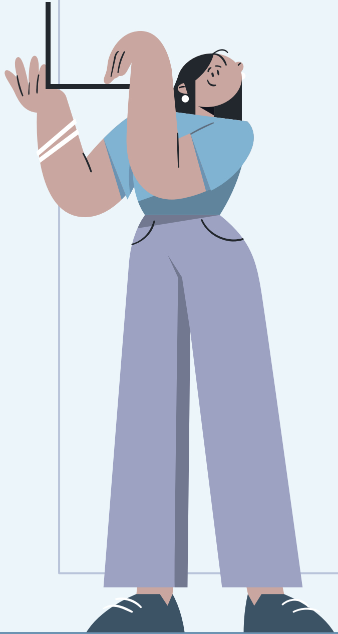
STA130H1

STA130 Course Project

Statistical Analysis of the Canadian Social Connection Survey



Ching-Chun (Cherrie) Wei
cc.wei@mail.utoronto.ca

Shimeng (Simone) Wang
simone.wang@mail.utoronto.ca

Songxuan (Jimmy) Wu
songxuan.wu@mail.utoronto.ca

TUT0111 TA: Christoffer Tan

Introduction

- **Overall Goal:** To help raise interest and awareness in the importance of social connection and community engagement for personal health and well-being.
- **CSCS Dataset** captures information on social connection, health practices, and demographic variables.
- **Target Audience**: Potential future collaborators, other teams associated genwell and CASCH, people interested in social health and community well-being.

Research topics:

1. The impact of **friendship** engagement on health practices.
2. The relationship between **family** interactions and loneliness.
3. The impact of **close social connections** on life satisfaction.

Data Wrangling

Data cleaning

- Missing data is removed for analysis 1 and 2, but not for analysis 3 (due to feasibility of performing the analysis)

Data transformation

- For data that is analyzed using simple linear regression and hypothesis testing, categorical variables are converted to numerical variables first

Data integration

- In analysis 2 and 3, some variable categories are combined to simplify the model for improved statistical power or better comparison

Analysis 1: Making Friends & Health Practices

Research Question:

- Is a high frequency of making new friends associated with better health practices?

Hypothesis:

- Higher social engagement positively influences personal health and well-being

Why it matters:

- The connection between social interactions and personal health is important

- It highlights the benefits of community engagement on health.

Key Variables

Independent Variable: CONNECTION_activities_new_friend_p3m

Frequency of making new friends

- "Not in the past three months": 0
- "Less than monthly": 1
- "Monthly": 2
- "A few times a month": 3
- "Weekly": 4
- "A few times a week": 5
- "Daily or almost daily": 6

Dependent Variable: HEALTH_hampson_good_health_practices_scale_score

Hampson Good Health Practices Score (continuous scale, 1–5). Measures behaviors like regular exercise, balanced diet, and adequate sleep.

Methodology: Linear Regression

Model: $y=\beta 0+\beta 1x+\epsilon$

- $\beta 0$: Intercept
- $\beta 1$: Effect of frequency of making new friends.

Steps:

- Data Subsetting:
  - Created a clean subset containing only the mapped independent variable and the dependent variable.
- Regression Analysis:
  - Predictor (X): Frequency of making new friends (numeric scale).
  - Response (y): Health practices score (continuous scale).
  - Added a constant to the predictor variable to account for the intercept.
  - Performed Ordinary Least Squares (OLS) regression to estimate the relationship between the predictor and response variables.

Results

- **Statistical Results**:

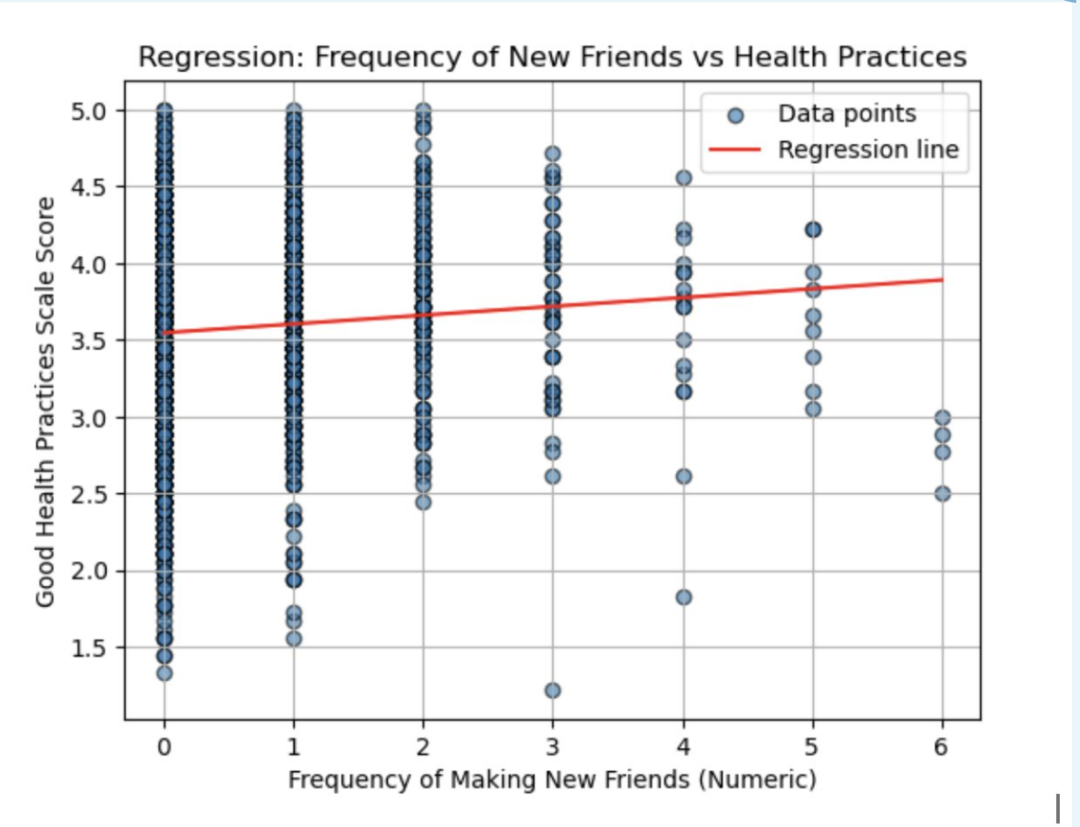
β1: 0.0573 (positive effect).

p-value: <0.001 (significant).

R-squared: 0.007 (weak explanatory power).

- **Interpretation**:

Making new friends is statistically associated with better health practices.

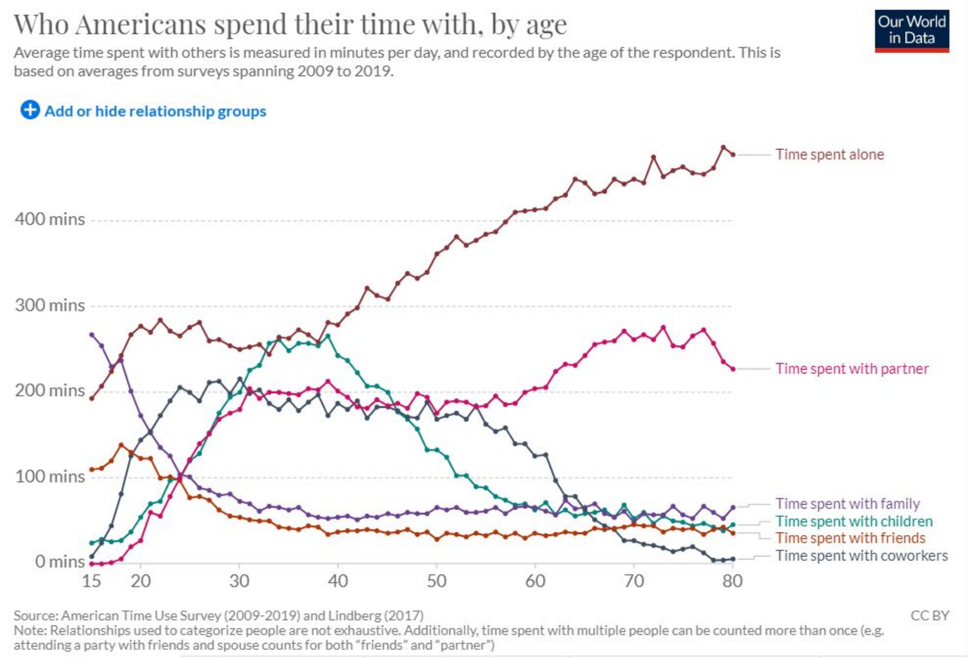


Analysis 2: Loneliness & Time Spent with Family

Research Question: Is there a statistically significant difference in *loneliness scores* between individuals who spend more *time with family* (4-7 days per week) and those who spend less time (0-3 days per week)?

Relevance:

- Time spent with family decline as people get older
- Help raise awareness about the importance of connection with family



Who Americans spend their time with, by age

Average time spent with others is measured in minutes per day, and recorded by the age of the respondent. This is based on averages from surveys spanning 2009 to 2019.

Source: American Time Use Survey (2009-2019) and Lindberg (2017)

CC BY

Note: Relationships used to categorize people are not exhaustive. Additionally, time spent with multiple people can be counted more than once (e.g. attending a party with friends and spouse counts for both "friends" and "partner")

Hypotheses

Independent variable: Number of days spent with family per week (Categorical)

- Less Group: 0 - 3 days (None & Some days)
- More Group: 4-7 days (Most days & Every day)

Dependent variable: DeJong Gierveld Loneliness Scale scores (Numerical)

- 0 - least lonely, 6 - most lonely

Null Hypothesis (H_0): There is no difference in the distribution of loneliness scores between the two types of social interaction groups.

Alternative Hypothesis (H_a): There is a statistically significant difference in the distribution of loneliness scores between the two types of social interaction groups.

Methodology: Hypothesis Testing & Box Plot

Statistical analysis: Mann-Whitney U test

- A non-parametric test used to compare the two groups without assuming normality
- Tests whether the difference between More Group and Less Group is statistically significant

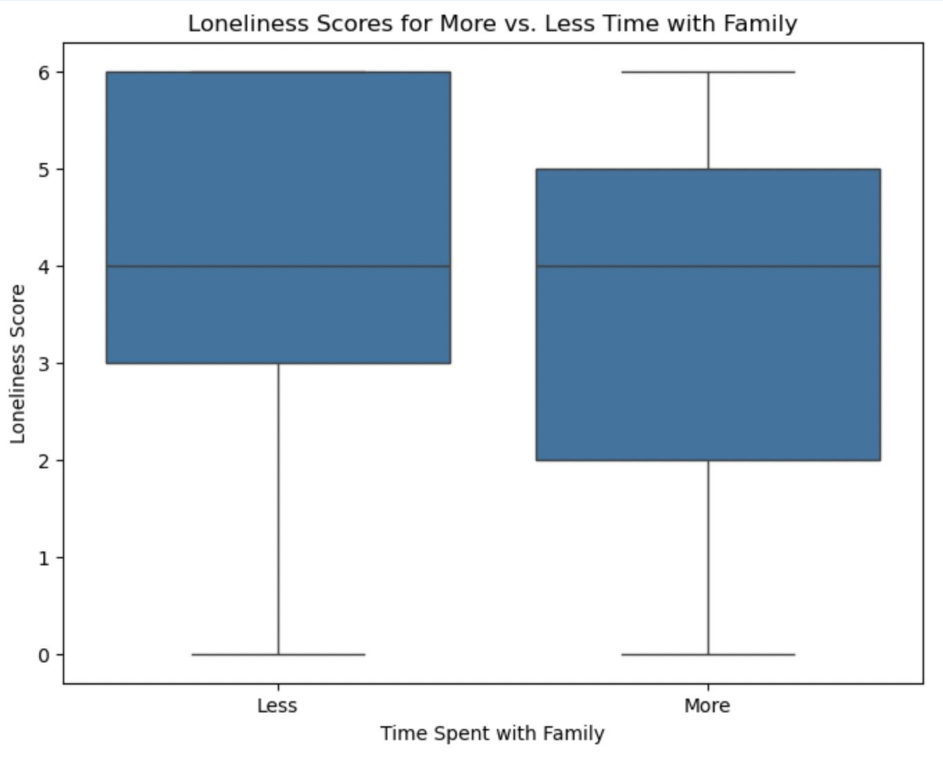
Normality Check (not normal if $p < 0.05$):

- More Group: W = 0.92314, p = 2.7536e-42
- Less Group: W = 0.890434, p = 0.0

Assumptions:

- Independence of groups.
- Ordinal or continuous dependent variable.
- Similar distribution shapes between groups (but doesn't require normality).
- No excessive ties in the rankings.
- Adequate sample size for statistical power.

Box Plot Visualization



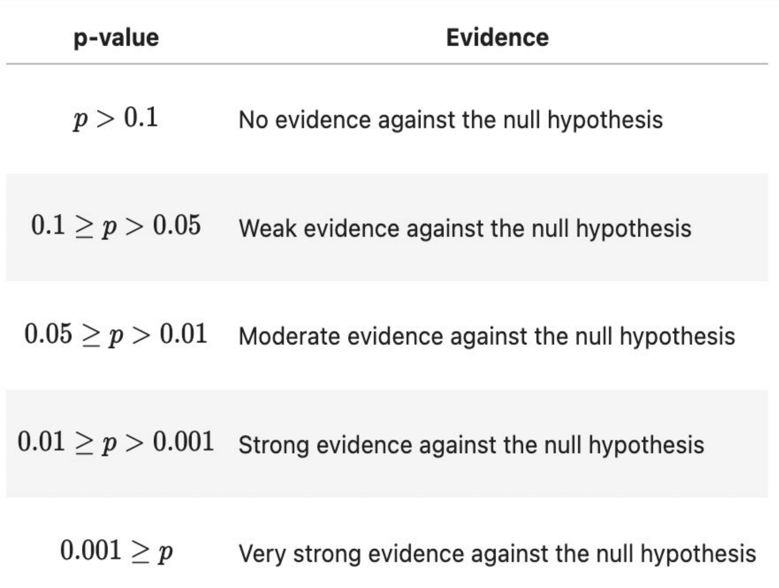
Key interpretations based on the visualization:

- Median: similar central tendencies
- IQR: The loneliness scores for individuals who spend more time with family show slightly less variability compared to those who spend less time.

Results

U Statistic: 7361751.5

P-value: 2.0596e-45



p-value	Evidence
$p > 0.1$	No evidence against the null hypothesis
$0.1 \geq p > 0.05$	Weak evidence against the null hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the null hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the null hypothesis
$0.001 \geq p$	Very strong evidence against the null hypothesis

- Both groups exhibit the full range of loneliness scores (0 through 6)
- P-value is extremely small → statistically significant difference
- **The amount of time spent with family has a significant effect on loneliness scores**
- More time spent with family contribute to higher levels of emotional well-being.

Analysis 3: Social Connections & Life Satisfaction

Research question:

- What patterns in perceived social connection best explain variations in life satisfaction?

Relevance:

- the findings may motivate individuals and communities to prioritize building stronger social networks.

Variables & Hypothesis

Independent variables: LONELY_dejong_emotional_social_loneliness_scale_close

- Whether respondents think there are enough people they feel close to (yes, no, more or less)

Dependent variable: WELLNESS_life_satisfaction

- Life satisfaction scores on a scale of 1 - 10

Null Hypothesis (H_0):

- The life satisfaction scores between people who have close social connections and people who don't do not differ significantly.

Alternative Hypothesis (H_a):

- People who feel they have close social connections will report higher life satisfaction compared to those who don't.

Methodology: Classification Decision Tree

- Use a **Decision Tree Regressor** because the DV is continuous.
- **Parameters**:
  - **Criterion**: squared_error (to minimize mean squared error for regression).
  - **Max Depth**: Limit depth to avoid overfitting.
- Fit the decision tree on the training dataset using the IV and the DV
- **Assumptions**:
  - Independence
  - Homogeneity within nodes
  - Sufficient data for training
  - Linearity of predictors

Results

Lack of Close Connections:

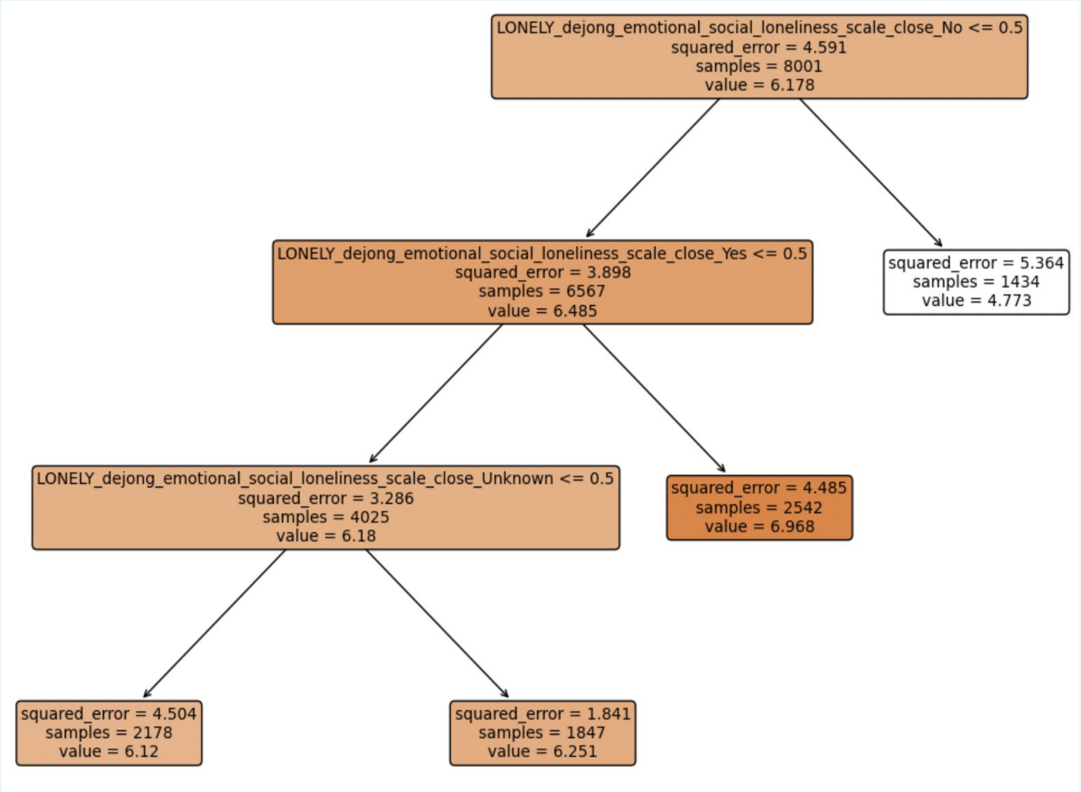
- significantly lower life satisfaction (4.773)

Strong Close Connections:

- highest life satisfaction (6.968)

Interpretation:

- **Perceived social connection is a strong predictor of life satisfaction.**



Limitations

Analysis 1:

- Low R-squared: social engagement explains only a small portion of health practices.
- Continuous independent variable might offer more detailed insights than categorical.
- May be other important predictors (e.g., income, mental health).

Analysis 2:

- Simplification of grouping
- Uncontrolled confounding variables
- Skewed distribution of data
- No consideration about the quality of interactions

Analysis 3:

- Ambiguous or missing data may reflect unique circumstances or noise in the data that require further exploration.

Overall Conclusion

Analysis 1: Higher frequency of making new friends is associated with improved health practices.
Analysis 2: More time spent with family contribute to decrease in loneliness.
Analysis 3: Having close social connections contribute to higher levels of life satisfaction.

Final conclusion:

Social connection and community engagement are crucial for enhancing personal health, mental well-being, and overall life satisfaction.

References & Acknowledgement

We would like to express our sincere gratitude to **Dr. Kiffer Card** and **Christine Ovcaric** for creating this project and providing the dataset, and to our amazing Professor **Scott Schwartz** for his guidance throughout the course. A special thank you to our wonderful TA, **Christoffer Tan**, for your support and insightful feedback.

Lastly, we are grateful to each other as group members for a rewarding experience and great collaboration.