Exploratory Data Analysis for Machine Learning

Project Overview

Dataset Summary: Big Five Personality Traits

Overview:

Code link in github:- https://github.com/ManikPandey/EDA_on_https-www.kaggle.com-datasets-tunguz-big-five-personality-test-Dataset

This dataset is taken from Kaggle (link:- https://www.kaggle.com/datasets/tunguz/big-five-personality-test).

Big five personality test dataset. This dataset contains over **1 million responses** (specifically, 1,015,342 entries) collected online between 2016 and 2018 via an interactive personality test. The test is based on the

Big Five personality traits (also known as OCEAN or the five-factor model), which groups personality into five broad dimensions:

1.Openness,

2.Conscientiousness,

3. Extraversion,

4.Agreeableness, and

5.Neuroticism.

Key Variables

- **Personality Trait Responses:**
  - 50 items (e.g., EXT1, AGR1, CSN1, EST1, OPN1) rated on a 1–5 scale.
  - Each set of 10 questions measures one of the five traits.
- **Timing Data:**
  - For each question, the time taken to respond (in milliseconds, variables ending in _E).
  - testelapse: Total time (in seconds) spent on the survey page.
- **User Metadata:**
  - dateload: Timestamp when the survey was started.
  - screenw, screenh: User's screen width and height.
  - introelapse, endelapse: Time spent on the intro and finalization pages.

    - IPC: Number of records from the user's IP address (used to filter for unique responses).
    - country: User's country (determined technically, not self-reported).
    - lat_appx_lots_of_err, long_appx_lots_of_err: Approximate latitude and longitude (with limited accuracy).

Dataset Size

- **Total records:** 1,015,342
- **Variables per record:**
    - 50 personality trait responses
    - 50 response times
    - 8+ metadata fields

Possible Target Variables

Depending on the project focus, potential target variables include:

- **Trait Scores:** Calculated scores for each of the Big Five traits (from the 50 responses)
- **Correlation between traits:** We can check for correlation between different traits .
- **Test Completion Time:** testelapse—useful for behavioral analysis.
- **Country or Region:** For demographic or cross-cultural studies.
- **Response Patterns:** Such as consistency or speed (e.g., unusually fast completions).

What Makes This Dataset Valuable

- **Large, diverse sample**: Over a million responses from around the world.
- **Rich behavioral data**: Includes both answers and response times.
- **Demographic context**: Country and technical metadata allow for subgroup analysis.

This dataset is well-suited for exploring relationships between personality traits, response behavior, and demographic factors.

1. Dataset Summary

The dataset contains responses to the Big Five personality test, collected through an online survey. Each row represents one participant's answers to 50 personality items and associated metadata. The dataset is large, with over 1 million entries and 110 columns.

Sample of the First 5 Records

EXT1	**EXT2**	**EXT3**	**EXT4**	**EXT5**	**...**	**dateload**	**screenw**	**screenh**	**introelapse**	**testelapse**
4.0	1.0	5.0	2.0	5.0	...	2016-03-03 02:01:01	768.0	1024.0	9.0	234.0

EXT1	EXT2	EXT3	EXT4	EXT5	...	dateload	screenw	screenh	introelapse	testelapse
3.0	5.0	3.0	4.0	3.0	...	2016-03-03 02:01:20	1360.0	768.0	12.0	179.0
2.0	3.0	4.0	4.0	3.0	...	2016-03-03 02:01:56	1366.0	768.0	3.0	186.0
2.0	2.0	2.0	3.0	4.0	...	2016-03-03 02:02:02	1920.0	1200.0	186.0	219.0
3.0	3.0	3.0	3.0	5.0	...	2016-03-03 02:02:57	1366.0	768.0	8.0	315.0

Note: Only a subset of columns is shown for brevity.

Dataset Structure

- **Total records:** 1,015,341
- **Columns:** 110
- **Data types:** 104 float64, 2 int64, 4 object
- **Memory usage:** ~852 MB

2. Descriptive Statistics

Below is a summary of key statistics for selected columns (all values are approximate):

Column	Mean	Std	Min	25%	50%	75%	Max
EXT1	2.65	1.26	0.00	1.00	3.00	4.00	5.00
EXT2	2.77	1.32	0.00	2.00	3.00	4.00	5.00
...	...	...	...	...	...	...	...
OPN10_E	5,336	440,822	-3.59e6	1,484	2,192	3,362	3.34e8
screenw	1,150	560	0	414	1,366	1,440	13,660
testelapse	675.4	20,179	1.0	171.0	224.0	313.0	1.19e7
endelapse	2,701	1.48e6	1.0	9.0	13.0	18.0	1.49e9
IPC	10.45	39.83	1.0	1.0	1.0	2.0	725

3. Missing Values

A summary of missing values in the original dataset:

Column	Null Count
EXT1	1,783
EXT2	1,783
...	...
country	77
endelapse	0
IPC	0
lat_appx_lots_of_err	0
long_appx_lots_of_err	0

- Most personality item columns have 1,783 missing values.
- The country column has 77 missing values.
- Some metadata columns are complete.

4. Data Cleaning and Preprocessing

Handling Missing Values

- **Row deletion:** Removing all rows with any null values reduced the dataset from 1,015,340 to 695,703 records. This approach ensures data completeness but results in data loss.
- **Forward fill:** Alternatively, missing values were filled using the last observed non-null value (fillna(method='ffill')). This can introduce bias or inaccuracies but preserves more data.

Reverse Scoring

Some items are negatively worded and require reverse scoring for accurate trait computation. The following items were reverse-scored:

- **Extraversion:** EXT2, EXT4, EXT6, EXT8, EXT10
- **Emotional Stability:** EST2, EST4
- **Agreeableness:** AGR1, AGR3, AGR5, AGR7
- **Conscientiousness:** CSN2, CSN4, CSN6, CSN8

- **Openness:** OPN2, OPN4, OPN6

Reverse-scoring formula:
new_value = 6 - original_value

Dropping Less Useful Columns

Columns like screenh, screenw, lat_appx_lots_of_err, and long_appx_lots_of_err were dropped to focus on core survey and personality data.

Aggregating Big Five Traits

For each participant, the average score for each Big Five trait was computed:

- **EXT_score:** Mean of EXT1–EXT10
- **EST_score:** Mean of EST1–EST10
- **AGR_score:** Mean of AGR1–AGR10
- **CSN_score:** Mean of CSN1–CSN10
- **OPN_score:** Mean of OPN1–OPN10

Encoding Categorical Variables

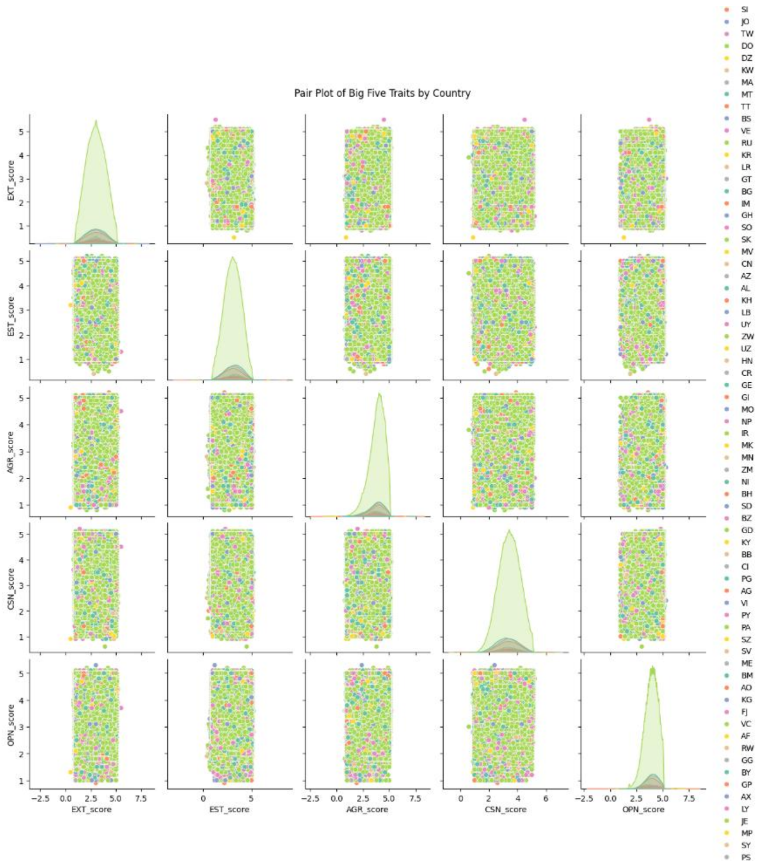
- **Country:** Encoded as country_code for easier analysis and modeling.

Date and Time Formatting

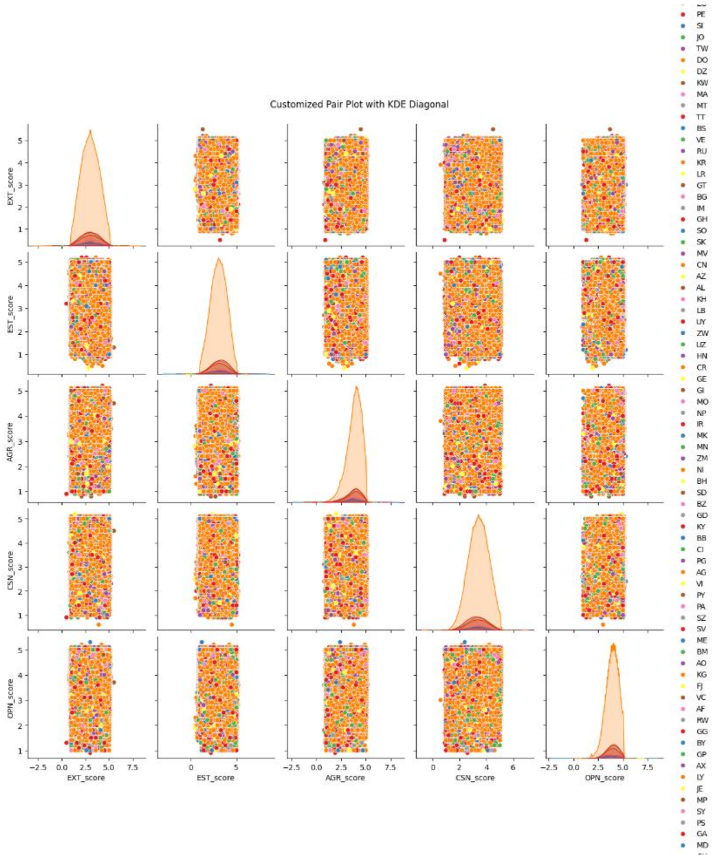
- **dateload:** Converted to datetime for time-based analysis, then reduced to only the time component for clarity.

Data visualization:

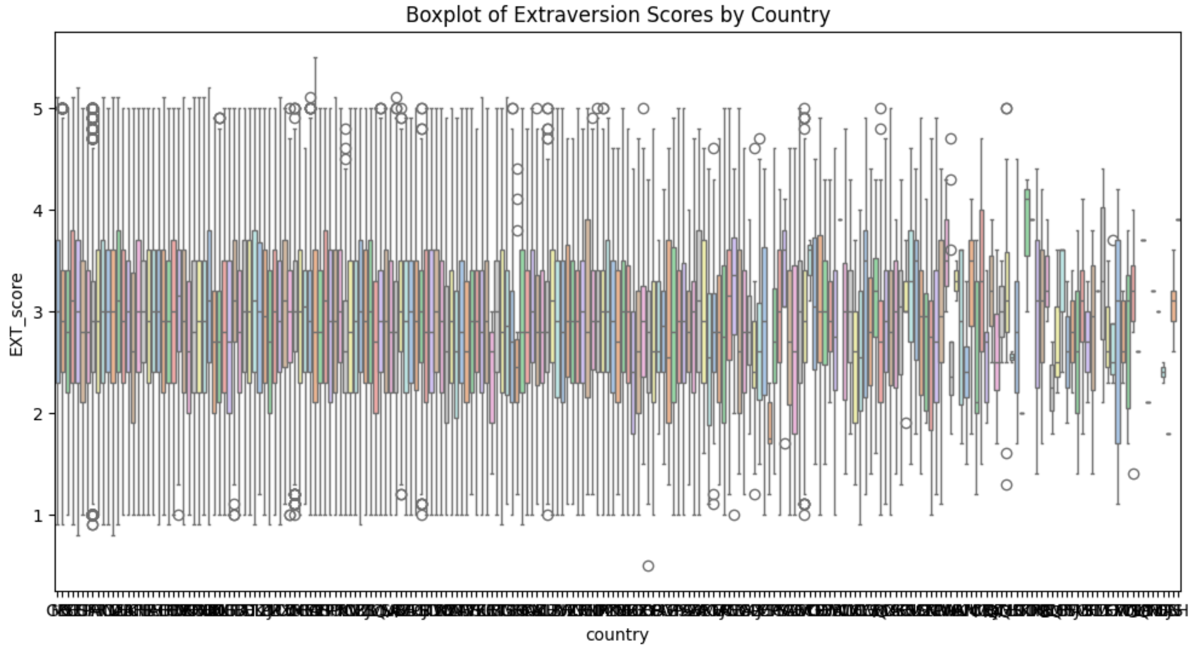
1. Scatter Plots and Pair Plots



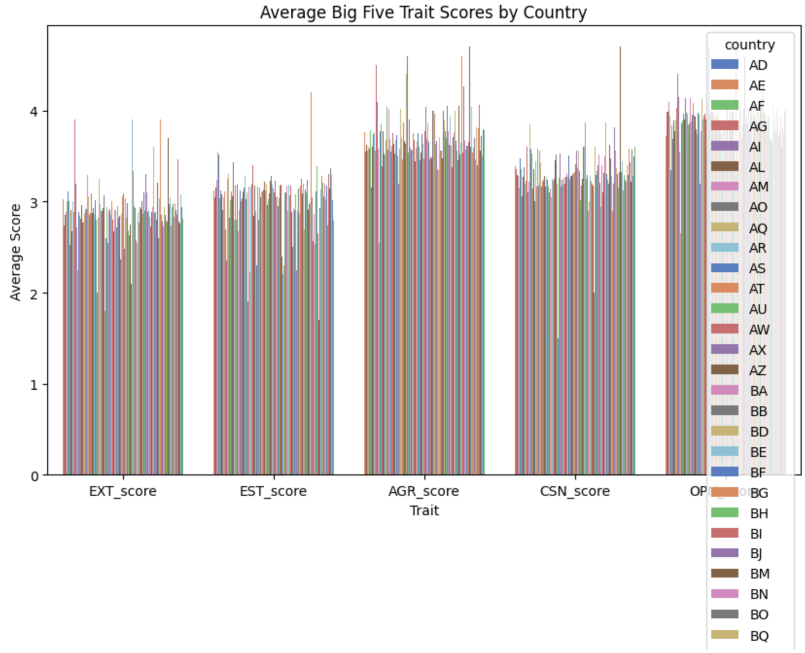
2. Customized Pair Plot



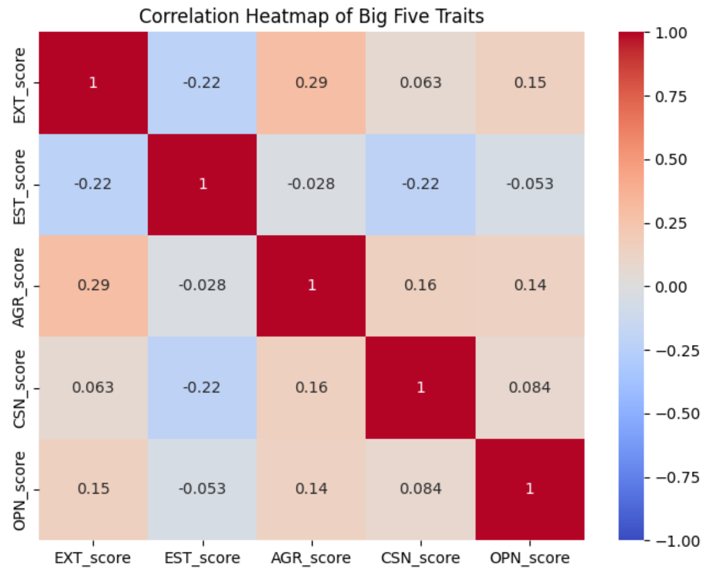
3. Boxplot: Extraversion by Country



4. Bar Plot: Average Trait Scores by Country

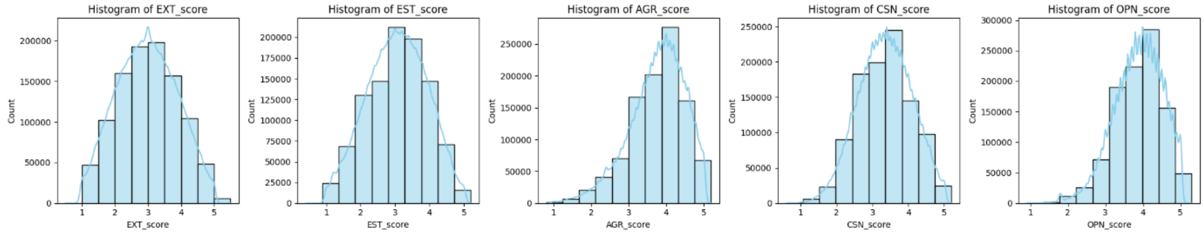


5. Heatmap: Correlation Matrix



Checking Skewness of BIG Five traits

Histograms for Each Trait



By Histogram of each trait we can infer that mostly the graph are near Normally distributed for EXT_score , EST_score , CSN_score . Where as AGR_score and OPN_score are little bit neatively skewed.

Checking Skewness of Big Five Trait Scores Measuring how much each trait score is skewed using the scipy.stats.skew function.

EXT_score: skewness = 0.029

EST_score: skewness = -0.089

AGR_score: skewness = -0.660

CSN_score: skewness = -0.090

OPN_score: skewness = -0.489

There is not much of skewness in our data so we can work with it easily.

Searching for outliers

Data info

```
RangeIndex: 1015341 entries, 0 to 1015340
Data columns (total 9 columns):
 #   Column        Non-Null Count    Dtype
---  ------        --------------    -----
 0   EXT_score     1015341 non-null  float64
 1   EST_score     1015341 non-null  float64
 2   AGR_score     1015341 non-null  float64
 3   CSN_score     1015341 non-null  float64
 4   OPN_score     1015341 non-null  float64
 5   country_code  1015341 non-null  int16
 6   country       1015341 non-null  object
 7   testelapse    1015341 non-null  float64
 8   total_minutes  1015341 non-null  int64
dtypes: float64(6), int16(1), int64(1), object(1)
memory usage: 63.9+ MB
```

Finding max and min values

```
Max values:
EXT_score              5.5
EST_score              5.2
AGR_score              5.2
CSN_score              5.2
OPN_score              5.3
testelapse       11892718.0
total_minutes        1439.0
dtype: float64

Min values:
EXT_score       0.5
EST_score       0.4
AGR_score       0.8
CSN_score       0.6
```

```
OPN_score       0.9
testelapse      1.0
total_minutes   0.0
dtype: float64
```

Searching for outliers:

from above mean , max and min values of column we can infer that there can be outlier values in testelapse as *mean(6.796953)* , *min(1.0)* values are close to eachother but the **max value 11892718** .

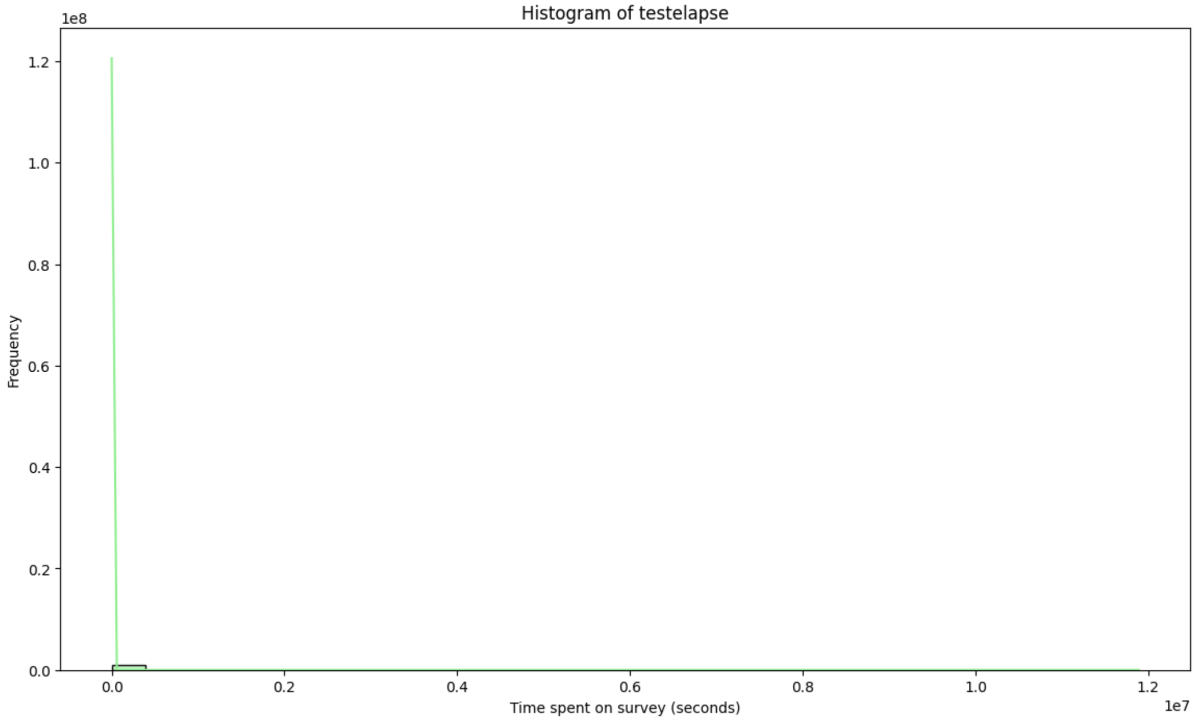
Inner quartile range and number of counts in inner quartile range -43.5 171.0 224.0 314.0 528.5

88666

by this we can say that interqueartile range is not perfect for detecting outliers in our data . As our data can be highly skewed

So lets visulize our *df_new['testelapse']* . To see if we can infer anything

1. Boxplot



from the above histogram we can say that data has an scaling issue or the data is highly skewed . Either of the way we can apply log transformation for better understanding of the data.

Checking for outliers

1. By comparing the 99th percentile, And Capping the data wiht 1 hour or 3600 seconds vs the original data.

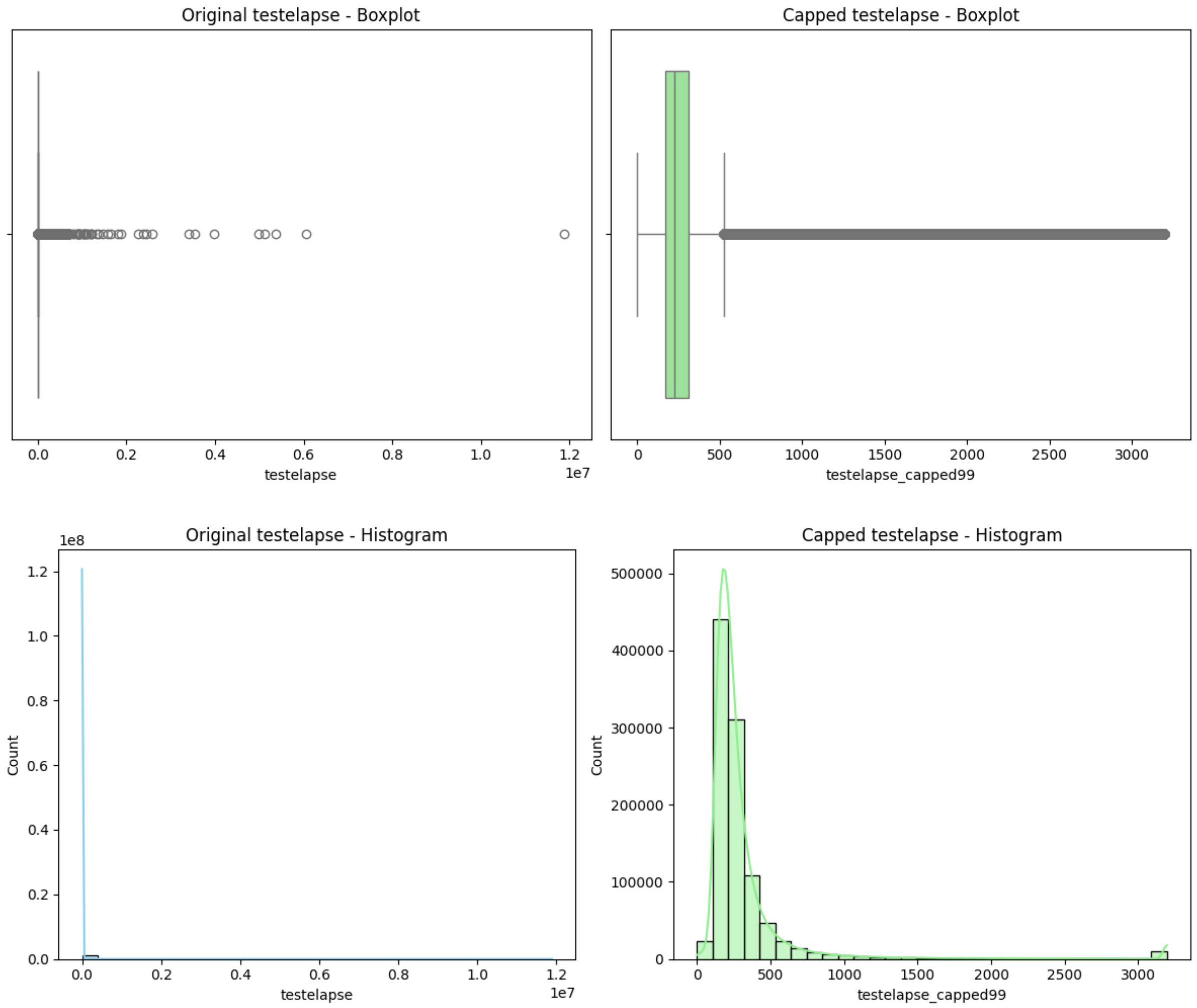
```
percentile_100(max_data) 11892718.0
```

```
percentile_99 3194.5999999999767
count of outliers:  10154
```

1. `total % of outliers from total rows 1.0000581085566327`

From this we can say that removing the 1% outliers can be a good choice as **99%** of **testelaspse** *is less than 1 hour(or 3600sec)*

Visualizing With and Without Outliers



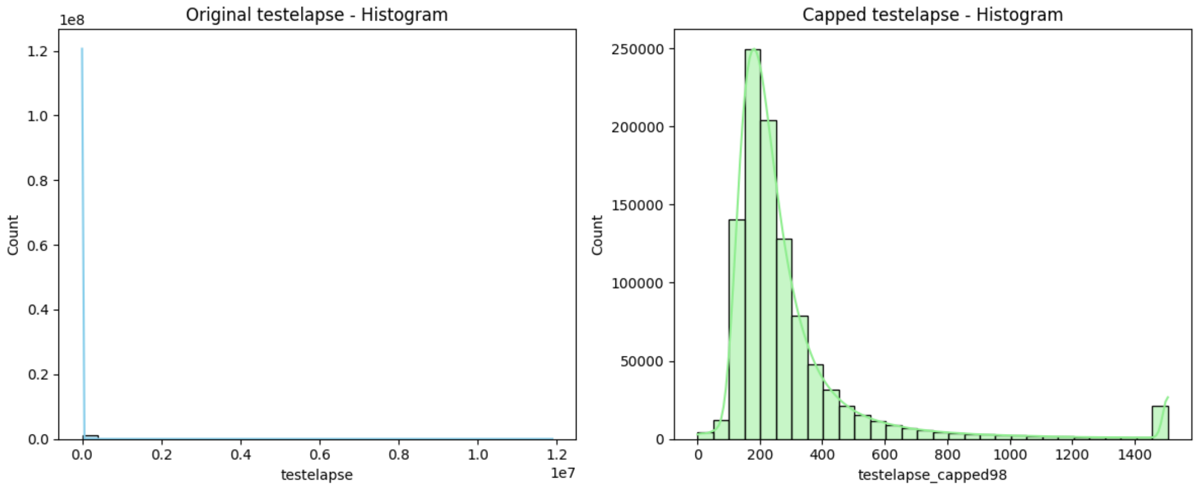
After removing the data, we can still see there are some outlier on the right side of the graph, So lets do the same thing by capping the data to 98percentile

percentile_100(max_data) 11892718.0

percentile_99 1506.0

count of outliers: 20296

total % of outliers from total rows 1.9989343481648036



Checking skewness of above data

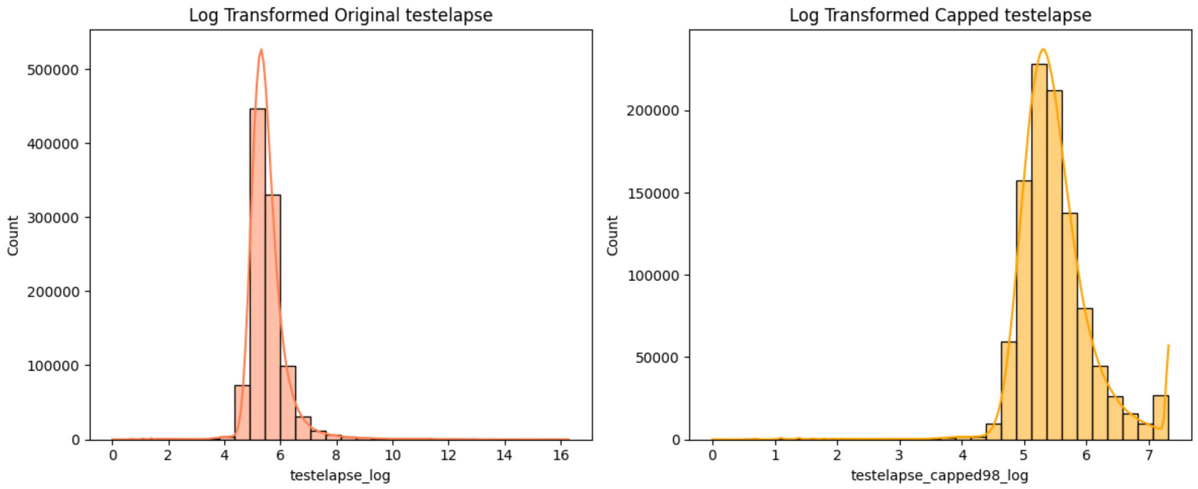
testelapse_capped98: skewness = 3.291

Transformation:

Transformation of Testelapse_capped98 data

Log Transformation

Apply a log transformation to both the original and capped variables to help normalize the distribution. By adding a small constant 1 to all values, including zero, you ensure that all values in the transformed data are valid for the logarithm function



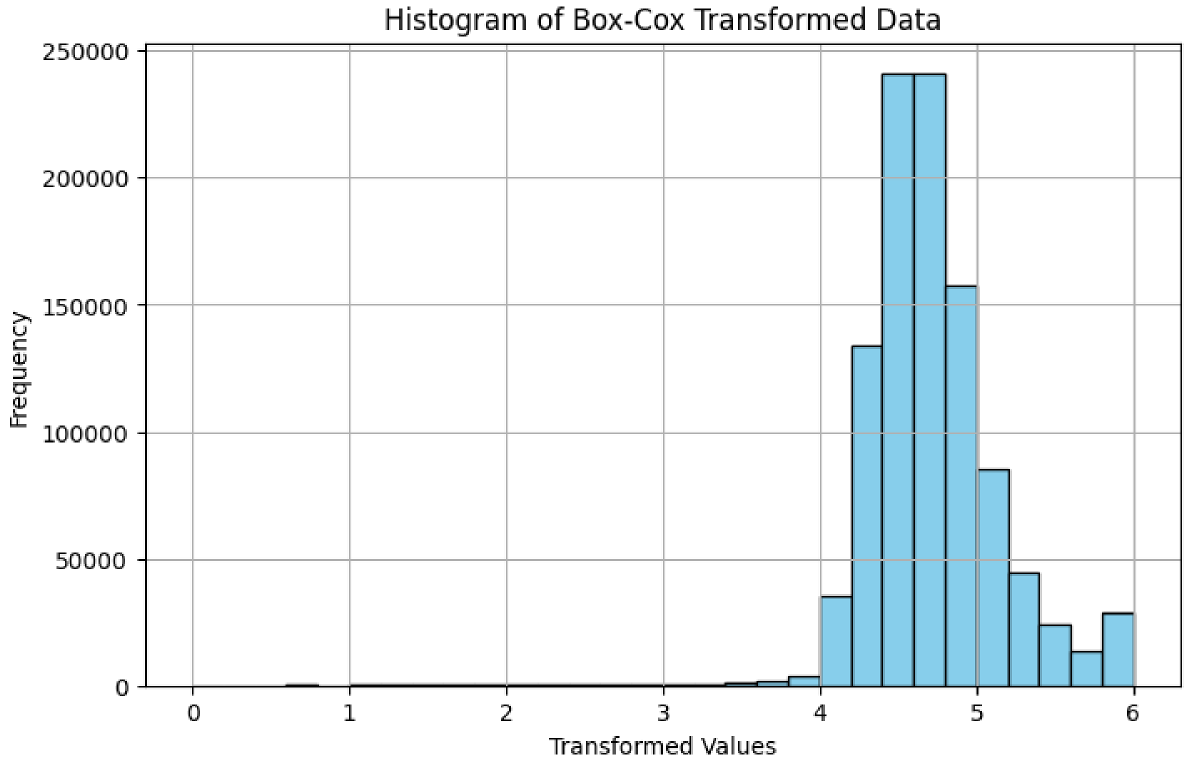
Still the log transformed data is highly skewed

testelapse_capped98_log: skewness = 3.291

Trying Box-Cox Transformation

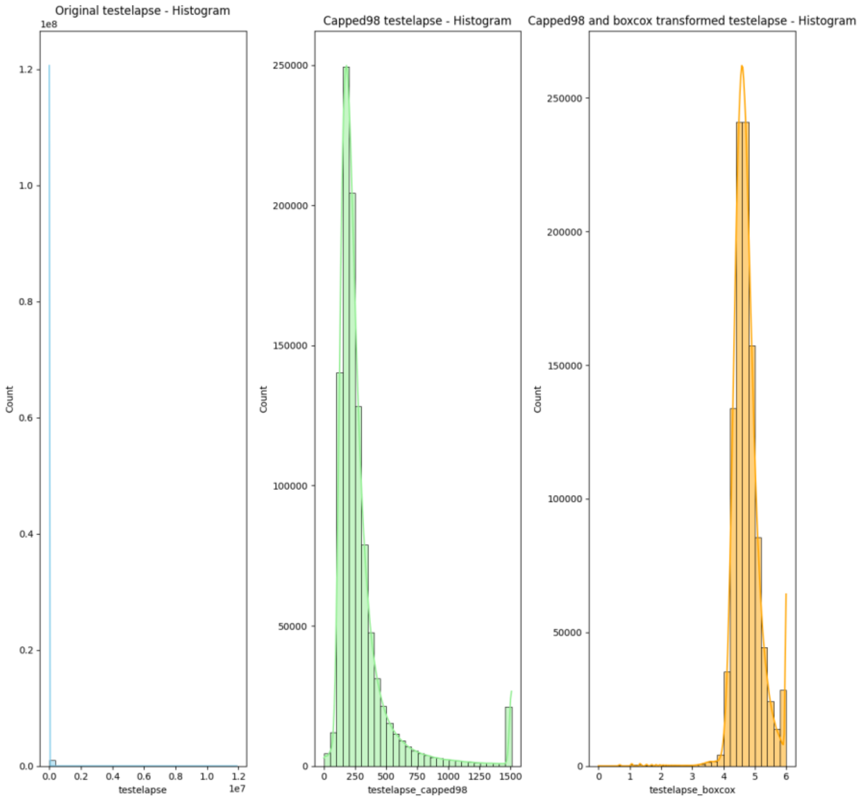
Skewness after Box-Cox: -0.07477958106275462

Lambda used: -0.0562376888333087



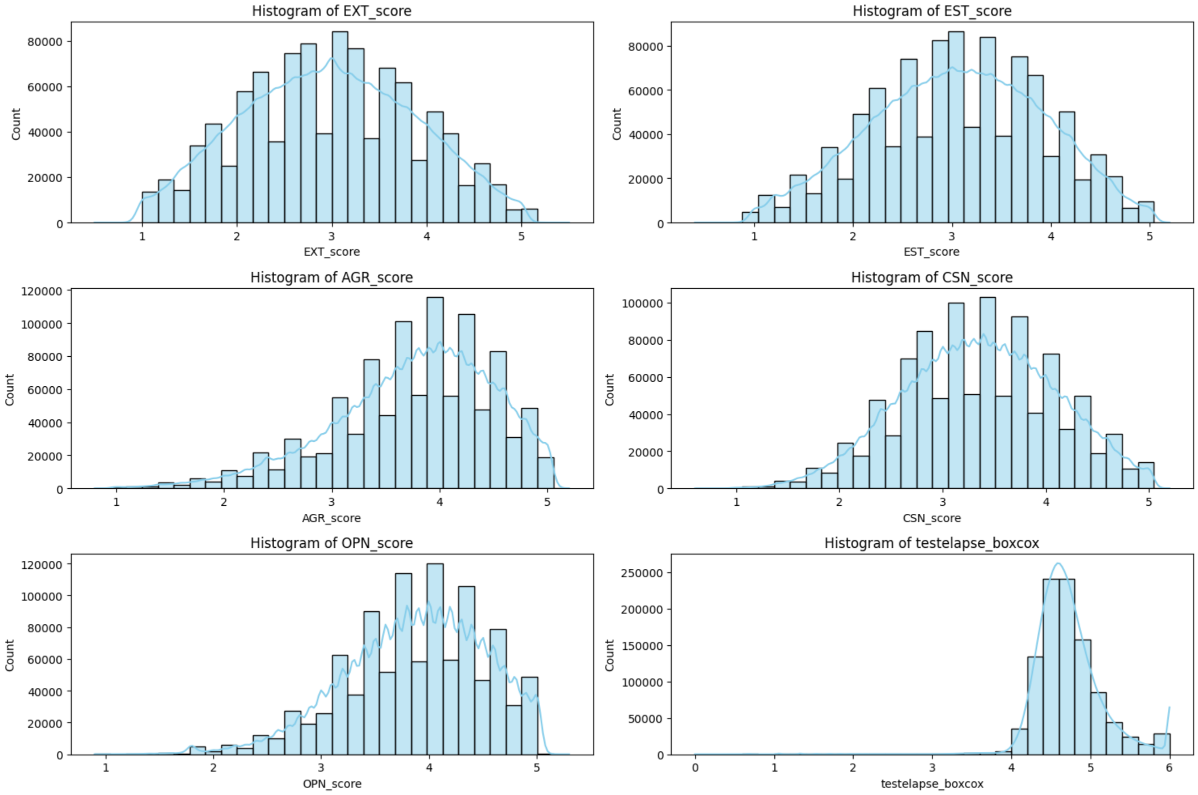
visualizing histogram with and wihtout outliers in Testelapse data

Original data vs capped98 vs Box cox transformed data



Visualizing Big Five Traits and Testelapse (Box-Cox Transformed)

1. Histograms for Big Five Traits and Testelapse_Boxcox



Hypotheses Testing

1. Personality Traits and Test Completion Time

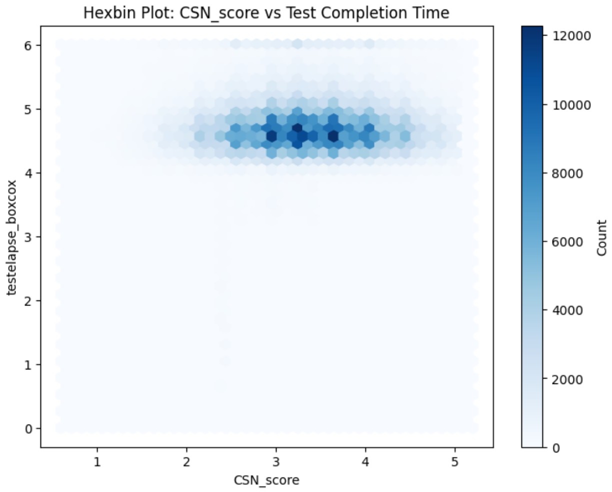
There is a significant association between Conscientiousness (CSN_score) and the time taken to complete the test (testelapse_log or testelapse_capped98_log). Rationale: More conscientious individuals may complete the survey more efficiently, resulting in shorter completion times

H0: There is no any significant association between Conscientiousness (CSN_score) and the time taken to complete the test (testelapse or testelapse_boxcox)

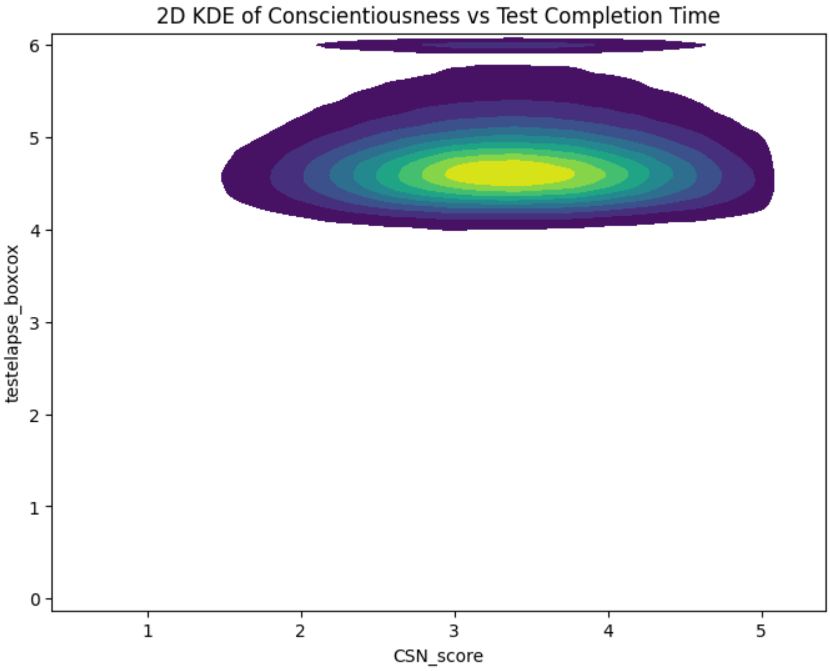
H1: There is a significant association between Conscientiousness (CSN_score) and the time taken to complete the test (testelapse or testelapse_boxcox).

Rationale: More conscientious individuals may complete the survey more efficiently, resulting in shorter completion times

1. Hexbin plot : Average Test Completion Time by Conscientiousness score



2. 2D KDE of Conscientiousness vs Test Completion Time



Correlation (log): 0.025, p-value: 0.0000

Correlation (Box-Cox): 0.034, p-value: 0.0000

The p-value is not significant, proves that there is no significant relation between testelapse and conscientiousness. Thus accepting our null hypothesis.

H0: There is no any significant association between Conscientiousness (CSN_score) and the time taken to complete the test (testelapse or testelapse_boxcox)

2. Inter-Trait Relationships

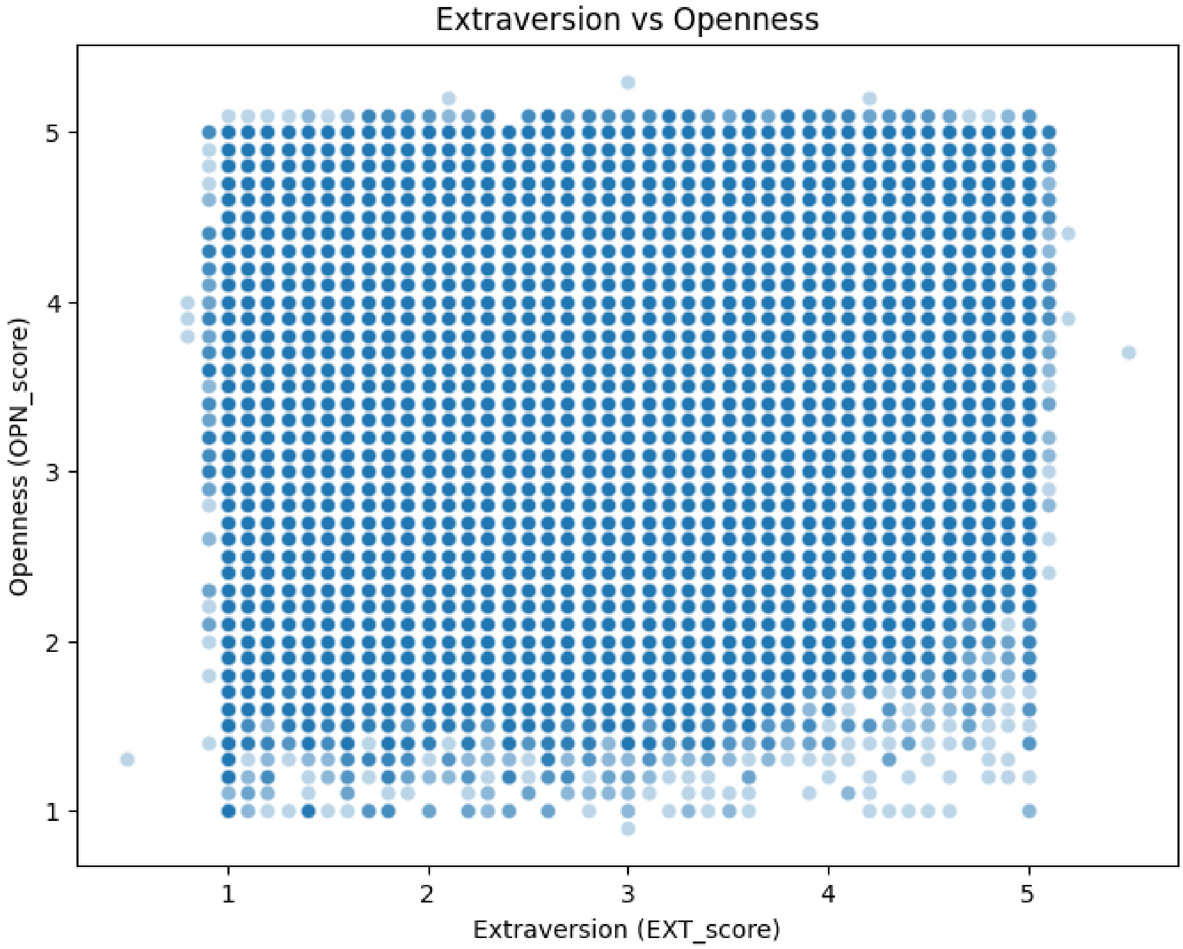
H5: Extraversion (EXT_score) and Openness (OPN_score) are positively correlated. Rationale: People who are outgoing may also be more open to new experiences.

H0: Extraversion (EXT_score) and Openness (OPN_score) are not positively correlated.

H1: Extraversion (EXT_score) and Openness (OPN_score) are positively correlated.

Rationale: People who are outgoing may also be more open to new experiences.

Visualize the Relationship (Scatter Plot)



Pearson correlation: 0.149

p-value: 0.0000

There is a significant positive correlation between Extraversion and Openness.

H1 is accepted

H1: Extraversion (EXT_score) and Openness (OPN_score) are positively correlated.

3. Group Differences

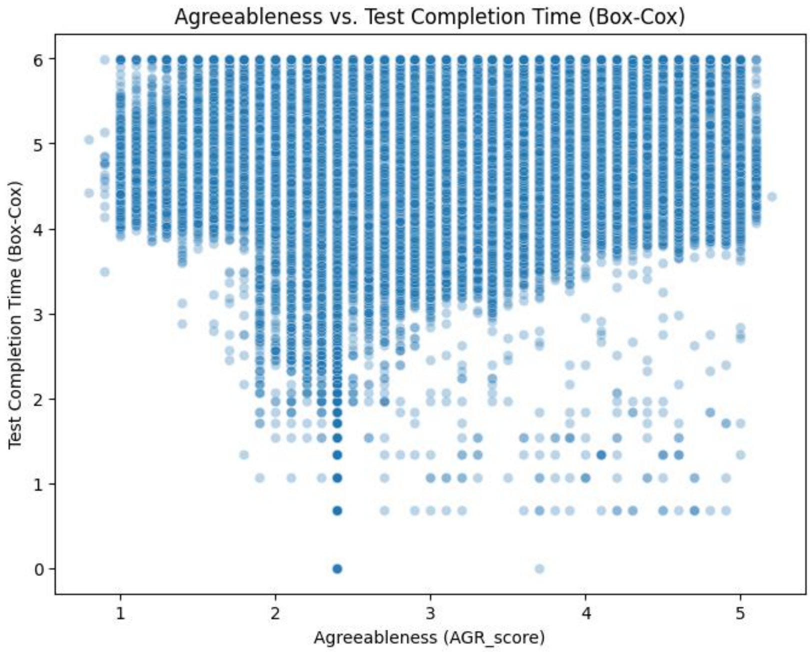
Participants with high Agreeableness (AGR_score) tend to have lower test completion times compared to those with low Agreeableness. Rationale: Agreeable individuals may be more cooperative and focused during surveys.

H0: There is a no realtion between test completion time and Agreeableness. Participants with high Agreeableness (AGR_score) don't have any signigicant affect on completion times compared to those with low Agreeableness.

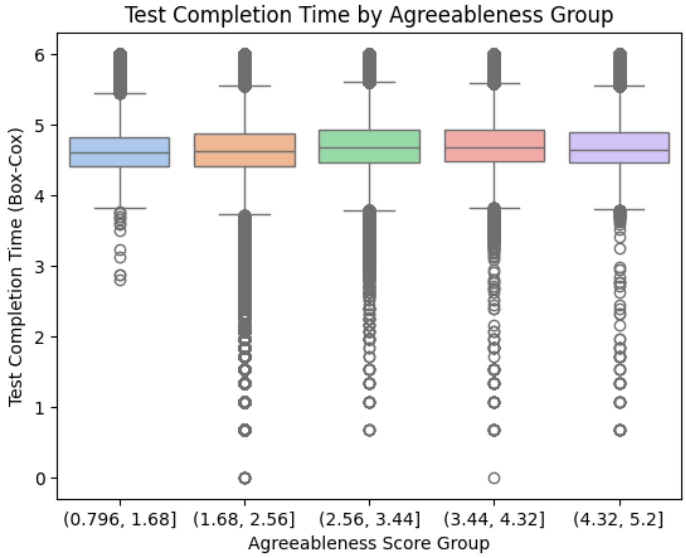
H1: Participants with high Agreeableness (AGR_score) tend to have lower test completion times compared to those with low Agreeableness. i.e there is a realtion betwen test completion time and Agreeableness

Relation: Agreeable individuals may be more cooperative and focused during surveys.

1. Visualize the Relationship



2. Boxplot by Agreeableness group



Statistical Test:

Pearson correlation: 0.041

p-value: 0.0000

Interpretation:

A negative and significant correlation ($p < 0.05$) supports the alternative hypothesis. If the correlation is not significant, there is no evidence of a relationship.

Results: Pearson correlation: 0.041 p-value: 0.0000 Result:

While the relationship is statistically significant, it is in the opposite direction of the hypothesis (higher Agreeableness is linked to slightly longer completion times), and the effect is so small as to be practically insignificant. Thus, there is no meaningful evidence that more agreeable individuals complete the survey more quickly.

Thus, Null Hypothesis is accepted. (H0): There is no relationship between Agreeableness and test completion time.

Summary

Big Five Personality Traits Dataset: EDA Summary & Analysis

1. Data Summary

Dataset Overview:

- **Source:** Kaggle, Big Five Personality Test (2016–2018)
- **Size:** 1,015,342 records, 110 columns
- **Variables:**
    - **Personality Trait Responses:** 50 items (e.g., EXT1, AGR1, etc.), each rated 1–5
    - **Response Times:** 50 variables (e.g., EXT1_E), time in milliseconds
    - **User Metadata:** Includes timestamp (dateload), screen dimensions, intro/finalization timing, IP-based counts, country, and approximate latitude/longitude
- **Possible Target Variables:**
    - Calculated trait scores (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism)
    - Test completion time (testelapse)
    - Country/region for demographic analysis
    - Response consistency or speed

2. Data Exploration Plan

Vision for Analysis:

- **Understand trait distributions:** Assess the overall and country-wise distributions of Big Five scores.
- **Behavioral patterns:** Analyze response times and completion behaviors.
- **Demographic insights:** Explore how traits and behaviors vary across countries.
- **Trait interrelations:** Investigate correlations between personality traits.
- **Outlier and anomaly detection:** Identify and address extreme values in responses and timings.

Steps:

1. Generate descriptive statistics for all variables.
2. Visualize trait distributions and relationships (histograms, scatter plots, pair plots).
3. Examine response time data for skewness and outliers.
4. Explore demographic patterns using country-level aggregations.
5. Formulate and test hypotheses about trait relationships and behavioral outcomes.

3. Exploratory Data Analysis (EDA) Results

Trait Distributions:

- Most trait scores (Extraversion, Emotional Stability, Conscientiousness) are nearly normally distributed.
- Agreeableness and Openness show slight negative skewness.

Summary Statistics Example:

Trait	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
EXT1	2.65	1.26	0.0	1.0	3.0	4.0	5.0
EXT2	2.77	1.32	0.0	2.0	3.0	4.0	5.0
testelapse	675	20179	1.0	171	224	313	1.19e7

Visualizations:

- **Scatter and pair plots:** Reveal relationships between traits, both overall and by country.
- **Boxplots:** Show trait score distributions by country and outlier presence in test completion times.
- **Histograms:** Confirm near-normal trait distributions; highlight skewness in timing data.
- **Heatmaps:** Display trait correlation matrices.

Outlier Detection & Response Time:

- Extreme outliers in testelapse (max: 11,892,718 sec) suggest data entry or behavioral anomalies.
- Capping at the 98th/99th percentile and applying Box-Cox transformation normalizes timing data.

4. Data Cleaning & Feature Engineering

Missing Values:

- Personality items: 1,783 missing per column.
- Country: 77 missing.
- Approach: Rows with any nulls removed (reduced to 695,703 records), or forward-fill used as alternative.

Reverse Scoring:

- Negatively worded items reverse-scored for accurate trait computation (e.g., EXT2, EST2, AGR1, etc.).
- Formula: new_value = 6 - original_value.

Feature Aggregation:

- Trait scores computed as the mean of their 10 respective items (e.g., EXT_score = mean(EXT1–EXT10)).

Encoding & Formatting:

- Country encoded as country_code.
- Timestamps converted to datetime, then reduced to time component.

Visualization Outputs:

- **Histograms:** For each trait and test completion time.
- **Boxplots:** For outlier detection in timing data.
- **Pair plots:** For trait interrelationships and country differences.

5. Key Findings & Insights

- **Trait Distributions:** Most traits are nearly normal; Agreeableness and Openness slightly skewed.
- **Outliers:** Test completion time contains extreme outliers; capping and transformation are necessary.
- **Trait Relationships:** Extraversion and Openness are positively correlated (Pearson r = 0.149, p < 0.001).
- **Behavioral Patterns:** No meaningful relationship between Conscientiousness or Agreeableness and test completion time, despite statistical significance due to large sample size.

- **Country Differences:** Visualizations show country-level variation in trait averages.

6. Hypotheses

1. **Conscientiousness and Completion Time:** Higher Conscientiousness (CSN_score) is associated with shorter test completion times.
2. **Extraversion and Openness:** Extraversion (EXT_score) and Openness (OPN_score) are positively correlated.
3. **Agreeableness and Completion Time:** High Agreeableness (AGR_score) leads to faster test completion.

7. Significance Test Discussion

Hypothesis Tested:
Conscientiousness is associated with shorter test completion time.

- **Null Hypothesis (H0):** No significant association between CSN_score and test completion time.
- **Alternative (H1):** Higher CSN_score correlates with shorter completion time.

Results:

- Correlation (log-transformed): 0.025, p-value < 0.001
- Correlation (Box-Cox transformed): 0.034, p-value < 0.001

Interpretation:

- The correlation is statistically significant but extremely small, indicating no practical relationship.
- The null hypothesis is accepted: Conscientiousness does not meaningfully predict test completion speed.

8. Conclusion & Next Steps

Key Takeaways:

- The dataset is robust, diverse, and suitable for advanced behavioral and demographic analyses.
- Trait distributions are mostly normal; outliers in timing data must be addressed for modeling.
- Some trait relationships exist (e.g., Extraversion–Openness), but behavioral predictions (e.g., completion speed) are weak.

Next Steps:

- Further investigate demographic and cultural influences on trait distributions.
- Apply advanced modeling (e.g., clustering, predictive analytics) to uncover deeper insights.
- Explore response patterns (e.g., consistency, speed) for potential quality control or behavioral research.