhe Li, Chloe Ngo, and Taoli Zhen

**Abstract**—Using a data set of nutrition facts for restaurants, this project aims to help people in choosing the best meals to minimize health risks when dining out. We worked on a dataset that has nutrition facts of each item in chain restaurants, notorious for being unhealthy, from 2008 to 2018. The questions we answered are which restaurants to avoid, what to avoid, and if there were healthy changes in the menus over the years. To answer these questions,we were required to set up and use a data warehouse cloud server to enable the analysts to analyze the data collected simultaneously. We chose a star schema as our data model and loaded the database from MySQL to the MariaDB cloud server. In this project, we used a Hybrid Database for both transactional operation and analytics operation, which simulates a real-time OLTP data input process and the auto-replication engine replicates the new data in the transactional database to the analytical database. This setup enables transactional data to feed to the data warehouse in real-time and enables OLAP possible for real-time data. The results will help people living in the USA to have healthy options. We can continue to update this database in order to find updated trends in the future.

**Index Terms**—Analysis, data, database design, database processing, data warehouse and repository, query design and implementation languages

————————————————

* *F.A. Author is with the San Jose State University, San Jose, CA 95192. E-mail: foroozan.akhavan@sjsu.edu.*
* *M.P. Author is with the San Jose State University, San Jose, CA 95192. E-mail: minh-tam.pham@ sjsu.edu.*
* *Z.L Author is with the San Jose State University, San Jose, CA 95192. E-mail: zhe.li01@ sjsu.edu.*
* *C.N. Author is with the San Jose State University, San Jose, CA 95192. E-mail: chau.ngo@ sjsu.edu.*
* *T.Z. Author is with the San Jose State University, San Jose, CA 95192. E-mail: taoli.zhen@ sjsu.edu.*

—————————— ◆ ——————————

# 1 Introduction

Nowadays dining out is an integral part of city life.

More than ⅓ of Americans dine out at chain restaurants. Studies [1] show that the fast-food chain restaurants' contribution to obesity is significant. For people who don’t have time to prepare meals by themselves, ordering meals that contain enough macronutrients while ensuring calories and unhealthy fat are under the daily consumption limit becomes the key to address the obesity issue. Although we know fast food restaurants are convenient, is it currently possible for Americans to eat nutritiously at these restaurants as well?

As more people start to care about the nutrition in the food they consume, a number of major chain restaurants responded by including some healthy options on their menu. Additionally, some restaurants voluntarily labeled their menus with calorie information and provided a fair amount of nutrition information on their website since 2008. In 2010, the U.S. government passed the Affordable Care Act, which includes a law requiring chain restaurants to display calories information on menus. With this requirement, there is a sufficient amount of menu nutrition data available from menustat.org [2] for us to explore this topic. The dataset enables us to investigate in what ways chain restaurants had altered their menu items in order to provide healthier food over the years.

In this paper, we will go through and explore the technologies that enable data analysis on our topic, as well as data cleansing procedures and operations that are needed to store and manage the menu nutrition data. These operations and database technologies enable us to perform efficient analytics on the menu dataset while keeping the cost at a minimum. We will employ a cloud database management platform to store and organize the data we gathered, which data can be replicated into an analytics platform to improve efficiency when performing On-Line Analytical Processing.

# 2 Methods and Implementation

## 2.1 Data Cleansing and Transform

The dataset files from menustat.org were in CSV format and the menu nutrition data was stored in separate files, by year. A star or galaxy schema would be suitable for the dataset. Detailed code for extracting the data, normalizing the data, and transforming the data is in the data\_transform.ipynb notebook file, located in the Appendix Github link.

Diagram

Description automatically generated

Fig. 1. Galaxy Schema for the Menu Nutrition Data. Serve as a data warehouse schema.

The 2018 (latest) dataset was used as the base case for all other years. All attributes were in one single large table, which was in the 1st normal form. The first step is to transform the table to 3rd normal form for the galaxy schema. A Python package Pandas was used in the cleansing and processing stage.

For our schema, there is a restaurant dimension, a food category dimension, an item information dimension, a nutrition fact table and a combo data fact table.

For the restaurant and food category dimensions, the first step was to extract the restaurant name and food category name from the main table. For each of the dimensions, we only retained the number of unique restaurant and food category names and assigned a primary key (surrogate key) for each unique name. These dimension tables were ready to load after the above processing.

For the item information dimension, we selected the item\_id, year, item\_name, and item\_description columns from the main table and retained the unique rows only. The item\_id and year attributes comprise the composite key for this dimension.

Once we have the metadata transformed into dimension tables, we can transform the main table into the nutrition fact table. First, we joined the restaurant dimension and food category dimension with the main table to obtain the restaurant id and food category id for each item. With the item\_id, year, restaurant\_id, and category\_id as the composite key in the fact table. Restaurant\_id and category\_id are also the foreign keys to the corresponding dimensions. Each transaction records the nutrition record for a menu item for the specific year.

Finally, for the combo meal fact table, we have combo\_id, year, item\_id, and category\_id as the composite primary key. The attribute, ‘builds’, records whether the item in a combo is the main item or an accompanying item. Each combo can have multiple main items and accompanying items, the logic for the combo is that people can only choose one main item and one accompanying item for each food category.

Text

Description automatically generated

Fig. 2. Example sql query to set up the replication filter function.

With the base case data extracted and transformed, we created a transform\_data(dat, year) function reusing the code for the 2018 year to transform the rest of the years’ data. Note that there was a small amount of NA values (under 200 rows) in the restaurant\_id column in the nutrition fact table in 2008, 2010, and 2012; which means that these restaurants were no longer in the menu data record in later years. Thus, these NA values can be ignored as we can’t compare the nutrition data with later years.

## 2.2 Loading Data to MySQL

Initially, we loaded the data to the local MySQL server using sqlalchemy. To enable the ability for team members to connect to the database remotely and optimize the data warehouse for analytics, we decided to employ MariaDB and SkySQL platforms to run our data warehouse. To migrate the database to MariaDB Cloud, we exported the menu nutrition database from MySQL to a dump file.

## MariaDB, SkySQL, and HTAP Platforms

According to [4], MariaDB is a fork of MySQL and the process of replacing MySQL with MariaDB can be as easy as importing a dump file to MariaDB from MySQL. When compared to MySQL, MariaDB has the capability to support a variety of storage engines. And in many cases, performance on MariaDB is better than MySQL.

SkySQL is the cloud service for MariaDB, and there are four platforms available in SkySQL. They include the following: the Transactions Platform, which is optimized for fast transaction processing (OLTP), the Analytics Platform, which is optimized for running ad hoc queries on data warehouses (OLAP), and the Hybrid Transactional-Analytical Processing Platform which supports both OLTP and OLAP using different storage engines for OLTP and OLAP [3]. Here in this project, we used the Hybrid Transactional-Analytical Processing Platform to simulate the real-life work environment. We can have a seamless data movement from the transaction database to the analytics data warehouse.

As documented in [5], transaction databases can handle OLTP queries “using row-based transactional storage engines, such as InnoDB or MyRocks,” and analytics data warehouse can handle OLAP queries “using the MariaDB ColumnStore storage engine.” The MariaDB server uses MariaDB MaxScale to handle client connections. Another function of MaxScale is to differentiate between SQL queries for RowStore and SQL queries for ColumnStore. In MariaDB, regular SQL queries can be used to query from the ColumnStore databases. MaxScale can route queries to the right server [5].

## HTAP Data Warehouse Setup and Loading Data

In this project, we created a RowStore database for transactional processing and a ColumnStore database to serve as the data warehouse to store historical data for analytics.

To enable auto replication, the MariaDB server replicates “writes from InnoDB tables to the MariaDB ColumnStore tables” using MariaDB Replication. According to the MariaDB Enterprise Documentation [5], “MariaDB Replication with MariaDB MaxScale configured as a Binlog Server, MariaDB Enterprise Server can host InnoDB and ColumnStore on the same Server.”

In our practice, we specified the Row Store tables with certain prefixes to automatically replicate to Column Store, which fed data directly from the transaction RowStore to the data warehouse ColumnStore. As demonstrated in the mariadb\_cloud\_load\_data.ipynb notebook file in the Github link under Appendix, we set up the replication filter using the set\_htap\_replication() UDF.

The next step was to create two databases with distinguishable names, one named as for menu\_rowstore, another one named as menu\_cstore. Create tables for our data model in the menu\_rowstore database and check if tables in the RowStore were auto replicated to the ColumnStore. We verified that the auto replication was working properly. Then we deleted the tables in the ColumnStore and re-created the same tables with the engine specified as ColumnStore in table creation queries. Finally, we imported the dump file exported from MySQL into the RowStore database and verified that the data was replicated to ColumnStore in nearly real time.

## DB Connectivity and Analytic Tools

The MariaDB Enterprise Cloud Server provided an easy solution for connecting to the data warehouse. With the username, password, host domain, and the certificate authority chain, any whitelisted IP addresses would be able to connect to the server and query the data warehouse. We can either use a graphical application such as MySQL Workbench or Python to run DML, DDL, DCL, and TCL to manage and control the data warehouse, define or alter database schemas, and manipulate the data in the warehouse.

We used Python as the main scripting language to perform data analytics. Because MariaDB is a fork of MySQL, we used the MySQL Connector for Python to connect with the data warehouse on the server.

## Data Aggregation

Our data warehouse is OLAP ready. We used the ROLLUP function to get aggregate data for both two-dimensional and three-dimensional cubes. Although MariaDB doesn’t support the CUBE function, we used the Pivot table function in the Pandas to make the ROLLUP results into a CUBE-like structure. With the GROUP BY clause, we sliced and diced to select the data we need to investigate and answer questions in our research topics.

## Filtering to Find the Best/Worst Restaurants in Certain Aspects

In our analysis, we applied filters based on daily nutrition requirements and daily intake limits to different nutrition facts features in order to build a recommendation list or restaurant to avoid list.

## Visualization

Python package Seaborn was used in visualizing the data. Even Though we used the ROLLUP function to perform data aggregation on the data warehouse server directly, there were still too many categories in the Restaurant and Food Category attribute. There were 96 restaurants and 12 categories in 2018. The CUBE structure was too hard to read and keep track of the information. Therefore, we used the seaborn package to visualize the query results such that it would be easier for us to interpret and compare.

## Clustering

We will find the worst restaurant to avoid from our analysis. However, It’s hard to set up solid criteria for building a bad-restaurant list since all restaurants have multidimensional nutrition factors. Clustering - an unsupervised machine learning technique is introduced here. It helps us identify groups of restaurants based on their nutrient pattern and segment these into a certain number of groups according to the level of inertia. After that, we are able to find out the avoid list from the group that has a bad nutritional pattern as in the worst restaurant we found before.

## Null Value Handling

When using aggregate function AVG() in queries, having Null values handled properly is crucial in making sure that the aggregation results are not biased. MariaDB server processes AVG() by summing up every transaction and dividing by the total number of transaction rows. For the nutrition attributes that contain Null values, we de-selected rows with Null Values. Because of the fact that some restaurants don't have records for every nutrition attribute, we ran queries on one nutrition attribute at a time to ensure that we didn’t include any Null values or dropped any non-Null values by accident.

# 3 Results

## 3.1 Trends

We have menu nutrition data from 2008 to 2018. Among the 10 years, there were only 55 restaurants recorded in 2008 while there were 96 restaurants recorded in the menu data in 2018. We decided to ignore years prior to 2013 and only focus on the most recent five years.

Fig. 3 shows the average calories of all menu items for 7-Eleven from 2013 to 2018. The result shows that for 7-Eleven, their average calories for all menu items were decreasing over the five-year period. The average calories decreased 55% in 2018 compared to the previous year.

Table

Description automatically generated

Fig. 3. Avg. calories for 7 Eleven from 2013 to 2018.

## 3.2 3 Dimension Cube Rollup for 7-Eleven

Fig. 4 is the table format of the 3-d cube structure representation of the average calories group by each restaurant and food category. There are much more details in this 3-dimensional cube. In 2017, 7-Eleven altered its menu and only remained beverages, salads and sandwiches. In 2018, they stopped serving salads. Table

Description automatically generated

Fig. 4. Cube structure table for 3-dimensional cube: average calories for each restaurant by food categories

Applebee’s have a lower average calorie across almost all food categories over the five-year period except for salads. The two examples offer a better overview of the aggregate data, and we will elaborate on the results more in the discussion section.

## 3.3 Aggregation Results - Determining Trends

Fig. 5 displays a line plot for average calories from protein in all food categories. We can see that there is an upward trend in the Burgers, the Sandwiches, the Appetizers & Sides, and the Soup category. Which is a good indicator of good ingredients. The average of calories from carbohydrates doesn’t seem to have a clear trend from line plots and so does the average calories and the average amount of dietary fibers.

Chart, line chart

Description automatically generated

Fig. 5. Line plot for the average protein for all categories from 2013 to 2018.

If we look at Fig 6, which is a grid of lm plots fitted with a simple linear regression line for the number of average calories across all categories, none of the categories have a down trend over the years. Appetizers & Sides, Entrees, Desserts, Sandwiches, Burgers, Pizza, and Fried Potatoes categories all have a mild to moderate increase over the years.

Fig. 7 is a grid of lm plots fitted with a simple linear regression model for the amount of average cholesterol across all categories. From the grid of visualization, it is obvious that there is a clear uptrend in Appetizers & Sides, Entrees, Desserts, and Pizza categories. Others seem to be consistent over the 5-year timeframe.

Based on the result from Fig. 6 and Fig. 7, there seems to be a correlation between the amount of cholesterol and calories. We will provide a correlation heat map in the following section and explore the correlation relationship in the discussion section.

Chart, calendar

Description automatically generated

Fig. 6. The grid of lm plots fitted with simple linear regression line for the amount of average calories across all categories

Chart

Description automatically generated

Fig. 7. The grid of lm plots fitted with simple linear regression line for the amount of average cholesterol across all categories

## 3.4 Worst Restaurant to Avoid

We want to view the restaurants that have the highest trans fat/fat ratio. And by grouping items by restaurant and filtering, we found that Long John Silver's and Arby's are the highest in this ratio. Of all restaurants, these two restaurants historically have more than 15% of trans fat in the total fat. Long John Silver's restaurant events have 27.45% of trans fat in the total fat.

For the sugar to total calories ratio, we find that two of the five worst restaurants are beverage stores: Jamba Juice and Starbucks. Long John Silver's, KFC, and Krispy Kreme are also in the top five in the sugar: calories ratio.

Restaurants that offer high amounts of fibers are Chipotle, Qdoba, Round Table Pizza, Jamba Juice, and Carrabba’s Italian Grill. While Long John Silver’s is one of the top five restaurants, it offers the least fibers in its items.

## 3.6 Clustering - A List of Worst Restaurants to Avoid

After handling NA values, data cleansing, and normalization, we ran clustering and found out k = 4 is the reasonable amount for groups since the decrease in distortion at 4 starts to level off.

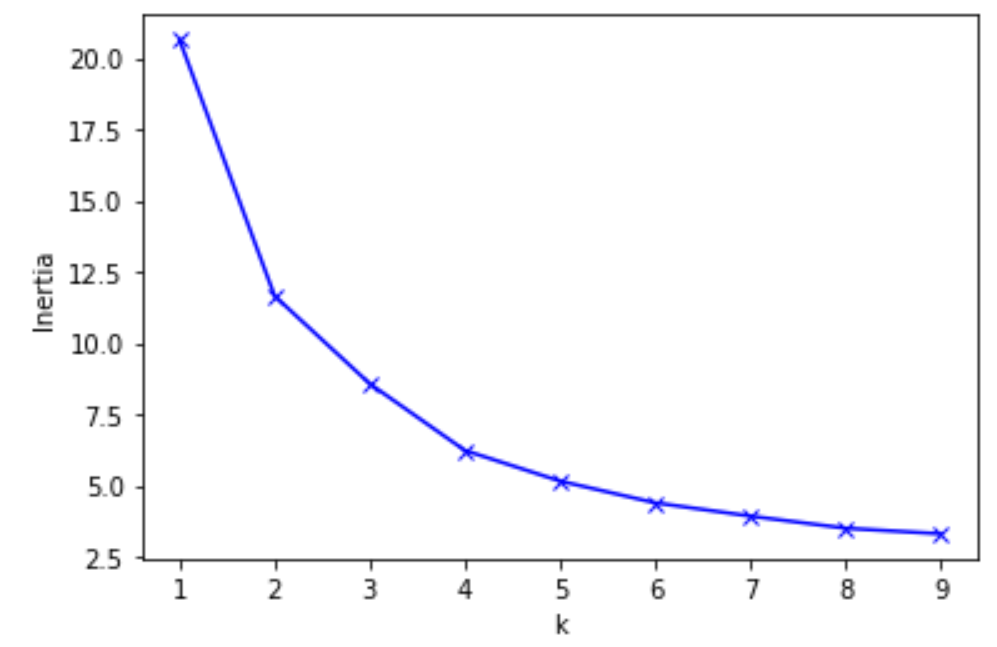


Fig. 13.1. Kmean Model: K vs Inertia.

From the previous analysis, we find out that Long John Silver's is the most unhealthy restaurant. We find out Long John Silver's and other fast-food chains like McDonald's, Starbucks are all in group1. Those are the restaurants with bad nutrient patterns that we would warn people not to eat for a healthy diet.

## 

Fig. 13.2. Group 1 from clustering: Restaurants to avoid.

From the 3-d cube, we also find that Applebee’s indeed have reduced their unhealthy nutrients over the years. With the same clustering model, we have located the group for Applebee’s. Fig. 13.3 displays a list of restaurants with similar nutrient patterns as in Applebee’s.

## 

Fig. 13.3. Group 0 from clustering: Restaurants to dine in.

## 3.5 Correlation

Fig. 8 is the heatmap for the correlation of average nutrition attributes grouped by the restaurant, category, and year. The result from the data warehouse is already included year, restaurant, and food category variances. From the correlation heatmap, calories and total fat have the highest correlation which is 0.93. Protein, carbohydrates, sodium, cholesterol, and trans fat all have a relatively strong correlation with calories between the range of 0.61 to 0.75.

With beverage items included, there is no correlation between sugar and other nutrients except for carbohydrates. One of the reasons for the low correlation between sugar and calories is that the weight of the beverage category is small while beverages have the highest amount of sugar in any given category. We will discuss the relationship of correlation and the implication on other results later in the discussion.

## Chart, treemap chart Description automatically generated

## Fig. 8. Correlation heatmap for nutrient features grouped by year, restaurant, and food category.

# 4 Discussion

## 4.1 Trends in Nutrition Composition

To answer the question of whether or not restaurants made their menu items healthier, we need to find trends in nutrition composition. We were not able to determine the trend of average calories, the average amount of carbohydrates, and the average amount of dietary fibers through the line plot for the average nutritional composition for all categories from 2013 to 2018. However, Fig. 5 showed an upward trend for the average protein content in all categories. This is a sign that chain restaurants made their items more nutritious by increasing the amount of protein (meat) in their food.

Having more protein can better satisfy people’s daily protein requirements. To consider a trend of healthier food in the industry, we need to check that calories and fat/cholesterol were maintained at the same level over the years. Based on Fig.6 and Fig.7, with a simple linear regression line drawn in each time vs average nutrition scatter plot by food category, the trend becomes distinguishable. Six main categories have an identifiable upward trend in the average category while the regression line for the remaining calories was flat. The same pattern also appeared in Fig.7 which is showing trends for the average cholesterol in all categories.

The correlation heatmap showed a strong positive correlation between protein, calories, total fat, cholesterol, and sodium. This further suggested the theory of increasing protein content doesn’t make the food healthier. One explanation for this result might be that the protein source is not clean enough (the proportion of fat in meat is high). To combine the findings in the Methods section 3.3 and 3.5, we can say that there isn’t a clear trend of chain restaurants making their food healthier. While they might be increasing the content of proteins and other nutrients in their food, the proportion of unhealthy nutrients such as cholesterol and trans fat also increased. Resulting in the overall calories increase over the years.

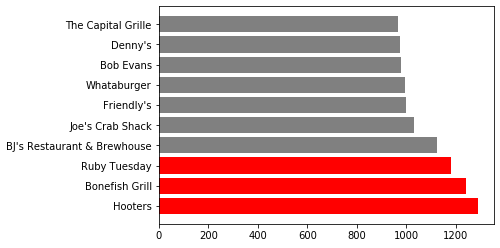
## 4.2 Unhealthy Categories

The reason behind this question is what people should avoid. Sometimes people order foods that have lower rates of calories and they think by ordering low-calorie foods they stay healthy and not gaining weight, but we can show from our data that this belief is not right all the time. First, we started to explore the data to find out what category of foods has higher average calories. In the table below we can see “Burgers” have higher average calories than any other categories.



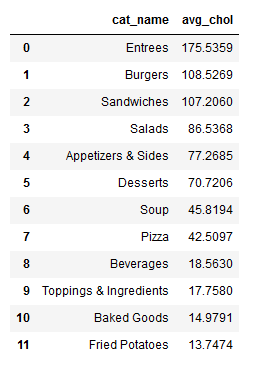
## Fig. 9.1. Result of the query which shows average calorie per food category.

The question that comes to mind immediately is what chain restaurants have high-calorie burgers so people can avoid ordering burgers from these restaurants. The chart below shows 10 restaurants that have average high-calorie burgers.



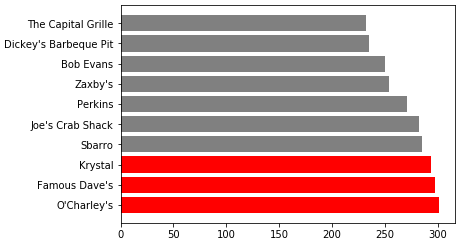
## Fig. 9.2. Result of the query which shows 10 restaurants that have higher-calorie burgers.

The next unhealthy factor is high cholesterol in meals. The table below shows the average cholesterol in each food category.



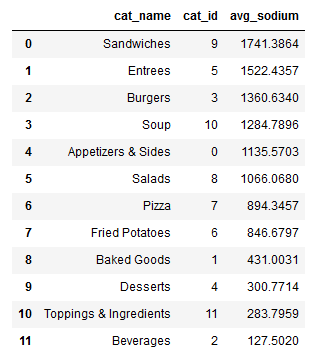
## Fig. 10.1. Result of the query which shows entrees have the highest average cholesterol.

In the next step, we go through the restaurants that have higher cholesterol in their entrees. The diagram below is a result of the query which is designed to answer this question.



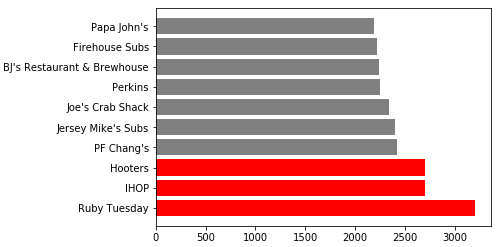
## Fig. 10.2. Result of the query which shows 10 top restaurants that have higher cholesterol in their entrees.

For the next unhealthy option, we explored the data through the sodium column in our database. The average daily sodium intake according to the FDA is 2300 mg. But some restaurants will give people high sodium rates more than the suggested amount. The table below shows the average sodium in milligrams each food category has.



## Fig. 11.1. Result of the query which shows sandwiches have the highest average sodium.

Let’s see what restaurants have the highest average sodium in their sandwiches in the diagram below.



## Fig. 11.2. Result of the query which shows 10 top restaurants that have higher sodium in their sandwiches.

## 4.3 Unhealthy and Healthy Restaurants (Filtering and Clustering)

Not all fats are created equal, as some foods such as olive oil, nuts, and fish contain good fats. However, trans fats are known to be artery-clogging, and so are bad for people who are suffering from heart diseases. We are interested to know which restaurants should be avoided by people who have/are prone to heart problems. We want to view the restaurants that have the highest trans\_fat/fat ratio. And we find that Long John Silver's and Arby's are the worst in the trans\_fat/fat ratio. Similarly, we would like to look at restaurants that should be avoided by those who have type 2 diabetes. We do this by examining restaurants that have the highest sugar/calories ratio. Interestingly, we find that two of the five worst restaurants are beverage stores: Jamba Juice and Starbucks. And Long John Silver's is still one of the five worst restaurants in high sugar levels. Now we are interested in restaurants that, on average, offer food with high amounts of fiber. Fiber is known to have the opposite effects of trans fats and sugar. It lowers blood sugars and lowers cholesterol levels. Long John Silver's is still one of the worst five.

Overall, it seems like Long John Silver's is the worst restaurant to eat at, especially for a person struggling with type 2 diabetes as well as heart problems. Their food doesn't seem to be nutritious (low fiber). In this analysis, we note that Long John Silver's is consistently the least nutritious for you.

With Long John Silver's satisfied the criteria for the worst restaurant, we classified restaurants with similar nutritional patterns as shown in Fig. 13.2. They are the restaurants that people should avoid. Although we determined that there isn’t a trend of chain restaurants making their food healthier, there are still restaurants that have signs of making their food healthier such as Applebee’s. In Fig. 13.3, we also provided a list of restaurants that have a similar pattern with Applebee’s and they are the restaurants we recommend.

# 5 Lessons Learned

## 5.1 The Effect of Null Values and Aggregation

Null values can have a huge impact on the aggregate results we obtained. In Fig.3, the average calories decreased 55% in 2018 for 7-Eleven. In this case, the aggregate data is misleading in the interpretation of the result. When we drilled down one level in the 3-d cube structure representation in Fig. 4, we discovered that the reason for the 55% average calories decrease in 2018 for 7-Eleven is because 6 out of 8 food categories were missing in the data. We couldn’t conclude that the trend for calories was decreasing because either 7-Eleven stopped selling these categories or the data collector failed to collect data for these categories.

The lesson we learned from this is that we have to be very careful in dealing with NA values. NA values can be dropped while the data is big enough. For aggregate data generated from only one dimension, e.g. Restaurants, there needs to be an exploration in the other dimensions to make sure the data in the later years are consistent. Also, examining 3-dimensional data cubes can be helpful in finding the root cause for missing values. To decide what to do with missing values, we have to drill down to see it in a finer grind level in order to determine what happened.

# 6 Conclusion

The population who consumes food regularly from chain-restaurants is still growing. Our analysis suggests that, in aggregate, chain-restaurants are not making their menu items healthier. Despite the fact that restaurants were trying to make their meals and items more nutritious, the accompanying increment in cholesterol, calories, total fat, and sodium didn’t make their food any healthier. Regardless, it is still possible to avoid the worst food categories to choose in specific restaurants and we have provided a list of relatively healthier restaurants. More data is needed in order to analyze the latest trends in chain-restaurants’ menu items nutrition. Furthermore, while the aggregate data didn’t show any healthier trends, there can be more in depth research to investigate the nutrition facts in newly introduced items in later years.

**References**

1. A. Tran, A. Moran, and S. N. Bleich, “Calorie changes among food items sold in U.S. convenience stores and pizza restaurant chains from 2013 to 2017,” Preventive Medicine Reports, vol. 15, p. 100932, 2019.
2. MenuStat. [Online]. Available: http://www.menustat.org/. [Accessed: 08-May-2021].
3. “Open Source Database (RDBMS) for the Enterprise,” MariaDB. [Online]. Available: https://mariadb.com/docs/. [Accessed: 08-May-2021].
4. T. Jankov, “MariaDB vs MySQL, a Database Technologies Rundown,” Kinsta, 30-Dec-2020. [Online]. Available: https://kinsta.com/blog/mariadb-vs-mysql/. [Accessed: 08-May-2021].
5. “Hybrid Transactional-Analytical Processing (HTAP),” MariaDB. [Online]. Available: https://mariadb.com/docs/multi-node/htap/. [Accessed: 08-May-2021].

# Appendices:

Presentation Skills, includes time management: Will demonstrate in the project presentation

Significance to the real world: In section 1 “Introduction” of this report, we emphasized the significance of dining out for people who are living in the cities. According to [this](https://www.insider.com/what-dining-out-does-to-your-body-2018-6#eating-out-for-lots-of-meals-increases-your-risk-of-heart-disease-or-stroke-3) report from an insider news webpage, dining out may increase your amount of food intake, raise sodium and cholesterol levels, increase the risk of heart disease, and in some cases increases weight gaining. But, we found ways to eat more healthy and avoid unhealthy foods to decrease the risks of dining out for our bodies.

Code Walkthrough: Will demonstrate in the project presentation

Report: This report is the proof.

Version Control: Our Github page is a great proof of this, it can be found on [this link](https://github.com/Feritaba/DATA225Project).

Discussion / Q&A: We have a discussion section in this report, and we will reserve time for Q&A at the end of the presentation.

Lessons learned: Can be found in section 5 of this report.

Innovation: Implemented the significant paper in our project. Explanation of the implementation can be found in 2.3 and 2.4 of this report and the paper can be found in [thi%8s link](https://doi.org/10.4236/jcc.2020.810008). The significant paper helped us to use Hybrid Transactional-Analytical Processing. The coding details about the transformation to this method can be found in [this link](https://github.com/TaoliZhen/DATA225Project/blob/main/data_transform.ipynb) in the project Github.

Teamwork: In this project every member had specific jobs to do, starting from the project we had meetings to distribute the work and make decisions on how to move forward with the project. Every member of the group worked on the code part from extracting the data, cleansing, and loading to the database to analyze the data. Afterward, each member worked on a specific part from preparing the documents and writings to making the slides and videos.

Technical difficulty: We used a new database and implemented a hybrid database/data warehouse design in our project.

Practiced pair programming?: The link to our github page can be found [here](https://github.com/TaoliZhen/DATA225Project).

Practiced agile / scrum (1-week sprints)?: Group chat history and meeting links and records are in [this link](https://github.com/TaoliZhen/DATA225Project/blob/main/Group_Chat_Records.txt).

Used Grammarly / other tools for language?: We used grammarly and google docs language tools to check for several types of errors.

Elevator pitch video:

Slides:

Demo: Will demo the connection to the cloud server and the python function used to access the query results.

Used unique tools: Jupyter Notebook, Cloud database and data warehouse, row store and column store techniques Hybrid Transactional-Analytical Processing), K-mean clustering.

Performed substantial analysis using database techniques:

Used aggregate functions, filtering and Rollup functions to analyze the data.

Used a new database or data warehouse tool not covered in the HW or class: MariaDB is an open-source database used in our project. HTAP platform, row-store for writes, and column-store for analytics.

Used appropriate data models: Can be found in section 2.1 of this report. Also, accessible from [this link](https://github.com/TaoliZhen/DATA225Project/blob/main/img/fig_1.png).

Used ETL tool: We didn’ use ETL tools, we did the ETL process in our Jupyter Notebook that should be found in our Github.

Data Cleansing: The explanation of the work can be found in section 2.1 of this report. Also, the step by step coding of cleansing the data is accessible from [this link](https://github.com/TaoliZhen/DATA225Project/blob/main/data_transform.ipynb).

Demonstrated how Analytics support business decisions: Our data warehouse platform supports real-time data monitoring and analysis. Anyone who wants to enter the chain-restaurant market can use our data warehouse and analysis to target customer pain points by providing healthy food, and showing what nutrients rivals currently provide.

Used NOSQL database: We didn’t use NOSQL in our project.

Used RDBMS: Our row store MariaDB database is evidence to this section. To prove the usage of RDBMS can be found in [this link](https://github.com/TaoliZhen/DATA225Project/blob/main/mariadb_cloud_load_data.ipynb).

Used Data Warehouse: Our star schema of data modeling shows the usage of data warehouses in our project. Evidence of such can be found in section 2 “Methods and Implementation”. Also, MariaDB roll up syntax example queries can be found in [this link](https://github.com/TaoliZhen/DATA225Project/blob/main/database_analysis_queries.ipynb) to the Jupyter Notebook of the project.

Includes DB Connectivity / API calls: Examples of connection to the MariaDB database can be found in [this Jupyter Notebook](https://github.com/TaoliZhen/DATA225Project/blob/main/database_analysis_queries.ipynb). There is a certificate needed to connect to the database under the name of “skysql\_chain.pem” in the [main repository](https://github.com/TaoliZhen/DATA225Project).