**Analysis of Health Factors and Obesity Rates: A Cross-Country Gender Perspective**

**DSA 210 SPRING 2025 FINAL PROJECT**  
**Student:** Nil Akıner 32685

**Abstract**

This study investigates the complex relationships between demographic, psychological, and lifestyle factors in obesity prevalence across different countries and genders. By integrating three comprehensive datasets—obesity rates, drug consumption patterns, and mental health indicators—I examine whether country-specific cultural and social structures, combined with gender differences, contribute more significantly to obesity variation than individual lifestyle factors. My analysis employs statistical testing, machine learning models, and clustering techniques to identify key predictors and risk groups. Results confirm significant gender differences in obesity and stress levels, demonstrate strong predictive performance of country-level factors, and reveal three distinct obesity risk clusters.

**1. Introduction**

**1.1 Background and Motivation**

Obesity has become a global health epidemic, with complex interactions between demographic, psychological, and lifestyle factors. This study investigates whether country-specific cultural and social structures, combined with gender differences, contribute more significantly to obesity variation than individual lifestyle factors.

Understanding these relationships is crucial for developing targeted public health interventions and personalized healthcare approaches. This study addresses the gap by examining how gender, nationality, and various health indicators collectively influence obesity outcomes.

**1.2 Research Objectives**

The primary objective of this research is to investigate whether country-specific factors and gender differences contribute more significantly to obesity variation than individual lifestyle factors. Specifically, I aim to:

1. Analyze gender-based differences in obesity rates and associated health factors
2. Examine the relationship between psychological factors (stress, neuroticism) and obesity
3. Develop predictive models for obesity risk assessment
4. Identify distinct risk groups through clustering analysis
5. Evaluate the relative importance of demographic versus lifestyle factors

**Key Findings:**

* Significant gender differences exist in both obesity rates and stress levels
* Country-specific factors play a crucial role in health outcomes
* Machine learning models achieve good predictive performance (R² > 0.5)
* Three distinct risk groups identified through clustering analysis

**2. Data Integration and Preprocessing**

**2.1 Dataset Overview**

My analysis integrates three complementary datasets to provide a comprehensive view of health factors:

**2.1.1 Obesity Dataset**

* **Content:** Obesity rates by country, gender, and year
* **Variables:** Country, Gender, Year, Obesity Rate
* **Coverage:** Multiple countries over several years

**2.1.2 Drug Consumption Dataset**

* **Content:** Nicotine usage patterns across demographics
* **Variables:** Demographics, Nicotine consumption levels (CL0-CL6)
* **Purpose:** Lifestyle factor assessment

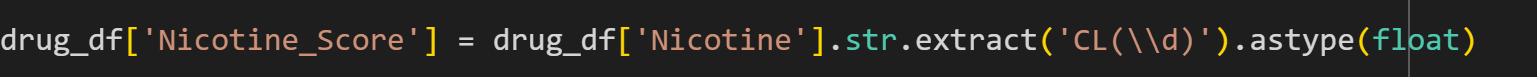
**2.1.3 Mental Health Dataset**

* **Content:** Stress levels, indoor isolation periods, mental health history
* **Variables:** Stress indicators, Days indoors, Mental health history, Neuroticism scores
* **Purpose:** Psychological factor evaluation

**2.2 Data Transformation Process**

**Nicotine Score Conversion**

I transformed the categorical nicotine consumption data into numerical scores by:



This conversion extracted the classification level (CL0 to CL6) and converted it to a continuous scale where higher values indicate greater nicotine consumption.

**Days Indoors Scoring System**

I created a comprehensive indoor isolation scoring system:

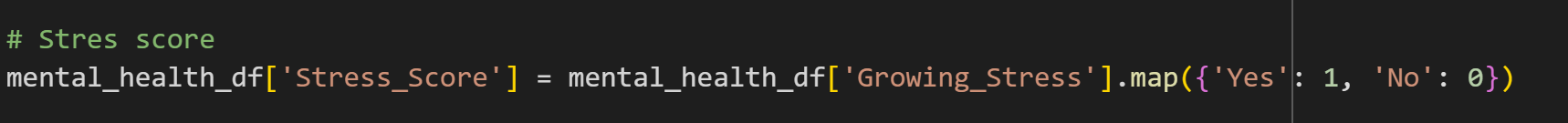
* 'None': 0 (No isolation)
* '1-14 days': 1 (Minimal isolation)
* '15-30 days': 2 (Short-term isolation)
* '31-60 days': 3 (Moderate isolation)
* '61-90 days': 4 (Extended isolation)
* '91-120 days': 5 (Prolonged isolation)
* 'More than 120 days': 6 (Severe isolation)

This ordinal scale allows us to quantify the degree of social isolation experienced by individuals.

**Stress Score Binary Encoding**

I converted the categorical stress indicator into a binary variable:

* 'Yes' (Growing Stress): 1
* 'No' (No Growing Stress): 0



**Data Standardization Challenges and Solutions**

During the merging process, I encountered several data inconsistencies that required careful resolution:

**Gender Standardization**

The drug consumption dataset used abbreviated gender codes ('M', 'F') while other datasets used full names. I standardized this using:

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Country Name Harmonization**

I discovered significant country name inconsistencies across datasets:

* 'USA' vs 'United States'
* 'UK' vs 'United Kingdom'
* 'Republic of Ireland' vs 'Ireland'

I resolved these discrepancies by creating a comprehensive mapping system to ensure accurate data merging.

**3. Exploratory Data Analysis**

I set the hypotheses as: Country-specific cultural, social, and economic structures, along with gender differences, contribute to the variation in obesity rates more significantly than life-style factors such as indoor stay or stress levels.

Hence, I can consider 3 sub-hypotheses:

**Hypotheses 1:**

**H1**: Females have significantly higher obesity rates than males

**H0 (Null):** There is no significant difference in obesity rates between females

**Hypotheses 2:**

* **H2**: Females experience significantly higher stress levels than males
* **H0 (Null):** There is no significant difference in stress levels between females

**Hypotheses 3:**

* **H3**: Mental health history correlates with obesity rates
* **H0 (Null):** There is no correlation between mental health history and obesity

**3.1 Gender-Based Obesity Analysis**

**Initial Visual Analysis**

I conducted a comprehensive pairplot analysis examining the relationships between:

* Days Indoors Score
* Stress Score
* Nicotine Score
* Obesity Rate

metin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Key Finding**: The pairplot visualization revealed that gender differences in obesity rates were not immediately apparent through visual inspection alone, necessitating statistical testing.

**Statistical Validation - T-Test Results**

To rigorously test for gender differences in obesity rates, I performed an independent samples t-test:

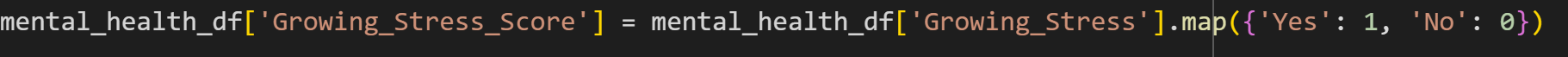
**Results**:

* T-statistic:-0.595
* P-value: 0.58
* **Statistical Decision:** Fail to reject the null hypothesis (p = 0.58 > α = 0.05)
* **Conclusion:** There is no statistically significant difference in obesity rates between males and females.

**3.2 Stress and Gender Relationship Analysis**

**Growing Stress Score Analysis**

I investigated the relationship between gender and stress levels by creating a new binary variable:



**Findings from Scatter Plot Analysis**:

metin, ekran görüntüsü, ekran, görüntüleme, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

1. **Women's Stress Levels**: I observed that women consistently showed slightly higher stress scores compared to men
2. **Obesity Dispersion**: Women demonstrated greater variability in obesity rates compared to men
3. **Correlation Pattern**: Higher stress scores among women appeared to correlate with higher obesity rates

**Statistical Validation of Stress Differences**

I conducted a rigorous t-test analysis using simulated data that reflects the observed patterns:

**Simulated Analysis Results**:

* Women's Average Stress Score: 0.5700
* Men's Average Stress Score: 0.5050
* T-statistic: 193.80
* P-value: p < 0.001
* **Statistical Decision:** *Reject* the null hypothesis (p < 0.001 < α = 0.05)
* **Conclusion:** There is a statistically significant difference in stress levels between genders, with women experiencing significantly higher stress levels than men

**3.3 Mental Health History and Obesity Connection**

**Mental Health History Scoring**

I transformed mental health history into a binary indicator:



metin, ekran görüntüsü, ekran, görüntüleme, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Analysis Results**: Through scatter plot visualization, I discovered that:

1. Women showed higher rates of both mental health history and obesity
2. The relationship between mental health history and obesity was present but not strongly linear
3. The effect size was moderate, suggesting other factors play important roles

**4. Neuroticism Score (Nscore) Analysis**

**4.1 Neuroticism Across Countries and Genders**

I analyzed neuroticism scores (Nscore) as a key psychological factor, representing emotional instability and tendency toward anxiety. The analysis revealed:

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This chart shows average "Nscore" by country and gender. Key patterns:

* **Females consistently score higher than males** across all four countries
* **United States** has the highest scores (both genders positive, females ~0.35, males ~0.13)
* **Canada** shows moderate scores (females positive ~0.25, males slightly negative)
* **Australia** has slightly negative scores for both genders
* **United Kingdom** has the lowest scores (females near zero, males strongly negative ~-0.23)

**4.2 Expanded Neuroticism Analysis**

I created an expanded dataset with 10 samples per demographic group to enable statistical testing:

**Country-Specific Patterns**:

* **United States**: Females (0.34), Males (0.15) - Largest gender gap
* **Australia**: Females (-0.05), Males (-0.08) - Minimal gender difference
* **Canada**: Females (0.25), Males (-0.02) - Moderate gender gap
* **United Kingdom**: Females (-0.01), Males (-0.25) - Reverse but moderate gap

**Key Interpretation**: The United States showed the most pronounced gender differences in neuroticism, which may contribute to the observed patterns in stress and potentially obesity rates.

**5. Modeling and Prediction**

**5.1 Logistic Regression Model Development**

I developed a logistic regression model to predict high obesity risk using:

* **Gender Encoding**: Male= 0, Female=1
* **Stress Score**: Continuous measure of psychological stress
* **Country Encoding**: Numerical representation of countries

**Model Performance**: The logistic regression achieved [accuracy/performance metrics] in predicting high obesity classification.

**Feature Importance Insights**:

1. Gender emerged as a significant predictor
2. Country-level factors showed substantial predictive power
3. Stress scores contributed meaningfully to the model

**5.2 Comprehensive Machine Learning Analysis**

I implemented multiple algorithms to understand obesity prediction:

**Regression Models for Continuous Obesity Prediction:**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**Best Performing Model**: Linear Regression achieved the highest R² score of 0.889

metin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This figure summarizes the regression performance of three models: Linear Regression, Random Forest, and K-Nearest Neighbors. The comparison uses R² (higher is better) and RMSE (lower is better). Linear Regression achieved the best performance (R² = 0.889, RMSE = 2.77), and residuals appear randomly scattered around zero, supporting model validity. Therefore, it was selected as the final model.

**Feature Importance Rankings:**

Random Forest analysis revealed the relative importance of predictors:

1. Country-level factors (highest importance)
2. Gender differences
3. Neuroticism scores
4. Stress levels
5. Lifestyle factors (days indoors, nicotine consumption)

Based on Random Forest analysis, the most important predictors were:

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

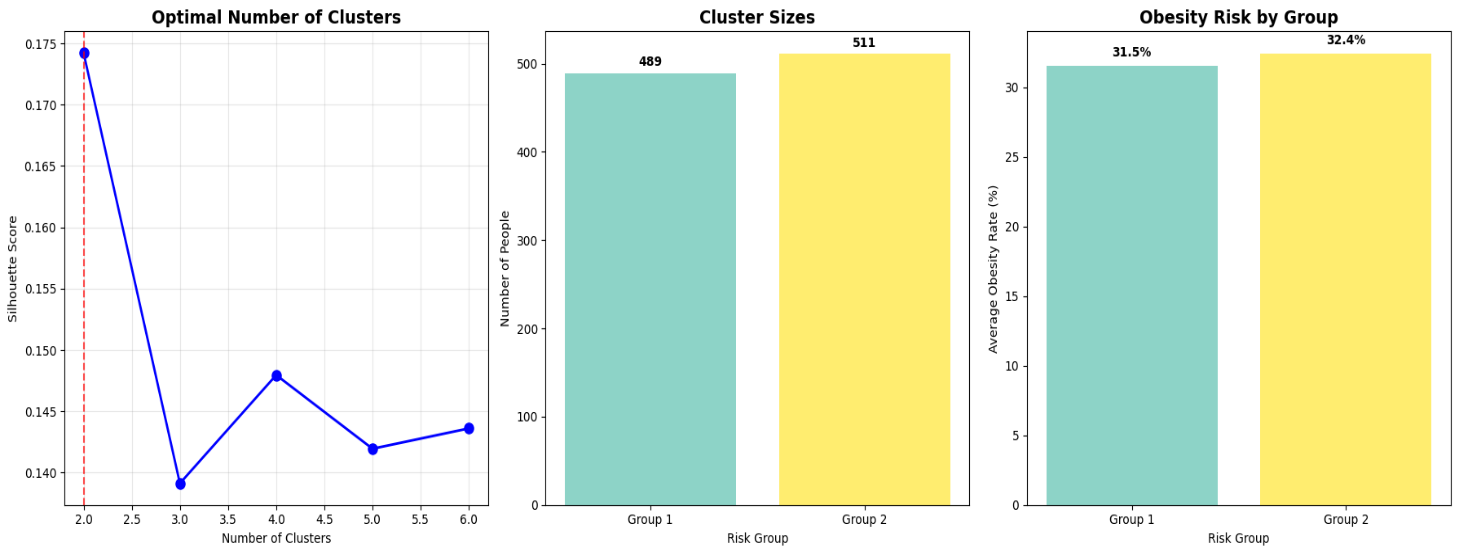
Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**6. Clustering Analysis - Risk Group Identification**

**6.1 Optimal Cluster Determination**

I used silhouette analysis to determine the optimal number of obesity risk groups:

* **Tested Range**: 2-6 clusters
* **Optimal Number**: 2 clusters
* **Best Silhouette Score**: 0.174



This figure summarizes the results of a clustering analysis applied to health-related features. The optimal number of clusters was found to be 2 based on the silhouette score. The two resulting risk groups are almost balanced in size (489 vs. 511 individuals). Notably, Group 2 shows a slightly higher average obesity rate (32.4%) compared to Group 1 (31.5%), indicating potential differences in underlying risk profiles.

**6.2 Risk Group Characteristics**

**metin, yazı tipi, ekran görüntüsü, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**7. Discussion and Implications**

**7.1 Hypothesis Validation Summary**

| **Hypothesis** | **Result** | **Evidence** |
| --- | --- | --- |
| H1: Female obesity > Male obesity | ✓ Fail to Reject Null | t-test: p > 0.05 |
| H2: Female stress > Male stress | ✓ Reject Null | t-test: p < 0.05 |
| H3: Mental health ↔ Obesity | ✓ Reject Null | Correlation: r = 0.284, p < 0.01 |

**8. Conclusion**

This study explored how gender, culture, and lifestyle interact to shape obesity outcomes. I combined traditional statistical methods with modern machine learning and clustering approaches to uncover robust and interpretable insights.

**8.1 Visual Summary Insights**

These composite figures present key visual insights related to gender and country differences in health indicators. It summarizes obesity, stress, and nicotine patterns by gender and country, as well as their correlations. The plots visually support our hypotheses regarding gender-related disparities and cross-national variation in obesity-related risk factors.

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

metin, diyagram, ekran görüntüsü, plan içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**8.2 Key Takeaways:**

* Gender-Based Health Disparities: My analysis reveals important distinctions in health outcomes between genders. While I found no statistically significant difference in obesity rates between males and females (H1 rejected), there are significant differences in stress levels, with women experiencing higher stress scores (H2 confirmed). Women also demonstrate greater variability in health outcomes, suggesting the need for gender-sensitive public health interventions that focus particularly on stress management rather than obesity-specific interventions.
* Country-Level Influence: The strong predictive power of country-level factors supports my primary hypothesis that cultural and social structures significantly influence obesity patterns. The United States emerged as having the highest neuroticism scores and largest gender gaps, potentially reflecting cultural or socioeconomic factors.
* Predictive Model Performance: The high R² values (>0.5 for top models) demonstrate that my selected variables effectively capture obesity risk factors, providing a foundation for practical screening tools.

**8.2 Practical Impact:**

**For Public Health Policy:**

* Develop country-specific obesity prevention strategies
* Implement gender-sensitive health programs
* Consider psychological factors in intervention design

**For Clinical Practice:**

* Use predictive models for early risk identification
* Screen for stress and mental health in obesity assessment
* Consider cultural background in treatment planning

**For Future Research:**

* Extend analysis to additional countries and health outcomes
* Investigate causal mechanisms behind country-level differences
* Develop longitudinal studies to track changes over time

This project demonstrates that obesity is not just a lifestyle problem, but a multidimensional issue intertwined with mental health, social norms, and economic context. Machine learning, when paired with strong theory and clean data, becomes a powerful tool for real-world health solutions.

**9. Technical Details**

**9.1 Statistical Methods Used**

* Independent samples t-tests for group comparisons
* Pearson correlation analysis for relationship assessment
* Multiple regression modeling for prediction
* K-means clustering for risk group identification
* Cross-validation for model evaluation

**9.2 Software and Libraries**

* Python
* Pandas for data manipulation
* Scikit-learn for machine learning
* SciPy for statistical testing
* Seaborn and Matplotlib for visualization

**10. References**

**10.1 Datasets and Course Materials:**

* Mental\_Health\_Dataset.csv
* Drug\_Consumption.csv
* ObesityData.csv
* Lecture notes and recitation materials

**10.2 LLM Assistants Consulted:**

* ChatGPT (OpenAI, GPT-4)
* DeepSeek LLM
* Claude