Phase 3

"Is Obesity an outcome of Financial stress?"

The topic of the report is related to Health and Finance domains.

Consumptions of food to compensate for stress is called stress eating and is a term commonly used nowadays. In this study, we are aiming to find the relationship between Financial stress and Obesity due to stress eating. In our preliminary investigation, we obtained interesting results hinting a relationship, we hope to do a deeper analysis and answer the question.

Datasets

The following Datasets are used:

- 2011 LGA Health Risk Factors (Mar/2011-July/2013): Modelled estimate of several health risk factors by LGA. This Dataset contains the estimated data of the percentage and count of persons who are experiencing health risk factors. It comes with a reliability rating for each estimation. Dataset downloaded from Aurin. https://portal.aurin.org.au/("personal_financial_stressors.csv").
- 2011 LGA Personal and Financial Stressors (Nov/2010). This dataset contains the percentage and count of persons who are having financial problems by LGA (e.g. Cash flow problems). It comes with a statistical significance(reliability) rating for each estimation. Dataset downloaded from Aurin. https://portal.aurin.org.au/("health_risk_factors_estimates.csv")

**The 2 datasets are from different years because stress cannot cause obesity immediately, it may take months or years for a person's mass to be in the obese zone. In this case, the financial stress caused in 2010 would affect the number of people who are obese in 2011,2012 and 2013. Ideally, we would like to have 2 datasets with its time periods overlapping each other but unfortunately such data is not available however, the current datasets we have found will be sufficient for the analysis.*

Pre-processing

The following 2 operations were done using Microsoft Excel:

1. 2011 LGA Health Risk Factors csv file contains data relating to all the health risk factors therefore, we created a new csv file named "new_health_risk.csv" with the columns only pertaining to obesity. Furthermore, we renamed the columns with shorter names with '_' instead of spaces.
2. 2011 LGA Personal and Financial Stressors csv file contains data with many different forms of Personal and Financial and personal stressors. We therefore, created a csv file with the columns we need called "new_stress.csv". This csv file contains columns count, percentage and the statistical significance of:
    - Cash flow problem in the last 12 months.
    - Could Raise 2000 dollars within a week.
    - Government support as main source of income in the last 2 years
    - At least one dissaving action in the last 6 months.

   Plus, columns were renamed with shorter names with '_' instead of spaces.

**The rows with "Unincorporated Vic" as the LGA name was removed since 2011 LGA Personal and Financial Stressors csv file did not contain data for "Unincorporated Vic" LGA. We did not want to impute data for this LGA as it will contain error. Therefore, having only 79 LGAs for comparison is a limitation.*
Following operations were done using python:

For us to compare these variables, we needed to have normalized values (e.g. percentages, rate per 100,000 people). Fortunately, both the above datasets contained columns containing the percentages of the data we need. Therefore, we did not have to calculate the percentage using the count and the total number of people in the LGAs. However, to make sure that all the percentage values are within its possible range, we calculated the number of percentage values greater than 100 and less than 0. There were no percentage values falling under this criterion. Therefore, we did not have to write any python code to correct any value out of range.

Integration

Using the merge attribute in pandas, we performed inner join on the 2 datasets ("new_health_risk.csv", "new_stress.csv") and created a new data frame. This newly created data frame was named "new_join". The LGA code was used as the key to join the 2 datasets. We did not experience any difficulties when joining these 2 datasets since both of them were downloaded from Aurin containing consistent LGA codes and the null record were removed ("*Unincorporated Vic*").

Next, we checked if the data we are analyzing is reliable. We used the reliability rating for each estimation to calculate the average reliability rating for each column we are going to use and the number of people used to obtain the percentages (data size). The following table shows the results.

Key	Column Name	Reliability Rating	% of Data Size < 10,000
C1	% of obese population (Persons)	3.00	21.52
C2	% of obese male population	2.97	43.04
C3	% of obese female population	3.00	43.04
R1	% of people with cash flow problems in the last 12 months	2.95	21.52
R2	% of people who has government support as the main source of income in the last 2 years	3.00	12.66
R3	% of people with at least one dissaving acti on in the last 6 months	3.00	20.25
R4	% of people who could raise 2000 dollars w ithin a week	3.00	16.46

TABLE 1(Key for the column names used for this analysis, its reliability rating and the data sizes)

Reliability rating ranges from 1 to 3 where 1 being unreliable and 3 being reliable. Since all the columns have an average reliability rating of 2.95 or more, we could conclude that the data we will be using is highly reliable. Furthermore, we checked the number of people used to obtain the data in the datasets (data size). Using this we calculated the percentage of LGAs with data sizes less than 10,000 in each column. 10,000 was selected because we believe that it should at least have 10,000 records to be substantial data to rely on. 43.0% of the LGAs in C2 and C3 are with data sizes less than 10,000, this is too high to provide a solid conclusion. However, since the reliability ratings are high for C2 and C3, we will use them excluding the main calculations (comparisons with R1, R2, R3 and R4). Other columns provide better percentages compared to C2 and C3. We did not remove the LGAs with data sizes less than 10,000 because we only have 79 LGAs, removing them would make this analysis extremely biased.

Results

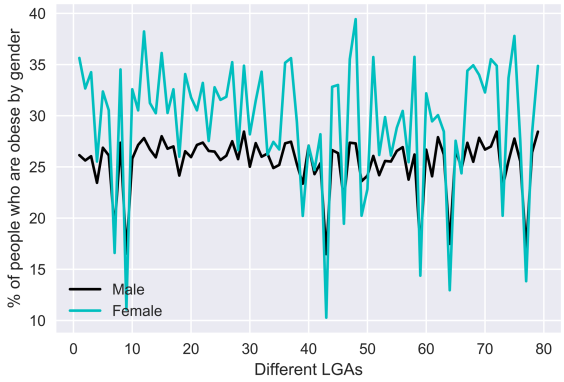


FIGURE 1(% of people who are obese by gender for each LGA) **FIGURE 2**(Pearson Correlations between Financial stress features and Obesity(C1))

FIGURE 1 shows us that, the percentage of male obese population behaves similarly to the percentage of female population in each LGA. We can identify that they are having similar patterns visually, therefore it is quite possible that the results we receive for the whole obese population will not differ greatly by gender.

See TABLE 1 for the key. R1, R2 and R3 are activities which display financial stress while R4 is an activity which displays less or no financial stress. For R1, R2 and R3, there is a strong positive linear relationship with C1(0.63-0.72). This means whenever R1, R2 or R3 increases, C1 will increase as well. This supports our theory that Financial Stress is proportional to Obesity. So, when financial stress decrease does obese percentage decrease as well? To answer this question, we checked R4's Pearson Correlation Coefficients against C1. According to Figure 1, they have a moderate negative linear relationship. This means that whenever R4 increases, C1 decreases. Limitation: It is impossible to receive a coefficient of 1 or -1 in real life due to exceptional events. In this case, it would be events such as genetic obesity and the presence of "lean gene". (Srivastava, A., Srivastava, N. & Mittal, B. Ind J Clin Biochem ,2016) These findings further support that financial stress is directly proportional to obesity in most of the cases.

Since the correlations are strong, we could use regression lines to predict continuous data. We drew regression lines for the R1, R2, R3 and R4 against C1(*see key*). The charts below show the predictions:

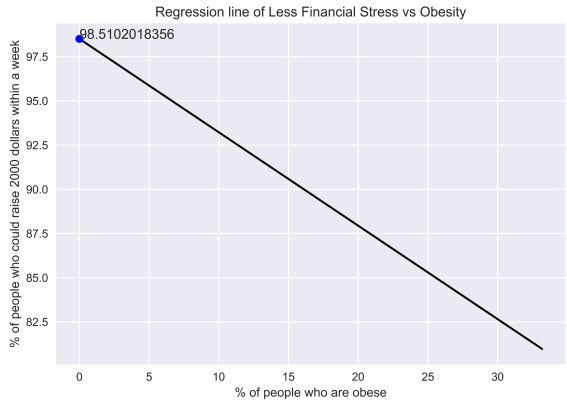


FIGURE 3(Regression line of Financial stress Vs Obesity) FIGURE 4(Regression line of *less* Financial stress Vs Obesity)

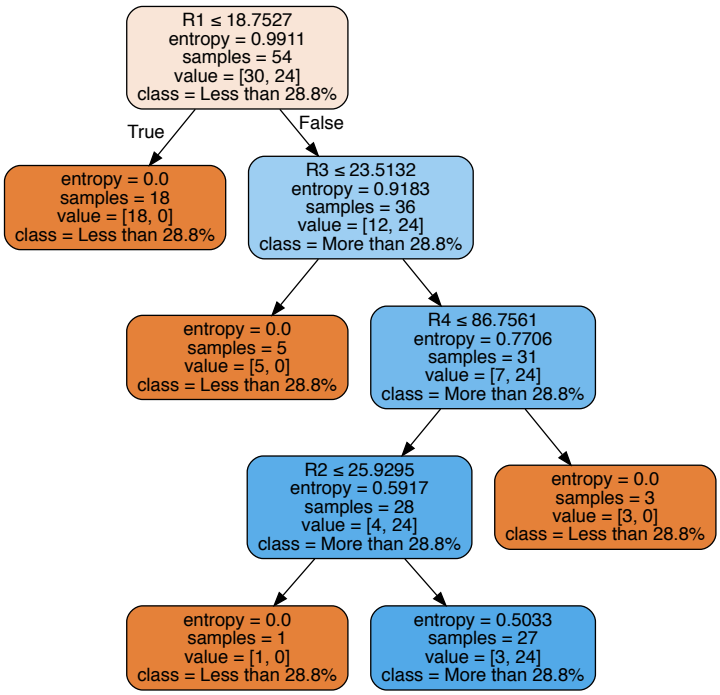
Figure 3 and Figure 4 points out the conditions which may cause 0% obese population. However, a 0% obese population is not achievable due to other factors which are causing obesity (genetic obesity/ lean). We will talk about external factors like this later in the report. The estimations provided by the graph is theoretical, but in real life there won't be a huge difference.

According to Figure 3 and Figure 4, if any of the following actions occur the % of people obese will reduce to 0%

1. % of people with cash flow problems in the last 12 months reduces to 4.81%.
2. % of people who has government support as the main source of income in the last 2 years reduces to 2.75%
3. % of people with at least one dissaving action in the last 6 months reduces to 19.52%
4. % of people who could raise 2000 dollars within a week increases to 98.51% (98.51%-100% genetically lean)

For further analysis, we break our dataset into 2 called the training dataset and testing dataset. The training dataset has 54 LGAs, while the testing dataset has 25 LGAs. **limitation: not enough* data* Furthermore, we divided the percentage of people who are obese in each LGA into 2 classes. We calculated and found out that half of the percentages are below or equal to 28.8% and half of the percentages are above 28.8%, therefore we classified it as below or equal to 28.8% and above 28.8%.

Firstly, we fit the **K-nearest neighbour** classifier with K=5 on the training dataset. Using the model created, we predict the classes for the % percentage of people who are obese for the testing dataset. 68% of the predictions were correct, however we ran an algorithm changing the K value checking which K value would give the best accuracy. We found out that K=9 gives the best accuracy with an accuracy of 84%. This shows that the relationship between Financial stress and Obesity can be used to estimate the class the percentage of obesity would fall on, successfully. We have identified that since the values above and below 28.8% are equal, there's a 50% chance of getting the test correct if the model estimated 1 class throughout the whole dataset, therefore obtaining an accuracy of 84% we can make sure that it has not been the case.



Next, we used the same training and testing datasets to build a model for a **decision tree**. The following figure shows the decision tree we used which provided the best accuracy of 71.9%The diagram shows the simple steps that needs to be taken to predict the class of the percentage of people who are obese. This shows that whenever R1, R2 and R3 is less than 18.75%,25.92% and 23.51% respectively, the predicted class is the percentage of obese people less than 28.8%. Moreover, we can deduce from the decision tree that whenever R4(low financial stress factor) is greater than 86.76% the predicted class is the percentage of obese people less than 28.8%. This shows that it's quite possible that the percentage of people with financial stress is directly proportional to the percentage of people obese.

Figure 5(Decision Tree to predict the class of the percentage of people who are obese)

Value

Although the datasets had the data we needed for this project, it was in a form which was only numerical, by simply looking at that data it was not possible to see any relationship. Pre-processing added a lot of value to the data in many ways, we were able to calculate and find the number of people who were used in each LGA to compile the data. This helped us understand the quality of the data we use, which added value to our processes later. Moreover, calculating the average reliability rating provided us with the level of reliability of the data. The integration of datasets into one helped us visually create a mind map to perform the rest of the analysis.

The use of visuals rather than just numeric data allowed to identify trends and patterns well. Furthermore, the regression line graphs showcase the similarities and differences between the financial stress features and obesity effortlessly. Percentage accuracy easy to understand as well. Moreover, the decision tree diagram is simple to read and easy to follow when classifying. It provides a visual path which explains the relationship. After all, visualization is the best value we added to the data to understand the relationship.

Challenges and Reflections

We had to change our question from **"**Will financial support reduce obesity?" to "Is obesity an outcome of Financial stress?" since we did not have sufficient data relating to financial support. Finding a dataset for financial stress with sufficient data was challenging at first. However, the dataset we found provided many features which allowed us to do a deep analysis. Unfortunately, the 2 datasets available weren't from the same year (see explanation under Datasets). When trying to inner join the 2 datasets, we came across a LGA with null values. We had to remove *"Unincorporated Vic"* from the 2 datasets and rejoin. Moreover, we attempted to find the correlation between obesity by gender and financial stress but hit a dead end when we realized that financial stress cannot be divided by gender with the available data. Therefore, we decided to compare percentage of obese male population with female population to check if there's a difference. There was another failed attempt trying to find the NMI, we were unable to divide the continuous data into classes properly. Creating the decision tree was challenging but after several tries, we were able to create a decision tree successfully.

Question Resolution

Using the result achieved, we are able to resolve the question we proposed in the beginning. However, there may be other factors which are not accounted for in the analysis. Factors such as the number of fast food restaurants in the area, number of sport complexes available in the area and means of transport could affect the percentage of obesity in a LGA. Furthermore, the average income in the LGA may affect the financial stress features we used in the analysis. It is not possible to consider all of these factors when analysing the relationship.

Moreover, there is a chance that the person who is obese may not be the same person who's experiencing financial stress. Therefore, we used various methods to understand the relationship. Since figure 1 gave us insight into the similarity between male and female obesity percentages, the rest of the results applies for both the genders equally. The strong positive linear relationship between financial stress and obesity, and the moderate negative linear relationship between less financial stress and obesity provided us results that hints that Financial stress and obesity is directly proportional. Afterwards, we used regression to check if it would provide any interesting results, the line graphs showed a clear difference between the less financial stress feature and financial stress features. The results from the K-nearest neighbour and the decision tree supported the aforementioned relationship as well by resulting high accuracies. Furthermore, the decision tree showed us clearly how the factors affect the prediction of the obesity class. Therefore, taking the above results into account we can say obesity is an outcome of Financial stress.

The government, health sector and finance sector will be interested in this relation since obesity is the main reason for many serious health conditions. Reducing the obesity population will be an excellent accomplishment itself, moreover it will also decrease the expenses at hospitals, thus benefiting the government as well.

Code

More than 350 lines of code was used to produce the results we achieved. Around 50% of the code was written from scratch and the rest was taken from workshops. To analyse the data we used Pandas, Numpy, Matplotlib, Scipy and Sklearn libraries. The code used to fit a K-nearest neighbour classifier and decision tree was obtained from the week 8 tutorial. We had no need to use any other language than python since all the operations provided by python was sufficient for our dataset analysis. The "readme" file lists the structure of the code used.

Bibliography

Srivastava, A., Srivastava, N. & Mittal, B. Ind J Clin Biochem (2016) 31: 361. doi:10.1007/s12291-015-0541-x