1. The client my company is representing is a restaurant chain in Melbourne, Australia. The business task assigned is to analyse the restaurant’s Food Delivery data to improve resource allocation as well as to identify which restaurant branch to improve their promotion of delivery service.
2. To achieve this business task. The key factors that I will be investigating are:
   1. The time of the day with the most demand for delivery orders.
   2. The restaurant branch with the least demand for delivery services.
   3. The restaurant branch with the most revenue generated from food deliveries.
3. The dataset is obtained from the Melbourne Restaurant Food Delivery data dataset on Kaggle https://www.kaggle.com/jianhanma/melbourne-restaurant-food-delivery-data
4. The audience for this data investigation are the business managers of the restaurant chain.

ASK:

* The topic I am exploring is the effective allocation of resources based on consumer data, as well as data-driven ideas on increasing revenue.
* The problem to be solved is the inefficient allocation of resources.
* The metrics used to measure the data will be:
  + Mean and total order count for all branches (per time category)
  + Total order count for each branch (per time category)
  + Mean and total order revenue for all branches.(per time category)
  + Mean and total order revenue for each branch.(per time category)
  + Mean and total branch revenue from deliveries for all branches.
  + Comparison of mean order revenue from customers with and without loyalty.
    - Where time category refers to Breakfast, Lunch or Dinner
* The stakeholders for this investigation are the business owners and business managers of the restaurant .
* The audience that my findings will be presented to are the business managers of the restaurant.
* The insights from my analysis will provide advice to the business client on which branch to allocate more delivery personnel, and which branch to promote their food delivery service more to.

PREPARE AND PROCESS:

As mentioned earlier, the dataset used for this investigation is the Melbourne Restaurant Food Delivery data dataset on Kaggle https://www.kaggle.com/jianhanma/melbourne-restaurant-food-delivery-data

The data files are organized in .csv files

The data is collected internally and is original. There is missing data and overall dirty data to be cleaned. The data is dated a year ago. However, for purpose of the case study, the data will be deemed as current.

The data used for this investigation does not contain information that requires addressing licensing as it is all internally collected. No information compromising customer privacy is used. The dataset is available publicly. However the analysis process and outcomes will be kept secure under my personal workspace until ready to be presented to the audience.

**The cleaning process:**

1. Using excel, I used conditional formatting and highlighted cells to identify duplicate order\_ids in the dirty, missing and outlier and discovered 0 duplicate order\_ids.
2. In the outlier data excel sheet, the data outliers are in the delivery fee attribute.
   1. I identified weekends and weekdays via conditional formatting.
   2. Delivery fee is calculated differently for each branch. The fee depends linearly on weekdays or weekends, time of day, whether the customer has loyalty and distance. Hence, I filtered the data by branches, weekend or weekday, loyalty and order\_type(BLD).
   3. For each set of filtered data, I calculated the delivery\_fee/distance\_to\_customer\_KM ratio and then the mean ratio. I then determined that ratios that strayed +-0.3 from the mean were outliers.
   4. I filtered and removed these outliers.
3. In the missing data excel sheet, I checked for missing data in each column using the same method in (1). There are missing values in the branch\_code, distance\_to\_customer\_KM and delivery\_fee columns.
   1. I used SQL to concatenate the latitudes and longitudes to form a combined\_location attribute in the missing\_data.csv, nodes.csv and branches.csv datasets.
   2. I then merged the new missing \_data and nodes datasets using R in R studio to form a mixeddf.csv dataset. I also merged the new branches and nodes datasets using R in R studio to form a branches\_mixed.csv dataset. These datasets are joined based on their combined\_location attribute
   3. As the delivery fee depends linearly on weekend or weekday (0,1), time of the day(0,1,2) and distance, using numpy in python, I solved for coefficients of the system of linear equations. Here however, I discovered that the coefficients are not constant for all values. <https://github.com/mraacus/melb_dataset_cleaning/blob/c96af718e311896a9bc4cb65359e58b329eeb6b3/missing%20values%20with%20np.linalg.ipynb>
   4. Hence, missing values of delivery fee could not be discovered. After consulting with the business managers, I omitted the rows with missing values of delivery fee in the excel.
   5. Using the Pandas and NetworkX packages in python, I created a graph from the edgelist dataframe. The shortest distance between customer and branch is calculated using Dijkstra’s algorithm. At the start, I learned to write the algorithm and subsequently solved via that method. However, I later discovered that NetworkX contained functions to perform the task similarly.
   6. While checking the results, I discovered that the shortest distance possible from the customer to any branch was not the same as the shortest distance from the customer to the branch where the delivery was ordered.
      1. ***Finding no.1: Most efficient delivery route not used***
   7. Since there was no way to find either the true source node(branch) or true distance, I consulted the business managers. We came to the conclusion of filling up the missing values using the shortest possible distance for this scenario.
   8. Using the methods mentioned, I found the missing values and inputted them into the excel sheet. <https://github.com/mraacus/melb_dataset_cleaning/blob/2e94e4096731e7bf9df24ec5673649c0316f9e5a/missing%20data%20input%20via%20Nx.ipynb>
   9. I checked the dataset in excel for any missing values left and found none.
4. In the dirty\_data excel sheet:
   1. I formatted all the dates to be in the dd/mm/yy format, as in accordance with the other data spreadsheets.
   2. I found and replaced all lower case branch codes to upper case.
5. Using SQL, I renamed the column header in the cleaned missing data from ‘customerHasLoyalty\_’ to ‘customerHasloyalty?’ to align with the other 2 datasets. I then merged all 3 datasets and exported it
6. After opening the exported file I excel, I noticed an accidental export of dates in the time column and used ‘RIGHT’ and ‘paste special as values’ to correct it.
7. I checked through the final dataset for blanks, lack of uniformness and other errors and found none. The dataset is ready for analysis.

ANALYSIS AND SHARING:

**Analysis results:**

Using SQL, I extracted some key statistics from the data:

* Total order revenue = 665114.40
* Total delivery revenue = 18837.55
* Mean order price = 491.58
* Total Breakfast orders = 443
* Total Lunch orders = 440
* Total Dinner orders = 470
* Total Breakfast revenue = 142949.90
* Total Lunch revenue = 231402.50
* Total Dinner revenue = 290762.00
* NS total order count = 468
* NS total revenue from orders = 236755.95
* NS mean revenue from orders = 505.89
* NS total revenue from deliveries = 6883.70
* BK total order count = 413
* BK total revenue from orders = 208822.55
* BK mean revenue from orders = 505.62
* BK total revenue from deliveries = 5760.15
* TP total order count = 472
* TP total revenue from orders = 219535.90
* TP mean revenue from orders = 465.12
* TP total revenue from deliveries = 6193.71
* Mean order price with loyalty = 368.29
* Mean order price without loyalty = 500.48

From these data statistics:  
Finding 2: The order time when there are the most delivery orders is dinner.

Finding 3: BK branch has significantly less orders than the other 2 branches.

Finding 4: NS branch has the most revenue generated from both orders and deliveries.

Finding 5: Customers with loyalty tend to have lower revenue generated per order.

Using excel pivot tables, I extracted the following graphics to show order count and sum of order prices over the time period of the year:

From the above 2 graphics:

Finding 6: The total order count each month has stayed relatively the same throughout the year, indicating a lack of increase in delivery order customers.

From the above 2 graphics:

Finding 7: Delivery order customers generally spent more towards the end of the year.

ACT:

The following are the findings from this data analysis:

Finding 1: Most efficient delivery route not used.

Finding 2: The order time when there are the most delivery orders is dinner.

Finding 3: BK branch has significantly less orders than the other 2 branches.

Finding 4: NS branch has the most revenue generated from both orders and deliveries.

Finding 5: Customers with loyalty tend to have lower revenue generated per order.

Finding 6: The total order count each month has stayed relatively the same throughout the year, indicating a lack of increase in delivery order customers.

Finding 7: Delivery order customers generally spent more towards the end of the year.

Therefore these are the recommendations:

1. Send delivery orders from the branch where there is the shortest distance from branch to customer, to increase manpower efficiency, as well as decreasing delivery times.
2. Deploy more delivery personnel in the evening to keep up with potential order overload.
3. Increase marketing efforts for the restaurant in regions near BK branch to attract more customers.
4. Continue to promote good product and service in NS branch region.
5. Provide incentives for customers with loyalty to purchase more orders.
6. Potential new delivery customers may not know of the availability of such a service. Hence, perhaps an increase in marketing efforts of the delivery service would attract more delivery customers.