1 Introduction

With the recent renewal of the MBTI's popularity, posts or analyses about it have become widespread on public platforms. It seems that MBTI has become a way people use at first meetings.

The Myers-Briggs Type Indicator (MBTI) is a psychological framework primarily used to describe and measure personality traits by categorizing individuals into four dichotomous dimensions. In essence, it aims to reflect how people perceive the world and make decisions based on their internal experiences(Yang, 2022).

And the test is conducted through asking a series of questions, and the type of question is kind of like asking about the tendency to do something. Based on the responses, individuals receive a result represented by four letters, each corresponding to one of the four bipolar dimensions. According to Boyle (1995), these four dichotomous dimensions classify individuals as either extraverted (E) or introverted (I), sensing (S) or intuitive (N), thinking (T) or feeling (F), and judging (J) or perceiving (P). Combinations of the four preferences generate one of 16 personality types (e.g., ESFJ, ENFP, INTP, ISFJ), each associated with distinct behavioral tendencies, reflecting differences in attitudes, orientation, and decision-making styles. Also, on the official website, it will provide the future career options for you, like INFP, which may be more suitable for being an author. The percentage of each standard will also show in the final result.

The MBTI provides a widely recognized and accessible way to understand personality, making it a useful foundation for further behavioral and data-driven analysis.

2 Objective

1. Evaluate the reliability of the MBTI from a statistical perspective.
2. Explore the potential application of MBTI in social media behavior analysis.
3. Help people better understand personality traits and behavioral patterns.
4. Helping people eliminate stereotypes caused by MBTI personality types.

3 Problem Statements

1. Are the results of the MBTI personality test statistically robust and reliable?
2. Do the four dimensions of MBTI work independently, or are they connected in some way?
3. Do people with different MBTI types have different levels of activity on the internet?
4. Do significant differences exist in the interest preferences and behavioral patterns of different MBTI personality types on social networking sites?

4 Data Collection

1. "MBTI Personality Type Twitter Dataset"

   Tweets were originally harvested from the public Twitter API by a third-party collector and later released on Kaggle by Mazlumi(https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset)

   - 8,600 Twitter users, and 1 million raw tweets.
   - Each user record contains the self-declared MBTI type (e.g., "ENFP") taken from the user's bio.
   - The text has not been further cleaned or filtered—URLs, emojis, hashtags, and retweets remain. Researchers must therefore perform their preprocessing (tokenisation, stop-word removal, emoji handling, etc.) before analysis.

2. "KPMIRU Questionnaires Data"

   Questionnaire responses compiled and shared on Kaggle by Pmenshih(https://www.kaggle.com/datasets/pmenshih/kpmiru-questionnaires-data)

   - Contains every participant's item-level answers to the full KPMIRU personality inventory (several dozen Likert-scale questions).
   - Provides the scored results for all four MBTI dimensions—reported as continuous scores (0–100 per axis) as well as the final type label (e.g., "ISTJ").

- Demographic fields (age range, gender, education) are included, enabling richer statistical controls.

Together, the Twitter dataset supplies large-scale, real-world language samples with self-reported types, while the KPMIRU dataset offers clean, psychometrically scored questionnaire data. The two sources complement each other for training and validating our emotion-aware MBTI models.

```
#reading the data sets
import pandas as pd
data1=pd.read_csv(r'E:\python\kpmi_ru_data.csv.zip')
print(data1)
```

```
       q1  t1  q2  t2  q3  t3  q4  t4  q5  t5  ...  t63   e   i   s   n   t  \
0       1   8   2   8   2   8   1   8   1   8  ...    8  32  18  16  19  26
1       1   8   1   8   1   8   2   8   1   8  ...    8  28  15  27  18  23
2       1   8   1   8   1   8   1   8   1   8  ...    8  25  15  25  16  19
3       2   8   1   8   1   8   1   8   1   8  ...    8  30  15  25  14  33
4       2   8   2   8   1   8   1   8   1   8  ...    8  22  22  23  18  33
...    ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ...  ...  ..  ..  ..  ..  ..
99995   1   8   2   8   1   8   2   8   2   8  ...    8  32  18  27  16  33
99996   1   8   2   8   2   8   1   8   2   8  ...    8  20  26  37  12  37
99997   1   8   2   8   1   8   1   8   1   8  ...    8  22  18  23  19  26
99998   2   8   2   8   1   8   1   8   1   8  ...    8  22  18  21  23  33
99999   1   8   1   8   1   8   1   8   2   8  ...    8  38   9  39  12  37

       f   j   p psychotype
0     24  19  44       ENTP
1     30  35  29       ESFJ
2     35  38  23       ESFJ
3     27  38  23       ESTJ
4     13  29  35       ISTP
...   ..  ..  ..        ...
99995 19  13  52       ESTP
99996 13  29  29       ISTP
99997 22  29  38       ESTP
99998 19  32  29       ENTJ
99999 22  45  23       ESTJ

[100000 rows x 133 columns]
```

Figure 1: reading data sets

```
import pandas as pd
data2=pd.read_csv(r"E:\python\twitter_MBTI.csv.zip")
print(data2)
```

```
      Unnamed: 0                                               text label
0              0  @Pericles216 @HierBeforeTheAC @Sachinettiyil T...  intj
1              1  @Hispanthicckk Being you makes you look cute||...  intj
2              2  @Alshymi Les balles sont réelles et sont tirée...  intj
3              3  I'm like entp but idiotic|||Hey boy, do you wa...  intj
4              4  @kaeshurr1 Give it to @ZargarShanif ... He has...  intj
...          ...                                                ...   ...
7806        7806  @sobsjjun God,,pls take care 🙄|||@sobsjjun Hir...  intp
7807        7807  @Ignis_02 wow last time i got intp https://t.c...  intp
7808        7808  @akupilled A 100%|||@akupilled That SOMEONE wi...  entp
7809        7809  If you're #INTJ this one is for you | What is ...  infj
7810        7810  @harry__lambert @gucci hey can you dm me a pic...  istp
```

As illustrated in the two Figures above, the combined Kaggle sources provide information on:

1. Self-reported MBTI types for each respondent
2. Raw Twitter posts and basic tweet metadata linked to those MBTI labels
3. Demographic and psychometric questionnaire answers (KPMIRU survey)
4. Behavioral metrics such as posting frequency and topic keywords extracted from the tweets

5 Data Cleaning and Pre-processing

5.1 Handling Missing Values

1. We use dropna() to eliminate any records with missing or null entries. Fortunately, the dataset had no missing psychotype labels or scoring data.

```
# Delete null values
df_clean = df.dropna()
```

5.2 Outlier Detection and Correction

1. Outliers in numeric scores were identified using the IQR (interquartile range) method.

2. For each numeric column, we computed Q1, Q3. IQR = Q3 - Q1 represent the middle 50% of the data distribution. The lower_bound and upper_bound represent the boundaries of the normal range under the Interquartile Range (IQR) rule. And then we replaced values outside [Q1 − 1.5×IQR, Q3 + 1.5×IQR] with the column median.

3. The median substitution method mainly serves to stabilize the overall distribution and avoid the influence of extreme values.

```
# Select all columns in df_clean with numeric data types (either int or float) in preparation for outlier processing.
numeric_cols = df_clean.select_dtypes(include='number').columns

# Replace the outliers (utilizing the Interquartile Range (IQR) rule) with the median value of the column.
df_processed = df_clean.copy()

for col in numeric_cols:
    Q1 = df_processed[col].quantile(0.25)
    Q3 = df_processed[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    median_value = df_processed[col].median()

    # Replace outliers
    df_processed[col] = df_processed[col].apply(lambda x: median_value if x < lower_bound or x > upper_bound else x)
```

5.3 Export the cleaned data

After completing the outlier replacement and data cleaning process, we used df.info() to verify the integrity and structure of the cleaned dataset. The cleaned DataFrame was then exported using the to_csv() method, which saved it as kpmi_ru_data(Cleaned).csv for downstream analysis. The index=False parameter ensured that row indices were not written to the CSV file.

```
print(df_processed.info())
print(df_processed)
df_processed.to_csv(r'E:\python\kpmi_ru_data(Cleaned).csv', index = False)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Columns: 133 entries, q1 to psychotype
dtypes: float64(75), int64(57), object(1)
memory usage: 101.5+ MB
None
       q1   t1  q2   t2  q3   t3  q4   t4  q5   t5  ...  t63     e     i  \
0       1  8.0   2  8.0   2  8.0   1  8.0   1  8.0  ...  8.0  32.0  18.0
1       1  8.0   1  8.0   1  8.0   2  8.0   1  8.0  ...  8.0  28.0  15.0
2       1  8.0   1  8.0   1  8.0   1  8.0   1  8.0  ...  8.0  25.0  15.0
3       2  8.0   1  8.0   1  8.0   1  8.0   1  8.0  ...  8.0  30.0  15.0
4       2  8.0   2  8.0   1  8.0   1  8.0   1  8.0  ...  8.0  22.0  22.0
...    ..  ...  ..  ...  ..  ...  ..  ...  ..  ...  ...  ...   ...   ...
99995   1  8.0   2  8.0   1  8.0   2  8.0   2  8.0  ...  8.0  32.0  18.0
99996   1  8.0   2  8.0   2  8.0   1  8.0   2  8.0  ...  8.0  20.0  26.0
99997   1  8.0   2  8.0   1  8.0   1  8.0   1  8.0  ...  8.0  22.0  18.0
99998   2  8.0   2  8.0   1  8.0   1  8.0   1  8.0  ...  8.0  22.0  18.0
99999   1  8.0   1  8.0   1  8.0   1  8.0   2  8.0  ...  8.0  38.0   9.0

          s     n     t     f     j     p psychotype
0      16.0  19.0  26.0  24.0  19.0  44.0       ENTP
1      27.0  18.0  23.0  30.0  35.0  29.0       ESFJ
2      25.0  16.0  19.0  35.0  38.0  23.0       ESFJ
3      25.0  14.0  33.0  27.0  38.0  23.0       ESTJ
4      23.0  18.0  33.0  13.0  29.0  35.0       ISTP
...     ...   ...   ...   ...   ...   ...        ...
99995  27.0  16.0  33.0  19.0  13.0  52.0       ESTP
99996  37.0  12.0  37.0  13.0  29.0  29.0       ISTP
99997  23.0  19.0  26.0  22.0  29.0  38.0       ESTP
99998  21.0  23.0  33.0  19.0  32.0  29.0       ENTJ
99999  39.0  12.0  37.0  22.0  45.0  23.0       ESTJ

[100000 rows x 133 columns]
```

The data in the figure 3.1 is the data after our data cleaning.

6 Model Building and Evaluation

Chi-square test is an on parametric statistical test to determine if the two or more classifications of the samples are independent or notZibran, 2007. We all know that MBTI has four dimensions(Extraversion–Introversion (E–I), Sensing–Intuition (S–N), Thinking–Feeling (T–F) and Judging–Perceiving (J–P)). But whether or not these four dimensions interrelated or independent of each other stays unknown. For explanation, let's consider the data presented in Figure 3.1 which comprising 100 000 respondents, providing information on the scoring fields of the four dimensions of MBTI. In order to find the answer, we use the chi-square test.

Each respondent's type label (e.g., ENTP) was decomposed into its four constituent letters, and the sample was cross-classified into a 2 × 2 × 2 × 2 contingency table (16 cells)(shown in

figure 3.2)

```
import pandas as pd
import numpy as np
from scipy.stats import chi2

# 1) 读入清理好的 CSV
FILE_PATH = r"E:\python\kpmi_ru_data(Cleaned).csv"
df = pd.read_csv(FILE_PATH)

# 2) Extract the columns of four dimensions. (IE, SN, TF, JP)
mbti = df['psychotype'].astype(str).str.upper().dropna()
df_dims = pd.DataFrame({
    'IE': mbti.str[0],   # E 或 I
    'SN': mbti.str[1],   # S 或 N
    'TF': mbti.str[2],   # T 或 F
    'JP': mbti.str[3]    # J 或 P
})

# 3)Construct a 2×2×2×2 contingency table: Obtain the observed frequencies of 16 cells. counts[i,j,k,l]
import pandas as pd
import numpy as np

# Define dimension labels
index = pd.MultiIndex.from_product(
    [['E','I'], ['S','N'], ['T','F'], ['J','P']],
    names=['IE','SN','TF','JP']
)

# Build a DataFrame, using the values of counts as columns.
counts_df = pd.DataFrame({
    'Count': counts.flatten()
}, index=index)

print(counts_df.head(10))
```

```
             Count
IE SN TF JP
E  S  T  J   19038
         P   11168
      F  J    9506
         P    6604
   N  T  J    2933
         P    2333
      F  J    1914
         P    1802
I  S  T  J   14935
         P    8426
```

Then we calculate the marginal distribution (the distribution of each dimension separately), for example, N_IE = [count(E), count(I)](shown in figure3.3)

```
# Calculate marginal distribution and total number
N_IE = counts.sum(axis=(1,2,3))
N_SN = counts.sum(axis=(0,2,3))
N_TF = counts.sum(axis=(0,1,3))
N_JP = counts.sum(axis=(0,1,2))
N = counts.sum()


marginals = pd.DataFrame({
    'Dimension': ['IE', 'IE', 'SN', 'SN', 'TF', 'TF', 'JP', 'JP'],
    'Category': ['E', 'I', 'S', 'N', 'T', 'F', 'J', 'P'],
    'Count': np.concatenate([N_IE, N_SN, N_TF, N_JP])
})


print(marginals)
```

```
  Dimension Category  Count
0        IE        E  55298
1        IE        I  44702
2        SN        S  81094
3        SN        N  18906
4        TF        T  64882
5        TF        F  35118
6        JP        J  60894
7        JP        P  39106
```

Under the null hypothesis of mutual independence, the expected frequency in each cell was computed as the product of the four one-dimensional marginal distributions multiplied by the sample size.(shown in figure3.4)

```
chi2_stat = ((counts - E)**2 / E).sum()
chi2_stat
```

```
np.float64(1342.93063615322)
```

Now, it's time to calculate the Pearson chi-square statistic, while O is observed value and E

is expected value.(shown in figure 3.5)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

```
chi2_stat = ((counts - E)**2 / E).sum()
chi2_stat
```

```
np.float64(1342.93063615322)
```

Finally, with χ^2 = 1342.93 and df = 11,we can compute the right - tailed probability corresponding to the chi - square statistic under 10 degrees of freedom, which is 0(shown in figure 3.6). Because the p-value falls far below the conventional α =0.05 threshold, the null hypothesis is decisively rejected: the observed joint distribution of MBTI preferences deviates dramatically from what would be expected if the four indices varied independently. In practical terms, substantial associations exist among the E–I, S–N, T–F and J–P scales, corroborating earlier psychometric critiques that the MBTI dimensions are not orthogonal factors but overlap to a non-trivial extent. Actually, some researchers had also proven that the four dimensions of MBTI are not independent. Fleenor (1997) investigated the intercorrelations among MBTI continuous scores and found that while most dimension pairs demonstrated relatively low correlations, the correlation between the Sensing–Intuition (SN) and Judging–Perceiving (JP) scales was notably higher. Specifically, the study reported a correlation coefficient of r = 0.41 between SN and JP, indicating a moderate positive relationship. This finding has been replicated in other research and suggests that individuals who prefer intuition are more likely to also prefer perceiving. As such, the assumption of strict statistical independence between MBTI preference axes, particularly between SN and JP, may not hold, which is in line with the findings of our study.

```
df_val = 10
p_val = 1 - chi2.cdf(chi2_stat, df_val)
print(p_val)
```

```
0.0
```

```
if p_val < 0.05:
    print("Conclusion: Reject H₀ — The four dimensions are not completely independent.")
else:
    print("Conclusion: H₀ cannot be rejected — the four dimensions can be regarded as independent.")
```

```
Conclusion: Reject H₀ — The four dimensions are not completely independent.
```

7 Exploratory Data Analysis (EDA)

7.1 Data Visualization

7.1.1 MBTI Type Frequency

```
# 2) Frequency bar chart
plt.figure(figsize=(10,4))
df['psychotype'].value_counts().plot(kind='bar')
plt.title('Frequency of 16 MBTI Types')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

A frequency bar chart of the 16 MBTI types was generated using the value_counts() function on the psychotype column. The result was plotted using matplotlib and is shown in Figure 4.1.1.

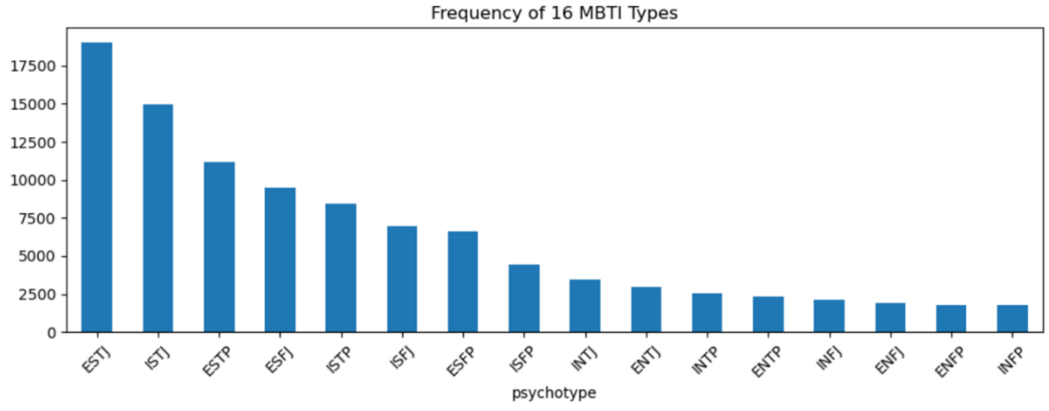

Figure 2: Frequency of 16 MBTI Types

As seen in Figure 4.1.1 above, the frequency of 16 MBTI types has been shown in a descending order. The type ESTJ ranks first, which shows that, based on our dataset, the proportion of ESTJ people is the highest. Conversely, the proportion of INFP people is the lowest. Additionally, it can be observed that the _S_J people rankings are relatively high in frequency, as the

_NF_ people rankings are relatively low.

Also, Figure 4.1.1 exhibits severe right oblique imbalance distribution, as the number of ESTJ people is nearly nine times as many as the number of INFP people. This phenomenon may be attributed to our database being originally from Kaggle, and certain MBTI type tends to participate in such investigations.

This contrasts with national MBTI distribution statistics reported in the MBTI® Manual, where ISFJ and ESFJ were found to be the most common types among U.S. adults (Myers et al., 1998), suggesting that our dataset, to a certain extent, fits the population-level trends. In distinct regions, the regional differences may influence MBTI type distributions in specific rankings.

7.1.2 MBTI Type Frequency

```
# 3) Histogram of scores in each dimension
score_cols = list('eisntfjp')
df[score_cols].hist(bins=20, figsize=(12,6))
plt.tight_layout()
plt.show()
```

Histograms of scores in each MBTI dimension were generated using pandas.DataFrame.hist and visualized with matplotlib as shown in Figure 4.1.2.

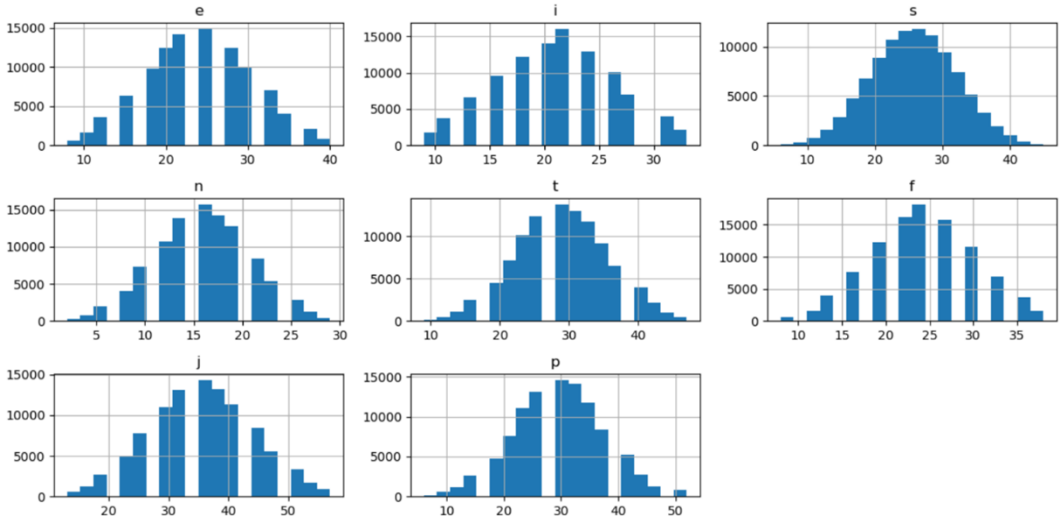

Figure 3: Histogram of MBTI Dimension Scores

As seen in Figure 4.1.2 above, from the score concentration trend, the eight MBTI poles vary in concentration. For example, E, S, and T scores are generally spread from 10 to 40,

while N and I scores are more narrowly concentrated between 5 to 30. And in detail, in the dimension of P and T, exhibit higher scores, which shows the preference toward Thinking and Perceiving. Compared to them, in the dimension of N and I, they exhibit lower scores, which indicates most people are more inclined toward Extraversion and Sensing.

From the data obtained from these diagrams, we can observe that most samples in these dimensions show distributional asymmetries, as not evenly distributed. Particularly for the high values of J and low values of N, this result matches the observation at 4.1.1 that _S_J types dominate the database.

7.1.3 MBTI Type Frequency

7.1.4 MBTI Type Frequency

```
# 4) Describe statistics (mean, std, quartile, skewness, kurtosis)
stats = df[score_cols].describe().T
stats['skew']  = df[score_cols].skew()
stats['kurt']  = df[score_cols].kurt()
print(stats.round(2))
```

Descriptive statistics of each MBTI dimension were computed and consolidated into a single summary table, including skewness and kurtosis, to facilitate score distribution analysis shown in Table 4.1.1

	count	mean	std	min	25%	50%	75%	max	skew	kurt
e	100000.0	24.00	6.52	8.0	20.0	25.0	28.0	40.0	0.02	-0.38
i	100000.0	21.00	5.43	9.0	18.0	22.0	24.0	33.0	-0.06	-0.44
s	100000.0	26.19	6.65	6.0	21.0	27.0	31.0	45.0	-0.09	-0.19
n	100000.0	16.12	4.71	2.0	12.0	16.0	19.0	29.0	-0.11	-0.05
t	100000.0	28.89	6.64	9.0	23.0	28.0	33.0	47.0	0.00	-0.25
f	100000.0	24.05	5.92	8.0	19.0	24.0	27.0	38.0	-0.08	-0.25
j	100000.0	35.41	8.80	13.0	29.0	35.0	41.0	57.0	0.01	-0.31
p	100000.0	29.34	7.93	6.0	23.0	29.0	35.0	52.0	0.05	-0.19

Figure 4: Descriptive Statistics of MBTI Dimension Scores

Based on Table 4.1.3, it can be observed that the dimensions of I, S, N, and F showcase a notable positive skew distribution, as the dimension of N, with a value of -0.11, demonstrates left skewness most. This phenomenon tends to show that most people are more likely to be Sensing(S) and so on. And in contrast, the dimension of P with the value of +0.05 exhibits right

skewness, implying a preference towards lower values, showing that Judging(J) dominates more in this database. Results of these also support the conclusion in Figure 4.1.1 that _S_J people occupy the majority.

And from the kurtosis value, it can be seen that most values are around 0 and negative, which suggests that all the dimension is distributed platykurtic. With relatively concentrated values, this exhibits the loss of extreme outliers. Also, the maximum value and minimum value that all dimensions have can be used in the form of Max-Min, which demonstrates that J and P dimensions have the largest difference, indicating that individual differences between them are the strongest.

		S		N	
		T n(%)	F n(%)	T n(%)	F n(%)
E	J	ESTJ	ESFJ	ENTJ	ENFJ
		30 (12.2)	32 (13.1)	12 (4.9)	11 (4.5)
	P	ESTP	ESFP	ENTP	ENFP
		12 (4.9)	30 (12.2)	14 (5.7)	20 (8.2)
I	J	ISTJ	ISFJ	INTJ	INFJ
		29 (11.8)	21 (8.6)	2 (0.8)	2 (0.8)
	P	ISTP	ISFP	INTP	INFP
		12 (4.9)	10 (4.1)	3 (1.2)	5 (2.0)

Figure 5: MBTI Distribution of 16 Personality Types

Figure 4.1.3 MBTI Distribution of 16 Personality Types from Jang and Kim (2014), supporting the dominance of _S_J types observed in our dataset.

7.1.5 MBTI Type Frequency

```
# 5) Correlation matrix heat map
plt.figure(figsize=(6,5))
sns.heatmap(df[score_cols].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation between dimension scores')
plt.tight_layout()
plt.show()
```

A Pearson correlation heatmap was constructed to visualize linear relationships among MBTI dimension scores, as shown in Figure 4.1.4.

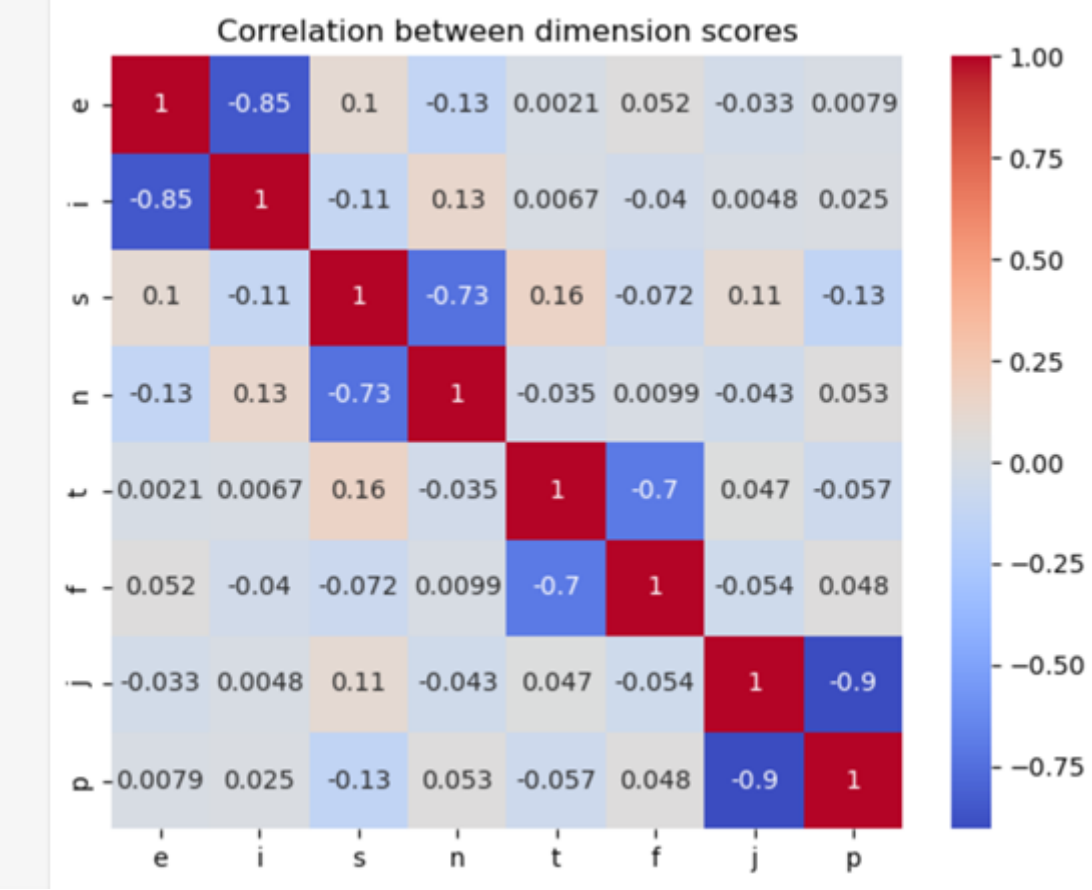

Figure 6: Correlation Heatmap of MBTI Dimension Scores

As seen in Figure 4.1.4 above, the correlation between the opposite dimensions, like E vs I, S vs N, T vs F, and J vs P. All of them are strongly negative, which is consistent with the design concept of MBTI, as the two poles of each dimension are opposed. As a result, the MBTI method has its rationality. This result aligns with (Li et al., 2024), who also observed that “the correlations between the four axes are generally weak, indicating that the personality traits on each axis are relatively independent” (p. 12).

Furthermore, in Figure 4.1.4, the correlation between other dimension pairs is low are nearly 0. This phenomenon supports that the dimensions are seemingly independent. While the correlation between S and T is +0.16, a bit larger than others, it indicates that someone who is Sensing(S) tends to be Thinking(T). In contrast, the correlation between S and P is -0.13, which indicates that those who are Sensing(S) tend to be Judging(J).

7.2 Sample Equation

Here is a displayed equation for illustration:

$$E = mc^2$$

8 Method

Continue writing your content here.

```
def quicksort(arr):
if len(arr) <= 1:
return arr
pivot = arr[0]
left = [x for x in arr[1:] if x < pivot]
right = [x for x in arr[1:] if x >= pivot]
return quicksort(left) + [pivot] + quicksort(right)
```

References

Boyle, G. J. (1995). Myers-briggs type indicator (mbti): Some psychometric limitations. *Humanities Social Sciences papers*, *30*. https://doi.org/10.1111/j.1742-9544.1995.tb01750.x

Fleenor, J. W. (1997). The relationship between the mbti and measures of personality and performance in management groups. *Developing leaders: Research and applications in psychological type and leadership development*, 115–138.

Yang, Y. (2022). Research on the application of mbti in organization. *2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)*, 1751–1754.

Zibran, M. F. (2007). Chi-squared test of independence. *Department of Computer Science, University of Calgary, Alberta, Canada*, *1*(1), 1–7.