Preprocessing

Charles J. Gomez, Harshvardhan Singh

2023-06-24

The country column in our dataset contains a more detailed information about the authors' affiliations, including first, middle, and last authors. However, this brings a level of complexity and potential confusion for machine learning models. Therefore, we propose a method to simplify the country column by retaining only the first and last author affiliations, ensuring a more straightforward representation of authorship.

```
library(stringr)

data = read.csv("INPUT_SQL_Text_Data_Astronomy_and_Astrophysics.csv")
data_CountryColumn <- data$country

#We removed all instances of middle authors' country codes. This step involved replacing the pattern "\

data$country <- gsub("\\w+\\+middle", "", data$country)
data$country <- trimws(data$country)
print(head(data$country, 50))
```

```
##  [1] "US+last US+first US+last US+first"
##  [2] "US+first"
##  [3] "US+first"
##  [4] "IN+first US+last"
##  [5] "US+last       US+first                US+last          US+first"
##  [6] "US+first"
##  [7] "US+first  RS+last"
##  [8] "US+last DE+first"
##  [9] "CH+first CH+first"
## [10] "US+first  US+first  US+last US+last"
## [11] "GB+first   GB+first US+last  GB+first  US+last GB+first"
## [12] "CN+first   CN+last CN+first CN+last"
## [13] "US+first"
## [14] "US+first US+last"
## [15] "US+last US+last  US+first    US+first    US+first  US+last US+last US+first"
## [16] "FR+last US+first"
## [17] "IT+first IT+first IT+first     IT+first"
## [18] "US+first US+first  US+last US+last"
## [19] "RU+last RU+first GB+first"
## [20] "US+first US+last US+first US+last"
## [21] "US+last US+first CA+first US+first US+last CA+first"
## [22] "US+last    US+first US+first        US+last"
## [23] "US+first US+first"
## [24] "US+first    US+last"
## [25] "US+last  GB+first                               US+last                             GB+first"
## [26] "GB+first NL+last GB+first NL+last"
## [27] "AU+last AU+last   AU+first  AU+first"
## [28] "US+last GB+first US+last GB+first"
## [29] "US+first US+last   US+last US+first"
## [30] "GB+last GB+first US+first GB+first  US+first GB+last"
## [31] "GB+first GB+last  IT+first  GB+last   GB+first IT+first"
## [32] "GB+last GB+last NL+first  NL+first"
## [33] "US+first"
## [34] "AU+first AU+first  AU+last  AU+last"
## [35] "GB+last"
## [36] "DK+last            US+first US+first     DK+last"
## [37] "CA+first SE+last"
## [38] "US+first US+last    US+first   US+last"
## [39] "US+last   US+last US+first    US+first"
## [40] "FR+last US+first"
## [41] "AU+first AU+first AU+first AU+first"
## [42] "US+last US+first"
## [43] "CL+first US+last"
## [44] "BR+first  BR+last BR+first BR+last"
## [45] "US+first US+first"
## [46] "AU+last   US+first    AU+last   US+first"
## [47] "IT+first  US+first US+last     US+last US+first   IT+first"
## [48] "US+last"
## [49] "US+first   US+last"
## [50] "SE+first  SE+last"
```

```
#We removed the "+first" and "+last" strings from the remaining country codes using another regular exp

data$country <- gsub("\\+last|\\+first", "", data$country)
# Remove any extra spaces resulting from the removal
data$country <- sapply(strsplit(data$country, "\\s+"), function(x) paste(x[x != ""], collapse = " "))

print(head(data$country, 50))
```

```
##  [1] "US US US US"             "US"
##  [3] "US"                      "IN US"
##  [5] "US US US US"             "US"
##  [7] "US RS"                   "US DE"
##  [9] "CH CH"                   "US US US US"
## [11] "GB GB US GB US GB"       "CN CN CN CN"
## [13] "US"                      "US US"
## [15] "US US US US US US US US" "FR US"
## [17] "IT IT IT IT"             "US US US US"
## [19] "RU RU GB"                "US US US US"
## [21] "US US CA US US CA"       "US US US US"
## [23] "US US"                   "US US"
## [25] "US GB US GB"             "GB NL GB NL"
## [27] "AU AU AU AU"             "US GB US GB"
## [29] "US US US US"             "GB GB US GB US GB"
## [31] "GB GB IT GB GB IT"       "GB GB NL NL"
## [33] "US"                      "AU AU AU AU"
## [35] "GB"                      "DK US US DK"
## [37] "CA SE"                   "US US US US"
## [39] "US US US US"             "FR US"
## [41] "AU AU AU AU"             "US US"
## [43] "CL US"                   "BR BR BR BR"
## [45] "US US"                   "AU US AU US"
## [47] "IT US US US US IT"       "US"
## [49] "US US"                    "SE SE"
```

```r
# calculate the percentage count of each country code in a vector
#calculate_percentage <- function(vec) {
#  counts <- table(vec)
#  percentages <- prop.table(counts) * 100
#  formatted <- paste0(round(percentages, 1), "%", names(percentages))
#  paste(formatted, collapse = " ")
#}

# Apply the calculate_percentage function to each row in the 'country' column
#data$country <- sapply(strsplit(data$country, "\\s+"), calculate_percentage)

# Print the modified 'country' column (first 20)
#print(head(data$country, 50))





# Calculate percentage count of each country code in a vector
calculate_percentage <- function(vec) {
  counts <- table(vec)
  percentages <- prop.table(counts) * 100
  return(percentages)
}

# Apply the calculate_percentage function to each row in the 'country' column
percentage_counts <- lapply(strsplit(data$country, "\\s+"), calculate_percentage)

# Get all unique countries
all_countries <- unique(unlist(lapply(percentage_counts, names)))

# Initialize a new data frame to hold the percentages
percentage_df <- data.frame(matrix(ncol = length(all_countries), nrow = length(percentage_counts)))
names(percentage_df) <- all_countries

# Fill the data frame with percentages
for (i in seq_along(percentage_counts)) {
  country_names <- names(percentage_counts[[i]])
  country_percentages <- percentage_counts[[i]]
  percentage_df[i, country_names] <- country_percentages
}

# Replace NA values with 0
percentage_df[is.na(percentage_df)] <- 0

# Combine the original data with the new percentage-encoded country columns
data <- cbind(data, percentage_df)

# Summary of dataframe
summary(data)
```

```
##   concept_id          work_id          publication_year    title
##  Length:63999       Length:63999       Min.   :1824     Length:63999
##  Class :character   Class :character   1st Qu.:2001     Class :character
##  Mode  :character   Mode  :character   Median :2005     Mode  :character
##                                         Mean   :2005
##                                         3rd Qu.:2010
##                                         Max.   :2022
##  paperabstract        country          year_concept             US
##  Length:63999       Length:63999       Length:63999       Min.   :  0.00
##  Class :character   Class :character   Class :character   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 50.00
##                                                            Mean   : 49.79
##                                                            3rd Qu.:100.00
##                                                            Max.   :100.00
##        IN                RS                 DE                CH
##  Min.   :  0.000   Min.   :  0.0000   Min.   :  0.000   Min.   :  0.0000
##  1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.0000
##  Median :  0.000   Median :  0.0000   Median :  0.000   Median :  0.0000
##  Mean   :  1.397   Mean   :  0.5605   Mean   :  5.642   Mean   :  0.8845
##  3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:  0.000   3rd Qu.:  0.0000
##  Max.   :100.000   Max.   :100.0000   Max.   :100.000   Max.   :100.0000
##        GB                CN                FR                IT
##  Min.   :  0.000   Min.   :  0.000   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :  0.000   Median :  0.000   Median :  0.000   Median :  0.000
##  Mean   :  9.964   Mean   :  2.149   Mean   :  1.947   Mean   :  3.778
##  3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.000
##  Max.   :100.000   Max.   :100.000   Max.   :100.000   Max.   :100.000
##        RU                CA                NL                AU
##  Min.   :  0.000   Min.   :  0.000   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :  0.000   Median :  0.000   Median :  0.000   Median :  0.000
##  Mean   :  1.822   Mean   :  2.628   Mean   :  1.975   Mean   :  2.777
##  3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.000
##  Max.   :100.000   Max.   :100.000   Max.   :100.000   Max.   :100.000
##        DK                 SE                CL                 BR
##  Min.   :  0.0000   Min.   :  0.000   Min.   :  0.0000   Min.   :  0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.0000
##  Median :  0.0000   Median :  0.000   Median :  0.0000   Median :  0.0000
##  Mean   :  0.2942   Mean   :  0.342   Mean   :  0.4229   Mean   :  0.7424
##  3rd Qu.:  0.0000   3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:  0.0000
##  Max.   :100.0000   Max.   :100.000   Max.   :100.0000   Max.   :100.0000
##        BE                 JP                VE                 MX
##  Min.   :  0.0000   Min.   :  0.000   Min.   : 0.00000   Min.   :  0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.: 0.00000   1st Qu.:  0.0000
##  Median :  0.0000   Median :  0.000   Median : 0.00000   Median :  0.0000
##  Mean   :  0.4449   Mean   :  2.551   Mean   : 0.00521   Mean   :  0.7465
##  3rd Qu.:  0.0000   3rd Qu.:  0.000   3rd Qu.: 0.00000   3rd Qu.:  0.0000
##  Max.   :100.0000   Max.   :100.000   Max.   :50.00000   Max.   :100.0000
##        PL                ES                IE                 FI
##  Min.   :  0.000   Min.   :  0.000   Min.   :  0.0000   Min.   :  0.0000
##  1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.0000
##  Median :  0.000   Median :  0.000   Median :  0.0000   Median :  0.0000
##  Mean   :  0.573   Mean   :  1.749   Mean   :  0.2064   Mean   :  0.2596
##  3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:  0.0000
##  Max.   :100.000   Max.   :100.000   Max.   :100.0000   Max.   :100.0000
##        KR                 TW                 AT                 IR
##  Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.0000
##  Median :  0.0000   Median :  0.0000   Median :  0.0000   Median :  0.0000
##  Mean   :  0.4737   Mean   :  0.2367   Mean   :  0.2039   Mean   :  0.1189
##  3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:  0.0000
##  Max.   :100.0000   Max.   :100.0000   Max.   :100.0000   Max.   :100.0000
##        HU                 IL                 LV                 UA
##  Min.   :  0.0000   Min.   :  0.0000   Min.   :0.00e+00   Min.   :  0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:0.00e+00   1st Qu.:  0.0000
##  Median :  0.0000   Median :  0.0000   Median :0.00e+00   Median :  0.0000
##  Mean   :  0.2299   Mean   :  0.9436   Mean   :4.69e-03   Mean   :  0.2302
##  3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:0.00e+00   3rd Qu.:  0.0000
##  Max.   :100.0000   Max.   :100.0000   Max.   :1.00e+02   Max.   :100.0000
##        GE                  GR                 AR                 NG
##  Min.   :  0.00000   Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.00000
##  1st Qu.:  0.00000   1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.00000
##  Median :  0.00000   Median :  0.0000   Median :  0.0000   Median :  0.00000
##  Mean   :  0.03542   Mean   :  0.2227   Mean   :  0.3265   Mean   :  0.04271
##  3rd Qu.:  0.00000   3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:  0.00000
##  Max.   :100.00000   Max.   :100.0000   Max.   :100.0000   Max.   :100.00000
##        HR                  TR                 AM                  CZ
##  Min.   :  0.00000   Min.   :  0.0000   Min.   :  0.00000   Min.   :  0.0000
##  1st Qu.:  0.00000   1st Qu.:  0.0000   1st Qu.:  0.00000   1st Qu.:  0.0000
##  Median :  0.00000   Median :  0.0000   Median :  0.00000   Median :  0.0000
##  Mean   :  0.02995   Mean   :  0.2077   Mean   :  0.09844   Mean   :  0.1352
##  3rd Qu.:  0.00000   3rd Qu.:  0.0000   3rd Qu.:  0.00000   3rd Qu.:  0.0000
##  Max.   :100.00000   Max.   :100.0000   Max.   :100.00000   Max.   :100.0000
##        KZ                  KH                  PT                 CO
##  Min.   :  0.00000   Min.   :  0.00000   Min.   :  0.0000   Min.   :  0.00000
##  1st Qu.:  0.00000   1st Qu.:  0.00000   1st Qu.:  0.0000   1st Qu.:  0.00000
##  Median :  0.00000   Median :  0.00000   Median :  0.0000   Median :  0.00000
##  Mean   :  0.01602   Mean   :  0.02526   Mean   :  0.1365   Mean   :  0.02904
##  3rd Qu.:  0.00000   3rd Qu.:  0.00000   3rd Qu.:  0.0000   3rd Qu.:  0.00000
##  Max.   :100.00000   Max.   :100.00000   Max.   :100.0000   Max.   :100.00000
##        MZ                  BW                 ZA                 SK
##  Min.   :  0.00000   Min.   : 0.00000   Min.   :  0.0000   Min.   :  0.00000
##  1st Qu.:  0.00000   1st Qu.: 0.00000   1st Qu.:  0.0000   1st Qu.:  0.00000
##  Median :  0.00000   Median : 0.00000   Median :  0.0000   Median :  0.00000
##  Mean   :  0.02552   Mean   : 0.00156   Mean   :  0.2787   Mean   :  0.08971
##  3rd Qu.:  0.00000   3rd Qu.: 0.00000   3rd Qu.:  0.0000   3rd Qu.:  0.00000
##  Max.   :100.00000   Max.   :50.00000   Max.   :100.0000   Max.   :100.00000
##        PK                  IS                 NO                  NZ
##  Min.   :  0.00000   Min.   : 0.00000   Min.   :  0.00000   Min.   :  0.00000
##  1st Qu.:  0.00000   1st Qu.: 0.00000   1st Qu.:  0.00000   1st Qu.:  0.00000
##  Median :  0.00000   Median : 0.00000   Median :  0.00000   Median :  0.00000
##  Mean   :  0.05391   Mean   : 0.00612   Mean   :  0.08399   Mean   :  0.08985
##  3rd Qu.:  0.00000   3rd Qu.: 0.00000   3rd Qu.:  0.00000   3rd Qu.:  0.00000
##  Max.   :100.00000   Max.   :66.66667   Max.   :100.00000   Max.   :100.00000
##        BG                  VN                MY                  GH
##  Min.   :  0.00000   Min.   : 0.0000   Min.   :  0.00000   Min.   : 0.00000
##  1st Qu.:  0.00000   1st Qu.: 0.0000   1st Qu.:  0.00000   1st Qu.: 0.00000
##  Median :  0.00000   Median : 0.0000   Median :  0.00000   Median : 0.00000
##  Mean   :  0.07552   Mean   : 0.0013   Mean   :  0.01042   Mean   : 0.00365
##  3rd Qu.:  0.00000   3rd Qu.: 0.0000   3rd Qu.:  0.00000   3rd Qu.: 0.00000
##  Max.   :100.00000   Max.   :50.0000   Max.   :100.00000   Max.   :50.00000
##        SI                  GT                EE                  EG
##  Min.   :  0.00000   Min.   :0.0e+00   Min.   :  0.00000   Min.   :  0.00000
##  1st Qu.:  0.00000   1st Qu.:0.0e+00   1st Qu.:  0.00000   1st Qu.:  0.00000
##  Median :  0.00000   Median :0.0e+00   Median :  0.00000   Median :  0.00000
##  Mean   :  0.02076   Mean   :7.8e-04   Mean   :  0.01563   Mean   :  0.01797
##  3rd Qu.:  0.00000   3rd Qu.:0.0e+00   3rd Qu.:  0.00000   3rd Qu.:  0.00000
##  Max.   :100.00000   Max.   :5.0e+01   Max.   :100.00000   Max.   :100.00000
##        RW                 UG                 UY                 JM
##  Min.   : 0.00000   Min.   : 0.00000   Min.   :0.00e+00   Min.   :0.00e+00
##  1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.:0.00e+00   1st Qu.:0.00e+00
##  Median : 0.00000   Median : 0.00000   Median :0.00e+00   Median :0.00e+00
##  Mean   : 0.00193   Mean   : 0.00052   Mean   :2.08e-03   Mean   :5.47e-03
##  3rd Qu.: 0.00000   3rd Qu.: 0.00000   3rd Qu.:0.00e+00   3rd Qu.:0.00e+00
##  Max.   :50.00000   Max.   :33.33333   Max.   :1.00e+02   Max.   :1.00e+02
##        SA                  RO                  UZ
##  Min.   :  0.00000   Min.   :  0.00000   Min.   :  0.00000
##  1st Qu.:  0.00000   1st Qu.:  0.00000   1st Qu.:  0.00000
##  Median :  0.00000   Median :  0.00000   Median :  0.00000
##  Mean   :  0.01979   Mean   :  0.01979   Mean   :  0.02349
##  3rd Qu.:  0.00000   3rd Qu.:  0.00000   3rd Qu.:  0.00000
##  Max.   :100.00000   Max.   :100.00000   Max.   :100.00000
##        TH                  NP                 TN                LK
##  Min.   :  0.00000   Min.   :0.00e+00   Min.   : 0.0000   Min.   :0.00e+00
##  1st Qu.:  0.00000   1st Qu.:0.00e+00   1st Qu.: 0.0000   1st Qu.:0.00e+00
##  Median :  0.00000   Median :0.00e+00   Median : 0.0000   Median :0.00e+00
##  Mean   :  0.01406   Mean   :6.25e-03   Mean   : 0.0013   Mean   :5.73e-03
##  3rd Qu.:  0.00000   3rd Qu.:0.00e+00   3rd Qu.: 0.0000   3rd Qu.:0.00e+00
##  Max.   :100.00000   Max.   :1.00e+02   Max.   :50.0000   Max.   :1.00e+02
##        PE                 LT                  CY                 ID
##  Min.   : 0.00000   Min.   :  0.00000   Min.   :0.00e+00   Min.   :0.00e+00
##  1st Qu.: 0.00000   1st Qu.:  0.00000   1st Qu.:0.00e+00   1st Qu.:0.00e+00
##  Median : 0.00000   Median :  0.00000   Median :0.00e+00   Median :0.00e+00
##  Mean   : 0.00495   Mean   :  0.01172   Mean   :8.91e-03   Mean   :9.11e-03
##  3rd Qu.: 0.00000   3rd Qu.:  0.00000   3rd Qu.:0.00e+00   3rd Qu.:0.00e+00
##  Max.   :66.66667   Max.   :100.00000   Max.   :1.00e+02   Max.   :1.00e+02
##        EC                 LU                CR                 TJ
##  Min.   : 0.00000   Min.   :0.0e+00   Min.   :0.00e+00   Min.   :0.00e+00
##  1st Qu.: 0.00000   1st Qu.:0.0e+00   1st Qu.:0.00e+00   1st Qu.:0.00e+00
##  Median : 0.00000   Median :0.0e+00   Median :0.00e+00   Median :0.00e+00
##  Mean   : 0.00182   Mean   :7.8e-04   Mean   :7.03e-03   Mean   :3.13e-03
##  3rd Qu.: 0.00000   3rd Qu.:0.0e+00   3rd Qu.:0.00e+00   3rd Qu.:0.00e+00
##  Max.   :50.00000   Max.   :5.0e+01   Max.   :1.00e+02   Max.   :1.00e+02
##        AE                 LB                SG                 BO
##  Min.   :0.00e+00   Min.   :0.0e+00   Min.   :0.00e+00   Min.   : 0.00000
##  1st Qu.:0.00e+00   1st Qu.:0.0e+00   1st Qu.:0.00e+00   1st Qu.: 0.00000
##  Median :0.00e+00   Median :0.0e+00   Median :0.00e+00   Median : 0.00000
##  Mean   :8.33e-03   Mean   :2.6e-03   Mean   :9.11e-03   Mean   : 0.00156
##  3rd Qu.:0.00e+00   3rd Qu.:0.0e+00   3rd Qu.:0.00e+00   3rd Qu.: 0.00000
##  Max.   :1.00e+02   Max.   :1.0e+02   Max.   :1.00e+02   Max.   :50.00000
##        JO                 AZ                 IQ                 OM
##  Min.   :0.00e+00   Min.   :0.00e+00   Min.   :0.00e+00   Min.   : 0.00000
##  1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.: 0.00000
##  Median :0.00e+00   Median :0.00e+00   Median :0.00e+00   Median : 0.00000
##  Mean   :6.25e-03   Mean   :4.38e-03   Mean   :7.81e-03   Mean   : 0.00195
##  3rd Qu.:0.00e+00   3rd Qu.:0.00e+00   3rd Qu.:0.00e+00   3rd Qu.: 0.00000
##  Max.   :1.00e+02   Max.   :1.00e+02   Max.   :1.00e+02   Max.   :50.00000
##        KW                PR                 TZ                 BD
##  Min.   :0.0e+00   Min.   :0.00e+00   Min.   : 0.00000   Min.   :0.00e+00
##  1st Qu.:0.0e+00   1st Qu.:0.00e+00   1st Qu.: 0.00000   1st Qu.:0.00e+00
##  Median :0.0e+00   Median :0.00e+00   Median : 0.00000   Median :0.00e+00
##  Mean   :7.8e-04   Mean   :7.81e-03   Mean   : 0.00104   Mean   :1.56e-03
##  3rd Qu.:0.0e+00   3rd Qu.:0.00e+00   3rd Qu.: 0.00000   3rd Qu.:0.00e+00
##  Max.   :5.0e+01   Max.   :1.00e+02   Max.   :33.33333   Max.   :1.00e+02
##        MU                 MT                 CU                MK
##  Min.   :0.00e+00   Min.   :0.00e+00   Min.   :0.0e+00   Min.   :0.00e+00
##  1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.:0.0e+00   1st Qu.:0.00e+00
##  Median :0.00e+00   Median :0.00e+00   Median :0.0e+00   Median :0.00e+00
##  Mean   :1.56e-03   Mean   :1.56e-03   Mean   :7.8e-04   Mean   :1.56e-03
##  3rd Qu.:0.00e+00   3rd Qu.:0.00e+00   3rd Qu.:0.0e+00   3rd Qu.:0.00e+00
##  Max.   :1.00e+02   Max.   :1.00e+02   Max.   :5.0e+01   Max.   :1.00e+02
##        VI                 PS                 ZW                 NA
##  Min.   :0.00e+00   Min.   :0.00e+00   Min.   :0.00e+00   Min.   : 0.00000
##  1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.: 0.00000
##  Median :0.00e+00   Median :0.00e+00   Median :0.00e+00   Median : 0.00000
##  Mean   :1.56e-03   Mean   :1.56e-03   Mean   :1.56e-03   Mean   : 0.00156
##  3rd Qu.:0.00e+00   3rd Qu.:0.00e+00   3rd Qu.:0.00e+00   3rd Qu.: 0.00000
##  Max.   :1.00e+02   Max.   :1.00e+02   Max.   :1.00e+02   Max.   :50.00000
##        QA                 HN                 SD
##  Min.   : 0.00000   Min.   :0.00e+00   Min.   :0.0e+00
##  1st Qu.: 0.00000   1st Qu.:0.00e+00   1st Qu.:0.0e+00
##  Median : 0.00000   Median :0.00e+00   Median :0.0e+00
##  Mean   : 0.00156   Mean   :3.13e-03   Mean   :7.8e-04
##  3rd Qu.: 0.00000   3rd Qu.:0.00e+00   3rd Qu.:0.0e+00
##  Max.   :50.00000   Max.   :1.00e+02   Max.   :5.0e+01
```

```
# structure of the dataframe
str(data)
```

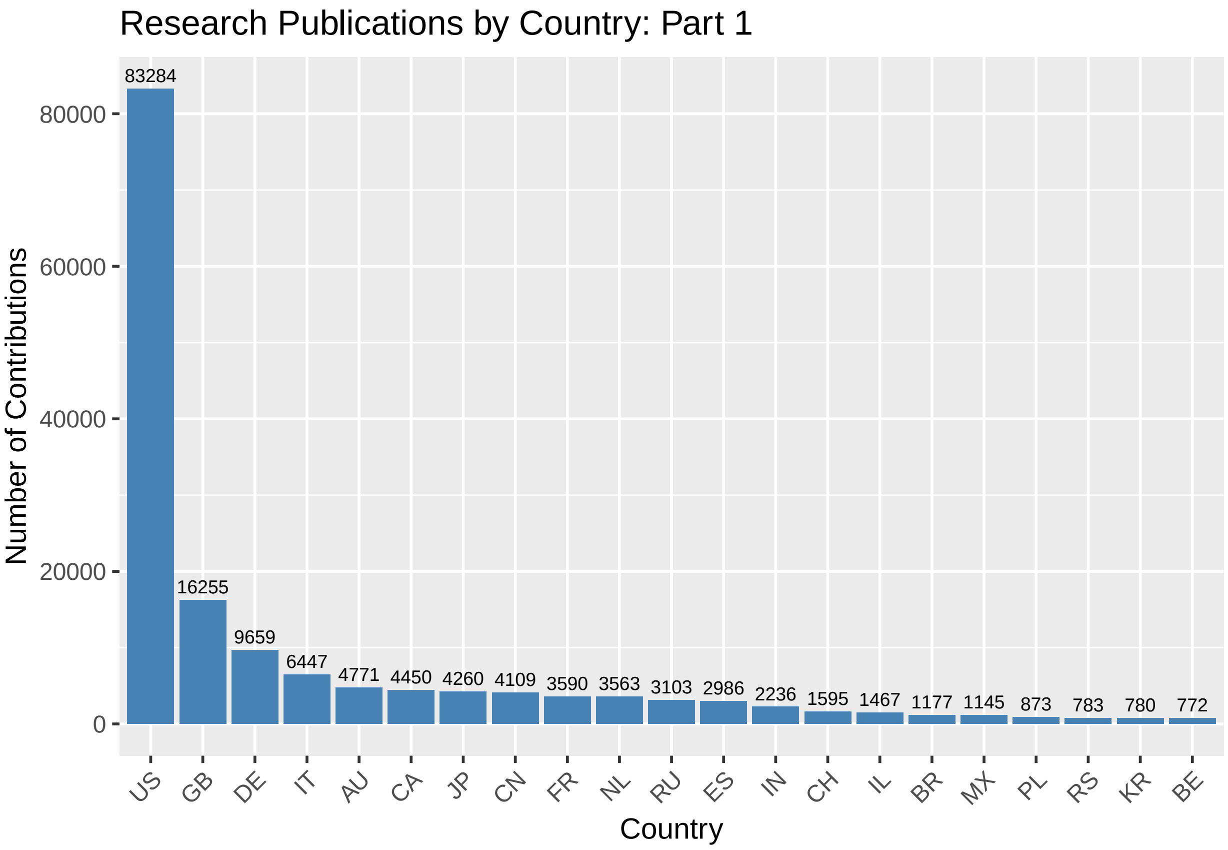
```
## 'data.frame':    63999 obs. of  110 variables:
##  $ concept_id      : chr  "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https:/
##  $ work_id         : chr  "https://openalex.org/W1993867637" "https://openalex.org/W2022503540" "htt
##  $ publication_year: int  2004 1991 2003 2003 2002 1999 2003 1997 2007 2002 ...
##  $ title           : chr  "KINEMATIC TREATMENT OF CORONAL MASS EJECTION EVOLUTION IN THE SOLAR WIND"
##  $ paperabstract   : chr  "We present a kinematic study of the evolution of coronal mass ejections (
##  $ country         : chr  "US US US US" "US" "US" "IN US" ...
##  $ year_concept    : chr  "2004+https://openalex.org/C44870925" "1991+https://openalex.org/C44870925
##  $ US              : num  100 100 100 50 100 100 50 50 0 100 ...
##  $ IN              : num  0 0 0 50 0 0 0 0 0 0 ...
##  $ RS              : num  0 0 0 0 0 0 50 0 0 0 ...
##  $ DE              : num  0 0 0 0 0 0 0 50 0 0 ...
##  $ CH              : num  0 0 0 0 0 0 0 0 100 0 ...
##  $ GB              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CN              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ IT              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RU              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CA              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NL              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AU              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DK              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CL              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JP              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MX              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PL              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ES              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ IE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FI              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ KR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TW              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AT              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ IR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HU              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ IL              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LV              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ UA              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NG              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AM              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ KZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ KH              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PT              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CO              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BW              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ZA              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SK              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PK              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ IS              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NO              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BG              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VN              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MY              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GH              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SI              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GT              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ EE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ EG              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RW              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ UG              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ UY              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JM              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SA              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RO              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ UZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TH              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NP              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TN              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LK              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LT              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CY              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ID              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ EC              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LU              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TJ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AE              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LB              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SG              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BO              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JO              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ IQ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ OM              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ KW              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PR              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TZ              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BD              : num  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```r
# Load the necessary libraries
library(ggplot2)
library(dplyr)
```

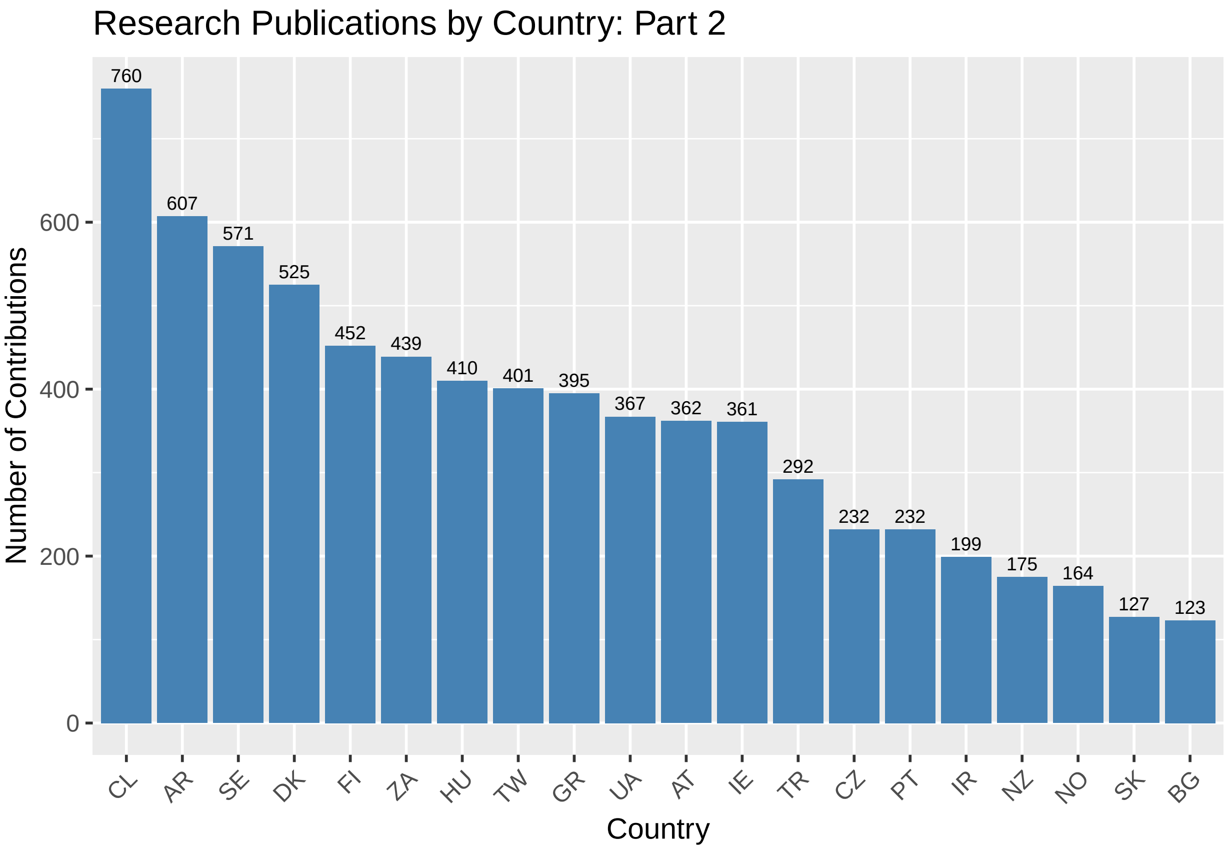
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#Histogram by number of articles each country has contributed to

# Calculate count of each country code in a vector
calculate_count <- function(vec) {
  counts <- table(vec)
  return(counts)
}

# Apply the calculate_count function to each row in the 'country' column
counts <- lapply(strsplit(data$country, "\\s+"), calculate_count)

# Get all unique countries
```

```r
all_countries <- unique(unlist(lapply(counts, names)))

# Initialize a new data frame to hold the counts
count_df <- data.frame(matrix(ncol = length(all_countries), nrow = length(counts)))
names(count_df) <- all_countries

# Fill the data frame with counts
for (i in seq_along(counts)) {
  country_names <- names(counts[[i]])
  country_counts <- counts[[i]]
  count_df[i, country_names] <- country_counts
}

# Replace NA values with 0
count_df[is.na(count_df)] <- 0

# Combine the original data with the new count-encoded country columns
data <- cbind(data, count_df)

# Combine all countries into one column for the histogram
all_countries_df <- stack(count_df)

# Rename the columns
colnames(all_countries_df) <- c("Count", "Country")

# Remove rows where count is zero
all_countries_df <- all_countries_df[all_countries_df$Count > 0,]

# Calculate the total counts for each country
all_countries_df <- all_countries_df %>%
  group_by(Country) %>%
  summarise(Total = sum(Count)) %>%
  arrange(desc(Total))

# Split the data frame into 5 equal parts
split_data <- split(all_countries_df, cut(seq(nrow(all_countries_df)), 5, labels = FALSE))

# Create a list to store the plots
plot_list <- list()

# Iterate over each subset of data and create a histogram with count
plot_list <- lapply(1:5, function(i) {
  ggplot(split_data[[i]], aes(x=reorder(Country, -Total), y=Total)) +
    geom_bar(stat="identity", fill="steelblue") +
    geom_text(aes(label=Total), vjust=-0.5, size=2.5) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x="Country", y="Number of Contributions",
         title = paste("Research Publications by Country: Part", i))
})

# View each plot by calling it from the list
plot_list[[1]]
```

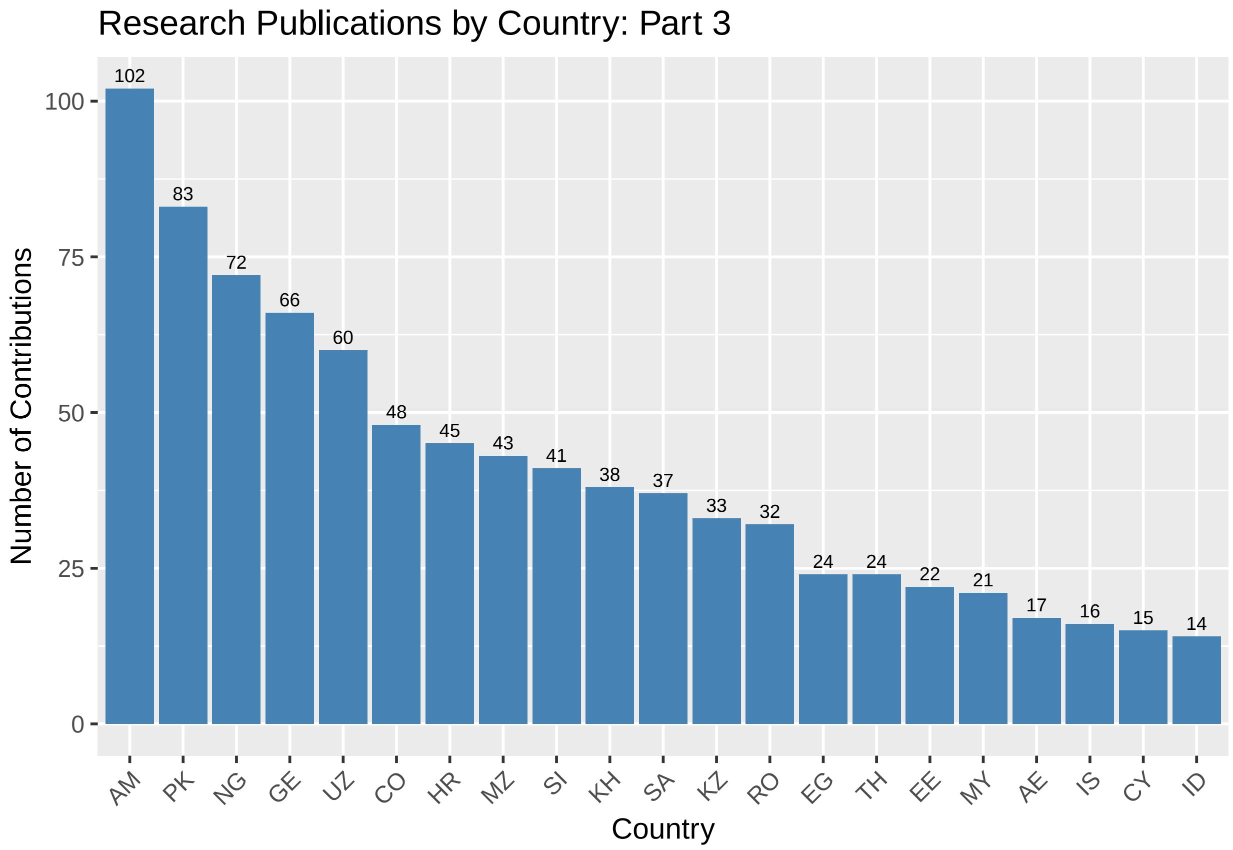

Research Publications by Country: Part 1
Number of Contributions
Country
0
20000
40000
60000
80000
83284
16255
9659
6447
4771
4450
4260
4109
3590
3563
3103
2986
2236
1595
1467
1177
1145
873
783
780
772
US
GB
DE
IT
AU
CA
JP
CN
FR
NL
RU
ES
IN
CH
IL
BR
MX
PL
RS
KR
BE

```
plot_list[[2]]
```

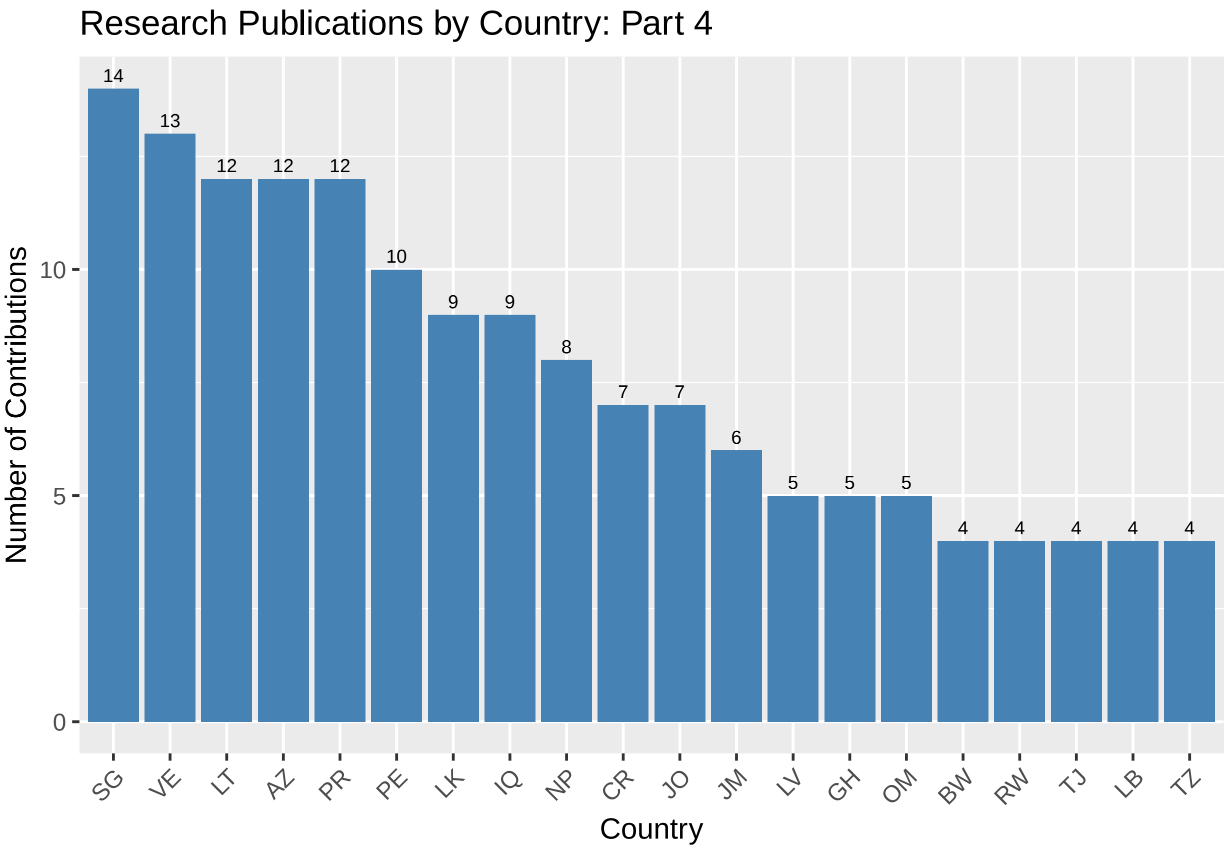


Research Publications by Country: Part 2

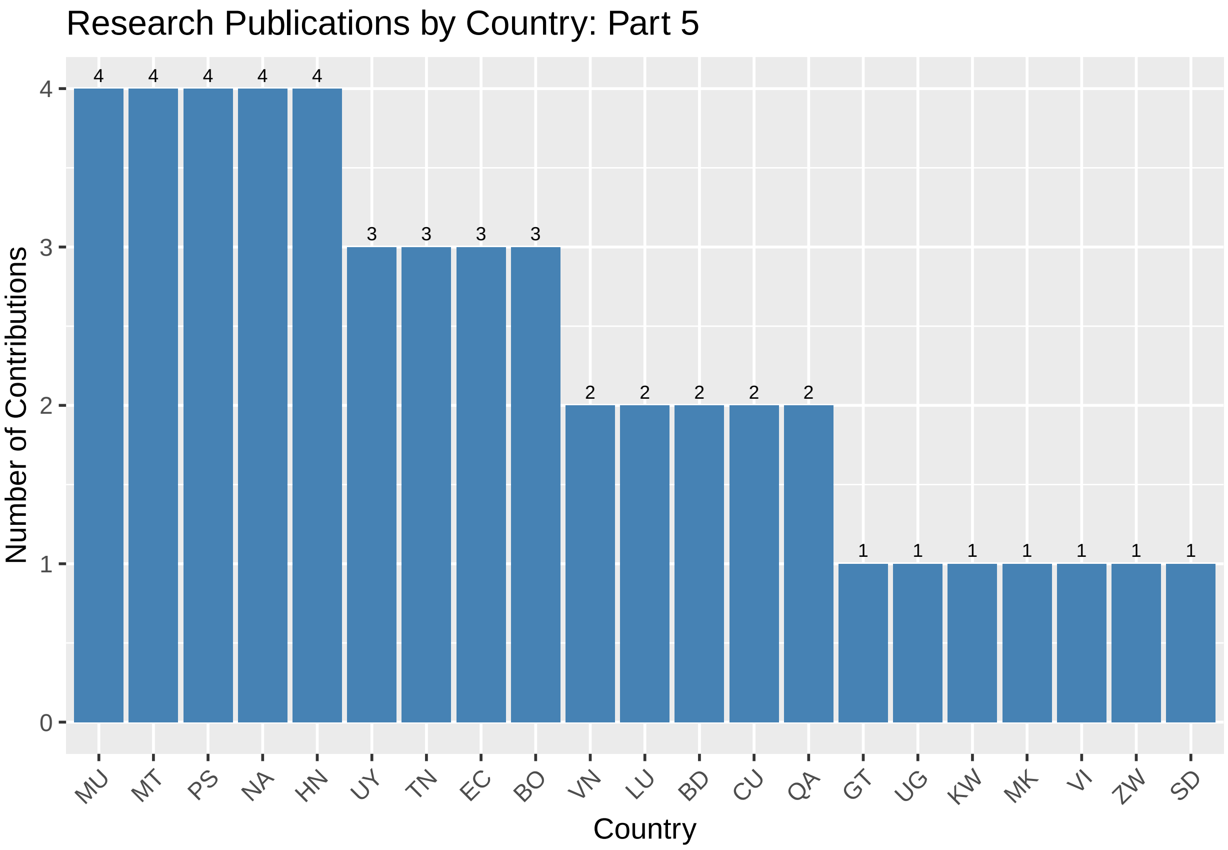
```
plot_list[[3]]
```


Research Publications by Country: Part 3
Number of Contributions
0
25
50
75
100
102
83
72
66
60
48
45
43
41
38
37
33
32
24
24
22
21
17
16
15
14
AM
PK
NG
GE
UZ
CO
HR
MZ
SI
KH
SA
KZ
RO
EG
TH
EE
MY
AE
IS
CY
ID
Country

```
plot_list[[4]]
```

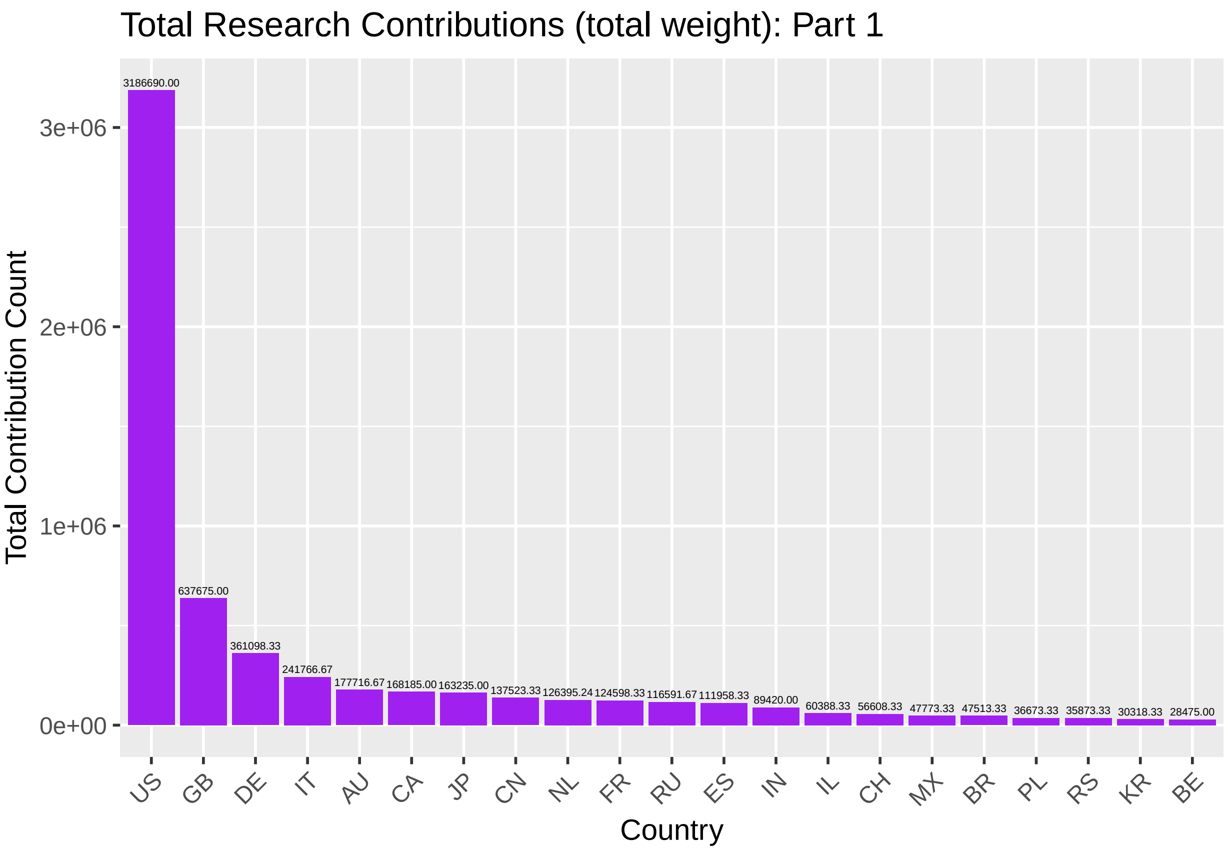

Research Publications by Country: Part 4
Number of Contributions
0
5
10
14
13
12
12
12
10
9
9
8
7
7
6
5
5
5
4
4
4
4
4
SG
VE
LT
AZ
PR
PE
LK
IQ
NP
CR
JO
JM
LV
GH
OM
BW
RW
TJ
LB
TZ
Country

```
plot_list[[5]]
```

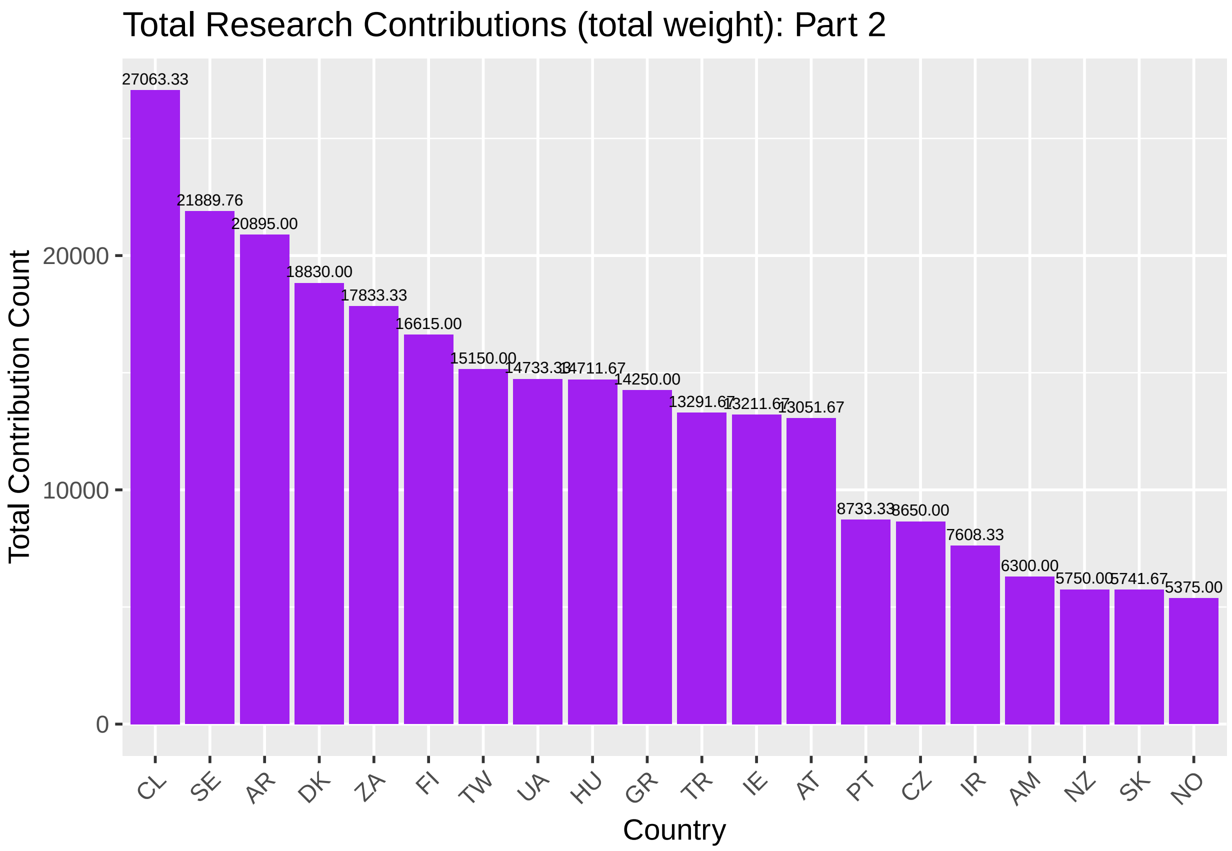


```r
#Histogram of total research contributions by country


# Combine all countries into one column for the histogram
all_countries_df <- stack(percentage_df)

# Rename the columns
colnames(all_countries_df) <- c("Percentage", "Country")

# Remove rows where percentage is zero
all_countries_df <- all_countries_df[all_countries_df$Percentage > 0,]

# Calculate the total percentage for each country
all_countries_df <- all_countries_df %>%
  group_by(Country) %>%
  summarise(Total = sum(Percentage)) %>%
  arrange(desc(Total))

# Split the data frame into 5 equal parts
split_data <- split(all_countries_df, cut(seq(nrow(all_countries_df)), 5, labels = FALSE))

# Create a list to store the plots
plot_list <- list()

# Define the text size for each plot
text_size <- c(1.5, 2, 2, 2.5, 2.5)
```

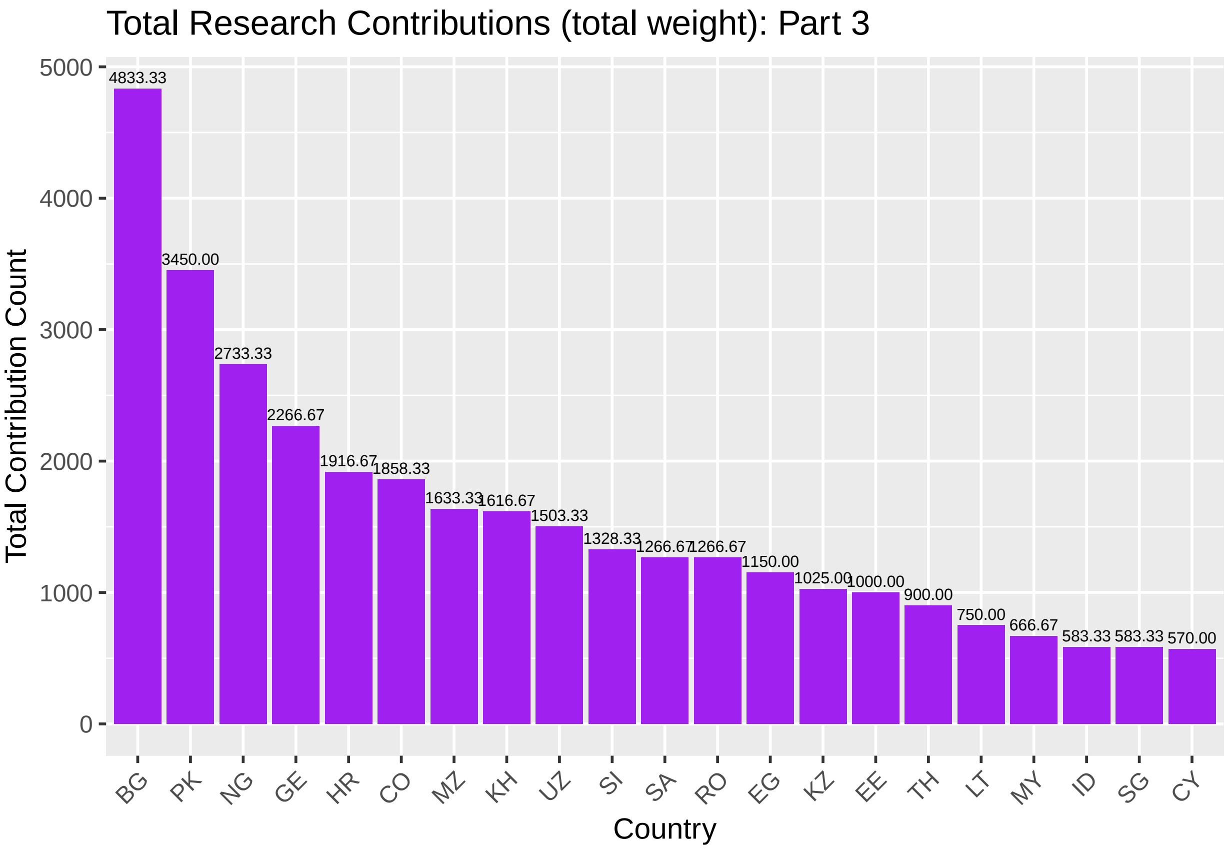
```r
# Iterate over each subset of data and create a histogram
plot_list <- lapply(1:5, function(i) {
  ggplot(split_data[[i]], aes(x=reorder(Country, -Total), y=Total)) +
    geom_bar(stat="identity", fill="purple") +
    geom_text(aes(label=sprintf("%.2f", Total)), vjust=-0.5, size=text_size[i]) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x="Country", y="Total Contribution Count",
         title = paste("Total Research Contributions (total weight): Part", i))
})

# View each plot by calling it from the list
plot_list[[1]]
```



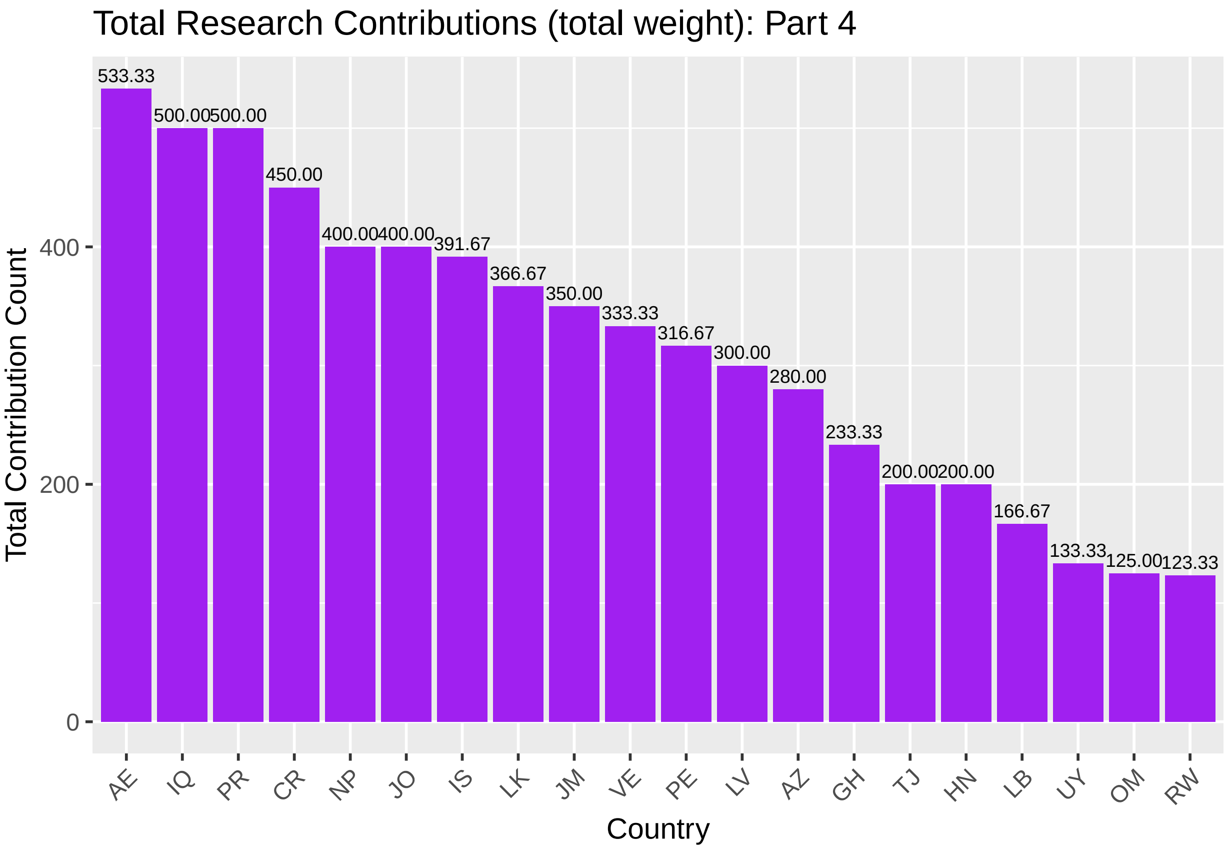
```r
plot_list[[2]]
```


Total Research Contributions (total weight): Part 2
Total Contribution Count
0
10000
20000
Country
CL
SE
AR
DK
ZA
FI
TW
UA
HU
GR
TR
IE
AT
PT
CZ
IR
AM
NZ
SK
NO
27063.33
21889.76
20895.00
18830.00
17833.33
16615.00
15150.00
14733.33
14711.67
14250.00
13291.67
13211.67
13051.67
8733.33
8650.00
7608.33
6300.00
5750.00
5741.67
5375.00

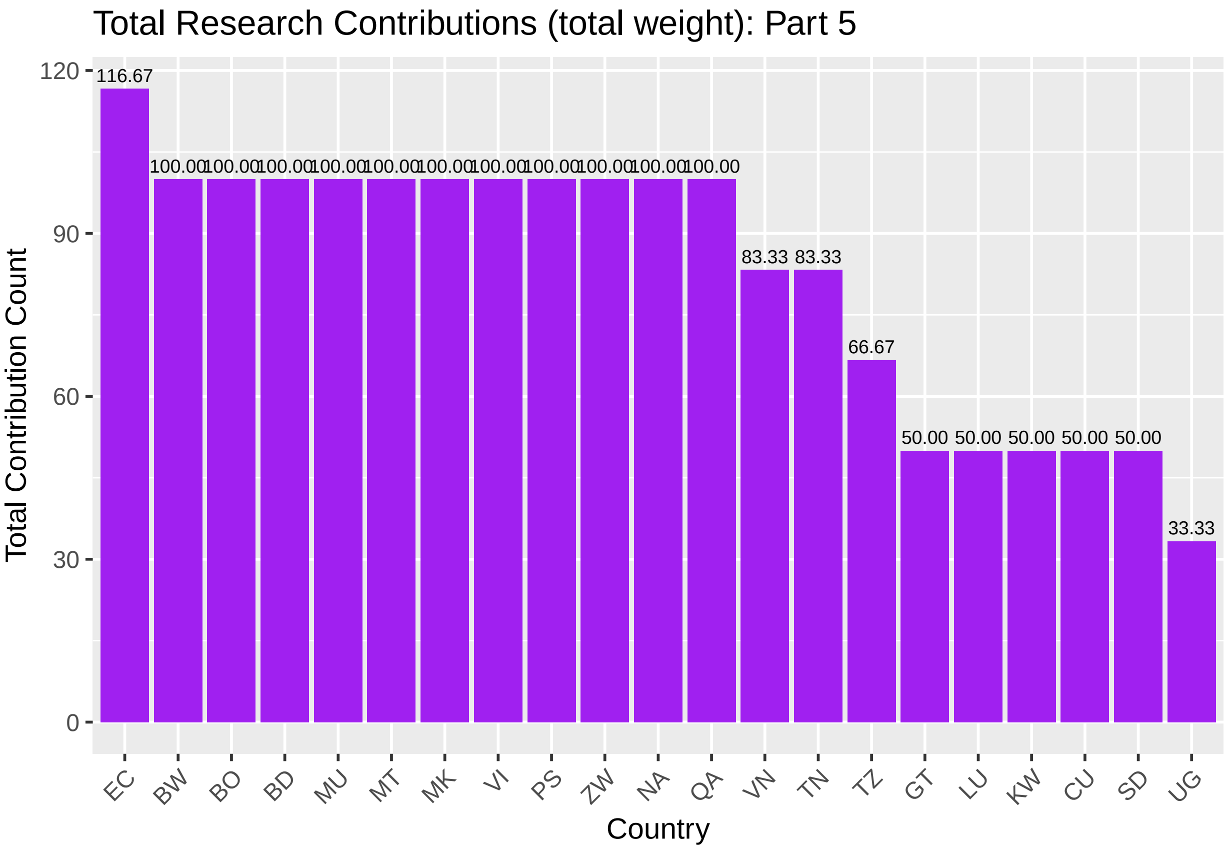
```
plot_list[[3]]
```


Total Research Contributions (total weight): Part 3
Total Contribution Count
5000
4000
3000
2000
1000
0
4833.33
3450.00
2733.33
2266.67
1916.67
1858.33
1633.33
1616.67
1503.33
1328.33
1266.67
1266.67
1150.00
1025.00
1000.00
900.00
750.00
666.67
583.33
583.33
570.00
BG
PK
NG
GE
HR
CO
MZ
KH
UZ
SI
SA
RO
EG
KZ
EE
TH
LT
MY
ID
SG
CY
Country

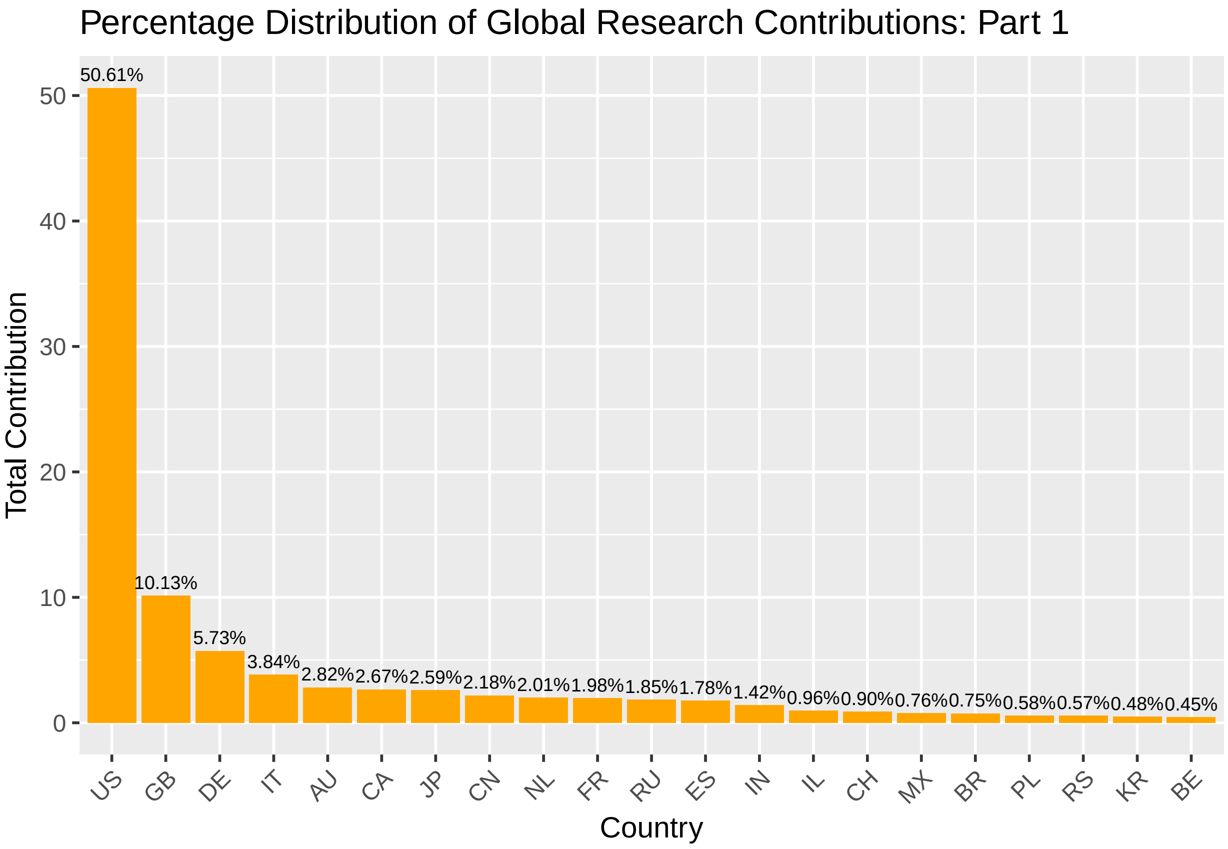
```
plot_list[[4]]
```


Total Research Contributions (total weight): Part 4
Total Contribution Count
0
200
400
Country
AE 533.33
IQ 500.00
PR 500.00
CR 450.00
NP 400.00
JO 400.00
IS 391.67
LK 366.67
JM 350.00
VE 333.33
PE 316.67
LV 300.00
AZ 280.00
GH 233.33
TJ 200.00
HN 200.00
LB 166.67
UY 133.33
OM 125.00
RW 123.33

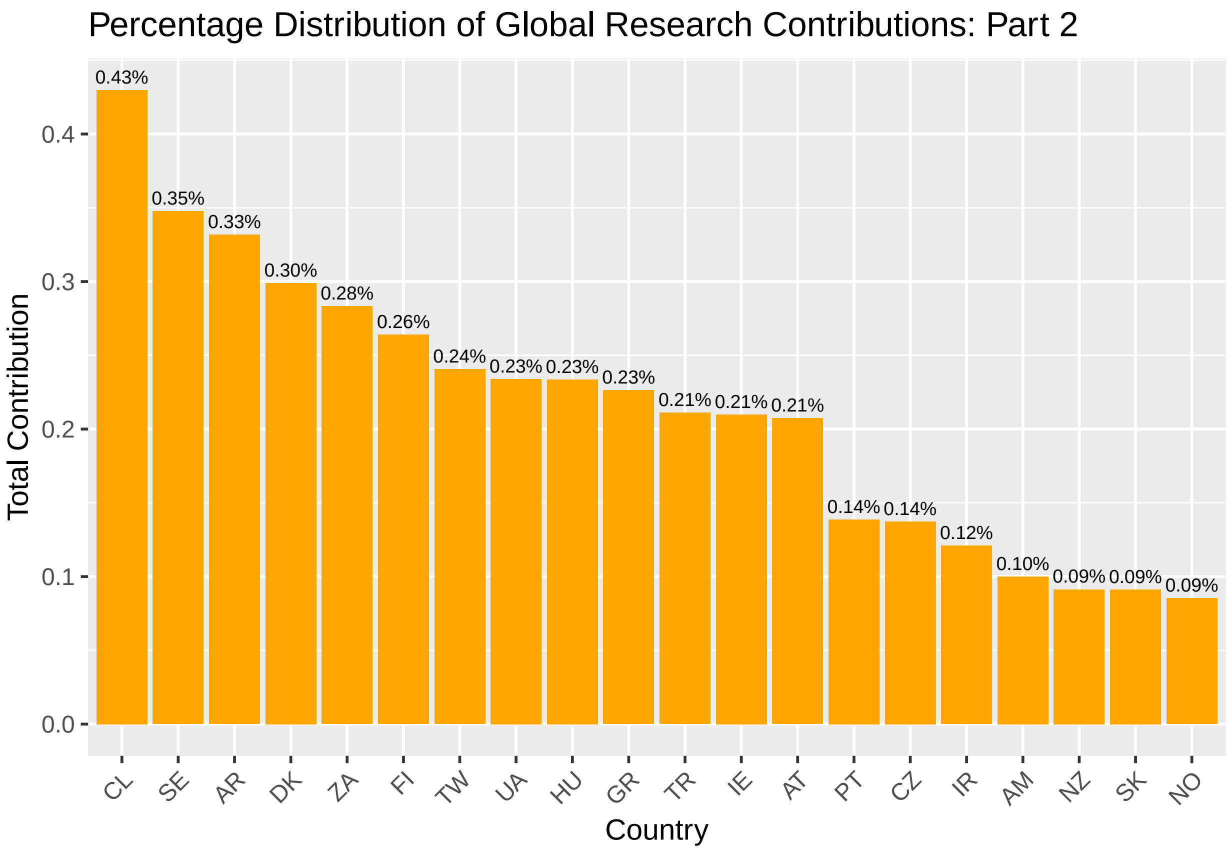
```
plot_list[[5]]
```



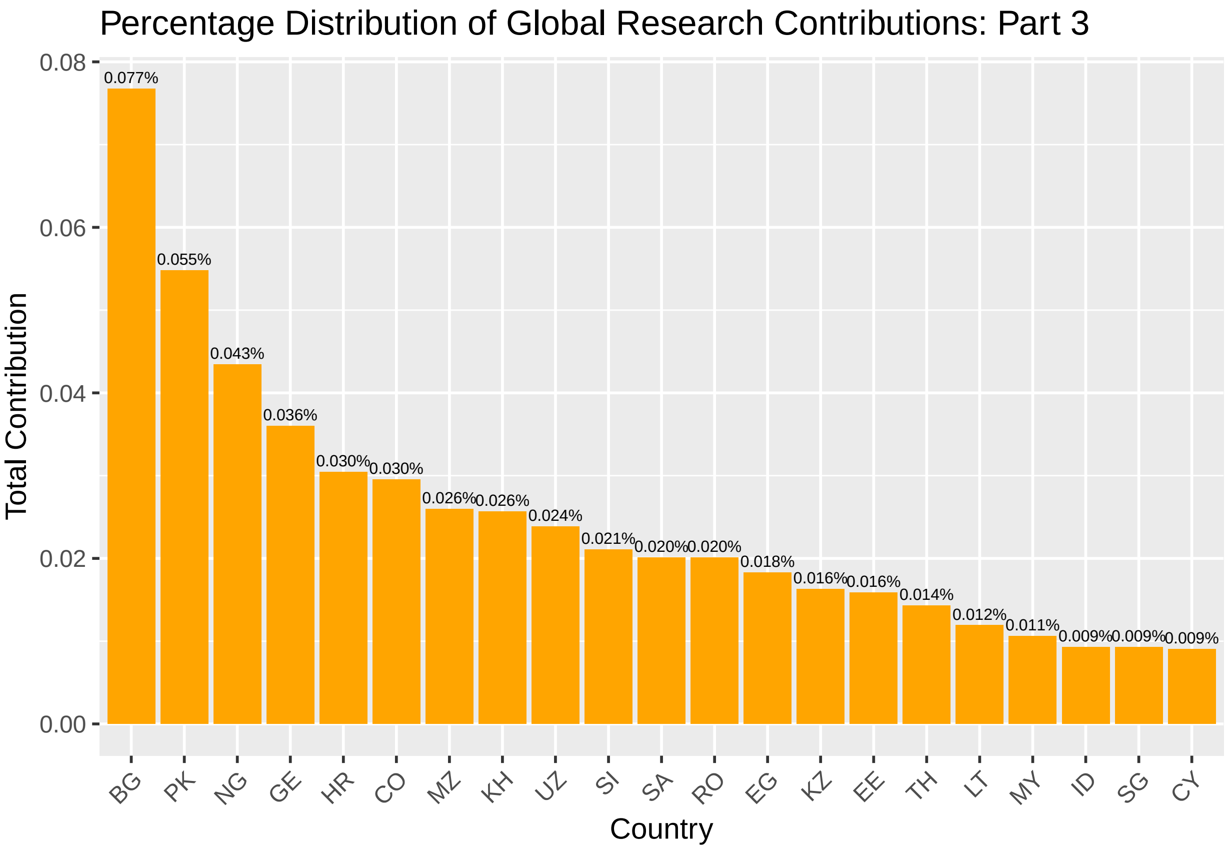
```r
# Histogram of Research Contributions % by Country

# Calculate total of all percentages
total_percentage <- sum(all_countries_df$Total)

# Express each country's contribution as a percentage of total
all_countries_df <- all_countries_df %>%
  mutate(Percentage_of_Total = Total / total_percentage * 100) %>%
  arrange(desc(Percentage_of_Total))

# Split the data frame into 5 equal parts
split_data <- split(all_countries_df, cut(seq(nrow(all_countries_df)), 5, labels = FALSE))

# Create a list to store the plots
plot_list <- list()

# Define the decimal places for each plot
formats <- c("%.2f%%", "%.2f%%", "%.3f%%", "%.4f%%", "%.4f%%")

# Define the text size for each plot
text_size <- c(2.5, 2.5, 2, 2, 2)

# Iterate over each subset of data and create a histogram
plot_list <- lapply(1:5, function(i) {
  ggplot(split_data[[i]], aes(x=reorder(Country, -Percentage_of_Total), y=Percentage_of_Total)) +
    geom_bar(stat="identity", fill="orange") +
    geom_text(aes(label=sprintf(formats[i], Percentage_of_Total)), vjust=-0.5, size=text_size[i]) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x="Country", y="Total Contribution",
        title = paste("Percentage Distribution of Global Research Contributions: Part", i))
})

# View each plot by calling it from the list
plot_list[[1]]
```


Percentage Distribution of Global Research Contributions: Part 1

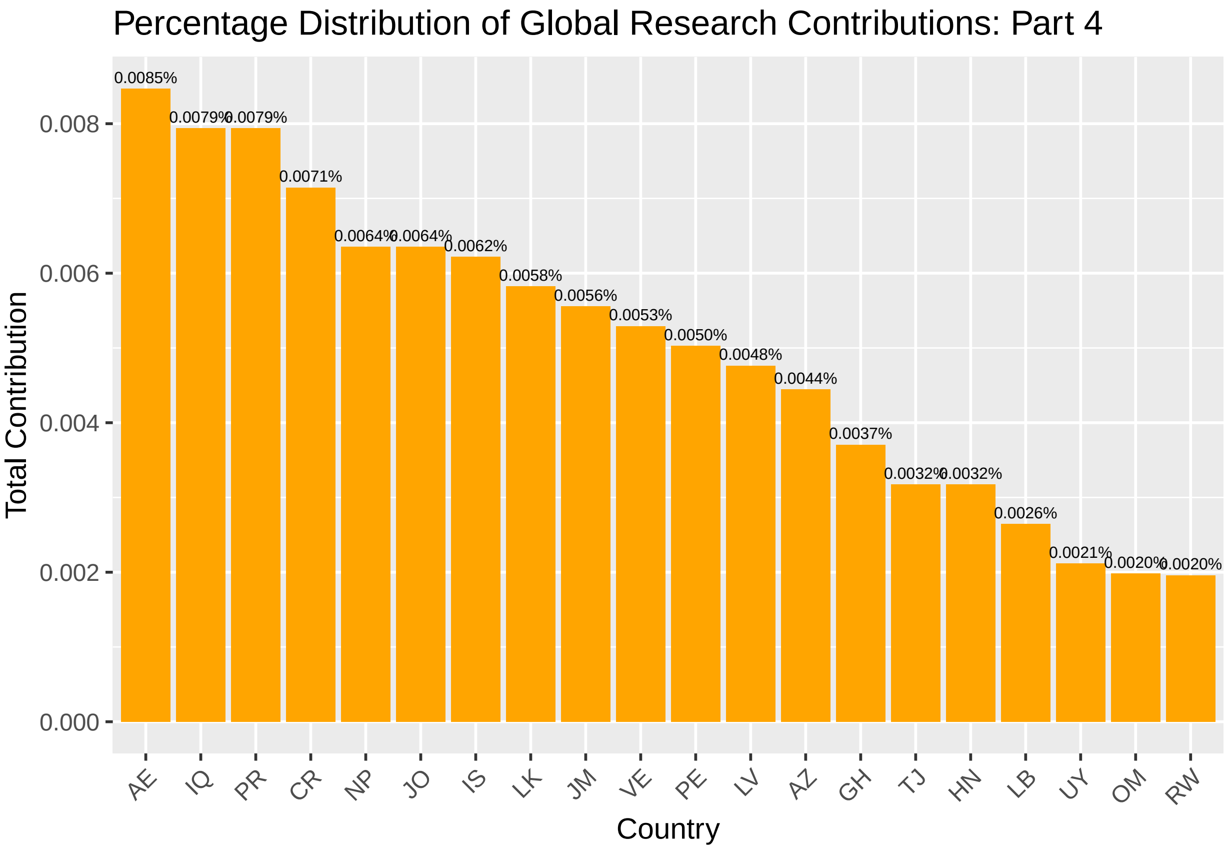
```
plot_list[[2]]
```


Percentage Distribution of Global Research Contributions: Part 2
Total Contribution
0.0
0.1
0.2
0.3
0.4
0.43%
0.35%
0.33%
0.30%
0.28%
0.26%
0.24%
0.23%
0.23%
0.23%
0.21%
0.21%
0.21%
0.14%
0.14%
0.12%
0.10%
0.09%
0.09%
0.09%
CL
SE
AR
DK
ZA
FI
TW
UA
HU
GR
TR
IE
AT
PT
CZ
IR
AM
NZ
SK
NO
Country

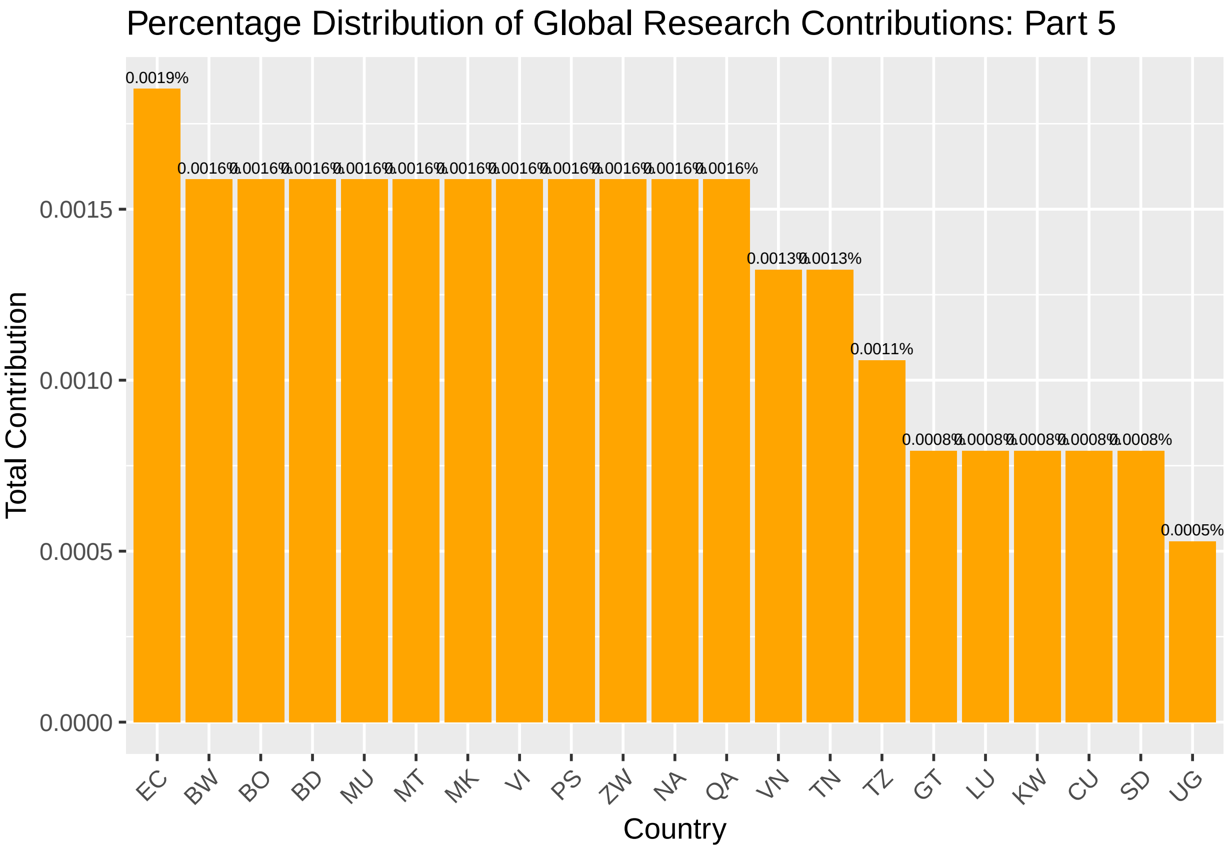
```
plot_list[[3]]
```


Percentage Distribution of Global Research Contributions: Part 3
Total Contribution
0.08
0.06
0.04
0.02
0.00
0.077%
0.055%
0.043%
0.036%
0.030%
0.030%
0.026%
0.026%
0.024%
0.021%
0.020%
0.020%
0.018%
0.016%
0.016%
0.014%
0.012%
0.011%
0.009%
0.009%
0.009%
BG
PK
NG
GE
HR
CO
MZ
KH
UZ
SI
SA
RO
EG
KZ
EE
TH
LT
MY
ID
SG
CY
Country

```
plot_list[[4]]
```


Percentage Distribution of Global Research Contributions: Part 4
Total Contribution
0.000
0.002
0.004
0.006
0.008
0.0085%
0.0079%
0.0079%
0.0071%
0.0064%
0.0064%
0.0062%
0.0058%
0.0056%
0.0053%
0.0050%
0.0048%
0.0044%
0.0037%
0.0032%
0.0032%
0.0026%
0.0021%
0.0020%
0.0020%
AE
IQ
PR
CR
NP
JO
IS
LK
JM
VE
PE
LV
AZ
GH
TJ
HN
LB
UY
OM
RW
Country

```
plot_list[[5]]
```



Percentage Distribution of Global Research Contributions: Part 5