ODDLE ASSIGNMENT

Upon receiving the assignment from Yong Xiang, I decided to set certain things up. These are my steps in attempting the assignment questions.

1: Downloaded the data (google drive from Yong Xiang's email link)
2: Loaded the data into sqlite to simulate a DB to execute SQL queries
3: Attempted assignment: Q1A

> 1) The database includes the transactional data of three merchants. Please provide the following analysis:
> a. Revenue, orders, and new signups report for Merchant 4 (menu_id 4) to see if its performance have improved from Jan 2016 to Jun 2017

- Ran SQL query: (Connection string put in to address the "Oddle" DB

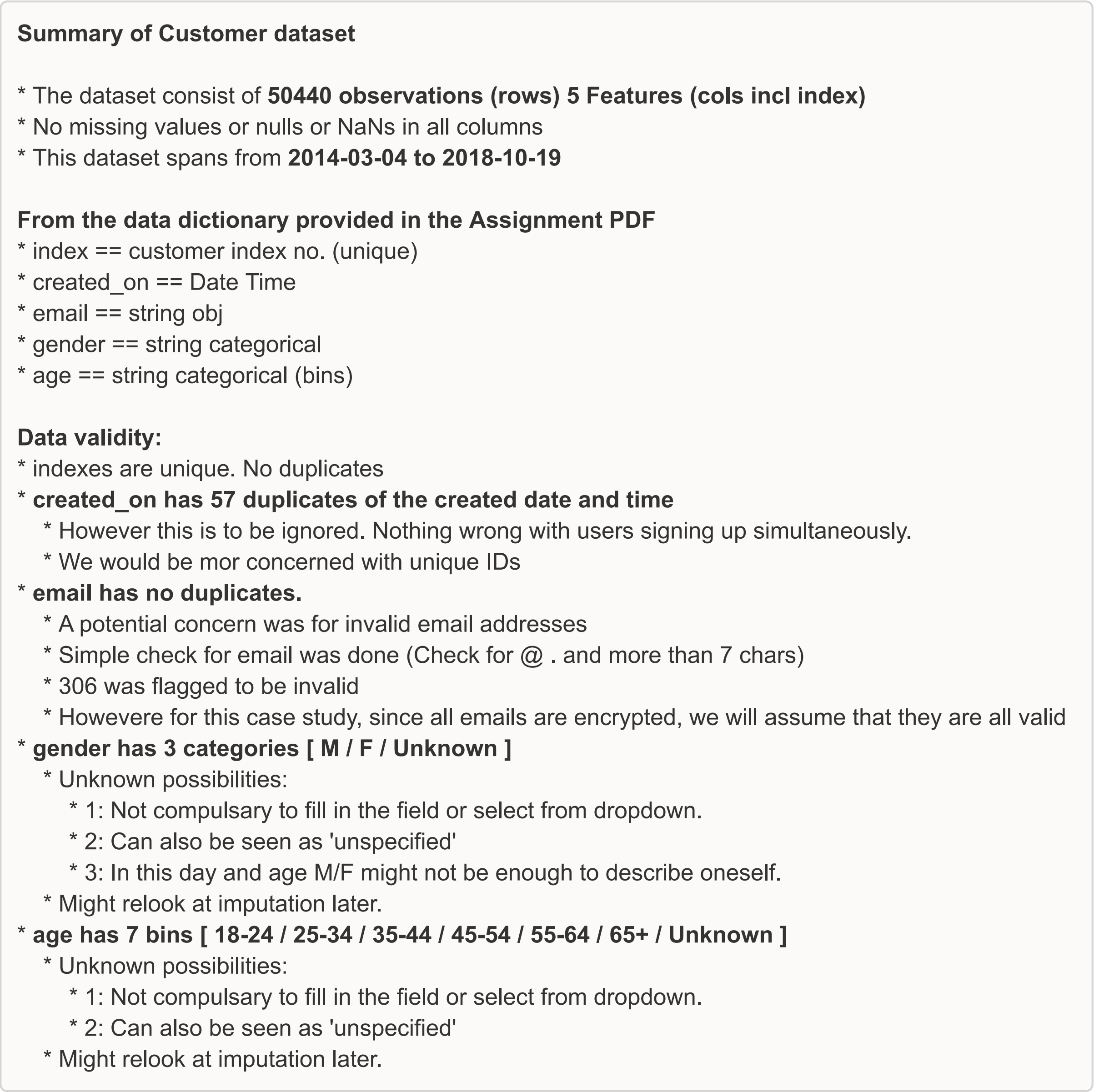
```
--Server=myServerAddress;Database=Oddle;Uid=myUsername;Pwd=myPassword;

SELECT orders_csv.'basket_size', orders_csv.'menu_id',
orders_csv.'submitted_on' as
'order_date',  orders_csv.'organisation_name_provided' as 'corporate'
,cust_csv.'index' as 'Customer ID', cust_csv.'created_on' as 'signedup_on'
FROM 'Oddle_Assignment Orders Data' AS orders_csv
LEFT JOIN 'Oddle_Assignment Customers Data' AS cust_csv
ON orders_csv.'customer_index' = cust_csv.'index'
WHERE orders_csv.'submitted_on'  BETWEEN '2015-12-31' AND '2017-07-01' --
between the dates (not including)
AND orders_csv.'menu_id' = 4 -- Select Merchant 4
ORDER BY datetime(orders_csv.'submitted_on') ASC; --want to see it sorted
by the earliest date
--ORDER BY datetime(orders_csv.'submitted_on') DESC; --want to see it
sorted by the latest date
```

- Attached in the email are the SQL queries used in the assignment.
- **Spotted that the date-range of the report (Jan 2016 to Jun 2017) was not able to be covered by the data provided.**
- Decided to inspect the table data to have an overview

4: Data Overviews + Clean up + Feature Engineering

- **Customer data:** **https://github.com/ziigyee/Assignment/blob/master/Data%20Overview%20-%20Customers.ipynb**



Summary of Customer dataset

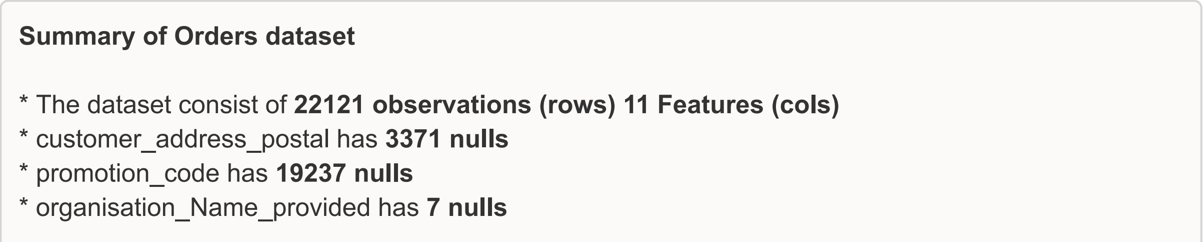
* The dataset consist of **50440 observations (rows) 5 Features (cols incl index)**
* No missing values or nulls or NaNs in all columns
* This dataset spans from **2014-03-04 to 2018-10-19**

From the data dictionary provided in the Assignment PDF
* index == customer index no. (unique)
* created_on == Date Time
* email == string obj
* gender == string categorical
* age == string categorical (bins)

Data validity:
* indexes are unique. No duplicates
* **created_on has 57 duplicates of the created date and time**
    * However this is to be ignored. Nothing wrong with users signing up simultaneously.
    * We would be mor concerned with unique IDs
* **email has no duplicates.**
    * A potential concern was for invalid email addresses
    * Simple check for email was done (Check for @ . and more than 7 chars)
    * 306 was flagged to be invalid
    * Howevere for this case study, since all emails are encrypted, we will assume that they are all valid
* **gender has 3 categories [M / F / Unknown]**
    * Unknown possibilities:
        * 1: Not compulsary to fill in the field or select from dropdown.
        * 2: Can also be seen as 'unspecified'
        * 3: In this day and age M/F might not be enough to describe oneself.
    * Might relook at imputation later.
* **age has 7 bins [18-24 / 25-34 / 35-44 / 45-54 / 55-64 / 65+ / Unknown]**
    * Unknown possibilities:
        * 1: Not compulsary to fill in the field or select from dropdown.
        * 2: Can also be seen as 'unspecified'
    * Might relook at imputation later.

- Built an overall viz based on the data to better understand it
- https://public.tableau.com/profile/ziig.yee#!/vizhome/Oddle_Customer_Overview/CustDataOverview

- **Orders data: https://github.com/ziigyee/Assignment/blob/master/Data%20Overview%20-%20Orders.ipynb**



Summary of Orders dataset

* The dataset consist of **22121 observations (rows) 11 Features (cols)**
* customer_address_postal has **3371 nulls**
* promotion_code has **19237 nulls**
* organisation_Name_provided has **7 nulls**

* This dataset spans from **2016-10-01 to 2018-09-30**

From the data dictionary provided in the Assignment PDF
* submitted_on == Date Time for order submission
* pickup == 1/0 binary
* delivery_date == Date Time
* customer_address_postal == String -> need to convert to 6 digit integer
* promotion_code == String obj
* lead_time == integer (No. of mins) -> Might convert to time object
* basket_size == float ($$ worth of basket)
* customer_index == customer ID, unique integer (Match from customer table)
* menu_id == Merchant ID
* merchant_type == String obj categorical
* organisation_name_provided == binary 1/0 (Corporate or not)

Data validity:
* customer_address_postal: had to check for valid postal codes. Did a clean up to extract 6 digit postal codes. Those invalid was replaced with nulls.
    * postal_good is the new feature to hold the valid postal codes
* promotion_codes: has to do corrections on the strings, this allowed for the correct groupings of the promo codes
    * promo_good is the new feature to hold the corrected promo codes
* lead_time: was in minutes. Was reformated to 'DAYS:HRS:MINS' format, allowed for binning of the lead times
    * lead_time_good is the new feature to hold the reformatted lead times
* **basket_size: has 199 entries of 99999.0** >> These entries are would be deemed problematic, however we can revisit this to either impute or drop.
    * If to be imputed: We would use the median to impute because the mean may be skewed by the rare-occasion large basket. Taking the median is a safer approach.
    * For now, we will keep it and ***EXCLUDE IT** for the Analysis (see tableau viz)
* menu_id and merchant_type: No issues here as they were checked to match up. (no clashing or errors)
* organisation_name_provided: nulls were imputed to 0

    - Built an overall viz based on the data to better understand it
    - https://public.tableau.com/profile/ziig.yee#!/vizhome/Oddle_Orders_Overview/SalesPerformanceDash

- Re-exported the data for orders: orders_cleaned.csv
- Ingested cleaned up data back into the sqlite DB to re-run the SQL query
- Re-run SQL queries on the cleaned up data for Q1A

```
--Server=myServerAddress;Database=Oddle;Uid=myUsername;Pwd=myPassword;

SELECT orders_csv.'basket_size', orders_csv.'menu_id',
orders_csv.'submitted_on' as
'order_date',  orders_csv.'organisation_name_provided' as 'corporate'
,cust_csv.'index' as 'Customer ID', cust_csv.'created_on' as 'signedup_on'
```

```
FROM 'orders_cleaned' AS orders_csv
LEFT JOIN 'Oddle_Assignment Customers Data' AS cust_csv
ON orders_csv.'customer_index' = cust_csv.'index'
WHERE orders_csv.'submitted_on'  BETWEEN '2015-12-31' AND '2017-07-01' --
between the dates (not including)
AND orders_csv.'menu_id' = 4 --Select Merchant 4
ORDER BY datetime(orders_csv.'submitted_on') ASC; --want to see it sorted
by the earliest date
--ORDER BY datetime(orders_csv.'submitted_on') DESC; --want to see it
sorted by the latest date
```

5: Re-attempted assignment: Q1A (caveat with the data starting from Oct 2016 instead of Jan 2016)

- Please see the visualisation here:
- https://public.tableau.com/profile/ziig.yee#!/vizhome/Oddle_Assignment_Q1A/Story1
- (Please view in full-screen)

> 1) The database includes the transactional data of three merchants. Please provide the following analysis:
>
> a. Revenue, orders, and new signups report for Merchant 4 (menu_id 4) to see if its performance have improved from Jan 2016 to Jun 2017

The 3 main metrics that are used in determining the performance are:

- Revenue (Sum of Basket Size)
- Number of Sign ups (Unique Counts of Customer IDs)
- Number of Orders (Sum of Orders)

From the visualisation we can see very similar graphs. The amount of signups are highly correlated to the amount of orders and revenue.

> We do see a performance improvement from May 2017 at $5.4k to $9.2k in June 2017. We see the lowest point in performance at March 2017.

Other info to understand the sales movements are the:

- Average Basket Size per Order
- Median Basket Size per Order
    - Median is a good indicator because its the most ‘common’ basket size to understand customer behaviour. Averages may be skewed by one-off large orders.

Performance improvement can be attributed to several factors such as better marketing lift -> More Sign ups -> More orders -> More Sales. However more data is needed to do a deeper dive.

In a bid to better understand how Merchant 4 (Cake) is performing, I created the second dashboard to compare it with the other bakery merchants (1 and 4). As they are serving up the same type of menu, doing a comparison would allow us to understand what and how better performing merchants are operating. Oddle would be able to provide better advice and consultation to the merchants in the same type. (ie. “These are what successful merchants are doing, better lead times, lower avg prices, more promo codes etc)

6: Question 1B

> b. Customer report based on the type of customers based on gender, age, and customer type (corporate/individual) for all merchants.

SQL query

```
--Server=myServerAddress;Database=Oddle;Uid=myUsername;Pwd=myPassword;

-- This view allows us to see a more wholistic perspective on each customer
id, if they placed orders or not.

SELECT orders_csv.'basket_size',
orders_csv.'menu_id',orders_csv.'merchant_type',
orders_csv.'organisation_name_provided' as 'corporate' ,cust_csv.'index' as
'Customer ID', cust_csv.'age', cust_csv.'gender', cust_csv.'created_on' as
'signedup_on', orders_csv.'submitted_on'
FROM 'Oddle_Assignment Customers Data' AS cust_csv
LEFT JOIN 'Oddle_Assignment Orders Data' AS orders_csv
ON  cust_csv.'index' = orders_csv.'customer_index'
--WHERE  (orders_csv.'basket_size' is null)  --if we want to see those
customers that signed up but no orders placed
WHERE  (orders_csv.'basket_size' is not null)  --if we want to see those
customers that have orders placed
```

- Please see the visualisation here:
- https://public.tableau.com/profile/ziig.yee#!/vizhome/Oddle_Assignment1B/CustomerAnalysis
- (Please view in full-screen)
- Dashboard allows us to have an overall view
  - You can browse and filter by merchant, date range, corporate/individual
  - Gender breakdown per age group
  - Gender Percentages breakdown by menu
  - Corporate / Individuals (organization_name_provided) breakdown
  - Customer baskets by menu (Avg basket size, Median, Total orders)
- We have many Unknowns for Age:
  - Too little data to impute;
  - Imputation method to try: KNN
  - Other Method:
    - Exclude the Unknowns.
    - Get the Percentage of the known age groups, distribute the

        Unknowns by that percentage
    - Reasons that Unknown is recorded in the data:
        - Not compulsary to fill in the field or select from dropdown.
        - Can also be seen as ‘unspecified’ when users fill in the form
        - Might be able to speak to engineering to understand how this data is collected and improve it

Customer Profiling:

- Females make up the larger portion of customers, about 70% across all merchants
- Although corporate sales are a lot less compared to individual sales, the average basket size is larger. (As expected for corporate)
- Menu 5’s Dim Sum is the overall best performer in revenue
- Bulk of the customers are in the 18-24, 25-34, and 35-44 age groups, 25-34 being the highest count of customers

Conclusions as points of engagement for Oddle to support its merchants:

- Boost marketing to corporate customers
    - Geographical marketing (CBD), “lunch time office specials” etc.
- Targeting customers in the 25-34 age range
    - Social media marketing audience
- Targeting by gender
    - Different ways. ie. keywords for digital can be ‘girls night out’ etc

Inactive Users: Have Signed up but have not make any transactions

- Amidst the active sign ups, we also should look at those inactive users. Marketing might be able to reach out to them and incentivise them to get back onto the platform/merchants and make purchases.
- They haver the same profile as the active audience
- https://public.tableau.com/profile/ziig.yee#!/vizhome/Oddle_Assignment1B_NoOrders/InactiveCustomerAnalysis

7: Question 1C:

> c. Monthly lifetime revenue cohort analysis of the three bakery merchants, split by customer signup month

- Please see the visualisation here:
- This is for merchant menu 1,3 and 4
- https://public.tableau.com/profile/ziig.yee#!/vizhome/Oddle_Assignment1C/CohortRevenueAnalysis
- (Please view in full-screen)
- Dashboard allows us to have an overall view

  - You can browse and filter by menuID and date range
  - Revenue and sign ups, split Sign up month

We see a slight upward trend (although not stable) for signups and revenue.

- the more merchants get customers to sign up, the better the revenue.
- Sales data that we have only started from Oct 2016 (All revenue generated is from Oct 2016 onwards)
- First active user's 'age' goes back till Sept 2014, however only transacted on January 2017.
  - Note: This graph is based on signup month, not month of spend
- Difference in number of customers and number of orders = repeat customers
- MenuID 1, and 3 are outperforming MenuID 4 by almost $100k in revenue
- Since then MenuID 4 is headed on an upward growth, however menu 1 and 3 are seeing a decline in their revenue.
- From June 2018 we see that the performance of all 3 menus converging.
- Menu Id 4 still generally performing lower possibly because (deeper dive with more data required for clearer analysis)
  - Factors such as capacity and menu item offering may be affecting sales
  - Lack of marketing efforts
  - Promo code usage etc

8: Question 2:

> Oddle has a new bakery, Baker World, that is using our solution. Based on the analysis you have done above, please form three hypotheses and describe how you would conduct the experiment to see if they are true for Baker World.

Some assumptions that I will make:

1. Baker World being new, is already seeing some traction in sales, however is still trying to figure out what works best for them (design, configuration, types of food to offer) and are working with Oddle to improve.
2. Having had some activity on the platform, we do have, although limited) data to start off.
3. There is constant online marketing done to drive traffic to the site, which should be stable or even in a decline now. (Post launch)

Hypothesis 1:
IF Baker World offers promotional codes, THEN it will increase its revenue BY 10% Over 2 months.

- We will run this experiment over 2 months
- We take the average sales over a period of time (ie. past 2 months) without promo codes
  - This sets the baseline that we can compare and measure from. (X %

increase from this)

- Baker World begins offering with promo codes with their marketing efforts.
  - We begin to measure, traffic and conversion revenues.
- We will compare the measurements week by week.
- If the experiment is successful, we should be able to see an increase in revenue by transactions that have promo codes applied to it.

Hypothesis 2:
IF Baker World's added a hero product feature on their landing page, THEN it will increase sales of that product by 15%

- We will run this experiment over 3 months
- If successful, this feature would be permanent, giving the merchant the option to change the hero product
- Analyse Baker World's current basket compositions to see Baker World's 'most added to basket' product.
- The baseline metric would be 'count of product in basket'
- Study other merchants that have this or similar feature, within Oddle's client base (data) and externally (design and use cases)
- We will propose this experiment and feature to the client, through the account manager with the supporting evidence from the data we initially collected.
- Upon agreement with the client, engineering and design will create the feature.
- Based on the monthly average traffic to the clients site, we will determine our population for the A/B test
- In this experiment: A = site with Hero feature, B = site without hero feature, product in catalog as per normal.
- Engineering would create a function to split the traffic, and tag the traffic as it enters
- Additional data to collect is 'basket_source' = A or B for us to know which site it came from.
- Over a period of 3 months, we will monitor traffic and track basket composition, review every month
- Measure and compare results to baseline.
- If successful, this feature would be permanent, giving the merchant the option to change the hero product

Hypothesis 3:
IF Baker World increases its product offering by X products , THEN it will see an increase in average basket size of Y% Over 6 months

- To qualify this hypothesis, we will compare its current amount of products offered by looking at other merchants in the similar category that have been using Oddles solution for a longer period of time (by at least 1 year)
- We will compare its amount of products offered and the average basket sizes.
- Comparing with Baker Worlds current average size, we will filter out merchants that are in the same category, with basket sizes that are at least 20% larger. (ie.

current basket size: $60. We will only look at same category merchants that have basket size of $72)

- Having filtered out those merchants that have larger basket sizes, we look at the amount of products offered and the average price of each product offered.
- We measure Baker World's current average basket size, and we note their current amount of products offered
- With this information, we can form the hypothesis:
- Ie: IF Baker World increases its product offering by 3 products , THEN it will see an increase in average basket size at least 10%
- We will propose this experiment and hypothesis to the client, through the account manager with the supporting evidence from the data we initially collected.
- Upon agreement with the client, we will await the clients response in preparing the additional product offering
- Setting a start date for the experiment, and begin monitoring a week before the experiment start date
- Upon start date, we will release the new products, we begin to measure, traffic and conversion revenues.
- We will compare the measurements week by week.
- If the experiment is successful, we will be able to see an increase in average basket size by 10%

END ASSIGNMENT

Ziig Yee.