Intro to Data Visualisation II

Dr. Gordon Wright

Refresher/Primer on Data Types

In statistics, we work with various types of data. Understanding these types is crucial for:

- Choosing appropriate statistical methods
- Interpreting results correctly
- Making informed decisions based on data

Let’s explore the four main levels of measurement…

Nominal Data

- **Definition**: Categories or groups without intrinsic order
- **Characteristics**:
  - Cannot be ordered
  - No numerical value
  - Only shows distinct groups
- **Examples from our dataset**:
  - MBTI (Myers-Briggs Type Indicator)
  - Coin (Heads or Tails)
  - DogCatBoth (Preference for pets)

Ordinal Data

- **Definition**: Categories with a meaningful order, but differences aren’t measurable
- **Characteristics**:
  - Can be ordered
  - Intervals between ranks aren’t necessarily equal
- **Examples from our dataset**:
  - 1-7 rating scales (Likert scale responses)
  - TIPI scores (Big Five - OCEAN)

Interval Data

- **Definition**: Numerical data with consistent intervals, but no true zero point
- **Characteristics**:
  - Ordered with equal intervals between values
  - Can be added or subtracted
  - No true zero point
- **Examples**:
  - Temperature in Celsius or Fahrenheit
  - Calendar years

Ratio Data

- **Definition**: Numerical data with equal intervals and a true zero point
- **Characteristics**:
  - Ordered with equal intervals
  - Has a true zero point (absence of the variable is possible)
  - Can be added, subtracted, multiplied, and divided
- **Examples from our dataset**:
  - LoginCount
  - CompTime (assuming it's recorded in minutes)
  - EyeContact (assuming it's a continuous measure)

Summary of Data Types

Data Type	Can Be Ordered?	Equal Intervals?	True Zero Point?	Example
Nominal	No	No	No	Gender, MBTI
Ordinal	Yes	No	No	Scale measures, TIPI scores
Interval	Yes	Yes	No	Temperature (°C)
Ratio	Yes	Yes	Yes	LoginCount, CompTime

Practical Application

Let’s look at some variables from our new, exciting dataset:

Column Name	Description	Data Type
PokeNumber	Unique identifier for each Pokémon	Categorical
PokeName	Name of the Pokémon	Text
PokeImage	URL to the Pokémon’s image	Text
LoginCount	Number of VLE logins (to date)	Ratio
Q_Scale	The Scale-Based Question you submitted for DS01	Text
Q_choice	The Choice-Based Question you submitted for DS01	Text
Q_option	The Open Question you submitted for DS01	Text
CompTime	DS02 Survey Completion Time	Ratio

Conclusion and Next Steps

- Understanding data types is crucial for proper statistical analysis
- In our dataset, we have a mix of nominal, ordinal, and ratio data
- Next, we'll explore how to summarize and visualize these different types of data

Introduction to Measures of Central Tendency

Measures of central tendency help us understand the typical or central value in a dataset. The three main measures are:

1. Mean
2. Median
3. Mode

Let’s explore each of these using our dataset…

Mean

- **Definition**: The average of all values in a dataset
- **Formula**: $\bar{x} = \frac{\sum X}{N}$
- **Best for**: Interval and ratio data
- **Example**: Let’s calculate the mean LoginCount

Let’s see all the values available, first

This is called a Stem and Leaf Plot and we will calculate the mean below it.

```
  The decimal point is 1 digit(s) to the right of the |

  0 | 112333444556668888999
  1 | 000000000111122223333455555556677788899
  2 | 00123333444456699
  3 | 0023355567
  4 | 34578
  5 | 0168
  6 | 1

Mean LoginCount: 19.43
```

Median

- **Definition**: The middle value when data is ordered
- **Best for**: Ordinal, interval, and ratio data
- **Example**: Let’s find the median LoginCount

```
Median LoginCount: 15
```

Mode

- **Definition**: The most frequently occurring value
- **Best for**: Any type of data, especially nominal
- **Example**: Let’s find the mode of DogCatBoth

```
Counts for DogCatBoth:


Cats Both Dogs
  38   20   12


Mode of DogCatBoth: Cats



Percentages for DogCatBoth:

Cats: 54.29%
Both: 28.57%
Dogs: 17.14%
```

Comparing Measures of Central Tendency

Let’s compare these measures for LoginCount:

```
Mean LoginCount: 19.43
Median LoginCount: 15
Mode LoginCount: 10
```

Introduction to Measures of Variance

Measures of variance help us understand the spread or dispersion of data. Key measures include:

1. Range
2. Interquartile Range (IQR)
3. Variance
4. Standard Deviation

Range and Interquartile Range (IQR)

- **Range**: Difference between the maximum and minimum values
- **IQR**: Range of the middle 50% of the data

Let’s calculate these for LoginCount:

```
Range of LoginCount: 60
IQR of LoginCount: 15
```

Variance and Standard Deviation

- **Variance**: Average squared deviation from the mean
- **Standard Deviation**: Square root of the variance

Let’s calculate these for LoginCount:

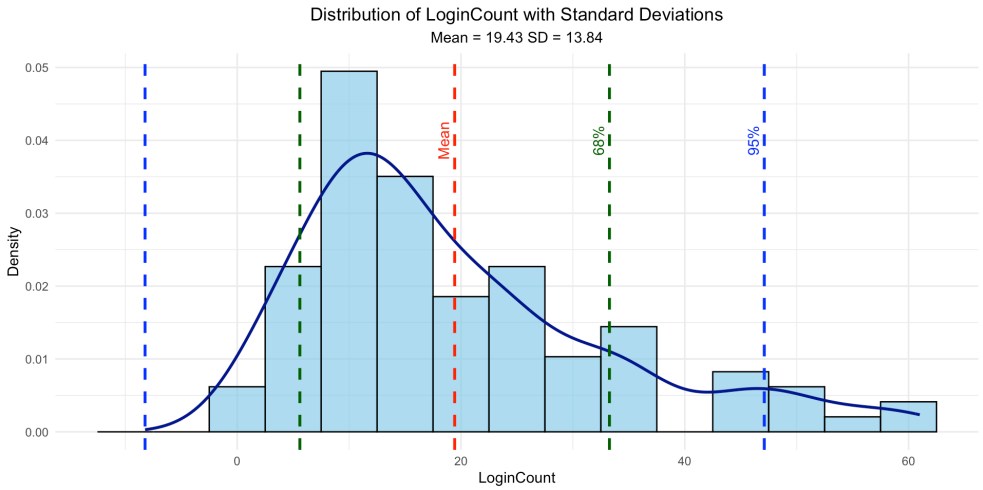
```
Variance of LoginCount: 191.44
Standard Deviation of LoginCount: 13.84
```

Interpreting Standard Deviation

- In a normal distribution:
  - About 68% of data falls within 1 SD of the mean
  - About 95% falls within 2 SD
  - About 99.7% falls within 3 SD

Let's visualize this for LoginCount:

```
Percentage within 1 SD: 73.2 %
Percentage within 2 SD: 93.81 %
Percentage within 3 SD: 98.97 %
```



Application to Different Data Types

Let’s apply these concepts to different types of data in our dataset:

```
LoginCount (Ratio):
Mean: 19.43
SD: 13.84
Q_Scale (Ordinal):
Mean: 3.04
Median: 3
IQR: 2
Frequency Table:

 1  2  3  4  5
 9 25 40 23 12
DogCatBoth (Nominal):

Both Cats Dogs
  20   38   12
```

Mi-Fin

- Measures of central tendency and variance provide crucial summaries of our data
- The choice of measure depends on the type of data and the shape of its distribution
- These measures form the basis for more advanced statistical analyses
- In the second half, we'll explore how to visualize these concepts and dive deeper into our dataset and preview the labs!

Data Exploration

Data exploration involves:

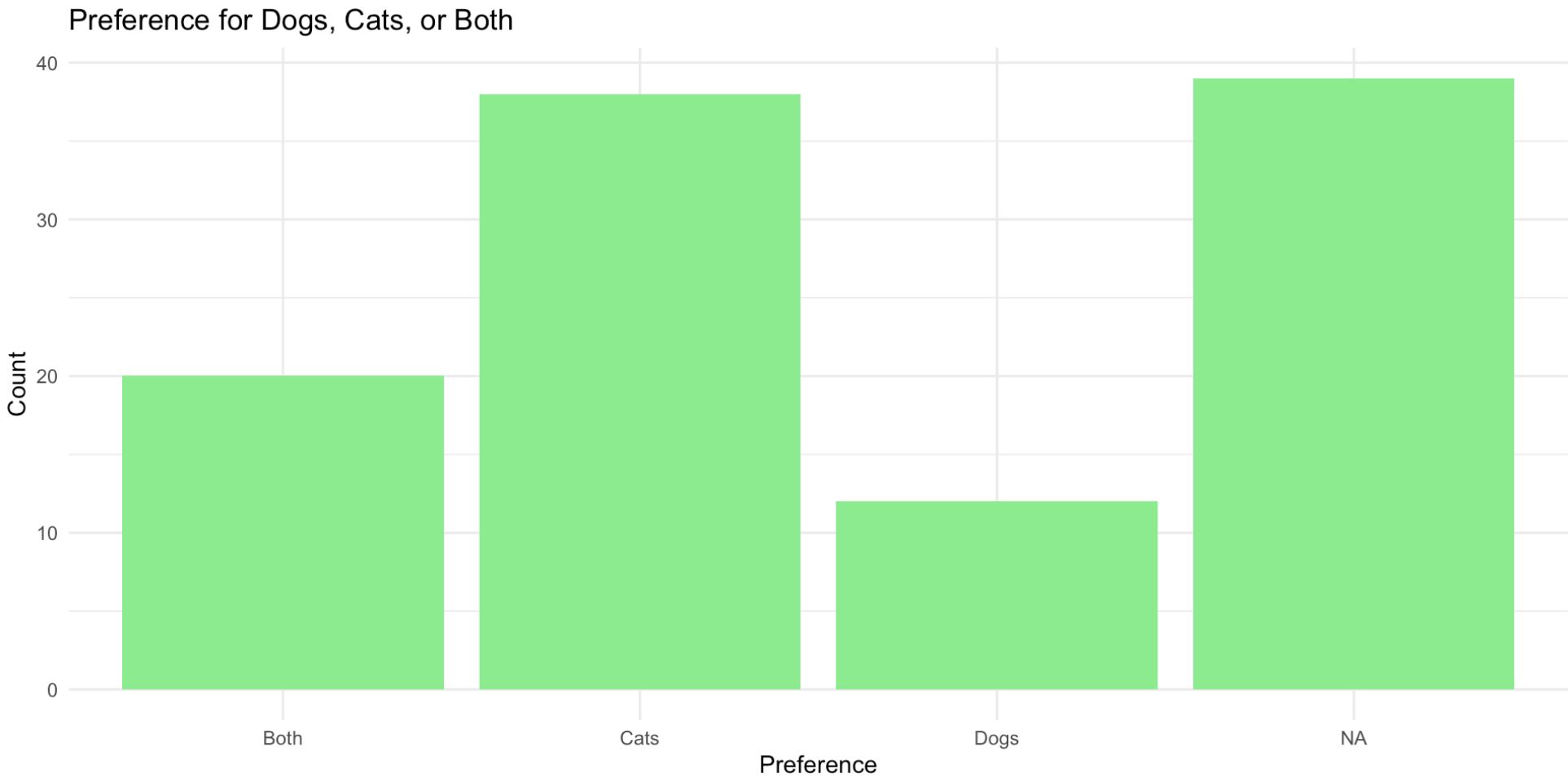
- Visualizing data distributions
- Identifying patterns and relationships
- Detecting outliers and anomalies

We'll use various chart types to explore our dataset.

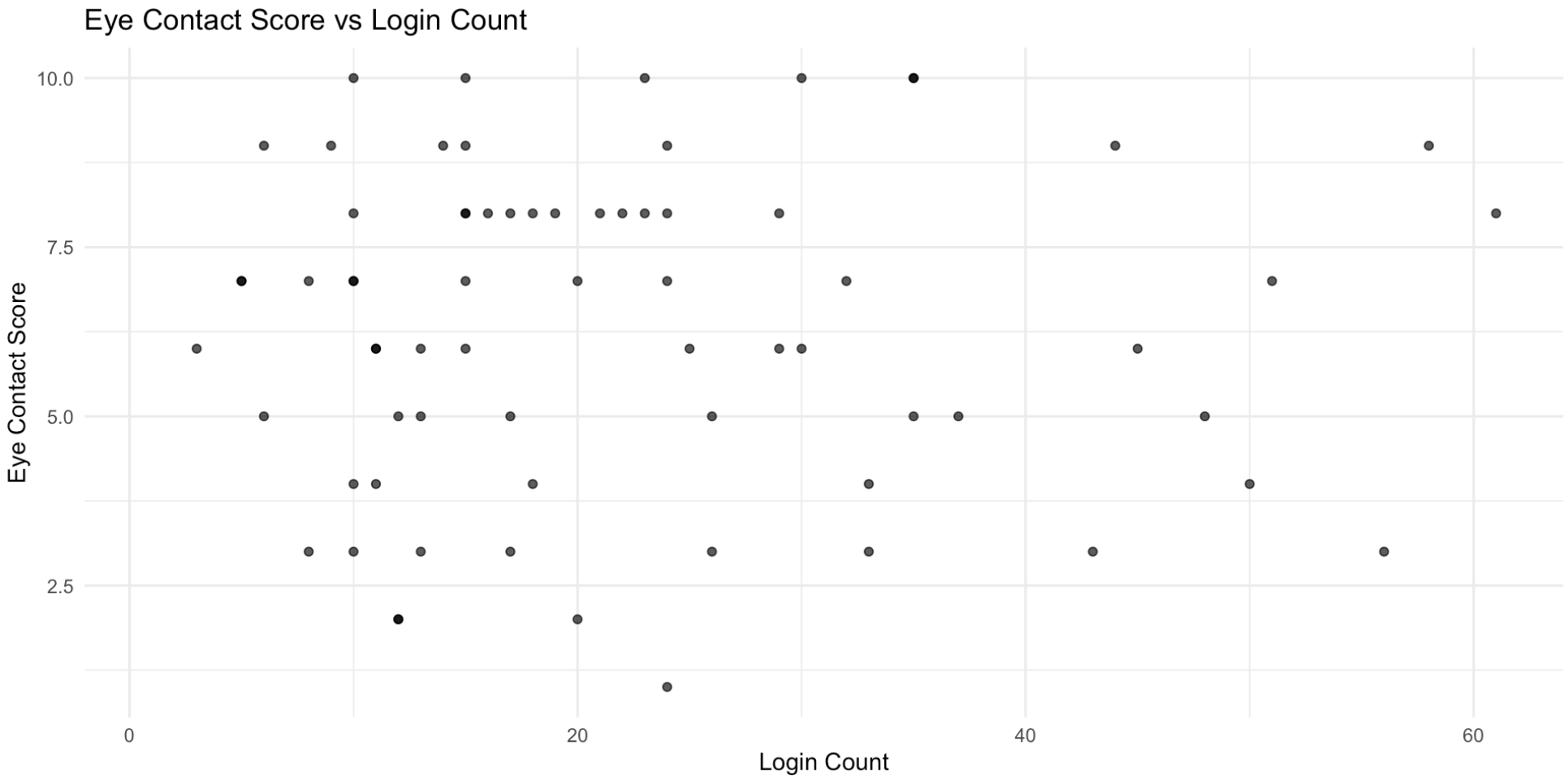
Histogram: Login Count Distribution

```
Interactive plot saved as 'login_count_distribution.html' in your working directory.
```

Bar Chart: Pet Preferences

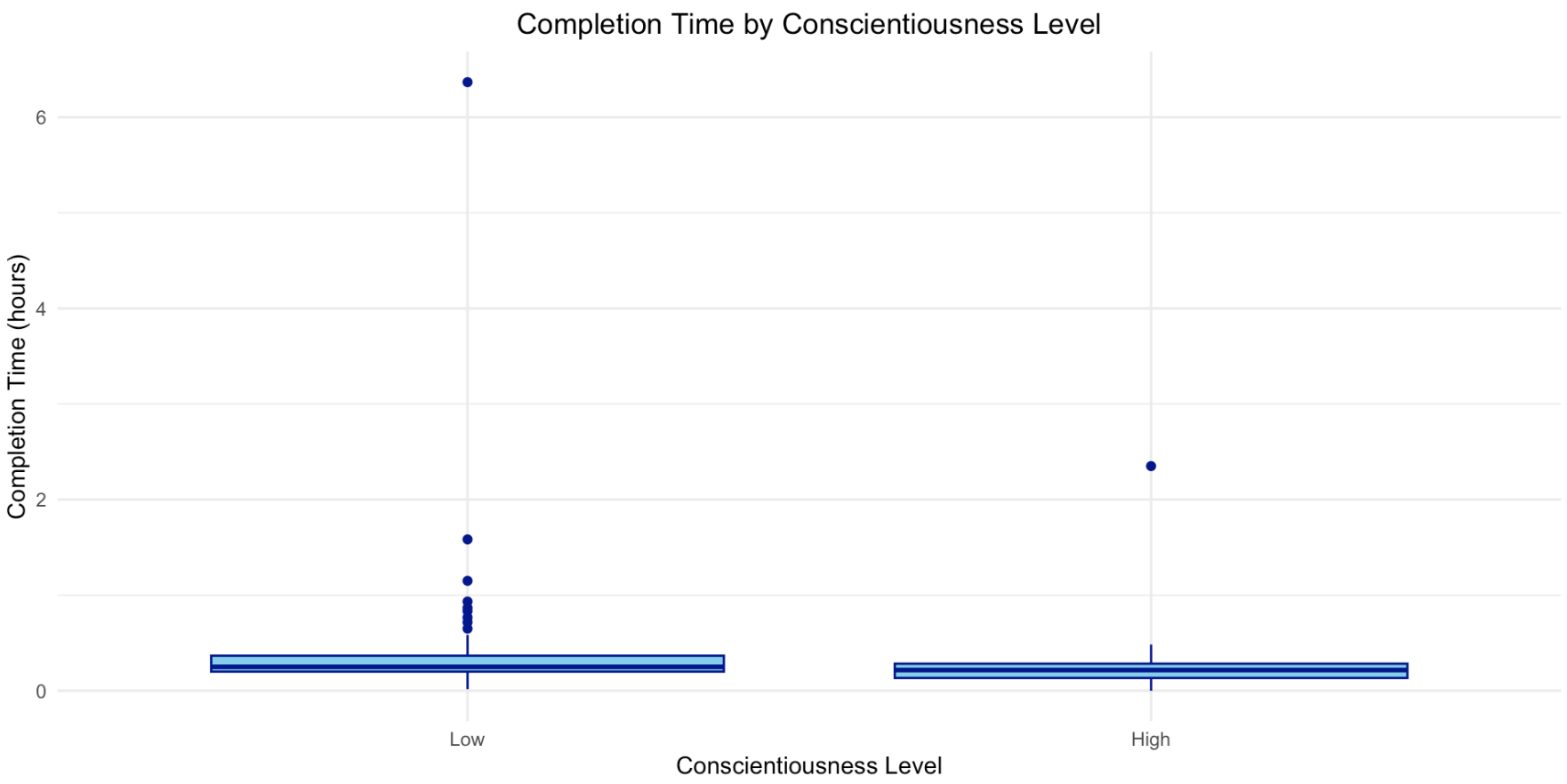


Scatter Plot: Eye Contact vs Login Count



Box Plot: Completion Time by Conscientiousness

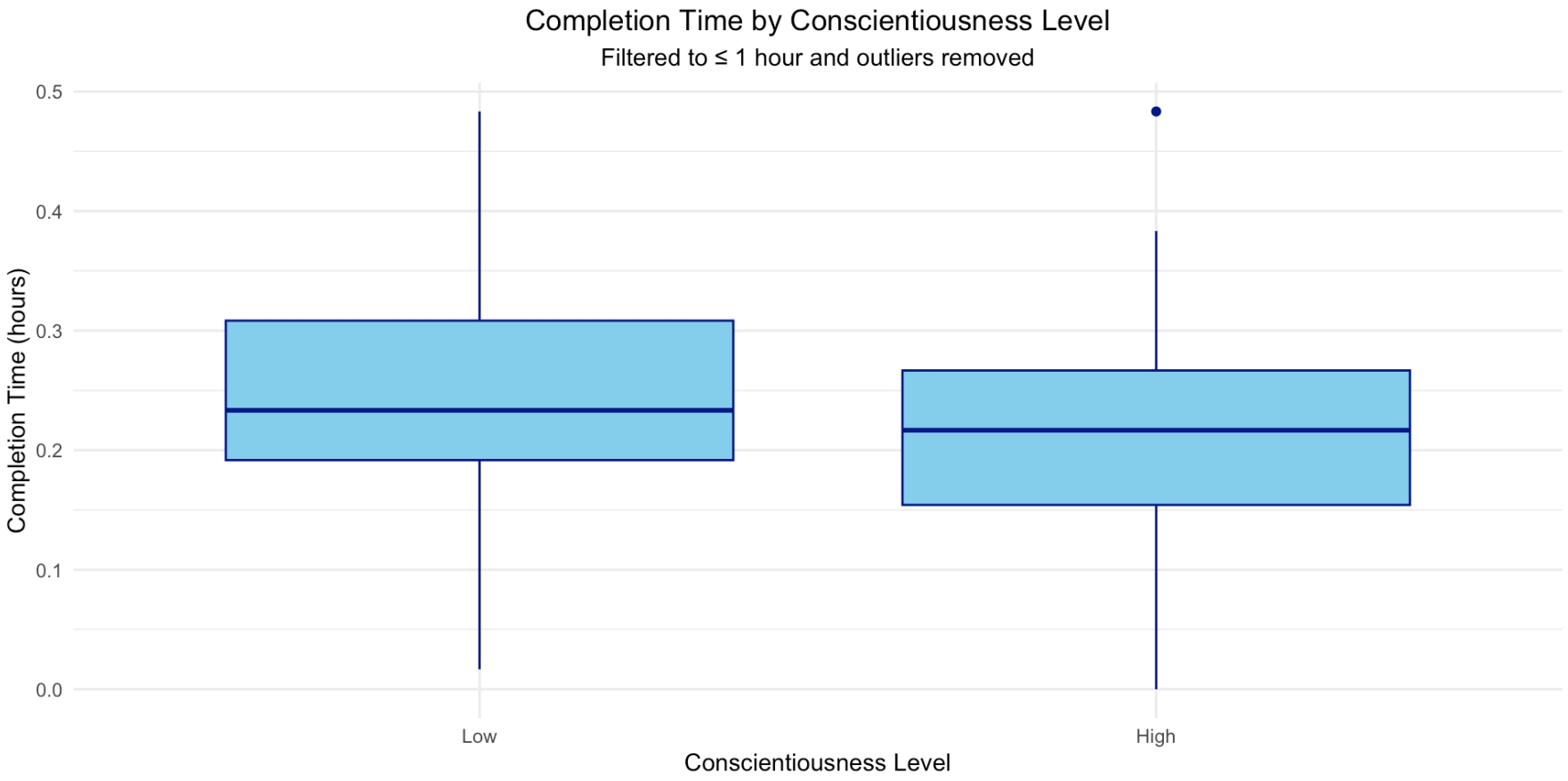
```
Low High
 61    9
```



Data Visualisation II

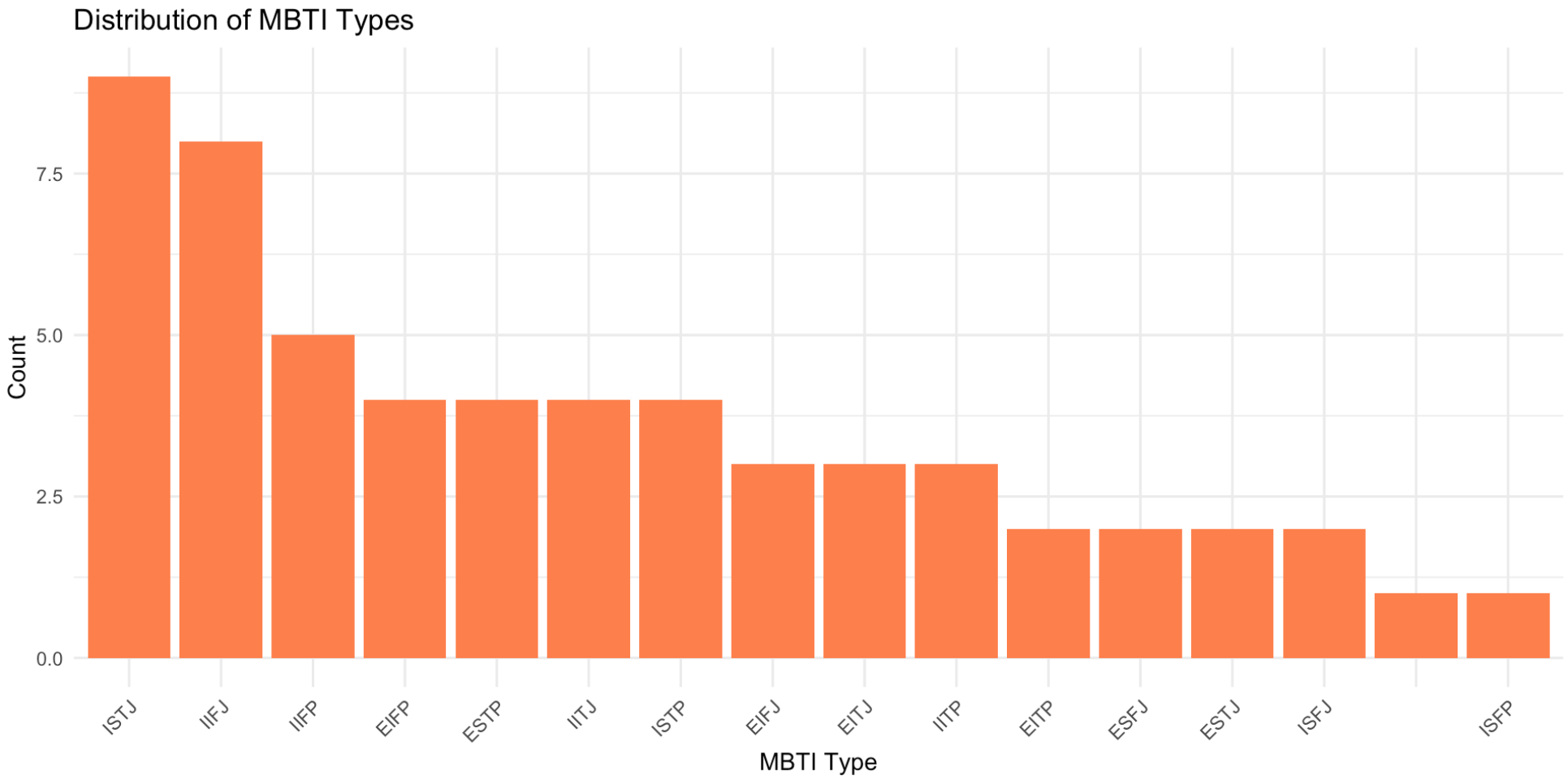
Outliers removed

```
Low High
 27   30
```



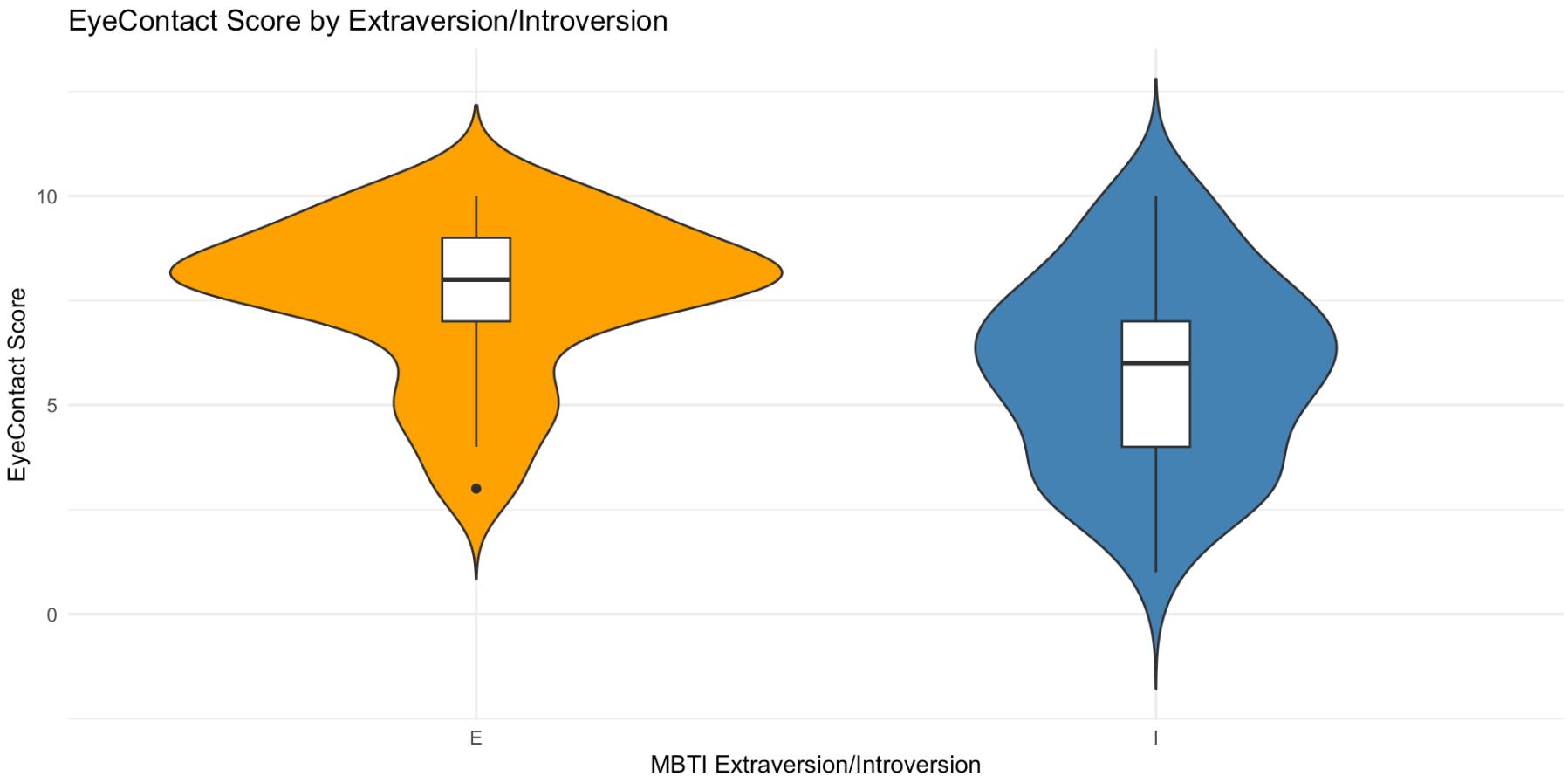
Data Visualisation II

MBTI Distribution



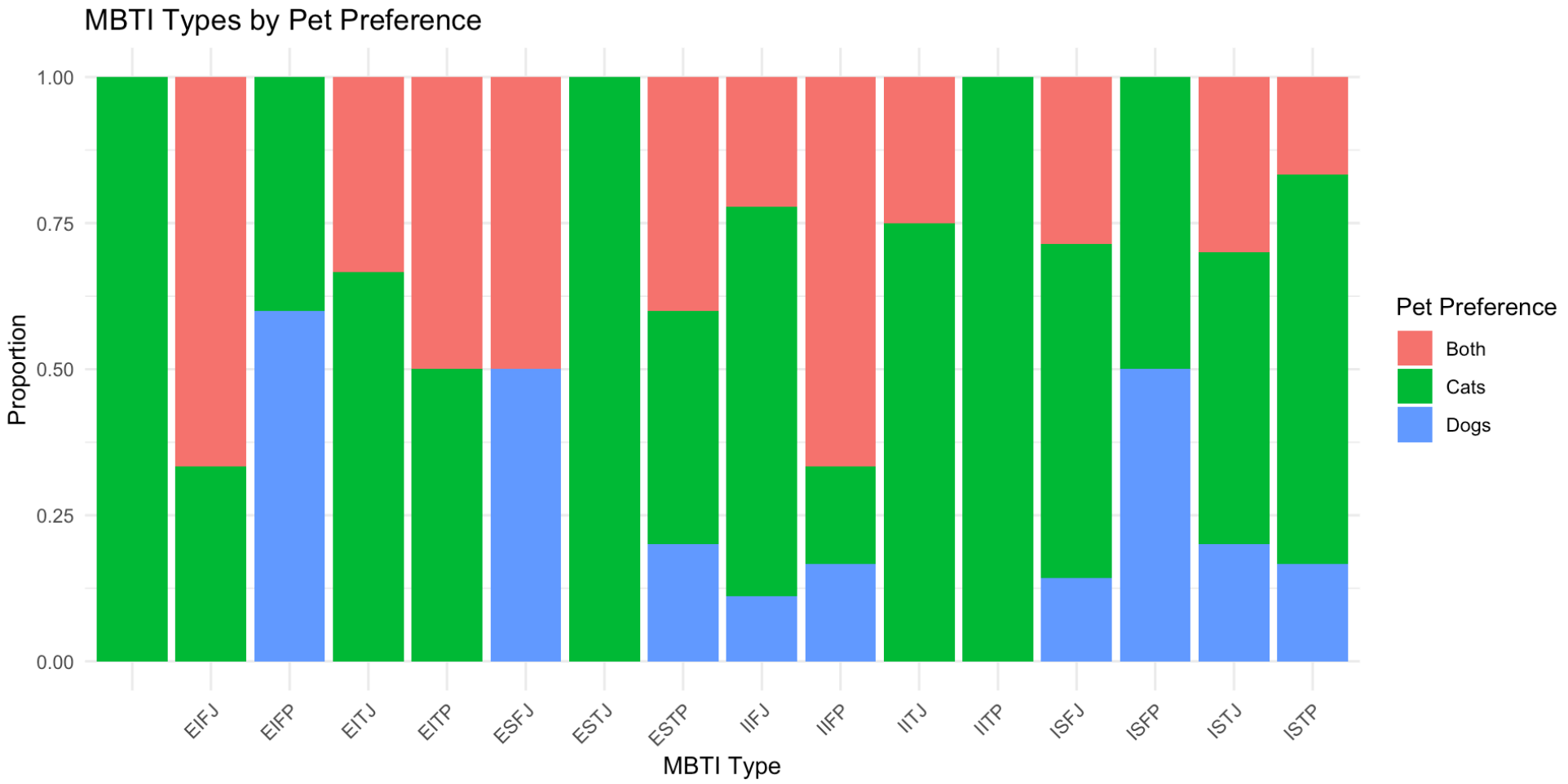
Data Visualisation II

Violin Plot: EyeContact by MBTI Extraversion/Introversion

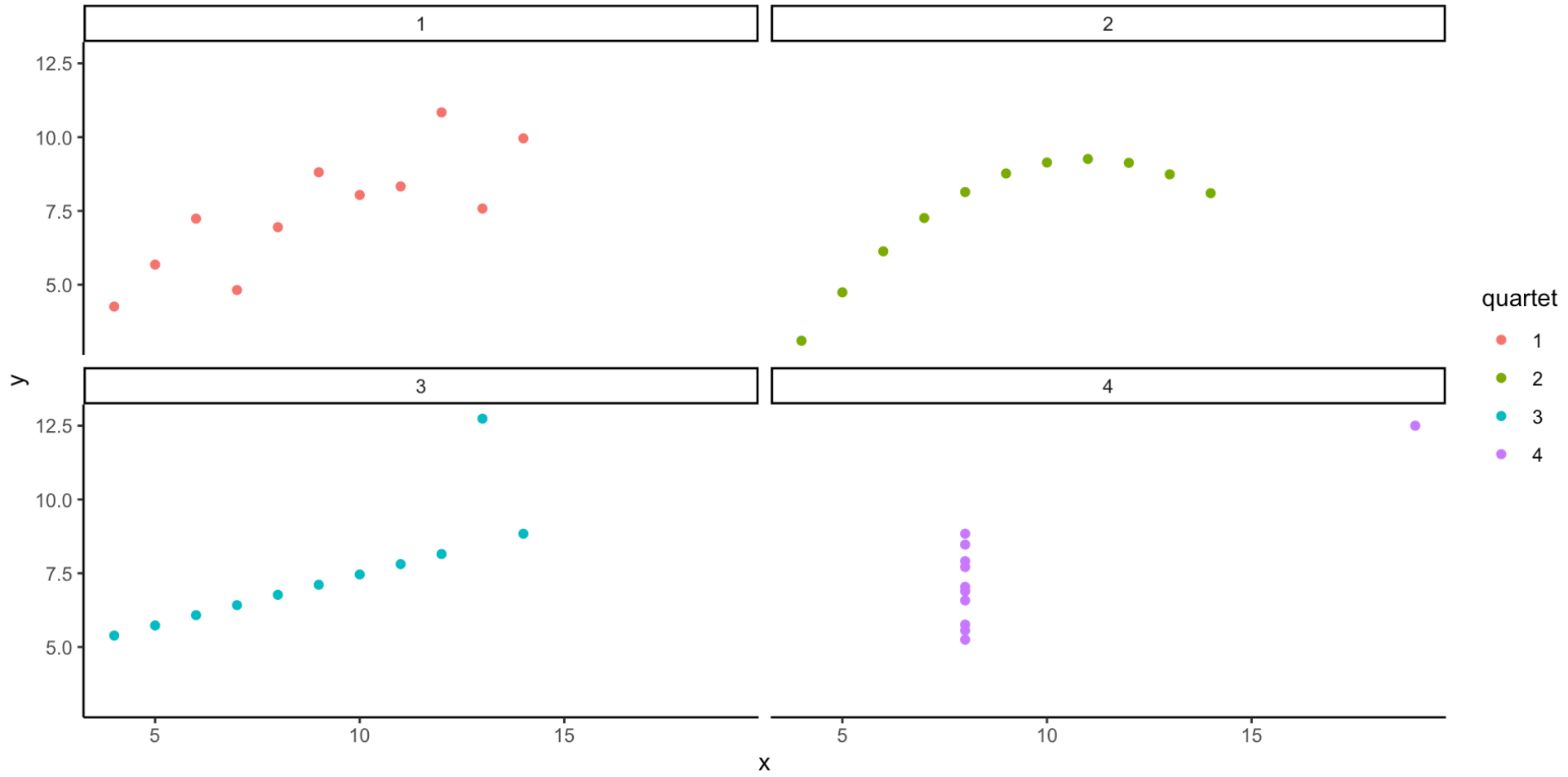


Data Visualisation II

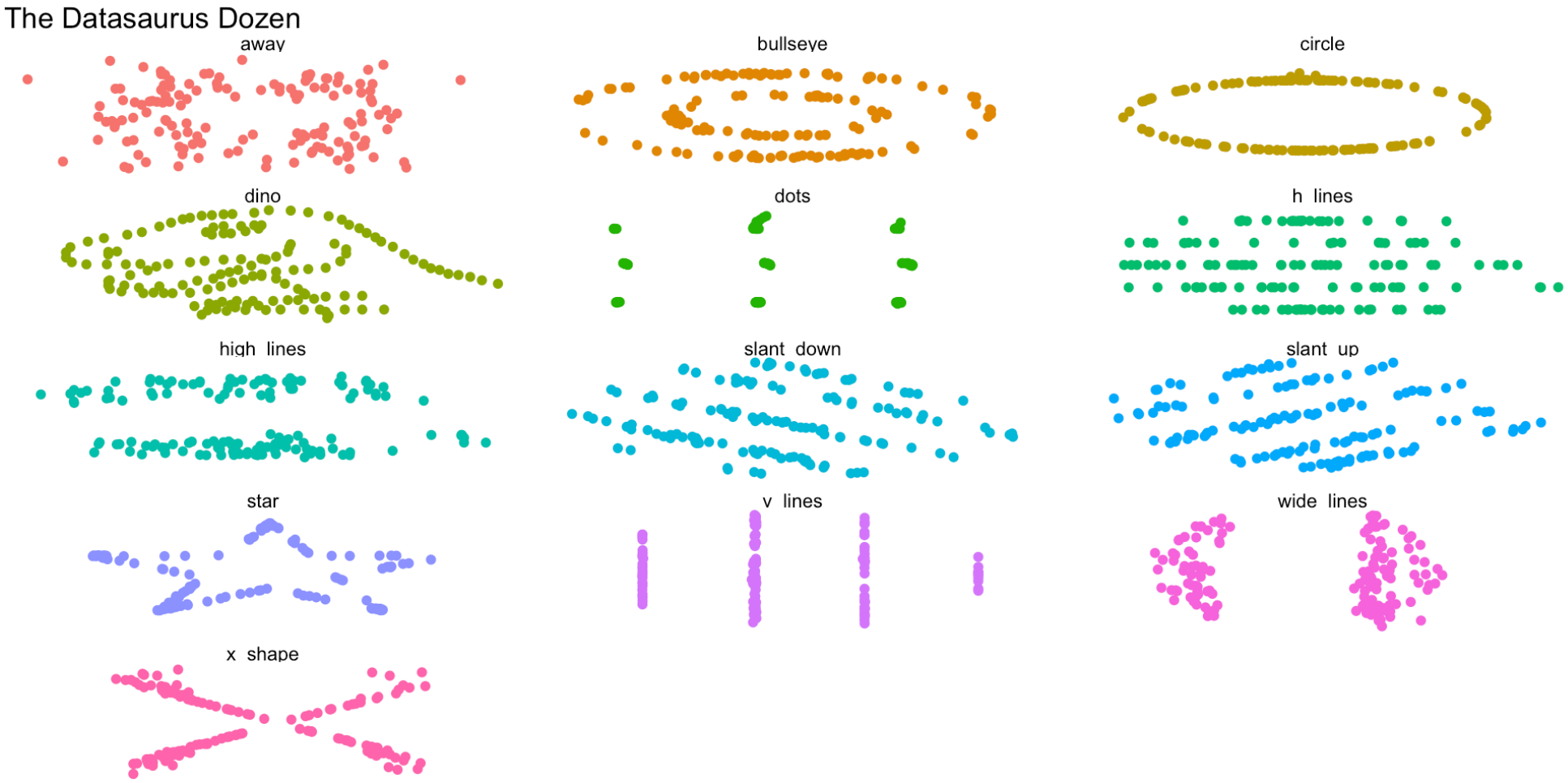
Stacked Bar Chart: MBTI Types by Pet Preference



Anscombe's Quartet



Datasuarus


The Datasaurus Dozen
away
bullseye
circle
dino
dots
h lines
high lines
slant down
slant up
star
v lines
wide lines
x shape

Summary DatasauRus Stats

```
# A tibble: 13 × 6
   dataset    mean_x mean_y  sd_x  sd_y correlation
   <chr>       <dbl>  <dbl> <dbl> <dbl>       <dbl>
 1 away         54.3   47.8  16.8  26.9     -0.0641
 2 bullseye     54.3   47.8  16.8  26.9     -0.0686
 3 circle       54.3   47.8  16.8  26.9     -0.0683
 4 dino         54.3   47.8  16.8  26.9     -0.0645
 5 dots         54.3   47.8  16.8  26.9     -0.0603
 6 h_lines      54.3   47.8  16.8  26.9     -0.0617
 7 high_lines   54.3   47.8  16.8  26.9     -0.0685
 8 slant_down   54.3   47.8  16.8  26.9     -0.0690
 9 slant_up     54.3   47.8  16.8  26.9     -0.0686
10 star         54.3   47.8  16.8  26.9     -0.0630
11 v_lines      54.3   47.8  16.8  26.9     -0.0694
12 wide_lines   54.3   47.8  16.8  26.9     -0.0666
13 x_shape      54.3   47.8  16.8  26.9     -0.0656
```

Plato’s Triad: The True, The Good, and The Beautiful

Plato believed in the intrinsic connection between **truth**, **goodness**, and **beauty** — a concept that has influenced Western thought for centuries.

For Plato, these three qualities are inseparable in the realm of **Forms**.

The **truth** is inherently **good**, and what is **good** is also inherently **beautiful**.

Thus, if something is **false**, it cannot be truly **beautiful** or **good**.

R Big Picture

Quarto Publishing System

What do you currently use?

How do you write your essays or lab reports?

- Microsoft Word?
- Google Docs?
- Markdown?

What do you currently use?

How do you write your essays or lab reports?

- Microsoft Word?
- Google Docs?
- Markdown?

How do you currently play with numbers?

- Excel?
- SPSS?
- R?
- Python?

What is Quarto?

Quarto is an open-source scientific and technical publishing system

What is Quarto?

Quarto is an open-source scientific and technical publishing system that allows you to combine text, images, code, plots, and tables in a fully-reproducible document.

What is Quarto?

Quarto is an open-source scientific and technical publishing system that allows you to combine text, images, code, plots, and tables in a fully-reproducible document.

Quarto has support for multiple languages including R, Python, Julia, and Observable.

What is Quarto?

Quarto is an open-source scientific and technical publishing system that allows you to combine text, images, code, plots, and tables in a fully-reproducible document.

Quarto has support for multiple languages including R, Python, Julia, and Observable.

It also works for a range of output formats such as PDFs, HTML documents, websites, presentations,...

Why use Quarto? Why use R?

- More journals require code to be submitted (for transparency and reproducibility). Keeping the code with the paper makes this easier.
- Copying and pasting is tedious (and a great source of accidental errors).
- If you fix an error in code or data, the results and figures in the paper update automatically.
- Easy to share publicly.
- Open source so anyone can use it.

What about R Markdown?

R Markdown isn’t going anywhere but...

- Quarto has better multi-language support
- More user-friendly
- Better control of the output layouts

Rendering a document

Within RStudio IDE: click **Render** (or Ctrl+Shift+K)

Rendering a document

Within RStudio IDE: click **Render** (or Ctrl+Shift+K)

Content

- Text, links, images
- Code, tables, plots
- Equations, references

Output types

Output types

- Documents: HTML, PDF, MS Word, Markdown

Output types

- Documents: HTML, PDF, MS Word, Markdown
- Presentations: Revealjs, PowerPoint, Beamer

Output types

- Documents: HTML, PDF, MS Word, Markdown
- Presentations: Revealjs, PowerPoint, Beamer
- Websites

Output types

- Documents: HTML, PDF, MS Word, Markdown
- Presentations: Revealjs, PowerPoint, Beamer
- Websites
- Books

Output types

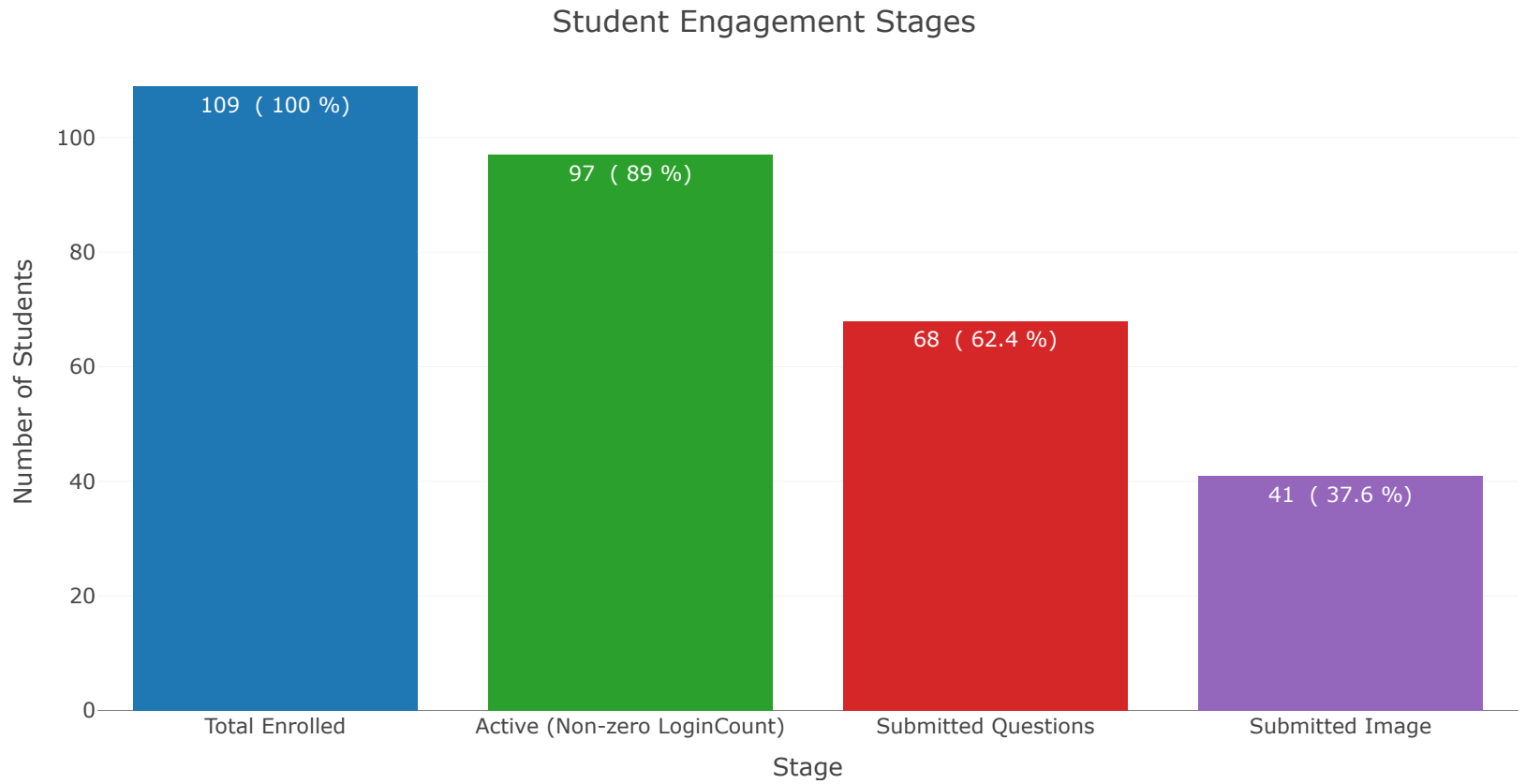
- Documents: HTML, PDF, MS Word, Markdown
- Presentations: Revealjs, PowerPoint, Beamer
- Websites
- Books
- ...

Introduction

Student Engagement Summary

Stage	Count	Percentage
Total Enrolled	109	100.00
Active (Non-zero LoginCount)	97	88.99
Submitted Questions	68	62.39
Submitted Image	41	37.61

Visual


Student Engagement Stages
109 (100 %)
97 (89 %)
68 (62.4 %)
41 (37.6 %)
Number of Students
0
20
40
60
80
100
Total Enrolled
Active (Non-zero LoginCount)
Submitted Questions
Submitted Image
Stage

Visualisation of Your Data

This analysis covers levels of measurement, measures of central tendency, variance, and various plot types using the student data collected.

1. Levels of Measurement

1. **Nominal**:
    - MBTI (Myers-Briggs Type Indicator)
    - Coin (Heads or Tails)
    - DogCatBoth (Preference for pets)
    - InsectApocalypse preference
2. **Ordinal**:
    - EyeContact (Likert or scale responses)
    - Ten Item Personality Inventory scores (OCEAN)
3. **Interval/Ratio**:
    - Count of VLE Logins so far [LoginCount]
    - Survey Completion Time [CompTime - minutes]

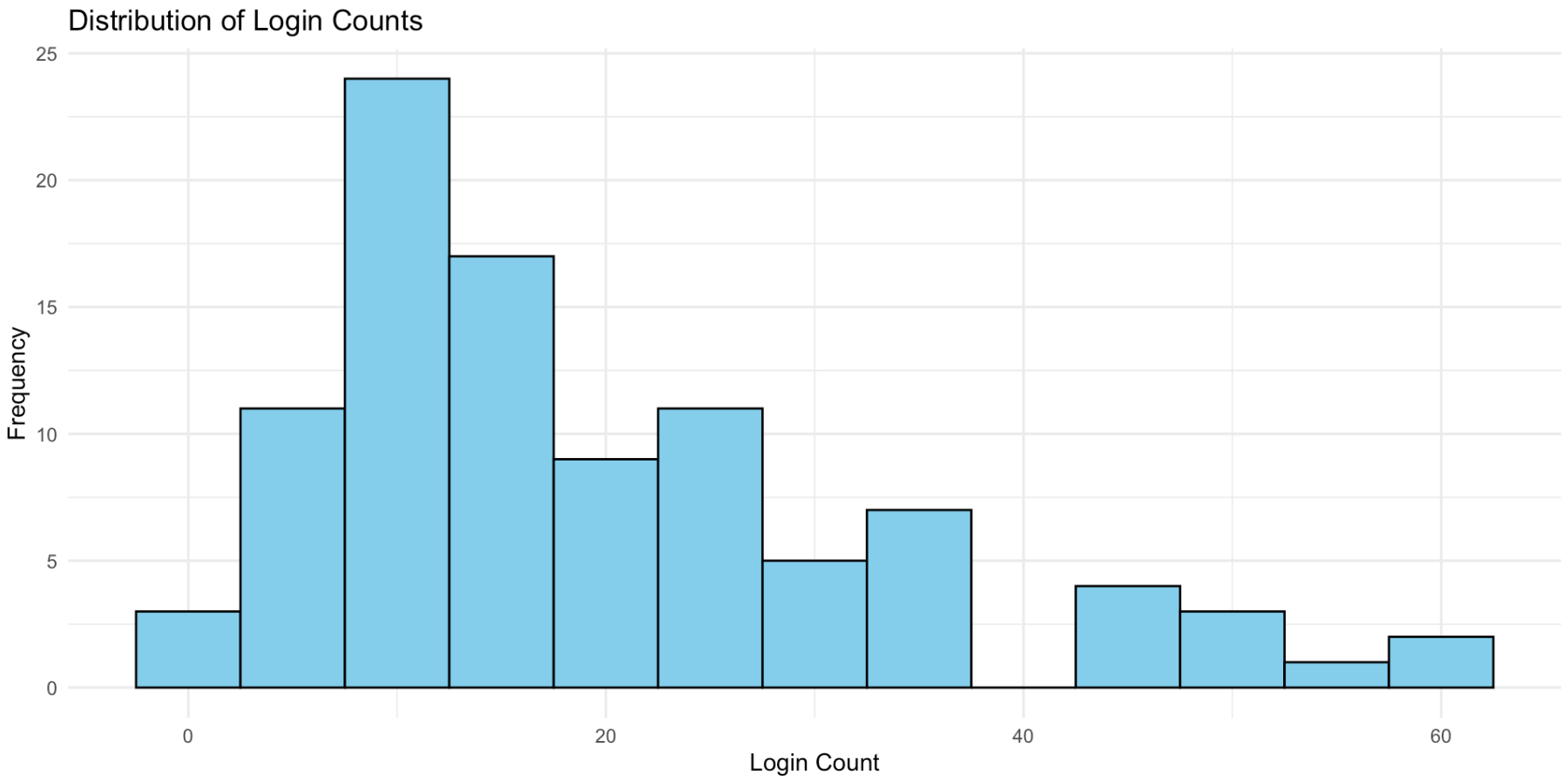
2. Central Tendency and Variance

Let’s calculate these for LoginCount and the Extraversion (TIPI) score:

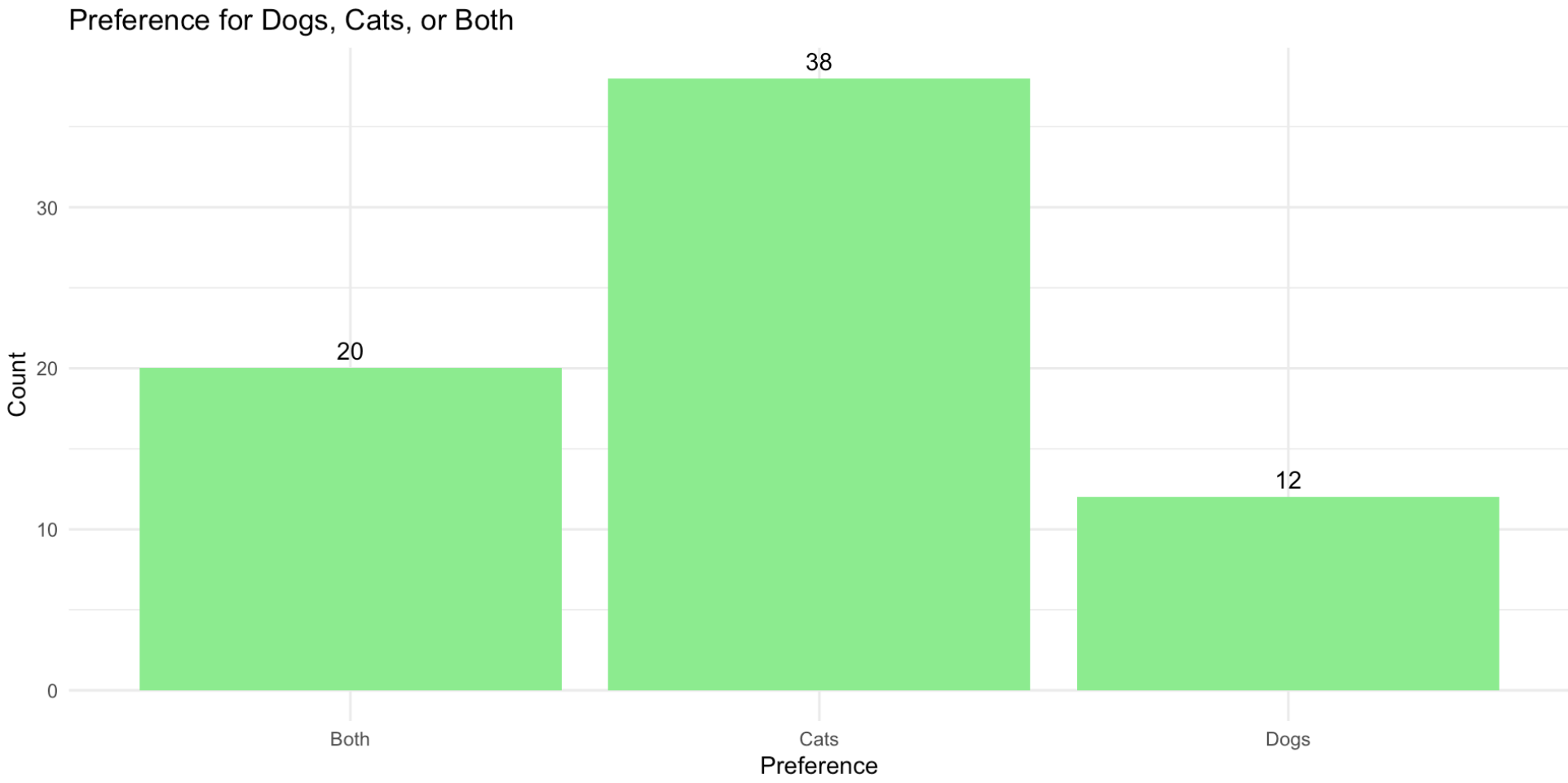
Summary Statistics	
LoginCount_Mean	22.29
LoginCount_Median	18.00
LoginCount_SD	13.93
LoginCount_Variance	193.95
E_TIPI_Mean	3.18
E_TIPI_Median	3.00
E_TIPI_SD	0.49
E_TIPI_Variance	0.24

Various Plot Types

1. Histogram: LoginCount

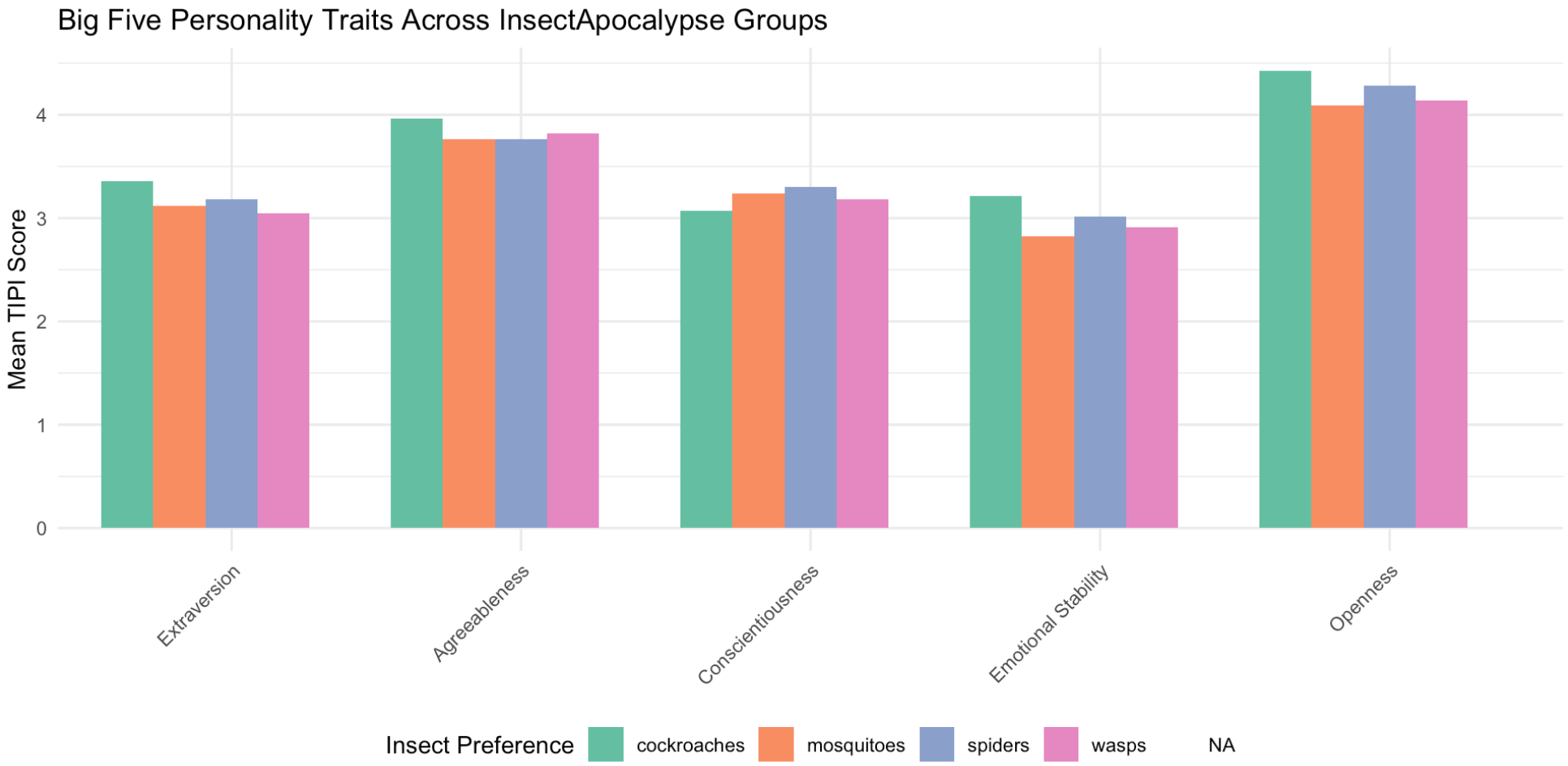


2. Bar Chart: DogCatBoth Preferences

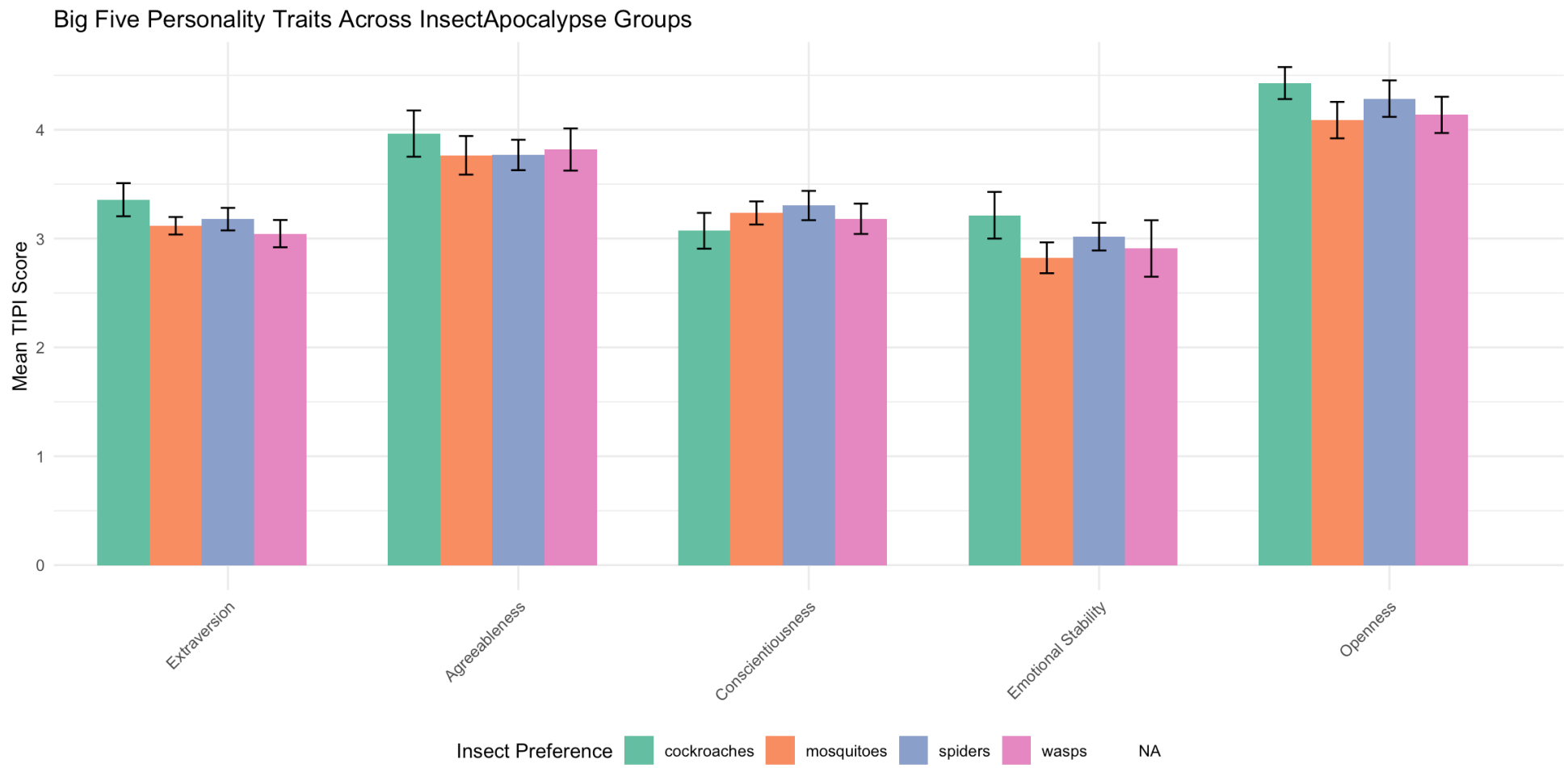


Data Visualisation II

3a. Insect Apocalypse

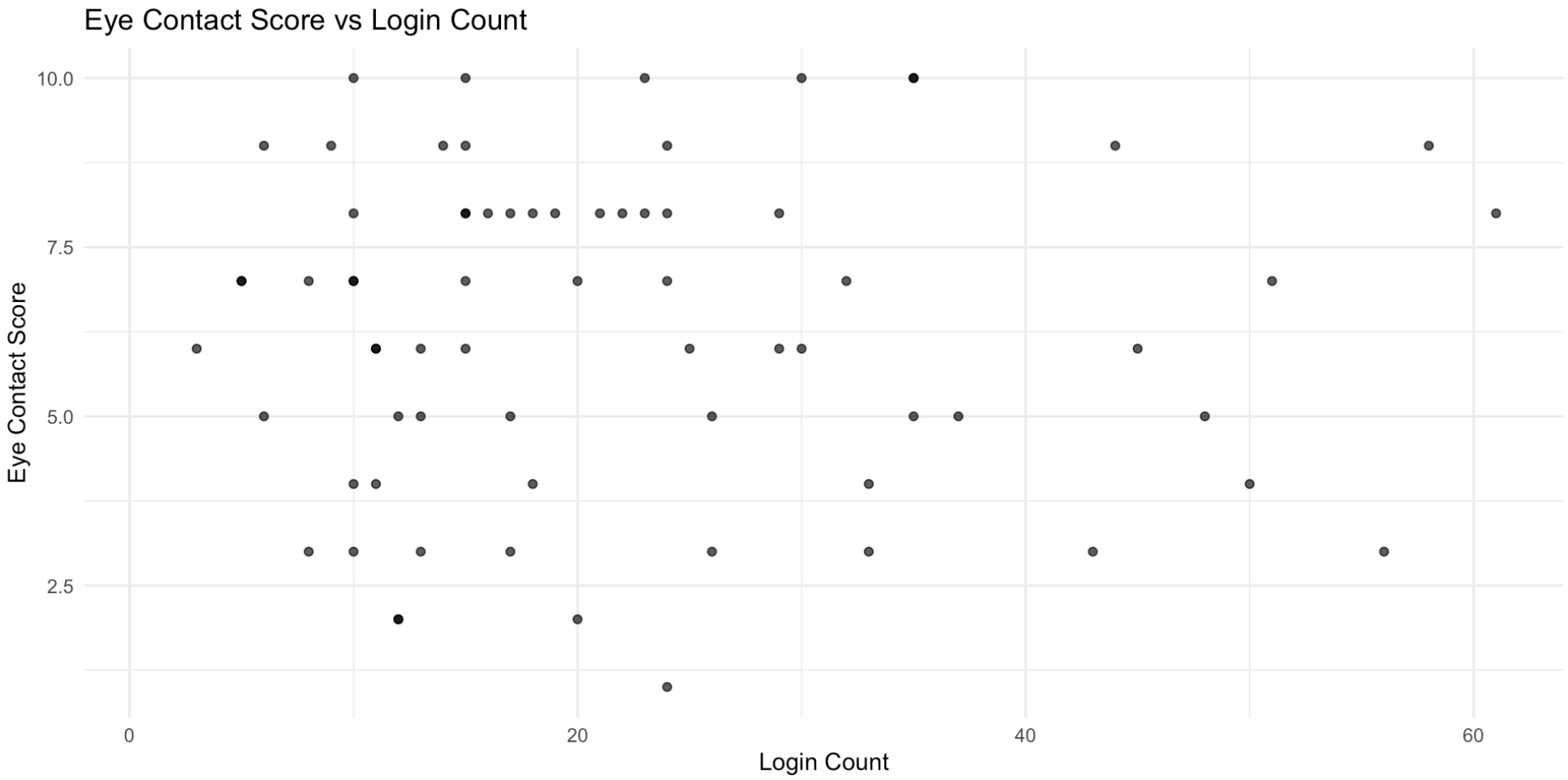


3b. Insect Apocalypse



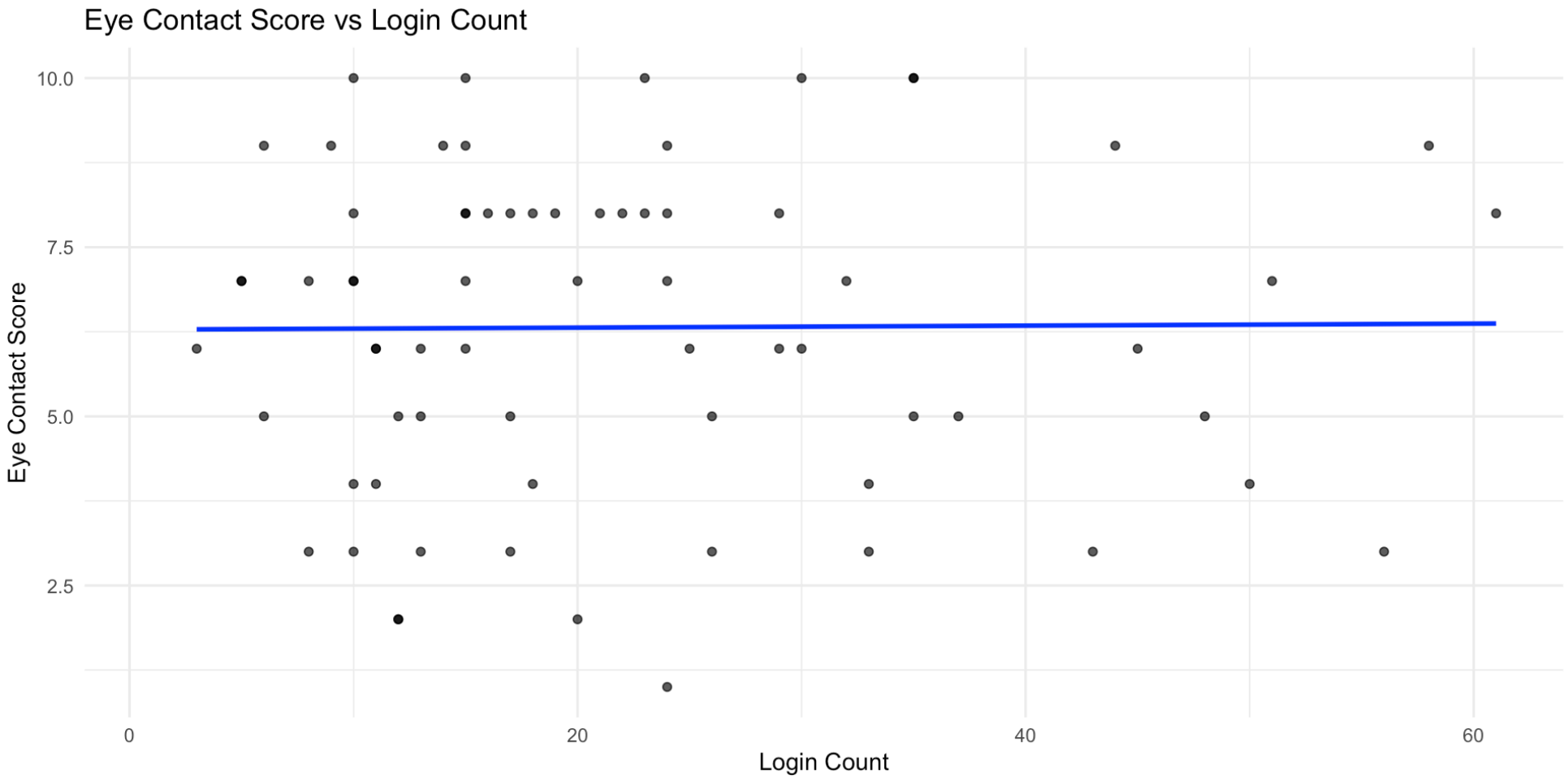
Data Visualisation II

4a. Scatter Plot: EyeContact vs LoginCount

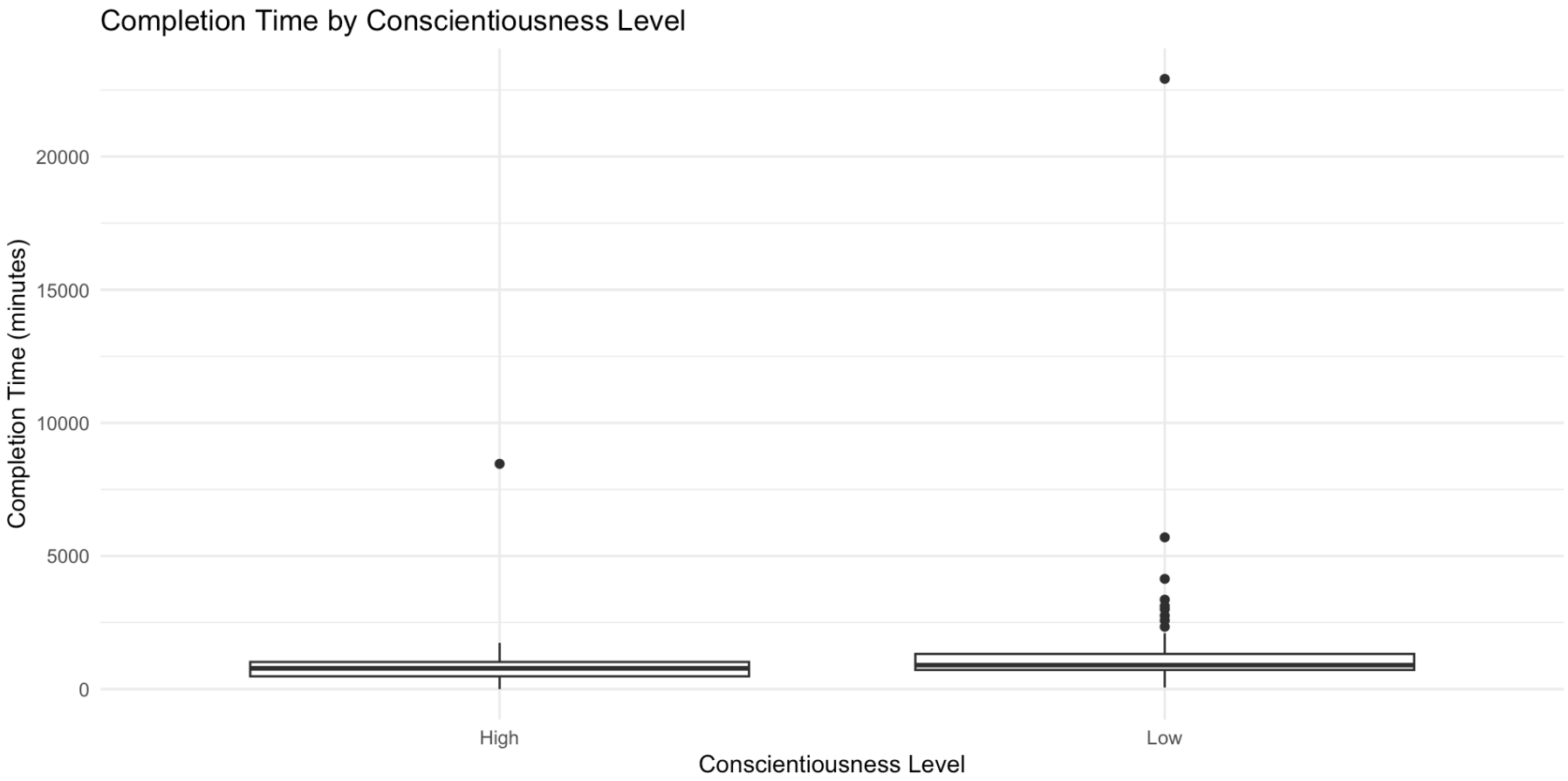


Data Visualisation II

4b. Scatter Plot: EyeContact vs LoginCount

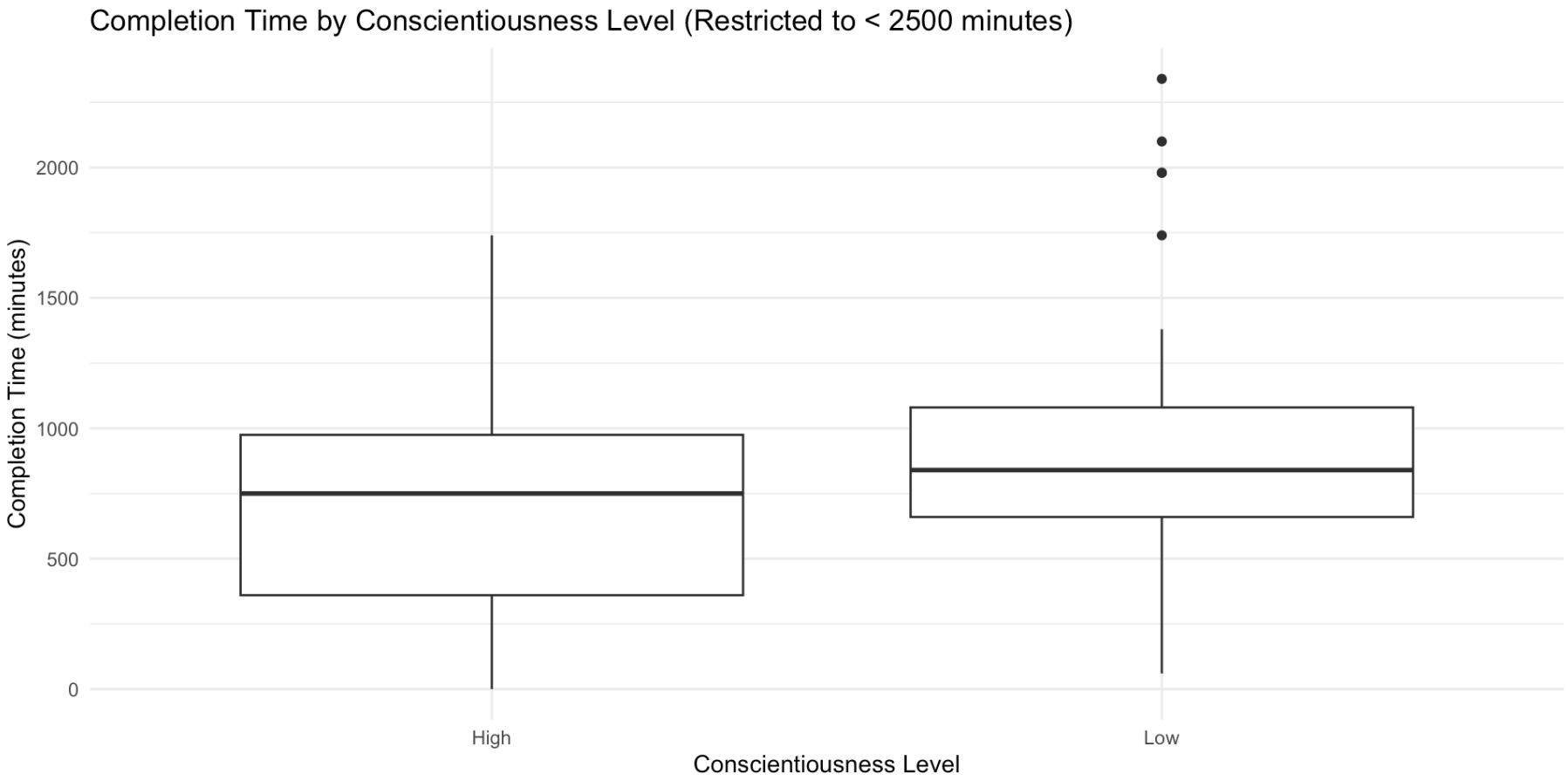


5a. Box Plot: CompTime by Conscientiousness Level



Data Visualisation II

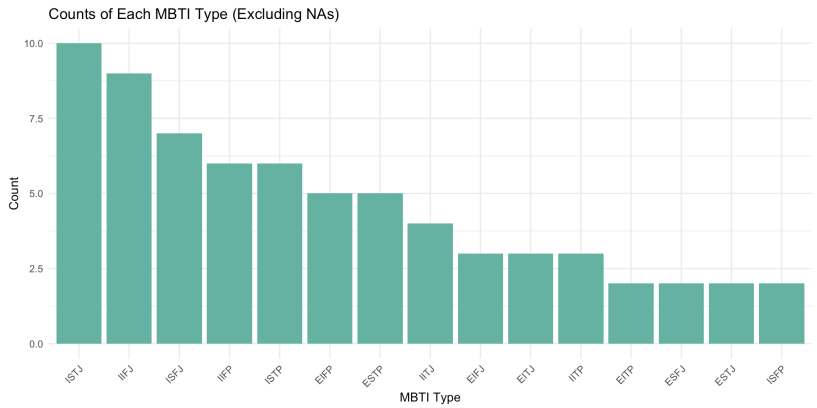
5b. Box Plot: CompTime by Conscientiousness Level



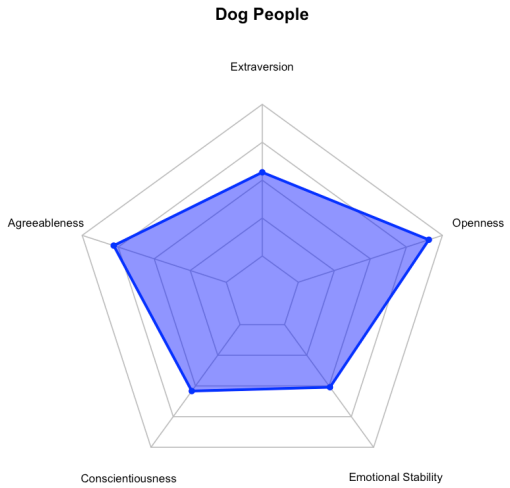
Data Visualisation II

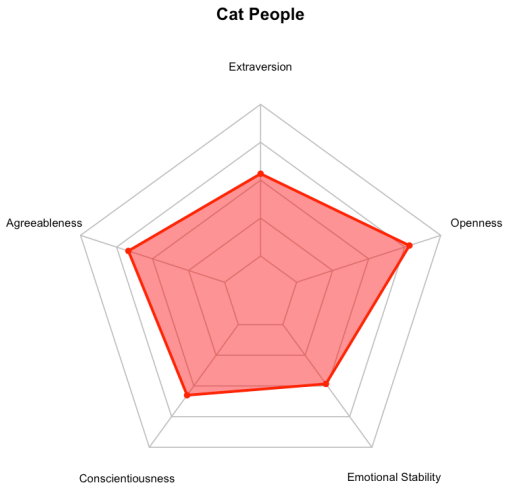
6. MBTI Count Matrix

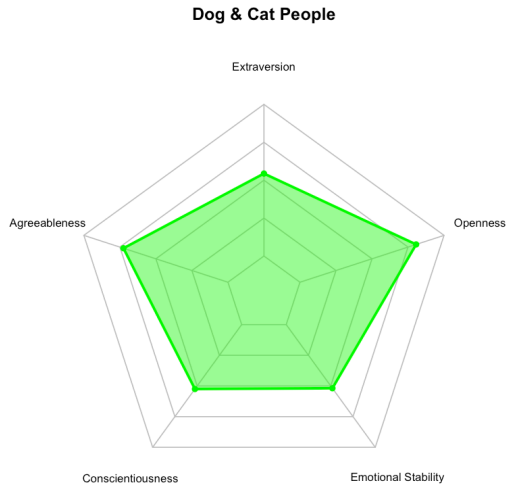
```
# A tibble: 15 × 2
   MBTI  Count
   <chr> <int>
 1 ISTJ     10
 2 IIFJ      9
 3 ISFJ      7
 4 IIFP      6
 5 ISTP      6
 6 EIFP      5
 7 ESTP      5
 8 IITJ      4
 9 EIFJ      3
10 EITJ      3
11 IITP      3
12 EITP      2
13 ESFJ      2
14 ESTJ      2
15 ISFP      2
```



7. TIPI Radar Charts by Pet Preference

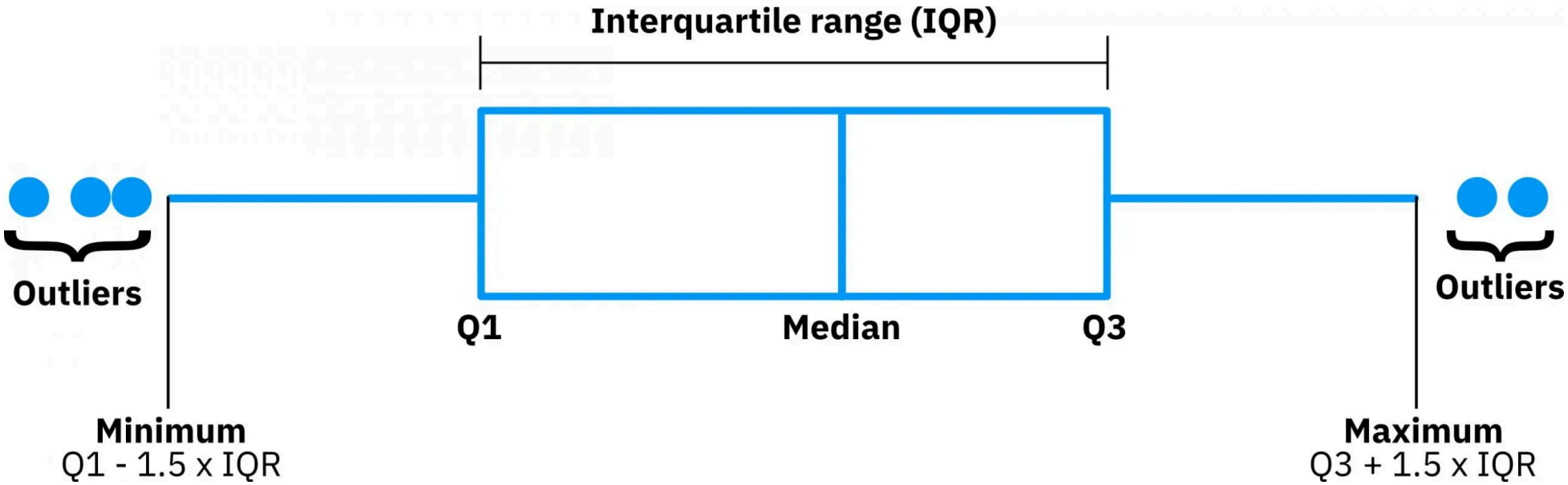




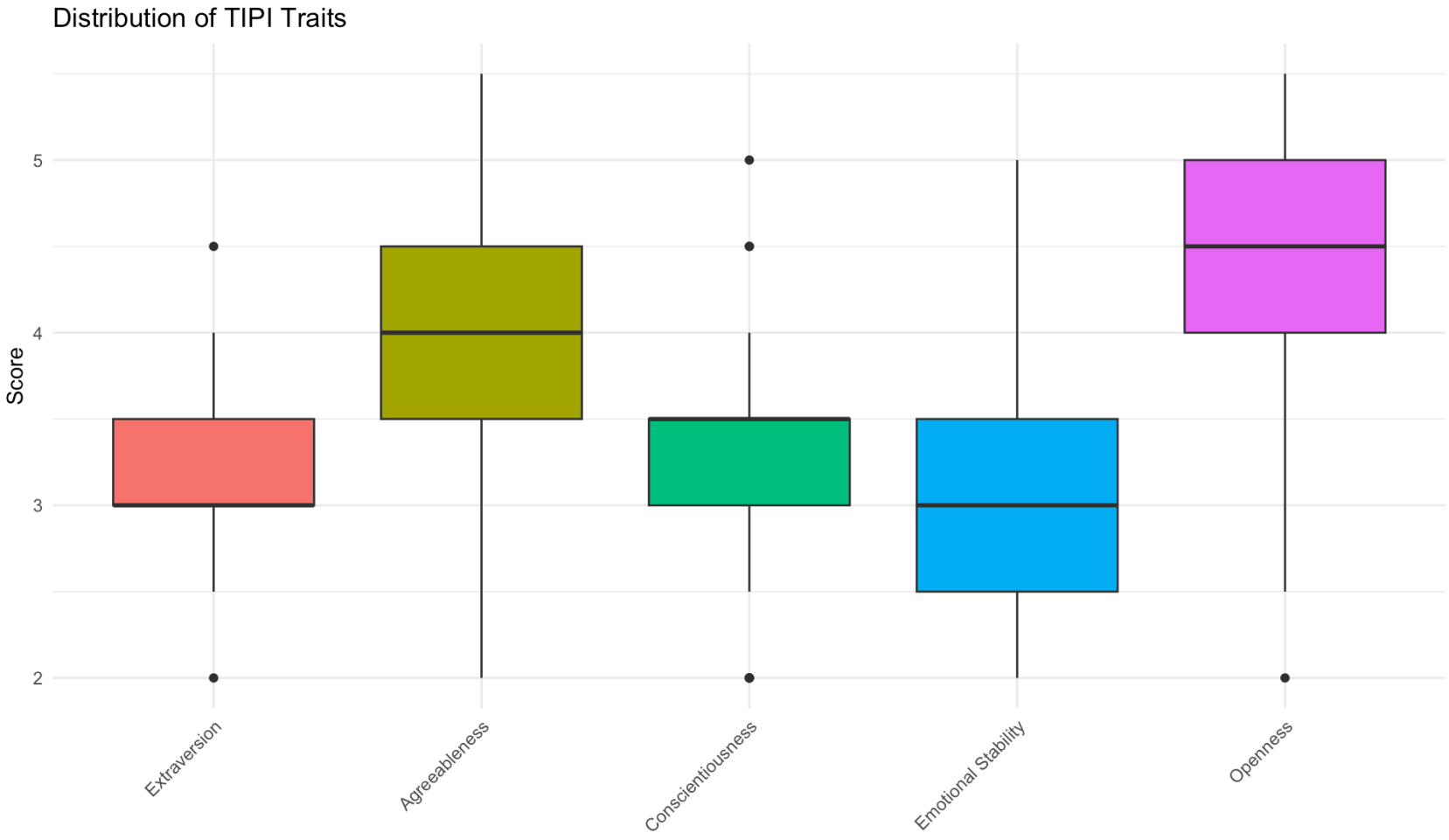


8. Heatmaps for Seating Preferences and TIPI Traits

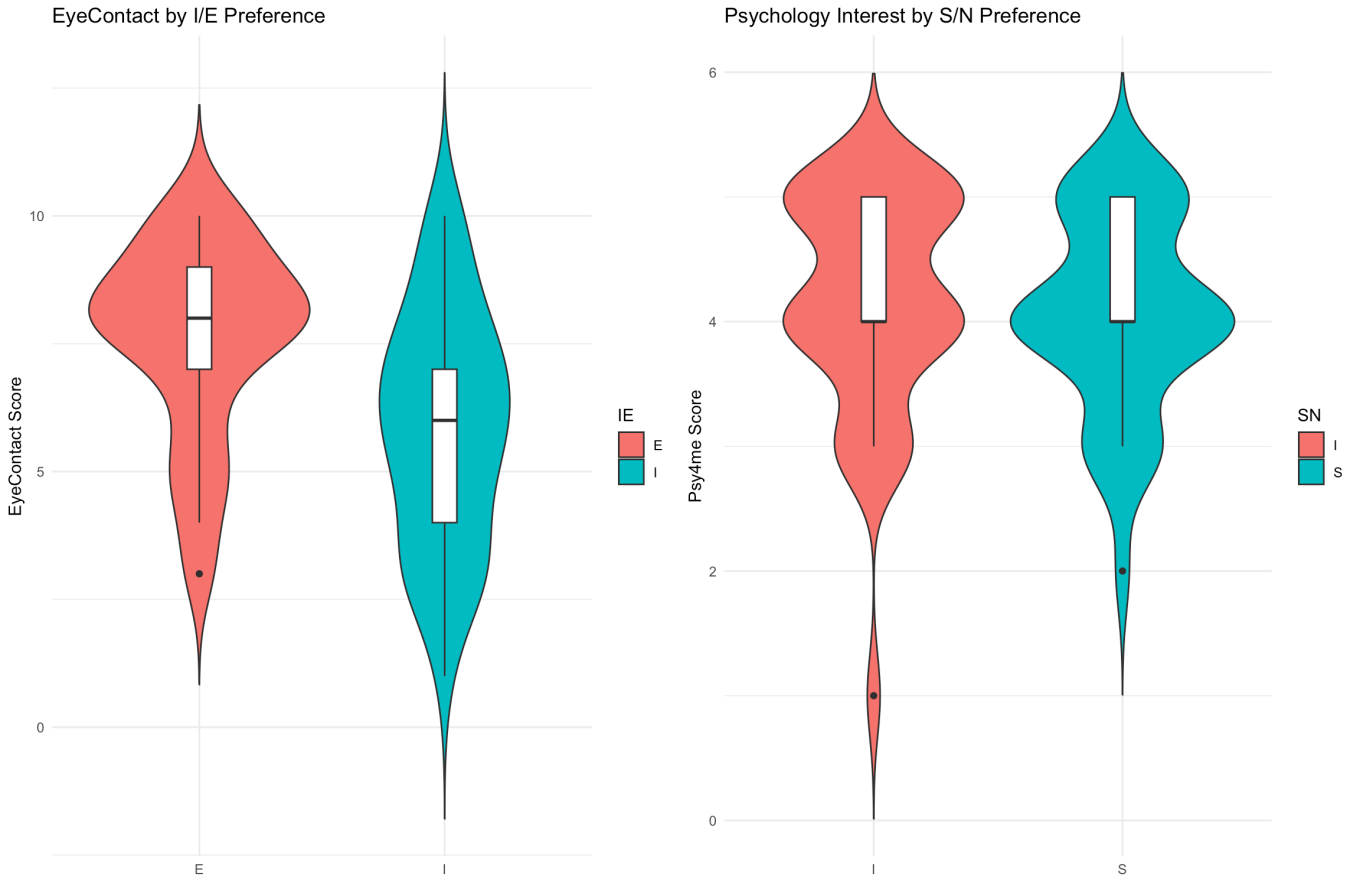
note: BoxPlot Basics



9. Compact Boxplots: TIPI Traits

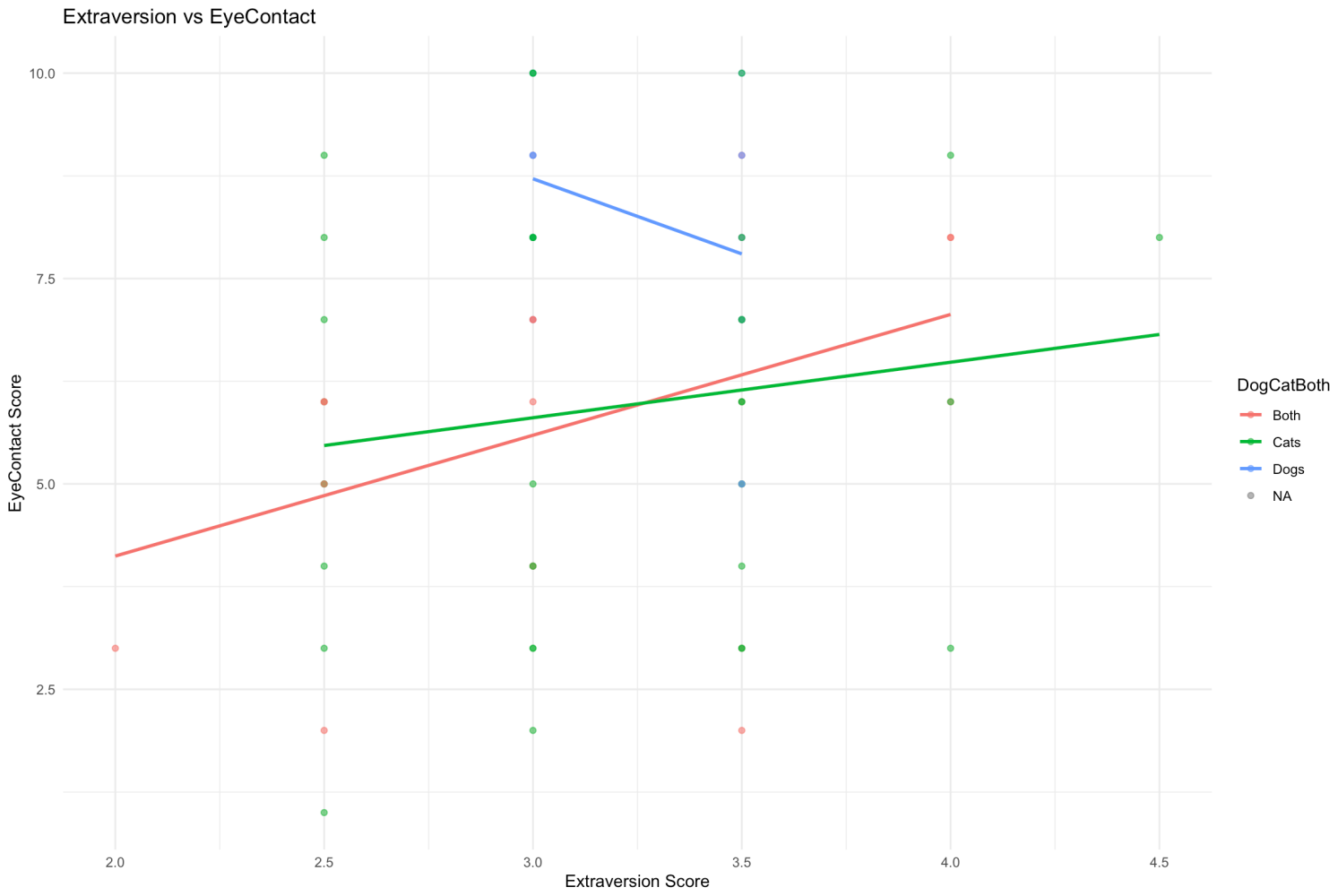


10. Violin Plots: EyeContact and Psy4me by MBTI Dimension



Data Visualisation II

11. Scatterplots: Exploring Relationships



Data Visualisation II