# Project Proposal

**Due November 17 at 11:59pm**

Arko Bhattacharya, Eric Ortega Rodriguez, Mu Niu, Nruta Choudhari

### Load Packages

```
library(tidyverse)
library(ggplot2)
library(dplyr)
```

## Dataset 1 (top choice)

**Data source:** **UFC Fight Stats**

**Brief description:** This data set includes information on fighters, their fights, and key statistics like fight outcomes, fighter attributes (e.g., height, weight), and fight-specific statistics (e.g., strikes landed, submission attempts). Each row represents a single fight, providing detailed insights into each bout's characteristics and outcomes.

**Load the data and provide a `glimpse()`:**

```
ufc = read_csv('ufc-master.csv', show_col_types = FALSE)

# glimpse on variables of interest
glimpse(ufc %>% select(BlueReachCms, RedReachCms, BlueAvgSigStrLanded,
                       RedAvgSigStrLanded, WeightClass, Winner,
                       RedAvgSubAtt, BlueAvgSubAtt,TotalFightTimeSecs))
```

```
Rows: 6,478
Columns: 9
$ BlueReachCms        <dbl> 172.72, 165.10, 205.74, 172.72, 190.50, 187.96, 16~
$ RedReachCms         <dbl> 177.80, 167.64, 198.12, 170.18, 187.96, 185.42, 17~
$ BlueAvgSigStrLanded <dbl> 2.72, 3.71, 3.16, 3.70, 3.47, 3.17, 5.38, 7.66, 2.~
```

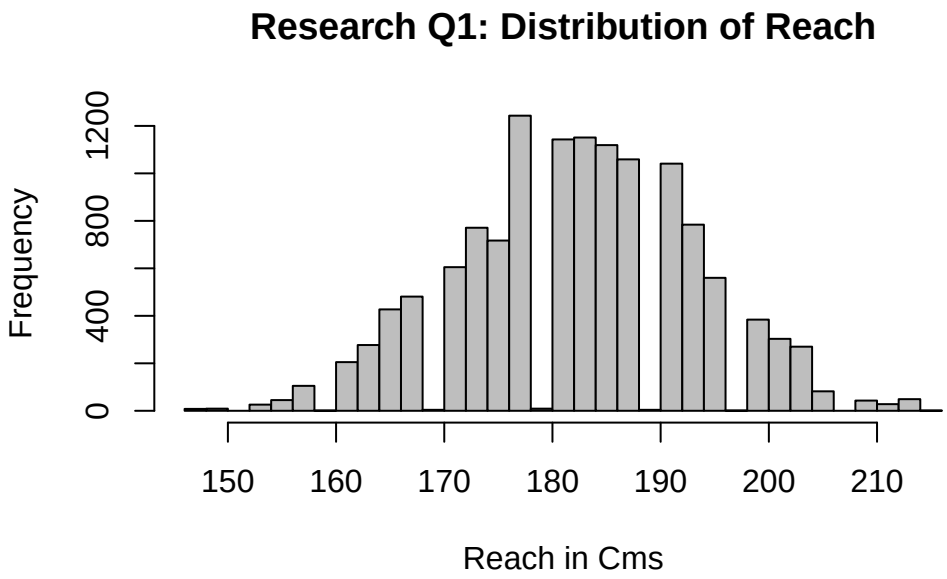
```
$ RedAvgSigStrLanded  <dbl> 3.99, 5.24, 5.82, 4.04, 5.98, 3.85, 4.06, 5.43, 3.~
$ WeightClass         <chr> "Flyweight", "Women's Flyweight", "Light Heavyweig~
$ Winner              <chr> "Red", "Red", "Blue", "Blue", "Blue", "Red", "Red"~
$ RedAvgSubAtt        <dbl> 0.4, 0.8, 0.0, 0.3, 0.1, 0.8, 0.0, 0.6, 0.6, 0.9, ~
$ BlueAvgSubAtt       <dbl> 0.5, 0.4, 0.4, 0.6, 1.1, 0.3, 0.5, 0.4, 1.5, 0.5, ~
$ TotalFightTimeSecs  <dbl> 1500, 1500, 900, 748, 268, 900, 900, 523, 359, 900~
```

**Research question 1:** How does the reach of the fighter relate to the total number of strikes landed during a fight?

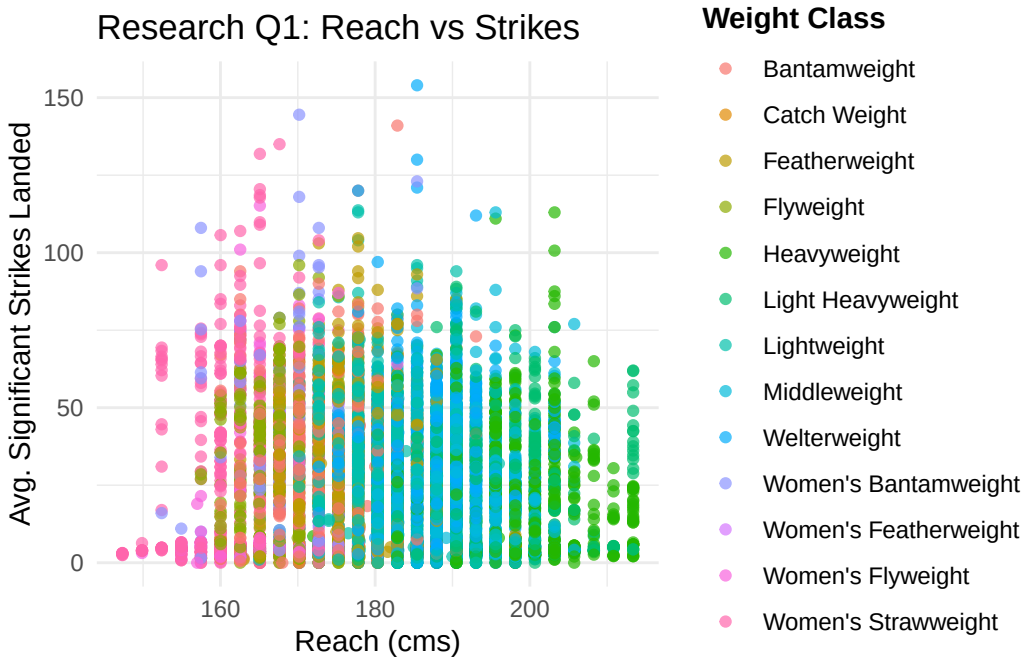
- *Outcome Variable (include the name/description and type of variable):* Reach (continuous)
- *Independent Variable:* Strikes landed (continuous variable)
- *Interaction Term:* Weight class, allowing analysis of the interaction between reach and weight class on striking effectiveness.

**Exploratory Plots:**

1. Outcome Variables




2. Relationship of Interest

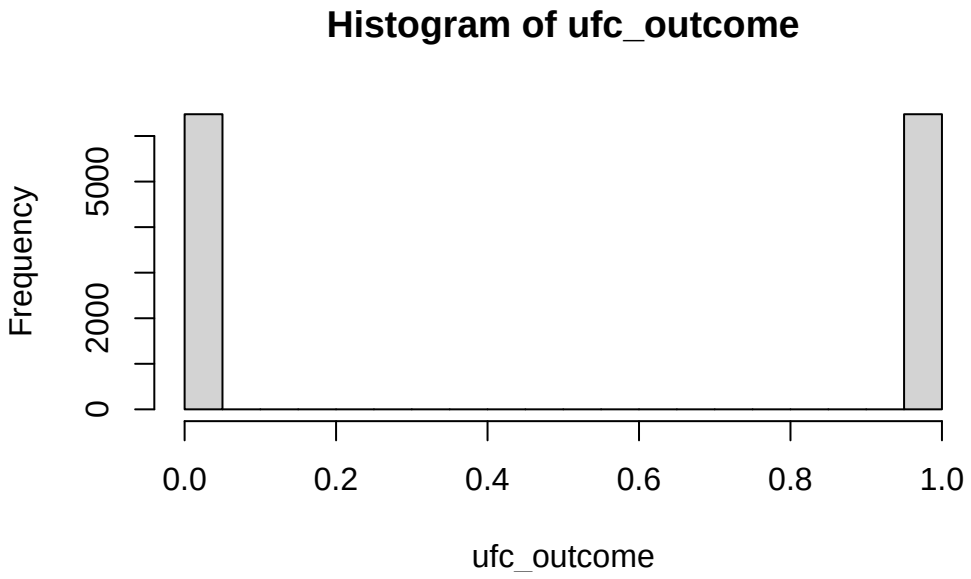



**Research question 2:** Is the fight outcome associated with the number of submission attempts made by a fighter?

- *Outcome Variable (include the name/description and type of variable):* Fight outcome (binary variable: Win or Loss)
- *Independent Variable:* Primary fight style (nominal variable)

**Exploratory Plots:**

1. Outcome Variables:



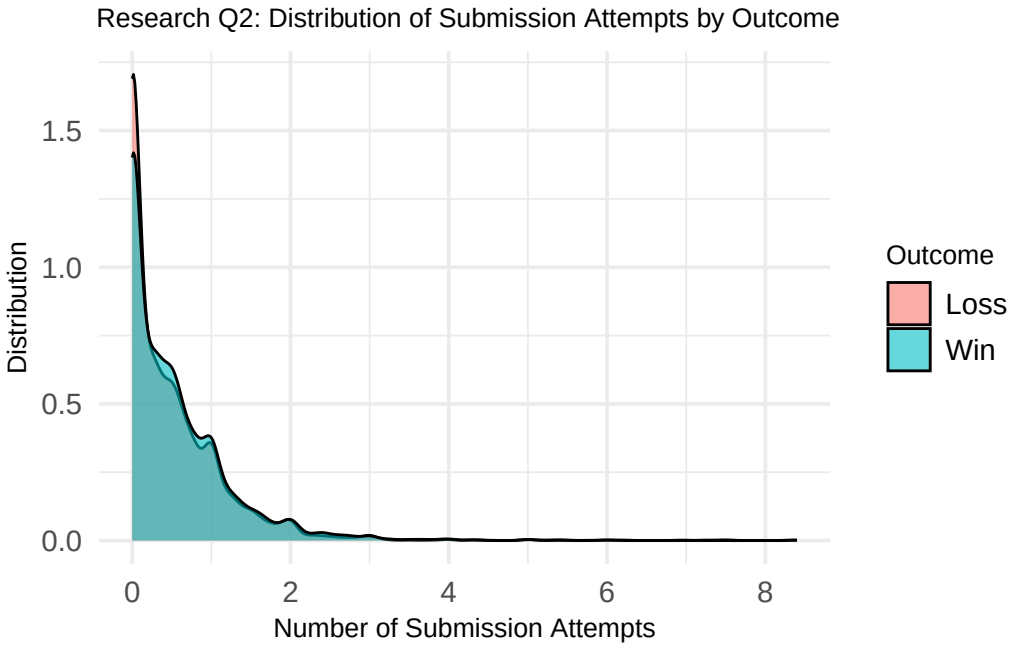

2. Relationship of Interest

```
[1] 0
```

```
[1] 357
```

```
[1] 832
```

```
# A tibble: 6 x 3
# Rowwise:
  AvgSubAtt     Win   Value
  <chr>         <fct> <dbl>
1 BlueAvgSubAtt Loss    0.5
2 RedAvgSubAtt  Win     0.4
3 BlueAvgSubAtt Loss    0.4
4 RedAvgSubAtt  Win     0.8
5 BlueAvgSubAtt Win     0.4
6 RedAvgSubAtt  Loss    0
```

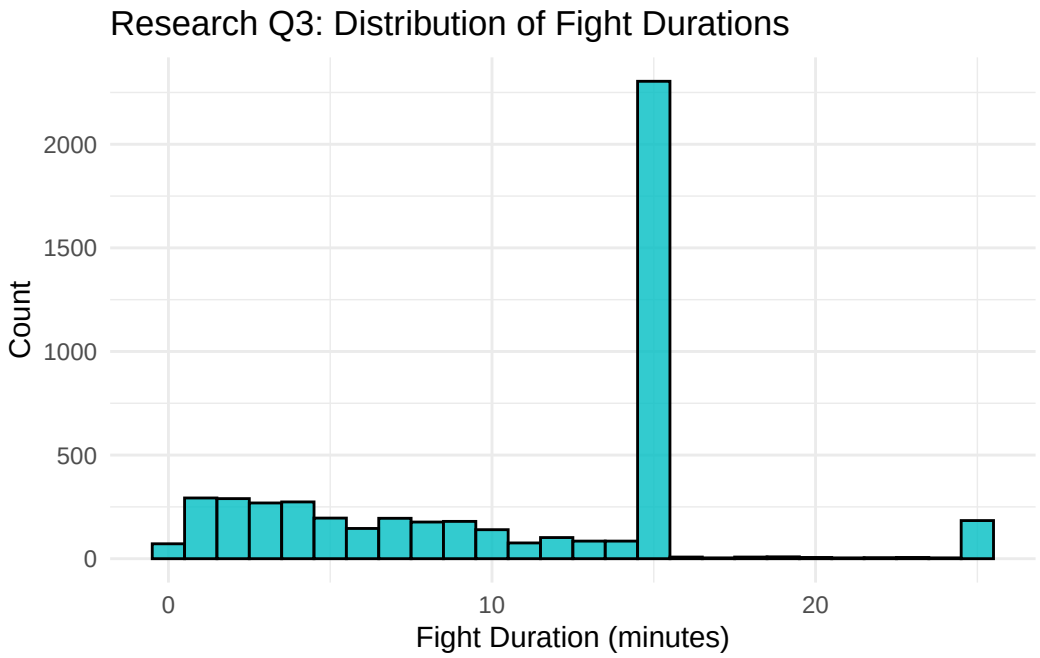



**Research question 3:** How does the average fight duration compare between fighters who primarily use striking versus grappling techniques?

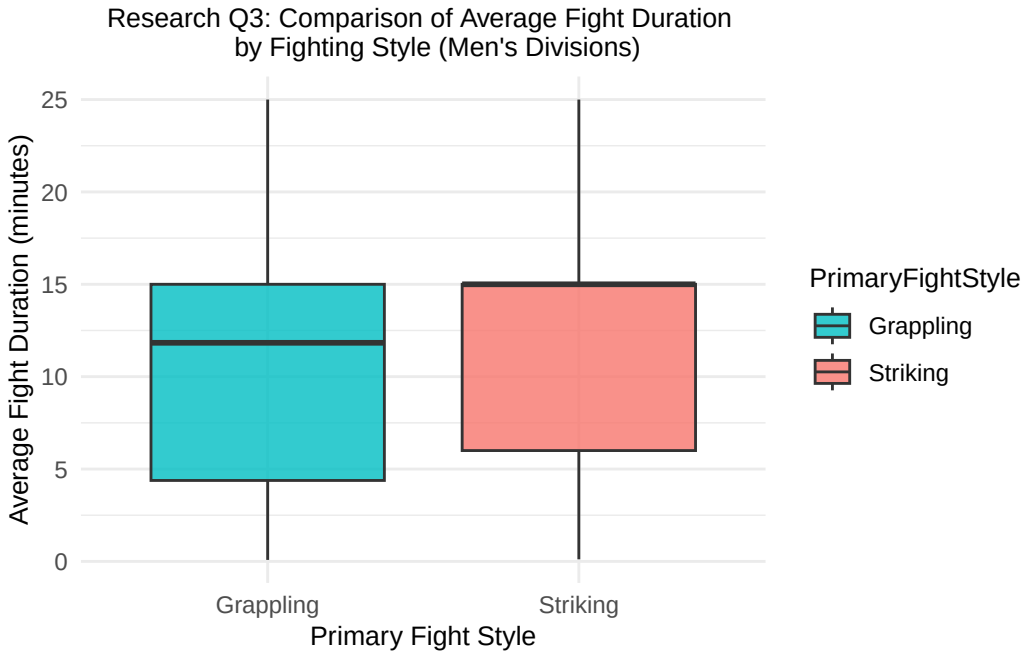
- *Outcome variable:* Average fight duration (continuous variable)
- *Independent Variable:* Primary fight style (nominal variable)

**Exploratory Plots:**

1. Outcome Variable




2. Relationship of Interest



## Dataset 2

**Data source:** **Superhero Power Analytics**

**Brief description:** The Superhero data set provides detailed information on 675 superheroes and villains from popular franchises. It contains attributes related to their powers, physical characteristics, and affiliations

**Load the data and provide a `glimpse()`:**

```
superheroes <- read_csv("superheroes_data.csv")
```

```
Rows: 731 Columns: 26
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (25): name, intelligence, strength, speed, durability, power, combat, fu...
dbl  (1): id

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(superheroes)
```

```
Rows: 731
Columns: 26
$ id                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
$ name               <chr> "A-Bomb", "Abe Sapien", "Abin Sur", "Abomination",~
$ intelligence       <chr> "38", "88", "50", "63", "88", "38", "63", "69", "n~
$ strength           <chr> "100", "28", "90", "80", "63", "80", "10", "10", "~
$ speed              <chr> "17", "35", "53", "53", "83", "25", "12", "33", "n~
$ durability         <chr> "80", "65", "64", "90", "100", "100", "100", "40",~
$ power              <chr> "24", "100", "99", "62", "100", "98", "100", "37",~
$ combat             <chr> "64", "85", "65", "95", "55", "64", "64", "50", "n~
$ `full-name`        <chr> "Richard Milhouse Jones", "Abraham Sapien", NA, "E~
$ `alter-egos`       <chr> "No alter egos found.", "No alter egos found.", "N~
$ aliases            <chr> "['Rick Jones']", "['Langdon Everett Caul', 'Abrah~
$ `place-of-birth`   <chr> "Scarsdale, Arizona", "-", "Ungara", "Zagreb, Yugo~
$ `first-appearance` <chr> "Hulk Vol 2 #2 (April, 2008) (as A-Bomb)", "Hellbo~
$ publisher          <chr> "Marvel Comics", "Dark Horse Comics", "DC Comics",~
$ alignment          <chr> "good", "good", "good", "bad", "bad", "bad", "good~
$ gender             <chr> "Male", "Male", "Male", "Male", "Male", "Male", "M~
$ race               <chr> "Human", "Icthyo Sapien", "Ungaran", "Human / Radi~
```

```
$ height              <chr> "[\"6'8\", '203 cm']", "[\"6'3\", '191 cm']", "[\"~
$ weight              <chr> "['980 lb', '441 kg']", "['145 lb', '65 kg']", "['~
$ `eye-color`         <chr> "Yellow", "Blue", "Blue", "Green", "Blue", "Blue",~
$ `hair-color`        <chr> "No Hair", "No Hair", "No Hair", "No Hair", "Black~
$ occupation          <chr> "Musician, adventurer, author; formerly talk show ~
$ base                <chr> "-", "-", "Oa", "Mobile", "-", "-", "U.S.; formerl~
$ `group-affiliation` <chr> "Hulk Family; Excelsior (sponsor), Avengers (honor~
$ relatives           <chr> "Marlo Chandler-Jones (wife); Polly (aunt); Mrs. C~
$ url                 <chr> "https://www.superherodb.com/pictures2/portraits/1~
```

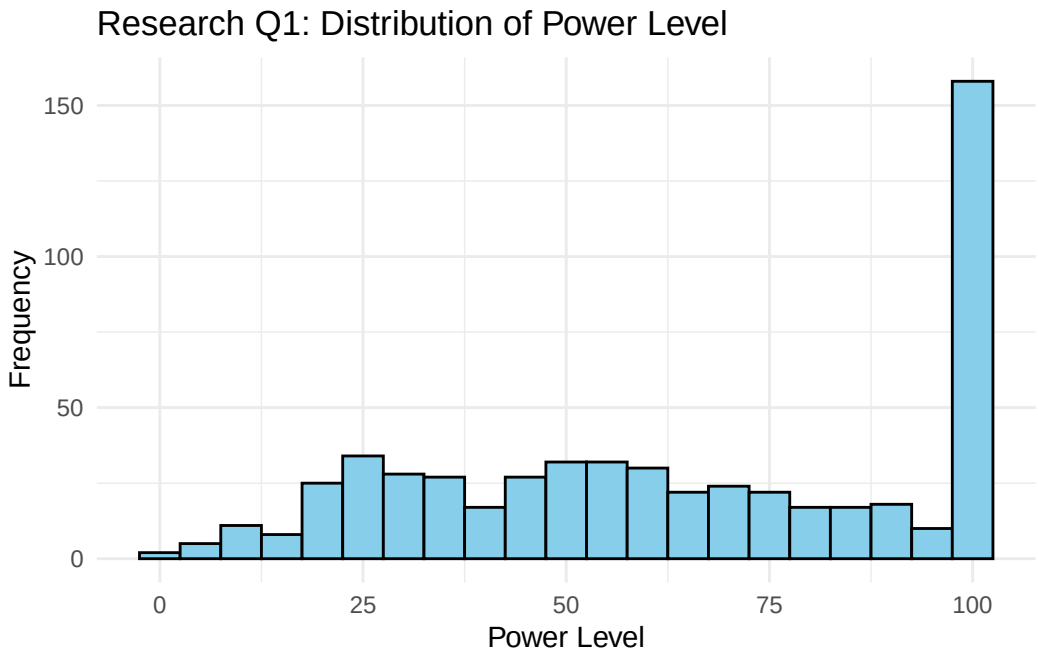
**Research question 1:** How do gender and alignment (good, neutral, evil) interact to influence the power level of superheroes

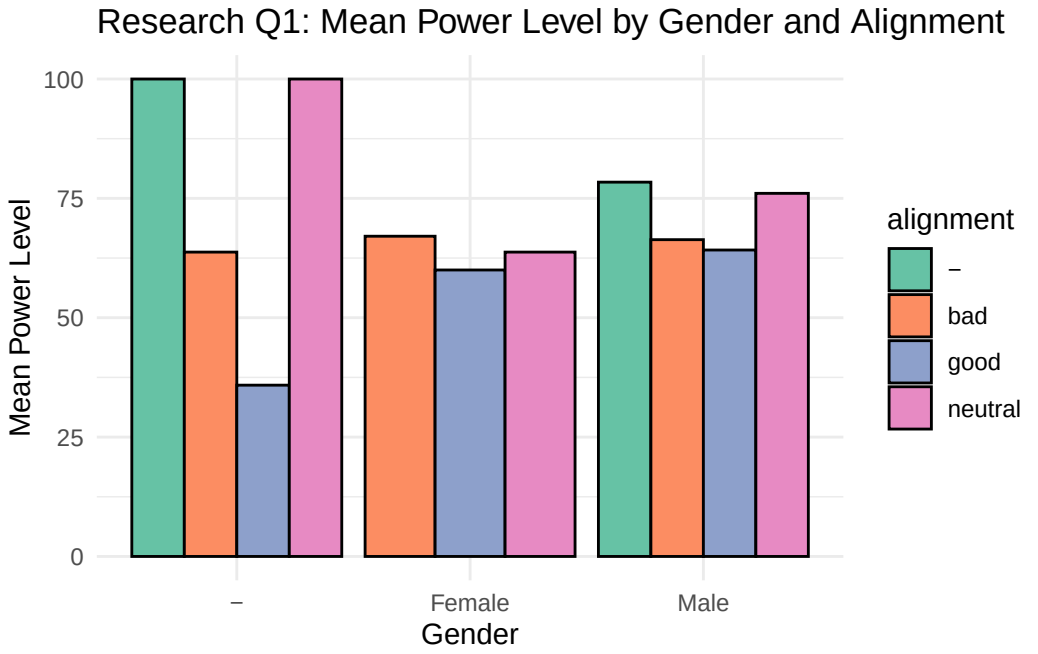
- *Outcome variable (include the name/description and type of variable):* Power level (continuous variable)
- *Independent Variable:* Gender, alignment
- *Interaction term:* Interaction between gender and alignment

**Research question 2:** How do intelligence, strength, and speed influence the likelihood of a superhero being classified as a hero versus a villain?

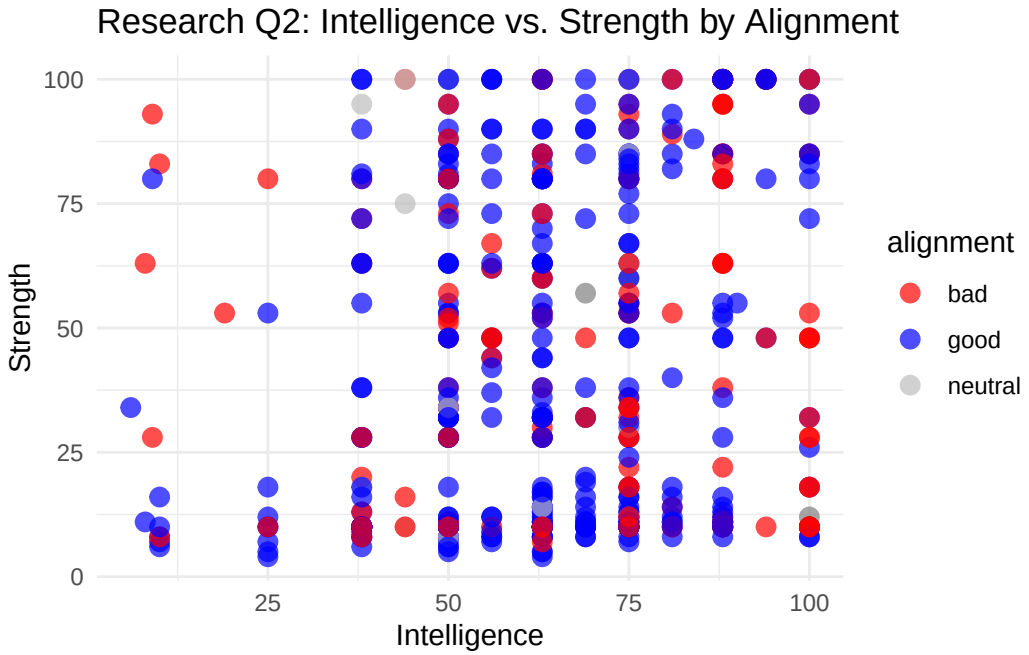
- *Outcome variable (include the name/description and type of variable):* Alignment (categorical variable)
- *Independent Variable:* Intelligence, strength, speed (continuous)
- *Interaction term:* Interaction between intelligence and strength to assess how these two traits impact alignment

**Exploratory Plots:**


Research Q1: Distribution of Power Level
Frequency
150
100
50
0
0
25
50
75
100
Power Level


Research Q1: Mean Power Level by Gender and Alignment
Mean Power Level
100
75
50
25
0
–
Female
Male
Gender
alignment
–
bad
good
neutral


Research Q2: Distribution of Alignment
Count
500
400
300
200
100
0
–
bad
good
neutral
Alignment


Research Q2: Intelligence vs. Strength by Alignment
Strength
100
75
50
25
0
25
50
75
100
Intelligence
alignment
bad
good
neutral

## Dataset 3 (Optional)

**Data source: Occupational Wages Around the World (OWW) Database**

**Brief description:** The Occupational Wages Around the World (OWW) Database is a comprehensive dataset that provides standardized information on occupational wage levels across 161 countries from the early 1980s to the early 2000s. Compiled from the International Labour Organization's Yearbook of Labour Statistics, the data covers various occupational categories, including clerical workers, agricultural laborers, and production workers, among others. The data set is designed to facilitate cross-country and longitudinal comparisons by adjusting for differences in currency, inflation, and cost of living. It is widely used for studying global labor markets, wage disparities, and economic development trends.

**Load the data and provide a `glimpse()`:**

```
oww <- read_csv("oww3.csv")
```

```
Rows: 125018 Columns: 37
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): y1, country, y3, curr, curr_conv
dbl (32): y0, y4, hw1, hw2wu, hw3wu, hw4wu, hw2wl, hw3wl, hw4wl, hw1us, hw2w...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(oww)
```

```
Rows: 125,018
Columns: 37
$ y0        <dbl> 1989, 1990, 1992, 1993, 1994, 1995, 1996, 1989, 1990, 1992, ~
$ y1        <chr> "AG", "AG", "AG", "AG", "AG", "AG", "AG", "AG", "AG", "AG", ~
$ country   <chr> "ATG", "ATG", "ATG", "ATG", "ATG", "ATG", "ATG", "ATG", "ATG~
$ y3        <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", ~
$ y4        <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 4, 5, 5, 5, 6, 6, ~
$ hw1       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 6.357692, NA~
$ hw2wu     <dbl> 9.418722, 6.350042, 7.134442, 9.047402, 11.975790, 15.325243~
$ hw3wu     <dbl> 9.418722, 6.350042, 7.134442, 9.047402, 11.975790, 15.325243~
$ hw4wu     <dbl> 9.146313, 6.182484, 6.928098, 8.846016, 11.631442, 14.884584~
$ hw2wl     <dbl> 9.418722, 6.350042, 7.134442, 9.047402, 11.975790, 15.325243~
$ hw3wl     <dbl> 9.418722, 6.350042, 7.134442, 9.047402, 11.975790, 15.325243~
$ hw4wl     <dbl> 9.146313, 6.182484, 6.928098, 8.846016, 11.631442, 14.884584~
```

```
$ hw1us     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2.354701, NA~
$ hw2wuus   <dbl> 3.488415, 2.351867, 2.642386, 3.350890, 4.435478, 5.676016, ~
$ hw3wuus   <dbl> 3.488415, 2.351867, 2.642386, 3.350890, 4.435478, 5.676016, ~
$ hw4wuus   <dbl> 3.387523, 2.289809, 2.565962, 3.276302, 4.307941, 5.512809, ~
$ hw2wlus   <dbl> 3.488415, 2.351867, 2.642386, 3.350890, 4.435478, 5.676016, ~
$ hw3wlus   <dbl> 3.488415, 2.351867, 2.642386, 3.350890, 4.435478, 5.676016, ~
$ hw4wlus   <dbl> 3.387523, 2.289809, 2.565962, 3.276302, 4.307941, 5.512809, ~
$ mw1       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1102, NA, 11~
$ mw2wu     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1102, NA, 11~
$ mw3wu     <dbl> 1635.3135, 1104.1616, 1238.7084, 1538.3588, 2079.7615, 2480.~
$ mw4wu     <dbl> 1563.2399, 1057.0548, 1184.1145, 1488.9290, 2003.4596, 2389.~
$ mw2wl     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1102, NA, 11~
$ mw3wl     <dbl> 1635.3135, 1104.1616, 1238.7084, 1538.3588, 2079.7615, 2480.~
$ mw4wl     <dbl> 1563.2399, 1057.0548, 1184.1145, 1488.9290, 2003.4596, 2389.~
$ mw1us     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 408.1481, NA~
$ mw2wuus   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 408.1481, NA~
$ mw3wuus   <dbl> 605.6716, 408.9487, 458.7809, 569.7625, 770.2820, 918.8571, ~
$ mw4wuus   <dbl> 578.9777, 391.5018, 438.5609, 551.4551, 742.0220, 885.1463, ~
$ mw2wlus   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 408.1481, NA~
$ mw3wlus   <dbl> 605.6716, 408.9487, 458.7809, 569.7625, 770.2820, 918.8571, ~
$ mw4wlus   <dbl> 578.9777, 391.5018, 438.5609, 551.4551, 742.0220, 885.1463, ~
$ curr      <chr> "dollar (EC)", "dollar (EC)", "dollar (EC)", "dollar (EC)", ~
$ exrt      <dbl> 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, 2.7, ~
$ conv      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ curr_conv <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

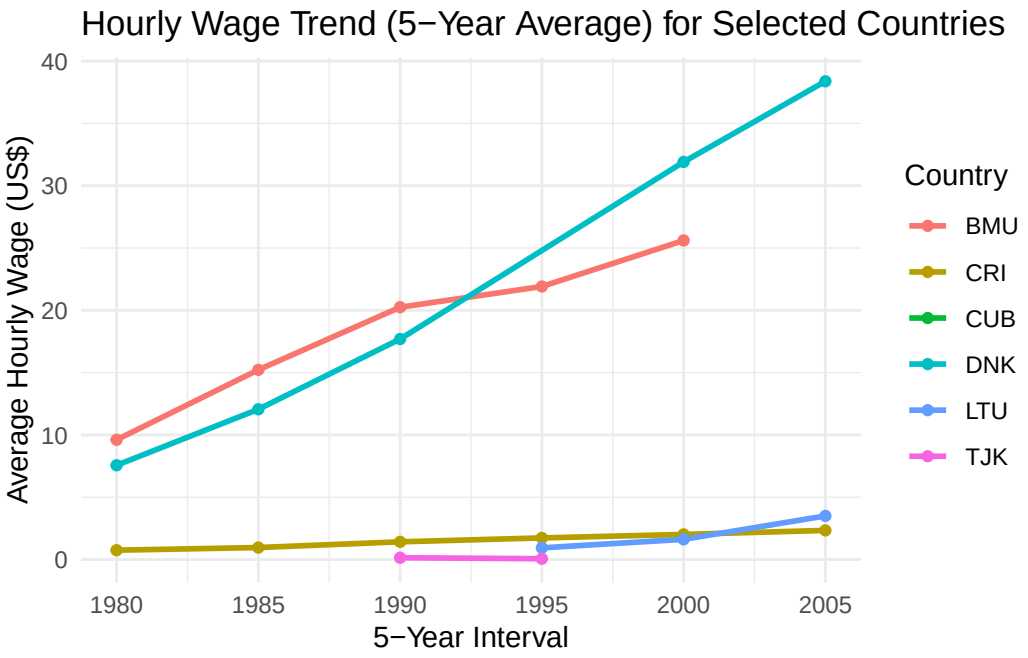
**Research question 1:** How to economic factors and time influence the hourly wage rates across different countries, and does this relationship vary by wage calculation method

- *Outcome variable (include the name/description and type of variable):* hw3wlus (Hourly wage in US dollars, lexicographically weighted) (continuous variable)
- *Independent Variable:* y0 (year), exrt (exchange rate), country
- *Interaction Term:* y0 * exrt

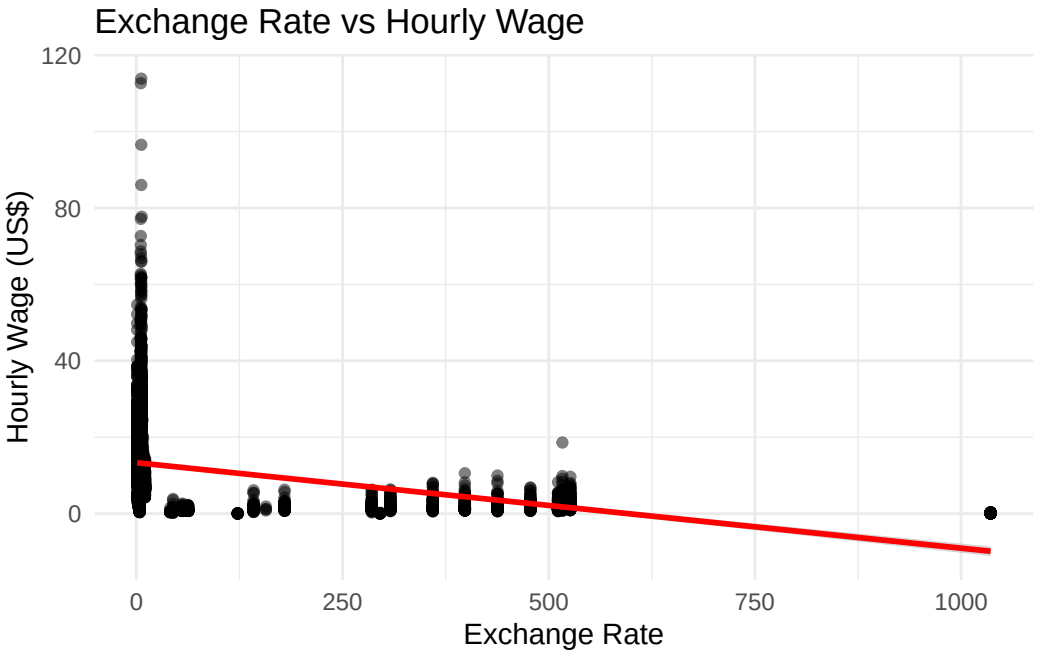
**Research question 2:** How do currency types and conversion factors affect the relative ranking of countries in terms of monthly wage levels

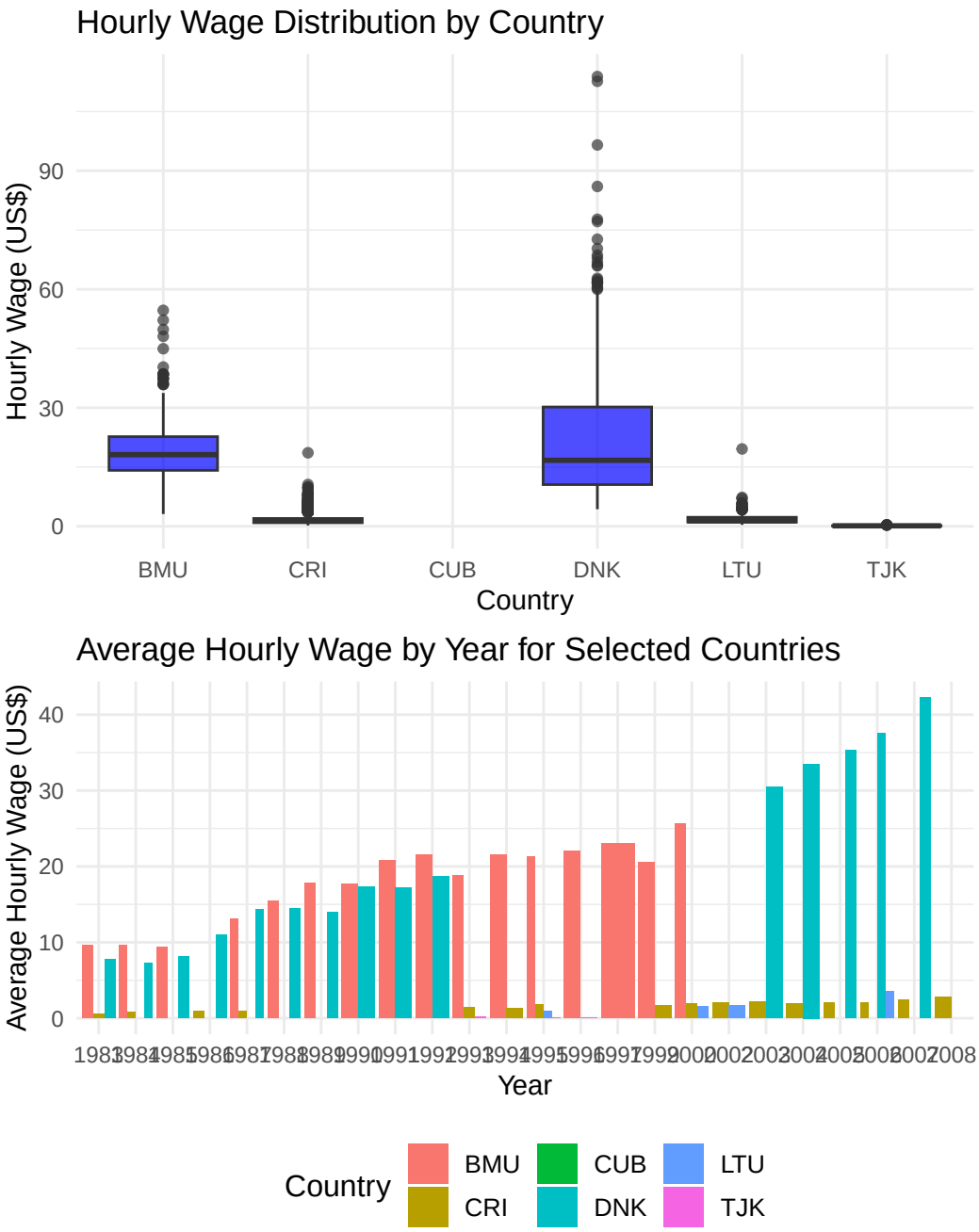
- *Outcome variable (include the name/description and type of variable): Country ranking based on monthly wages (to be derived from mw3wlus)*
- *Independent Variable:* curr (Currency type), conv (Conversion Factor), y0 (Year)

**Exploratory Plots:**




```
`geom_smooth()` using formula = 'y ~ x'
```




Hourly Wage Distribution by Country
Hourly Wage (US$)
90
60
30
0
BMU
CRI
CUB
DNK
LTU
TJK
Country
Average Hourly Wage by Year for Selected Countries
Average Hourly Wage (US$)
40
30
20
10
0
Year
Country
BMU
CRI
CUB
DNK
LTU
TJK

## Team Charter

**When will you meet as a team to work on the project components? Will these meetings be held in person or virtually?**

We will have weekly meetings - preferably in person but this can be held virtually based on everyone's availability.

**What is your group policy on missing team meetings (e.g., how much advance notice should be provided)?**

The policy is to notify the rest of the team on the Slack/WhatsApp channel a day in advance. In case of an emergency, no advanced communication needed.

**How will your team communicate (email, Slack, text messages)? What is your policy on appropriate response time (within a certain number of hours? Nights/weekends?)?**

The team will communicate through Slack and WhatsApp. The appropriate response time is a few hours, all days of the week.