Study of Tesco Grocery 1.0 dataset [2] vs Income

Overview

This study, presents the Tesco Grocery 1.0 dataset [1] which contains multiple datasets recording 420 million food items purchases by the Tesco Clubcard holders who have shopped over the course of the entire year of 2015. The datasets are aggregated at different levels of census areas (LSOA, MSOA, Borough and Ward) in accordance with the ONS requirements to preserve anonymity. The information for each area includes the number of transactions as well as the nutritional qualities of the typical food item purchased across several food categories such as sweets, fruits and vegetables, eggs, and fish. Additionally, the correlations between health issues such as diabetes and obesity and food products and their nutritional composition are investigated.

Dataset Bias and Limitations:

1. The representativeness of the data may be limited to Clubcard owners, possibly introducing a bias towards certain socioeconomic groups who are more likely to own and use loyalty cards and not representing the overall population.
2. The dataset collects data of a specific timeframe and geographical areas. This might not generalize well to other regions or periods.
3. Nutrient data are averages and may not accurately reflect the diversity of individual products.
4. Health-related outcomes in the dataset cannot be used to represent genuine health issue-causing causes because many other variables such as lifestyle, genetics, and socioeconomic status are not addressed, all of which have a significant impact on health outcomes.

Dataset Assumptions:

1. The data collected by Tesco is done in compliance to privacy laws and regulations.
2. The data is collected from its own database and it is as accurate as it can be.

Questions to research:

1. Do demographic factors like gender and age influence food purchasing patterns across different areas?
2. Are there correlations between food choices/nutrients and obesity-related hospitalizations?
3. How does income data trend over years and is there a correlation with grocery purchasing patterns of customer?

Methods & Insights:

1. **Method:** Male-female and age ratios were estimated, and two new columns were added to the dataset, categorizing the regions as Male-dominated, Female-dominated, Balanced for Gender and Low-youth areas or High-youth areas for age based on their young population concentration.
   Insights: Male-dominated areas show a preference for more energy-dense food categories like dairy, grains and eggs. Female-dominated areas tend to consume more readymade foods and sweets. Areas with a higher number of younger populations consume more grains and sweets, which may reflect lifestyle or energy choices. Areas with a lower proportion of younger residents had a higher intake of fruits and vegetables, and ready-made dishes, showing a preference for convenience and health. This adheres to the results of a Cambridge study [3].
2. **Method:** The obesity – related hospitalization data of London borough region is merged with the Tesco dataset and cleaned for any null values. Spearman correlation method was used to find the strength of correlation between these two.
   Insights: The data shows counterintuitive and unexpected correlations between dietary intake and obesity-related hospitalizations. Higher intakes of fat and sugar are associated with lower hospitalization rates, which might suggest issues with the data's completeness [4]. Conversely, higher protein intake correlates with increased hospitalization rates, possibly reflecting unmeasured dietary or lifestyle factors. On the other hand, a higher fibre intake is associated with lower hospitalization rates, consistent with advice that fibre is beneficial for health.
3. **Dataset:** Gross Disposable Household Income Dataset [2] - The average income of all individuals in the borough for 15 years is sourced from ONS (Office for National Statistics). This dataset does not take into account the income of self-employed and pensioners. For exploration, the total gross disposable household income (in millions) table and total resident population numbers is used.
   Method: Created Line charts displayed trends in income and population, while a scatter plot is used to examine the relationship between GDHI and population growth. The Spearman correlation method was used to analyse the correlation between 2015 income levels across boroughs and customer purchase patterns.
   Insights:
   - Most regions, including Wandsworth, Kensington and Chelsea, and Westminster, show rising GDHI levels, indicating increased income. Newham and barking and Dagenham have seen significant population growth. There's a clear link between GDHI and population growth in areas like Southwark and Greenwich, where both economic and demographic growth are aligned.
   - Higher income groups often prefer premium or health-oriented products, such as organic grains and bottled water. This preference also extends to high-quality proteins like fish and poultry [5]. Conversely, the negative trends in dairy, wine, and fruit/vegetable consumption might reflect dietary preferences for non-dairy alternatives or other lifestyle choices typical of wealthier demographics.

References:

1. Aiello, Luca Maria; Schifanella, Rossano; Quercia, Daniele; Del Prete, Lucia (2020): Tesco Grocery 1.0. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.4769354.v2
2. GDHI_Income: - https://www.ons.gov.uk/economy/regionalaccounts/grossdisposablehouseholdincome/datasets/regionalgrossdisposablehouseholdincomelocalauthoritiesbyitl1region
3. Bender AE. Food preferences of males and females. Proceedings of the Nutrition Society. 1976;35(2):181-189. https://doi.org/10.1079/PNS19760031
4. Scheelbeek PFD, Cornelsen L, Marteau TM, Jebb SA, Smith RD. Potential impact on prevalence of obesity in the UK of a 20% price increase in high sugar snacks: modelling study. BMJ. 2019 Sep 4;366:l4786. doi: 10.1136/bmj.l4786. PMID: 31484641; PMCID: PMC6724407.
5. RHE Global – Smarter Public Protection: - https://www.linkedin.com/pulse/how-poor-eat-hatchett-rheglobal/