THE INVISIBLE MAJORITY

UNVEILING THE ECOLOGY AND EMOTION OF INTERNET PERSONAS

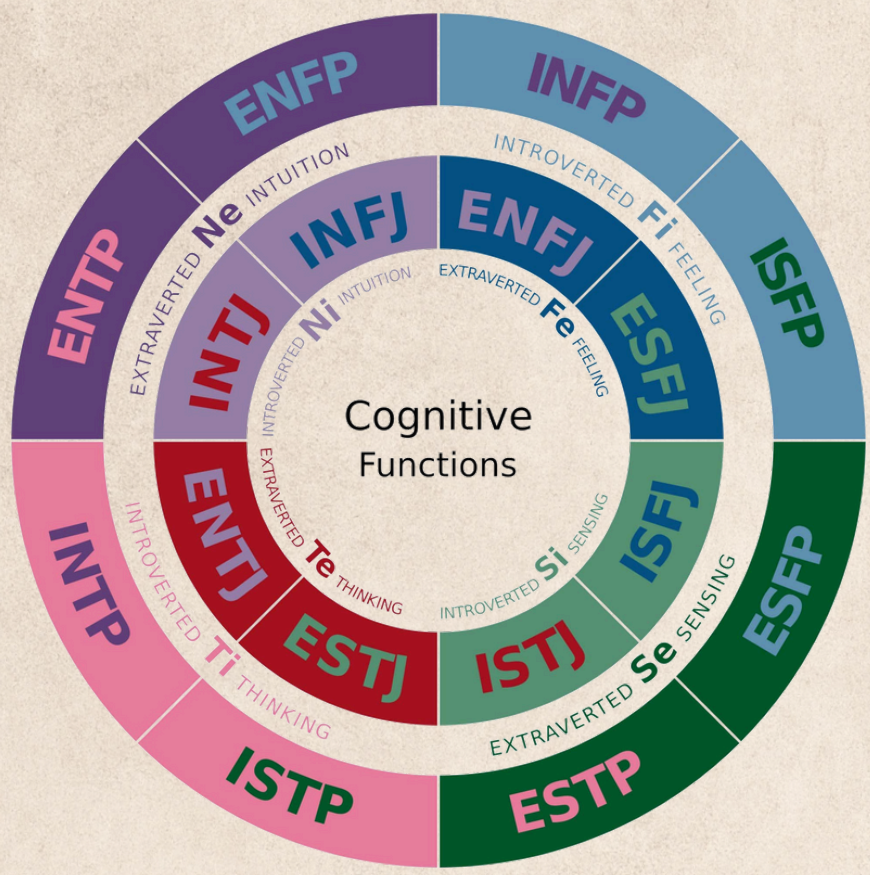
✳ DATA SCIENCE AND COMPUTER PROGRAMMING ✳

MBTI

PREFACE

Hi, I am Rou Zhen from the Department of Technology Application and Human Resource Development at National Taiwan Normal University. My interests lie in web architecture and human behavior analysis.

This project demonstrates how to use the data processing techniques I learned in this course to analyze unstructured social media text. While MBTI is a complex psychological theory, I will skip the theoretical details to focus on the primary goal of this project: Exploratory Data Analysis (EDA) and Data Storytelling. I aim to use raw data to verify—or debunk—common stereotypes about personality types online.



DATASET & PRE-PROCESSING

Dataset: MBTI Type Dataset (Kaggle) The dataset consists of 8,675 users, containing their personality labels and their last 50 social media posts.

The Problem with Raw Data: The raw data is unstructured. As shown below, the posts column is a single long string where comments are separated by a specific delimiter (|||). This format is not compatible with direct statistical analysis.

Data Pre-Processing Steps:

1. Splitting: I noticed the ||| delimiter, so I defined a function to split the text to isolate individual comments.
2. Feature Engineering: My computer cannot "read" emotions directly. Therefore, I converted the text into numerical features:
   - Activity Level: Calculated by total Word Count.
   - Emotional Intensity: Calculated by the frequency of exclamation marks (!).
   - Inquisitiveness: Calculated by the frequency of question marks (?).

type	posts
INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||http:
ENTP	'I'm finding the lack of me in these posts very alarming.|||
INTP	'Good one _____ https://www.youtube.com/watch?v=f
INTJ	'Dear INTP, I enjoyed our conversation the other day. E
ENTJ	'You're fired.|||That's another silly misconception. That ap
INTJ	'18/37 @.@|||Science is not perfect. No scientist claims t
INFJ	'No, I can't draw on my own nails (haha). Those were do
INTJ	'I tend to build up a collection of things on my desktop tha
INFJ	I'm not sure, that's a good question. The distinction betwe
INTP	'https://www.youtube.com/watch?v=w8-egj0y8Qs|||I'm in
INFJ	'One time my parents were fighting over my dad's affair a
ENFJ	'https://www.youtube.com/watch?v=PLAaiKvHvZs|||51 :
INFJ	'Joe santagato - ENTP|||ENFJ or ENTP? I'm not too sur
INTJ	'Fair enough, if that's how you want to look at it. Like I st
INTP	'Basically this... https://youtu.be/1pH5c1JkhLU|||Can I ha
INTP	'Your comment screams INTJ, bro. Especially the useless
INFJ	'some of these both excite and calm me: BUTTS bodies
INFP	'I think we do agree. I personally don't consider myself A
INFJ	'I fully believe in the power of being a protector, to give a
INFP	'That's normal, it happens also to me. If I am in high moo
INTP	'Steve Job's was recognized for his striving for efficiency
INFJ	'It is very annoying to be misinterpreted. Especially with

CSV DATA

```
【DATASET OVERVIEW】
Total Users (Rows): 8675
Total Features (Cols): 6
Columns: ['type', 'posts', 'word_count', 'link_count', 'exclaim_count', 'ques

【1. POPULATION DISTRIBUTION (By Personality Type)】
type
INFP    1832
INFJ    1470
INTP    1304
INTJ    1091
ENTP     685
ENFP     675
ISTP     337
ISFP     271
ENTJ     231
ISTJ     205
ENFJ     190
ISFJ     166
ESTP      89
ESFP      48
ESFJ      42
ESTJ      39

Most Common: INFP (1832 users)
Least Common: ESTJ (39 users)

【2. ACTIVITY & BEHAVIOR STATISTICS】
       Word Count (Activity)  Link Count  Exclamation (!)  Question (?)
count                 8675.0      8675.0           8675.0        8675.0
mean                  1262.7         3.3              8.5          10.7
std                    317.3         5.8             11.6           7.0
min                      4.0         0.0              0.0           0.0
25%                   1081.0         0.0              2.0           6.0
50%                   1314.0         1.0              5.0          10.0
75%                   1497.0         4.0             11.0          14.0
max                   2212.0        91.0            219.0         121.0

Complete
```

DATASET OVERVIEW & STATISTICS

OBJECTIVES & ANALYSIS

1. THE ECOLOGICAL DISTRIBUTION

I want to identify which personality types are the most active and prevalent in the online community.

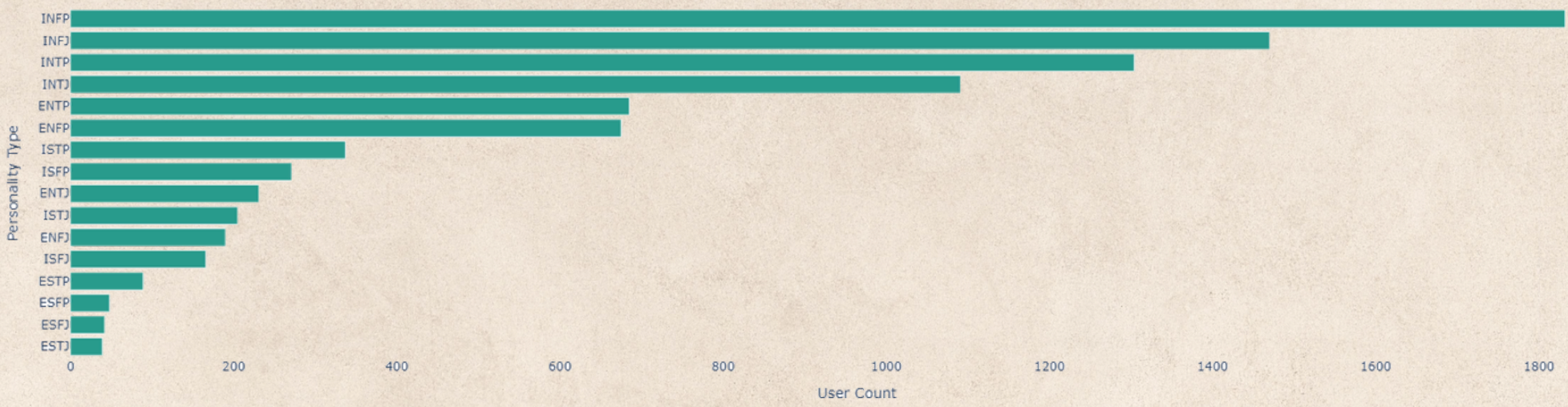
- Pseudocode:
  - Load the dataset and extract the type column.
  - Count the frequency of each unique MBTI type.
  - Sort the data from highest to lowest count.
  - Plot a horizontal bar chart to visualize the imbalance.

```
# PLOT 1: POPULATION DISTRIBUTION
 POP_DATA = DF['TYPE'].VALUE_COUNTS().SORT_VALUES(ASCENDING=TRUE)
 FIG1 = PX.BAR(X=POP_DATA.VALUES, Y=POP_DATA.INDEX, ORIENTATION='H',
        TITLE="FIG 1: ECOLOGICAL DISTRIBUTION (INTROVERTS VS EXTROVERTS)",
        LABELS={'X': 'USER COUNT', 'Y': 'PERSONALITY TYPE'},
        TEMPLATE="PLOTLY_WHITE")
 FIG1.UPDATE_TRACES(MARKER_COLOR='#2A9D8F')
 FIG1.SHOW()
```

- Algorithm:
  - Using the aggregation function value_counts(), we can obtain the population distribution and identify the "Long Tail" phenomenon.

Analysis Results:(Insert your "MBTI Sample Distribution" chart here) The data reveals a significant Sampling Bias. The number of users classified as INFP and INFJ is significantly higher than any other type. In contrast, ESTJ (a common real-world type) is the minority online. This suggests that the internet acts as a sanctuary for Introverts (INxx), who may be quieter in physical settings but are the loudest majority online.

Fig 1: Ecological Distribution (Introverts vs Extroverts)



OBJECTIVES & ANALYSIS

2. THE "TALKATIVE EXTROVERT" STEREOTYPE

I want to test the stereotype that "Extraverts (E) talk more than Introverts (I)."

- Pseudocode:
  - Create a binary label E/I based on the personality type.
  - Calculate the total word_count for each user.
  - Group the data by E/I.
  - Compute the Median and Interquartile Range (IQR) for both groups.
  - Visualize the comparison using a Box Plot.

```
  # PLOT 2: ACTIVITY LEVEL
  FIG2 = PX.BOX(DF, X='E/I', Y='WORD_COUNT',
         TITLE="FIG 2: ACTIVITY LEVEL (WORD COUNT DISTRIBUTION)",
         LABELS={'WORD_COUNT': 'TOTAL WORDS', 'E/I': 'PERSONALITY DIMENSION'},
         COLOR='E/I',
         TEMPLATE="PLOTLY_WHITE",
         COLOR_DISCRETE_MAP={'INTROVERSION (I)': '#264653', 'EXTRAVERSION (E)':
'#E9C46A'})
  FIG2.SHOW()
```

- Algorithm:
  - By mapping textual attributes to numerical length, we use a Box Plot to compare the statistical distribution of activity levels, eliminating the influence of outliers.

Analysis Results:(Insert your "Activity Level Box Plot" chart here)This is the most counter-intuitive finding. The graph shows that the median word counts for Extraverts and Introverts are nearly identical (approx. 1,300 words). The distribution shape is also highly similar. This debunks the myth: Introverts are just as active and "talkative" as Extraverts in digital text communication.

Fig 2: Activity Level (Word Count Distribution)

OBJECTIVES & ANALYSIS

3. DECODING LOGIC VS. EMOTION

I want to see if punctuation usage can predict whether a person is a "Thinker (T)" or a "Feeler (F)."

- Pseudocode:
  - Define a function to count occurrences of ! and ?.
  - Group the dataset by Thinking (T) vs Feeling (F).
  - Calculate the average usage of these symbols per person.
  - Compare the means using a Grouped Bar Chart.

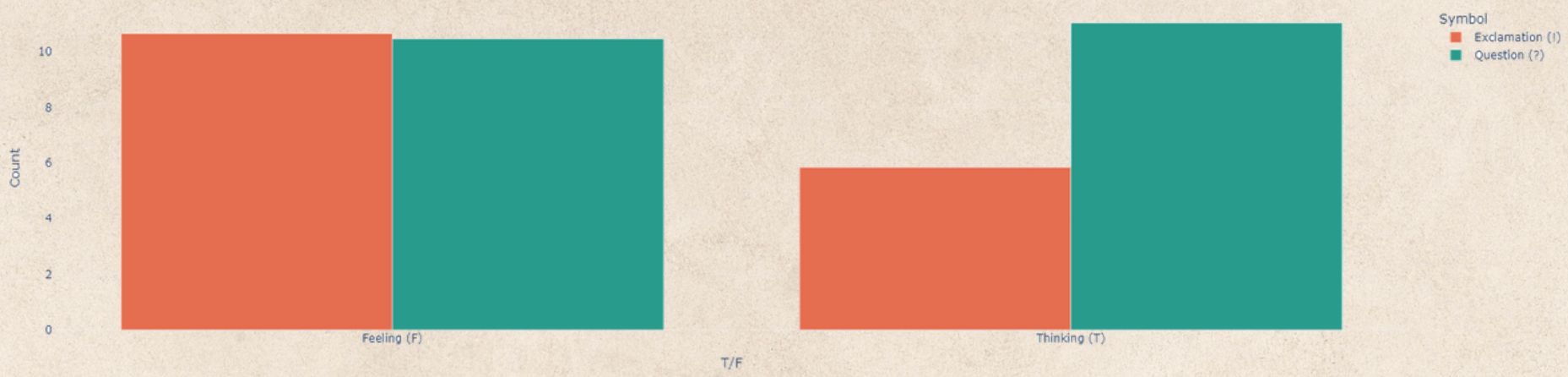
```
 # PLOT 3: EMOTION ANALYSIS
TF_STATS = DF.GROUPBY('T/F')[['EXCLAIM_COUNT', 'QUESTION_COUNT']].MEAN().RESET_INDEX()
TF_MELTED = TF_STATS.MELT(ID_VARS='T/F', VAR_NAME='SYMBOL', VALUE_NAME='COUNT')
TF_MELTED['SYMBOL'] = TF_MELTED['SYMBOL'].REPLACE({'EXCLAIM_COUNT': 'EXCLAMATION (!)',
'QUESTION_COUNT': 'QUESTION (?)'})

FIG3 = PX.BAR(TF_MELTED, X='T/F', Y='COUNT', COLOR='SYMBOL', BARMODE='GROUP',
       TITLE="FIG 3: LOGIC VS EMOTION (PUNCTUATION USAGE)",
       TEMPLATE="PLOTLY_WHITE",
       COLOR_DISCRETE_MAP={'EXCLAMATION (!)': '#E76F51', 'QUESTION (?)': '#2A9D8F'})
FIG3.SHOW()
```

- Algorithm:
  - We use specific characters as proxies for psychological traits. The aggregation of means allows us to observe group-level behavioral differences.

Analysis Results:(Insert your "Logic vs Emotion" chart here) The analysis confirms the stereotype. Feeling (F) types use exclamation marks (!) significantly more often than Thinking (T) types, indicating a higher tendency for emotional expression. However, both groups use question marks (?) at a relatively high rate, suggesting that curiosity is a shared trait across all personalities.

Fig 3: Logic vs Emotion (Punctuation Usage)



CONCLUSION

Through this project, I successfully transformed unstructured text into meaningful insights. The data tells a story that contradicts our intuition:

1. Population: The internet is dominated by Introverts.
2. Activity: Introverts are not "quiet" online; they are as active as Extraverts.
3. Behavior: Punctuation is a reliable indicator of emotional traits.

This project demonstrates that Data Storytelling is not just about making charts, but about using rigorous logic to uncover hidden truths in human behavior.

APPENDIX

- Data Source: Kaggle MBTI Type Dataset:
  - https://www.kaggle.com/datasets/datasnaek/mbti-type?resource=download
- Tools Used:
  - Python
  - Pandas
  - Plotly
- Github Repo:
  - https://github.com/41371125h-chinrouzhen/DSCP-Final-Project