Applied Data Science Executive Summary

The datasets consist of data from 420 million food items purchased by 1.6 million loyalty card owners at Tesco in Greater London in 2015. This includes the average nutrient content of fat, sugar, protein, fibre, saturated fats, carbohydrates, and salt of food and drink items segmented by Borough, Ward, MSOA, and LSAO, with a focus on yearly averages for Boroughs and Wards. Additional data covers child obesity, diabetes estimates, general obesity, and obesity hospitalization rates across London, specifically using diabetes data for Wards and obesity rates for Boroughs for the case study.

The initial analysis revealed the following correlation coefficients between obesity and nutrient weights: fat (0.06), sugar (0.40), protein (-0.25), fibre (-0.15), saturated fats (-0.01), carbohydrates (0.45), and salt (0.11). These coefficients, which range from -1 to +1, measure the strength of linear relationships between variables, where +1 indicates a perfect positive relationship and -1 a perfect negative one. The results suggest an inverse relationship between obesity and both protein and fibre, with protein showing a lesser effect. Meanwhile, fat, sugar, carbohydrates, and salt exhibited positive correlations with obesity, indicating a direct relationship. However, saturated fats, with a correlation close to zero, appear to have a negligible impact on obesity rates. This suggests that, except for saturated fats, increased consumption of these nutrients may be linked to higher obesity rates.

The second analysis focused on Type-2 Diabetes, revealing that the correlation coefficients between diabetes and both fibre and protein were -0.31 and -0.52, respectively. This suggests that both nutrients are inversely related to Type-2 Diabetes, with protein having a more pronounced effect. Conversely, carbohydrates, sugar, saturated fat, and fat demonstrated positive correlations with Type-2 Diabetes, suggesting a direct relationship with values of 0.65, 0.47, 0.36, and 0.34 respectively, listed from the most to the least impactful. The correlation with salt was a negligible 0.01, indicating it has a minimal impact on Type-2 Diabetes.

Another dataset was imported from the office for national statistics [1], which included average income, taxes, and benefits of decile groups of all households in London boroughs from the years 2002 up until 2022. This dataset was cleaned, and the 2015 weekly income values were merged with the Tesco nutritional dataset for London boroughs from the year 2015. The dataset included a margin of error for the incomes, this was assumed negligible for ease of analysis. The analysis performed on these datasets showed that over the years, the average income across households increased non-linearly from £392.0 to £610.0 from the years 2002 till 2022 with a plateau from the years 2009 till 2015.

The correlation between average income and nutrient purchases showed positive values for protein (0.42) and fibre (0.23), and negative values for saturated fat (-0.34), salt (-0.35), fat (-0.37), sugar (-0.46), and carbohydrates (-0.66), indicating that higher incomes increase protein and fibre intake, but decrease other nutrients. The linear regression model, validated with an 80/20 training-validation split and an MSE of 0.0606, reliably predicted nutrient weight changes from 2002 to 2022. It was then trained on the entire dataset and projected increases in protein from 5.21g to 5.40g and fibre from 1.60g to 1.64g. Conversely, the model forecasted decreases in fat from 9.12g to 8.84g, sugar from 10.63g to 9.47g, saturated fats from 3.59g to 3.48g, carbohydrates from 19.37g to 16.74g, and salt from 0.60g to 0.57g, with all nutrients having a plateau from 2009 to 2015, affirming the initial analysis that higher incomes impact dietary changes over the years.

In conclusion, out of all the nutrients, protein and fibre were the only two that had a negative correlation with obesity and Type-2 Diabetes and the foods that contain a large amount of these nutrients are among the more expensive foods [2]. This in turn is reflected in the correlation and predicted nutrient weight as it showed that when people started earning more money over the years, their intake of protein and fibres increased while the rest decreased as they started to eat healthier with protein and fibres considered to be among the healthier array of food.

Bibliography

[1] (2023) annual survey of hours and earnings - resident analysis – *Nomis, Office of National Statistics*. Available at: https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=30#

[2] Rao, M. et al. (2013) Do healthier foods and diet patterns cost more than less healthy options? A systematic review and meta-analysis, BMJ open. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3855594/