Louis_individual_analysis

2024-10-11

```
# Load the necessary libraries
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
```

```
library(tidyr)
library(ggplot2)
```

```
# Load the datasets
df_basic <- read.csv("Individual-level basic variable.csv")
df_calculate <- read.csv("Individual-level with calculate.csv")

# View the structure of the data
str(df_basic)
```

```
## 'data.frame':    109 obs. of  36 variables:
##  $ ID                          : int  577 895 894 925 403 681 971 578 646 957 ...
##  $ Age                         : int  12 7 12 7 8 9 13 9 13 6 ...
##  $ Sex                         : chr  "M" "F" "M" "F" ...
##  $ Date                        : chr  "11-Jan" "10-Jan" "10-Jan" "10-Jan" ...
##  $ HaveYou                     : chr  "Yes" "No" "No" "No" ...
##  $ WouldYou                    : chr  "Yes" "No" "Yes" "Yes" ...
##  $ SchoolSimplified            : chr  "Public" "Public" "Public" NA ...
##  $ Neighborhood                : chr  "Sandy Bay" "Camponado/Campolancho" "Camponado/Campolancho" NA .
##  $ UserLanguage                : chr  "ES-ES" "ES-ES" "ES-ES" NA ...
##  $ CaregiverImmigrate          : chr  "Yes" "Yes" "Yes" NA ...
##  $ CaregiverImmigrateYear      : int  2017 2022 2022 NA 1989 NA NA 2017 2020 2015 ...
##  $ CaregiverImmigrateMonth     : int  2 7 7 NA 1 NA NA 2 12 11 ...
##  $ CaregiverDOB_1              : int  1994 1991 1991 NA 1985 1993 1978 1994 1970 1985 ...
##  $ CaregiverDOB_2              : int  8 4 4 NA 4 9 5 8 3 9 ...
##  $ CaregiverDOB_3              : int  21 10 10 NA 22 29 7 21 26 17 ...
##  $ CaregiverResidencyDuration: int  83 18 18 NA 420 364 548 83 37 98 ...
##  $ Partners_count              : int  5 5 3 3 2 4 7 3 2 1 ...
```

```
##  $ Partners_no_adults        : int  5 5 3 3 2 4 7 3 2 1 ...
##  $ Count_adults              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Prop_adults               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partners_male             : int  4 3 3 1 0 4 5 0 0 1 ...
##  $ Partners_female           : int  1 2 0 2 2 0 2 3 2 0 ...
##  $ Count_same_sex            : int  4 2 3 2 2 4 5 3 2 1 ...
##  $ Count_other_sex           : int  1 3 0 1 0 0 2 0 0 0 ...
##  $ Count_immediate           : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ Count_extended            : int  0 1 0 0 0 0 2 0 0 0 ...
##  $ Count_related             : int  0 1 0 0 0 0 3 0 0 1 ...
##  $ Count_unrelated           : int  5 4 3 3 2 4 4 3 2 0 ...
##  $ Partners_male_noadult     : int  4 3 3 1 0 4 5 0 0 1 ...
##  $ Partners_female_noadult   : int  1 2 0 2 2 0 2 3 2 0 ...
##  $ Count_same_sex_noadult    : int  4 2 3 2 2 4 5 3 2 1 ...
##  $ Count_other_sex_noadult   : int  1 3 0 1 0 0 2 0 0 0 ...
##  $ Count_immediate_noadult   : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ Count_extended_noadult    : int  0 1 0 0 0 0 2 0 0 0 ...
##  $ Count_related_noadult     : int  0 1 0 0 0 0 3 0 0 1 ...
##  $ Count_unrelated_noadult   : int  5 4 3 3 2 4 4 3 2 0 ...
```

```
str(df_calculate)
```

```
## 'data.frame':    109 obs. of  54 variables:
##  $ ID                              : int  577 895 894 925 403 681 971 578 646 957 ...
##  $ Age                             : int  12 7 12 7 8 9 13 9 13 6 ...
##  $ Sex                             : chr  "M" "F" "M" "F" ...
##  $ Date                            : chr  "11-Jan" "10-Jan" "10-Jan" "10-Jan" ...
##  $ HaveYou                         : chr  "Yes" "No" "No" "No" ...
##  $ WouldYou                        : chr  "Yes" "No" "Yes" "Yes" ...
##  $ OpenToNew                       : int  1 0 1 1 1 1 1 1 1 0 ...
##  $ SchoolSimplified                : chr  "Public" "Public" "Public" NA ...
##  $ Neighborhood                    : chr  "Sandy Bay" "Camponado/Campolancho" "Camponado/Campolancho"
##  $ UserLanguage                    : chr  "ES-ES" "ES-ES" "ES-ES" NA ...
##  $ CaregiverImmigrate              : chr  "Yes" "Yes" "Yes" NA ...
##  $ CaregiverImmigrateYear          : int  2017 2022 2022 NA 1989 NA NA 2017 2020 2015 ...
##  $ CaregiverImmigrateMonth         : int  2 7 7 NA 1 NA NA 2 12 11 ...
##  $ MonthsSinceCaregiverImmigration: int  83 18 18 NA 420 NA NA 83 37 98 ...
##  $ CaregiverDOB_1                  : int  1994 1991 1991 NA 1985 1993 1978 1994 1970 1985 ...
##  $ CaregiverDOB_2                  : int  8 4 4 NA 4 9 5 8 3 9 ...
##  $ CaregiverDOB_3                  : int  21 10 10 NA 22 29 7 21 26 17 ...
##  $ CaregiverResidencyDuration      : int  83 18 18 NA 420 364 548 83 37 98 ...
##  $ AverageAgeDifference            : num  -0.2 0.6 4 0.333 0 ...
##  $ VarianceAgeDifference           : num  1.7 0.3 0 2.33 0 ...
##  $ StDevAgeDifference              : num  1.304 0.548 0 1.528 0 ...
##  $ AvgAbsAgeDifference             : num  1 0.6 4 1 0 ...
##  $ VarAbsAgeDifference             : num  0.5 0.3 0 1 0 ...
##  $ StDevAbsAgeDifference           : num  0.707 0.548 0 1 0 ...
##  $ Partners_count                  : int  5 5 3 3 2 4 7 3 2 1 ...
##  $ Partners_no_adults              : int  5 5 3 3 2 4 7 3 2 1 ...
##  $ Count_adults                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Prop_adults                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partners_male                   : int  4 3 3 1 0 4 5 0 0 1 ...
##  $ Partners_female                 : int  1 2 0 2 2 0 2 3 2 0 ...
##  $ Prop_male                       : num  0.8 0.6 1 0.333 0 ...
```

```
##  $ Count_same_sex                : int  4 2 3 2 2 4 5 3 2 1 ...
##  $ Count_other_sex               : int  1 3 0 1 0 0 2 0 0 0 ...
##  $ Prop_same_sex                 : num  0.8 0.4 1 0.667 1 ...
##  $ Count_immediate               : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ Count_extended                : int  0 1 0 0 0 0 2 0 0 0 ...
##  $ Count_related                 : int  0 1 0 0 0 0 3 0 0 1 ...
##  $ Count_unrelated               : int  5 4 3 3 2 4 4 3 2 0 ...
##  $ Prop_immediate                : num  0 0 0 0 0 ...
##  $ Prop_extended                 : num  0 0.2 0 0 0 ...
##  $ Prop_related                  : num  0 0.2 0 0 0 ...
##  $ Partners_male_noadult         : int  4 3 3 1 0 4 5 0 0 1 ...
##  $ Partners_female_noadult       : int  1 2 0 2 2 0 2 3 2 0 ...
##  $ Prop_malenoadult              : num  0.8 0.6 1 0.333 0 ...
##  $ Count_same_sex_noadult        : int  4 2 3 2 2 4 5 3 2 1 ...
##  $ Count_other_sex_noadult       : int  1 3 0 1 0 0 2 0 0 0 ...
##  $ Prop_same_sex_noadult         : num  0.8 0.4 1 0.667 1 ...
##  $ Count_immediate_noadult       : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ Count_extended_noadult        : int  0 1 0 0 0 0 2 0 0 0 ...
##  $ Count_related_noadult         : int  0 1 0 0 0 0 3 0 0 1 ...
##  $ Count_unrelated_noadult       : int  5 4 3 3 2 4 4 3 2 0 ...
##  $ Prop_immediate_noadult        : num  0 0 0 0 0 ...
##  $ Prop_extended_noadult         : num  0 0.2 0 0 0 ...
##  $ Prop_related_noadult          : num  0 0.2 0 0 0 ...
```

handle the missing data

```r
# Check for missing values
colSums(is.na(df_basic))
```

```
##                         ID                        Age
##                          0                          0
##                        Sex                       Date
##                          0                          0
##                   HaveYou                  WouldYou
##                         27                         27
##           SchoolSimplified               Neighborhood
##                          7                         10
##               UserLanguage         CaregiverImmigrate
##                         10                         10
##     CaregiverImmigrateYear    CaregiverImmigrateMonth
##                         43                         78
##             CaregiverDOB_1             CaregiverDOB_2
##                         11                         11
##             CaregiverDOB_3 CaregiverResidencyDuration
##                         11                         10
##             Partners_count         Partners_no_adults
##                          0                          0
##               Count_adults               Prop_adults
##                          0                          1
##              Partners_male            Partners_female
##                          0                          0
##             Count_same_sex            Count_other_sex
```

```
##                           0                           0
##            Count_immediate             Count_extended
##                           0                           0
##              Count_related           Count_unrelated
##                           0                           0
##      Partners_male_noadult    Partners_female_noadult
##                           0                           0
##     Count_same_sex_noadult    Count_other_sex_noadult
##                           0                           0
##    Count_immediate_noadult     Count_extended_noadult
##                           0                           0
##      Count_related_noadult    Count_unrelated_noadult
##                           0                           0
```

```
colSums(is.na(df_calculate))
```

```
##                             ID                            Age
##                              0                              0
##                            Sex                           Date
##                              0                              0
##                        HaveYou                       WouldYou
##                             27                             27
##                     OpenToNew                 SchoolSimplified
##                             27                              7
##                   Neighborhood                   UserLanguage
##                             10                             10
##              CaregiverImmigrate          CaregiverImmigrateYear
##                             10                             43
##         CaregiverImmigrateMonth MonthsSinceCaregiverImmigration
##                             78                             43
##                 CaregiverDOB_1                 CaregiverDOB_2
##                             11                             11
##                 CaregiverDOB_3      CaregiverResidencyDuration
##                             11                             10
##            AverageAgeDifference           VarianceAgeDifference
##                              3                              3
##              StDevAgeDifference             AvgAbsAgeDifference
##                              3                              3
##             VarAbsAgeDifference          StDevAbsAgeDifference
##                              3                              2
##                 Partners_count              Partners_no_adults
##                              0                              0
##                   Count_adults                   Prop_adults
##                              0                              1
##                  Partners_male                Partners_female
##                              0                              0
##                       Prop_male                 Count_same_sex
##                              1                              0
##                 Count_other_sex                  Prop_same_sex
##                              0                              1
##                 Count_immediate                 Count_extended
##                              0                              0
##                   Count_related                 Count_unrelated
##                              0                              0
```

```
##                  Prop_immediate                     Prop_extended
##                               1                                 1
##                    Prop_related             Partners_male_noadult
##                               1                                 0
##         Partners_female_noadult                  Prop_malenoadult
##                               0                                 2
##          Count_same_sex_noadult           Count_other_sex_noadult
##                               0                                 0
##           Prop_same_sex_noadult           Count_immediate_noadult
##                               2                                 0
##          Count_extended_noadult             Count_related_noadult
##                               0                                 0
##         Count_unrelated_noadult            Prop_immediate_noadult
##                               0                                 2
##           Prop_extended_noadult              Prop_related_noadult
##                               2                                 2
```

```
df_basic_clean <- df_basic %>% drop_na()
df_calculate_clean <- df_calculate %>% drop_na()

summary(df_basic_clean)
```

```
##        ID              Age             Sex                Date
##  Min.   :403.0   Min.   : 5.000   Length:26          Length:26
##  1st Qu.:652.8   1st Qu.: 8.000   Class :character   Class :character
##  Median :882.5   Median : 9.500   Mode  :character   Mode  :character
##  Mean   :776.9   Mean   : 9.538
##  3rd Qu.:955.5   3rd Qu.:11.750
##  Max.   :978.0   Max.   :13.000
##    HaveYou            WouldYou         SchoolSimplified   Neighborhood
##  Length:26          Length:26          Length:26          Length:26
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  UserLanguage       CaregiverImmigrate CaregiverImmigrateYear
##  Length:26          Length:26          Min.   :1989
##  Class :character   Class :character   1st Qu.:2010
##  Mode  :character   Mode  :character   Median :2015
##                                        Mean   :2013
##                                        3rd Qu.:2018
##                                        Max.   :2022
##  CaregiverImmigrateMonth CaregiverDOB_1 CaregiverDOB_2   CaregiverDOB_3
##  Min.   : 1.000          Min.   :1970   Min.   : 1.000   Min.   : 4.00
##  1st Qu.: 2.250          1st Qu.:1982   1st Qu.: 3.250   1st Qu.:12.00
##  Median : 5.000          Median :1986   Median : 5.500   Median :17.00
##  Mean   : 5.808          Mean   :1987   Mean   : 5.654   Mean   :17.42
##  3rd Qu.: 8.500          3rd Qu.:1991   3rd Qu.: 7.750   3rd Qu.:25.50
##  Max.   :12.000          Max.   :1997   Max.   :11.000   Max.   :29.00
##  CaregiverResidencyDuration Partners_count  Partners_no_adults  Count_adults
##  Min.   : 18.00             Min.   :1.000   Min.   :1.000      Min.   :0.0000
##  1st Qu.: 70.25             1st Qu.:2.000   1st Qu.:2.000      1st Qu.:0.0000
##  Median :103.00             Median :3.000   Median :3.000      Median :0.0000
```

```
##  Mean   :122.12            Mean   :3.385   Mean   :3.154      Mean   :0.2308
##  3rd Qu.:157.00            3rd Qu.:5.000   3rd Qu.:4.750      3rd Qu.:0.0000
##  Max.   :420.00            Max.   :7.000   Max.   :6.000      Max.   :3.0000
##   Prop_adults     Partners_male   Partners_female Count_same_sex
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000
##  Median :0.0000   Median :1.000   Median :1.500   Median :2.000
##  Mean   :0.0533   Mean   :1.577   Mean   :1.808   Mean   :2.615
##  3rd Qu.:0.0000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :0.6000   Max.   :5.000   Max.   :6.000   Max.   :6.000
##  Count_other_sex  Count_immediate  Count_extended   Count_related
##  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   :0.7692   Mean   :0.5385   Mean   :0.3846   Mean   :0.9231
##  3rd Qu.:1.0000   3rd Qu.:0.7500   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :4.0000   Max.   :4.0000   Max.   :2.0000   Max.   :6.0000
##  Count_unrelated Partners_male_noadult Partners_female_noadult
##  Min.   :0.000   Min.   :0.00          Min.   :0.000
##  1st Qu.:1.000   1st Qu.:0.00          1st Qu.:0.000
##  Median :2.000   Median :1.00          Median :1.000
##  Mean   :2.462   Mean   :1.50          Mean   :1.615
##  3rd Qu.:4.000   3rd Qu.:2.75          3rd Qu.:2.000
##  Max.   :6.000   Max.   :5.00          Max.   :6.000
##  Count_same_sex_noadult Count_other_sex_noadult Count_immediate_noadult
##  Min.   :0.0            Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.0            1st Qu.:0.0000          1st Qu.:0.0000
##  Median :2.0            Median :0.0000          Median :0.0000
##  Mean   :2.5            Mean   :0.6154          Mean   :0.3462
##  3rd Qu.:4.0            3rd Qu.:1.0000          3rd Qu.:0.0000
##  Max.   :6.0            Max.   :3.0000          Max.   :2.0000
##  Count_extended_noadult Count_related_noadult Count_unrelated_noadult
##  Min.   :0.0000         Min.   :0.0000        Min.   :0.000
##  1st Qu.:0.0000         1st Qu.:0.0000        1st Qu.:1.000
##  Median :0.0000         Median :0.0000        Median :2.000
##  Mean   :0.3077         Mean   :0.6538        Mean   :2.462
##  3rd Qu.:0.7500         3rd Qu.:1.0000        3rd Qu.:4.000
##  Max.   :2.0000         Max.   :3.0000        Max.   :6.000
```

```
summary(df_calculate_clean)
```

```
##        ID              Age             Sex                Date
##  Min.   :403.0   Min.   : 5.000   Length:26          Length:26
##  1st Qu.:652.8   1st Qu.: 8.000   Class :character   Class :character
##  Median :882.5   Median : 9.500   Mode  :character   Mode  :character
##  Mean   :776.9   Mean   : 9.538
##  3rd Qu.:955.5   3rd Qu.:11.750
##  Max.   :978.0   Max.   :13.000
##    HaveYou            WouldYou           OpenToNew      SchoolSimplified
##  Length:26          Length:26          Min.   :0.0000   Length:26
##  Class :character   Class :character   1st Qu.:0.2500   Class :character
##  Mode  :character   Mode  :character   Median :1.0000   Mode  :character
##                                        Mean   :0.7308
##                                        3rd Qu.:1.0000
```

```
##                                     Max.   :1.0000
##  Neighborhood       UserLanguage       CaregiverImmigrate
##  Length:26          Length:26          Length:26
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  CaregiverImmigrateYear CaregiverImmigrateMonth MonthsSinceCaregiverImmigration
##  Min.   :1989           Min.   : 1.000          Min.   : 18.00
##  1st Qu.:2010           1st Qu.: 2.250          1st Qu.: 70.25
##  Median :2015           Median : 5.000          Median :103.00
##  Mean   :2013           Mean   : 5.808          Mean   :122.12
##  3rd Qu.:2018           3rd Qu.: 8.500          3rd Qu.:157.00
##  Max.   :2022           Max.   :12.000          Max.   :420.00
##  CaregiverDOB_1 CaregiverDOB_2   CaregiverDOB_3  CaregiverResidencyDuration
##  Min.   :1970   Min.   : 1.000   Min.   : 4.00   Min.   : 18.00
##  1st Qu.:1982   1st Qu.: 3.250   1st Qu.:12.00   1st Qu.: 70.25
##  Median :1986   Median : 5.500   Median :17.00   Median :103.00
##  Mean   :1987   Mean   : 5.654   Mean   :17.42   Mean   :122.12
##  3rd Qu.:1991   3rd Qu.: 7.750   3rd Qu.:25.50   3rd Qu.:157.00
##  Max.   :1997   Max.   :11.000   Max.   :29.00   Max.   :420.00
##  AverageAgeDifference VarianceAgeDifference StDevAgeDifference
##  Min.   :-5.000       Min.   : 0.0000       Min.   :0.0000
##  1st Qu.:-0.425       1st Qu.: 0.0000       1st Qu.:0.0000
##  Median : 0.350       Median : 0.3333       Median :0.5774
##  Mean   : 0.125       Mean   : 2.5865       Mean   :0.9696
##  3rd Qu.: 1.188       3rd Qu.: 1.3583       3rd Qu.:1.1655
##  Max.   : 4.000       Max.   :22.0000       Max.   :4.6904
##  AvgAbsAgeDifference VarAbsAgeDifference StDevAbsAgeDifference Partners_count
##  Min.   :0.0000      Min.   : 0.0000     Min.   :0.0000        Min.   :1.000
##  1st Qu.:0.6167      1st Qu.: 0.0000     1st Qu.:0.0000        1st Qu.:2.000
##  Median :1.0833      Median : 0.3333     Median :0.5774        Median :3.000
##  Mean   :1.4045      Mean   : 2.0429     Mean   :0.9104        Mean   :3.385
##  3rd Qu.:1.9500      3rd Qu.: 1.1250     3rd Qu.:1.0428        3rd Qu.:5.000
##  Max.   :5.0000      Max.   :19.0000     Max.   :4.3589        Max.   :7.000
##  Partners_no_adults  Count_adults     Prop_adults     Partners_male
##  Min.   :1.000       Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:2.000       1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :3.000       Median :0.0000   Median :0.0000   Median :1.000
##  Mean   :3.154       Mean   :0.2308   Mean   :0.0533   Mean   :1.577
##  3rd Qu.:4.750       3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:3.000
##  Max.   :6.000       Max.   :3.0000   Max.   :0.6000   Max.   :5.000
##  Partners_female   Prop_male      Count_same_sex  Count_other_sex
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1.500   Median :0.5000   Median :2.000   Median :0.0000
##  Mean   :1.808   Mean   :0.5248   Mean   :2.615   Mean   :0.7692
##  3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :1.0000   Max.   :6.000   Max.   :4.0000
##  Prop_same_sex    Count_immediate  Count_extended   Count_related
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.5000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000
```

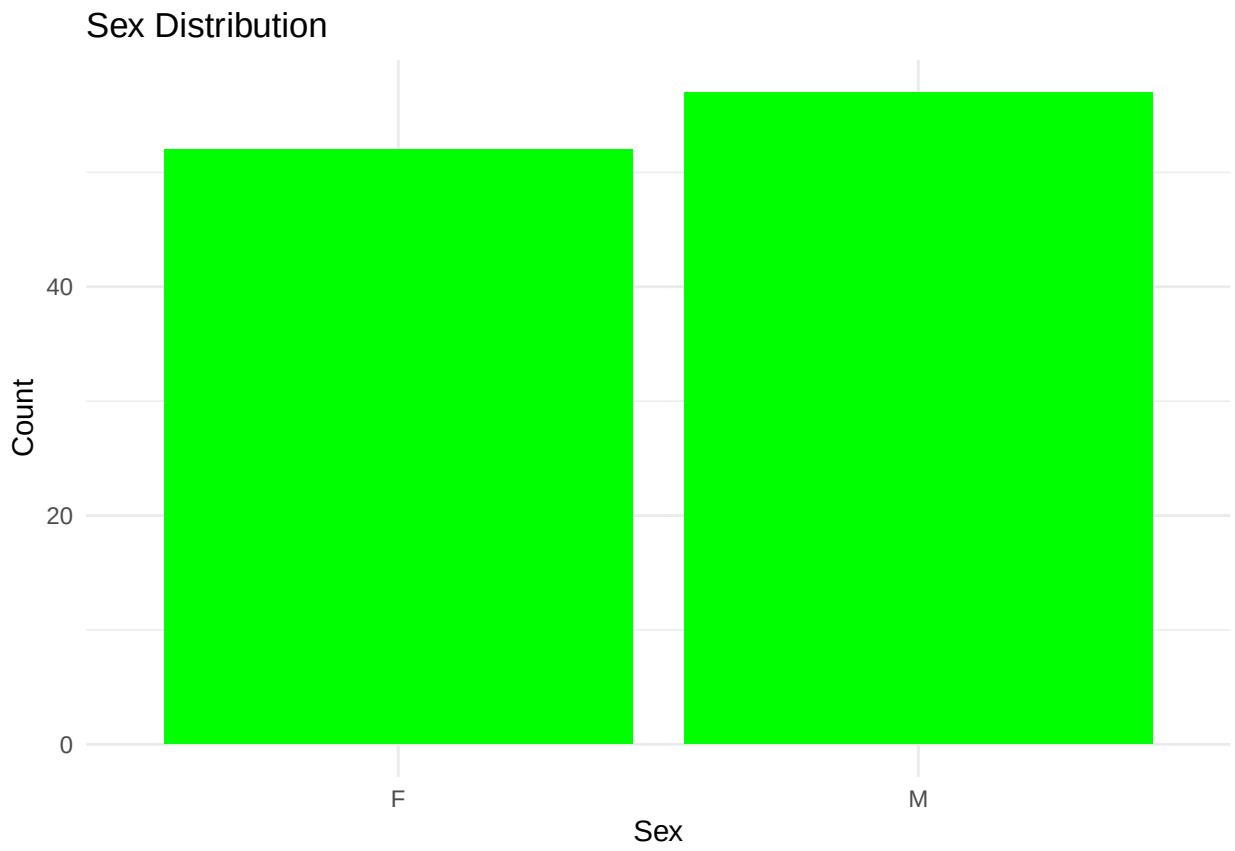
```
##  Mean   :0.7799   Mean   :0.5385   Mean   :0.3846   Mean   :0.9231
##  3rd Qu.:1.0000   3rd Qu.:0.7500   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :4.0000   Max.   :2.0000   Max.   :6.0000
##  Count_unrelated Prop_immediate   Prop_extended      Prop_related
##  Min.   :0.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :2.000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :2.462   Mean   :0.1970   Mean   :0.09496   Mean   :0.2919
##  3rd Qu.:4.000   3rd Qu.:0.1875   3rd Qu.:0.20000   3rd Qu.:0.5000
##  Max.   :6.000   Max.   :1.0000   Max.   :0.50000   Max.   :1.0000
##  Partners_male_noadult Partners_female_noadult Prop_malenoadult
##  Min.   :0.00          Min.   :0.000           Min.   :0.0000
##  1st Qu.:0.00          1st Qu.:0.000           1st Qu.:0.0000
##  Median :1.00          Median :1.000           Median :0.5500
##  Mean   :1.50          Mean   :1.615           Mean   :0.5506
##  3rd Qu.:2.75          3rd Qu.:2.000           3rd Qu.:1.0000
##  Max.   :5.00          Max.   :6.000           Max.   :1.0000
##  Count_same_sex_noadult Count_other_sex_noadult Prop_same_sex_noadult
##  Min.   :0.0            Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.0            1st Qu.:0.0000          1st Qu.:0.5000
##  Median :2.0            Median :0.0000          Median :1.0000
##  Mean   :2.5            Mean   :0.6154          Mean   :0.7673
##  3rd Qu.:4.0            3rd Qu.:1.0000          3rd Qu.:1.0000
##  Max.   :6.0            Max.   :3.0000          Max.   :1.0000
##  Count_immediate_noadult Count_extended_noadult Count_related_noadult
##  Min.   :0.0000          Min.   :0.0000         Min.   :0.0000
##  1st Qu.:0.0000          1st Qu.:0.0000         1st Qu.:0.0000
##  Median :0.0000          Median :0.0000         Median :0.0000
##  Mean   :0.3462          Mean   :0.3077         Mean   :0.6538
##  3rd Qu.:0.0000          3rd Qu.:0.7500         3rd Qu.:1.0000
##  Max.   :2.0000          Max.   :2.0000         Max.   :3.0000
##  Count_unrelated_noadult Prop_immediate_noadult Prop_extended_noadult
##  Min.   :0.000           Min.   :0.0000         Min.   :0.0000
##  1st Qu.:1.000           1st Qu.:0.0000         1st Qu.:0.0000
##  Median :2.000           Median :0.0000         Median :0.0000
##  Mean   :2.462           Mean   :0.1827         Mean   :0.1051
##  3rd Qu.:4.000           3rd Qu.:0.0000         3rd Qu.:0.1500
##  Max.   :6.000           Max.   :1.0000         Max.   :1.0000
##  Prop_related_noadult
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2878
##  3rd Qu.:0.5000
##  Max.   :1.0000
```

Age and Sex Distribution

```
ggplot(df_basic, aes(x = Age)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```

A little right skewed.

```
ggplot(df_basic, aes(x = Sex)) +
  geom_bar(fill = "green") +
  theme_minimal() +
  labs(title = "Sex Distribution", x = "Sex", y = "Count")
```



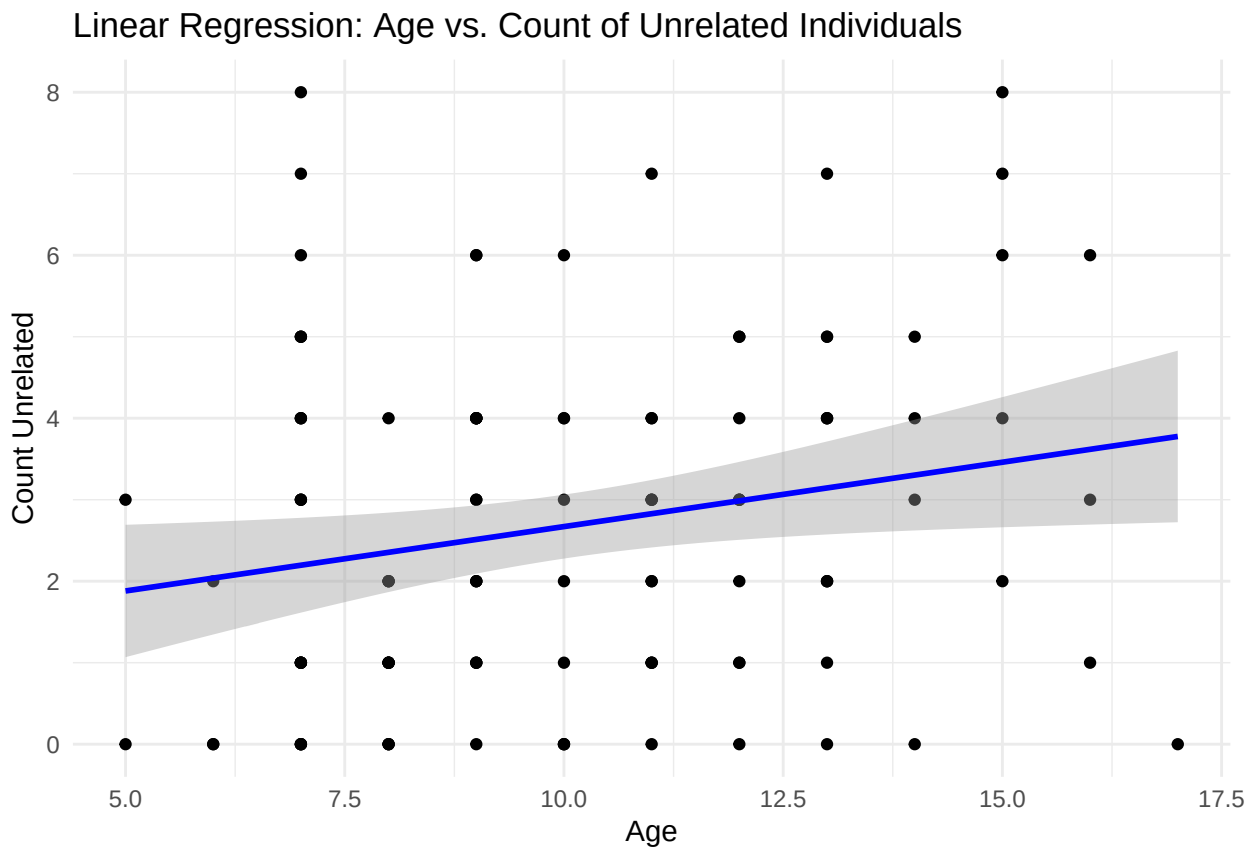
Number of female and male playmates are close.

```
# Simple linear regression using Age as the predictor for Count_unrelated
model <- lm(Count_unrelated ~ Age, data = df_basic)
summary(model)
```

```
##
## Call:
## lm(formula = Count_unrelated ~ Age, data = df_basic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7762 -1.5122 -0.3542  1.4878  5.8038
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.09026    0.74035   1.473   0.1438
## Age          0.15799    0.07095   2.227   0.0281 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.066 on 107 degrees of freedom
## Multiple R-squared:  0.04429,    Adjusted R-squared:  0.03536
## F-statistic: 4.959 on 1 and 107 DF,  p-value: 0.02806
```

```r
ggplot(df_basic, aes(x = Age, y = Count_unrelated)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  theme_minimal() +
  labs(title = "Linear Regression: Age vs. Count of Unrelated Individuals", x = "Age", y = "Count Unrela
```

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



check for other possible predictors

```r
# Select relevant numeric and factor variables for the analysis
df_basic_clean <- df_basic %>% select(Age, Sex, SchoolSimplified, Neighborhood, Count_unrelated)

# Check for correlations (only for numeric variables)
cor(df_basic_clean %>% select_if(is.numeric))
```

```
##                       Age Count_unrelated
## Age             1.0000000       0.2104495
## Count_unrelated 0.2104495       1.0000000
```

```r
summary(df_basic_clean)
```

```
##       Age             Sex            SchoolSimplified   Neighborhood
##  Min.   : 5.00   Length:109         Length:109         Length:109
##  1st Qu.: 8.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.06
##  3rd Qu.:12.00
##  Max.   :17.00
##  Count_unrelated
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.679
##  3rd Qu.:4.000
##  Max.   :8.000
```

```r
# Convert categorical variables to factors
df_basic_clean$Sex <- as.factor(df_basic_clean$Sex)
df_basic_clean$SchoolSimplified <- as.factor(df_basic_clean$SchoolSimplified)
df_basic_clean$Neighborhood <- as.factor(df_basic_clean$Neighborhood)

# Fit a multiple linear regression model
model_multi <- lm(Count_unrelated ~ Age + Sex + SchoolSimplified + Neighborhood, data = df_basic_clean)

summary(model_multi)
```

```
##
## Call:
## lm(formula = Count_unrelated ~ Age + Sex + SchoolSimplified +
##     Neighborhood, data = df_basic_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6477 -1.3565  0.0000  0.9232  4.6435
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -1.4360     1.5301  -0.938  0.35092
## Age                                 0.2559     0.0811   3.155  0.00229 **
## SexM                                0.2150     0.4422   0.486  0.62818
## SchoolSimplifiedAdventist          -0.3435     2.1184  -0.162  0.87162
## SchoolSimplifiedKinder              1.4159     1.5870   0.892  0.37510
## SchoolSimplifiedMethodist           1.2511     0.8658   1.445  0.15252
## SchoolSimplifiedMethodist           2.2434     1.6880   1.329  0.18776
## SchoolSimplifiedOther               2.6737     1.8552   1.441  0.15359
## SchoolSimplifiedPublic              0.2925     0.8238   0.355  0.72349
## NeighborhoodCamponado/Campolancho   1.4936     1.0213   1.462  0.14769
## NeighborhoodCentro                 -0.9955     1.4552  -0.684  0.49598
## NeighborhoodCola de Mica/Lozano    -0.2101     1.0643  -0.197  0.84400
## NeighborhoodCountryside             6.3936     2.2166   2.884  0.00508 **
## NeighborhoodDonkey City             2.1313     2.2225   0.959  0.34058
## NeighborhoodJericho                -1.6064     2.2166  -0.725  0.47081
```

```
## NeighborhoodLa Loma                 1.0962     2.1783   0.503  0.61623
## NeighborhoodMonte Fresco            0.1135     1.6860   0.067  0.94650
## NeighborhoodPumpkin Hill           -2.5467     1.6609  -1.533  0.12931
## NeighborhoodRocky Hill              1.7142     1.1838   1.448  0.15166
## NeighborhoodSandy Bay               0.4701     1.0223   0.460  0.64691
## NeighborhoodThe Point               2.0171     1.3136   1.536  0.12873
## NeighborhoodTreasure Hill              NA         NA      NA       NA
## NeighborhoodWestern Path            3.3935     2.2166   1.531  0.12987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.93 on 77 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.3556, Adjusted R-squared:  0.1798
## F-statistic: 2.023 on 21 and 77 DF,  p-value: 0.01373
```

1. Age is statistically significant.
2. Neighborhood: Children living in the Countryside are expected to know around 6.39 more unrelated individuals than children in the reference neighborhood, indicating that the Countryside neighborhood has a strong positive effect on the number of unrelated playmates.

improve the model

```r
# Fit a refined model with only Age and Neighborhood as predictors
model_refined <- lm(Count_unrelated ~ Age + Neighborhood, data = df_basic_clean)

summary(model_refined)
```

```
##
## Call:
## lm(formula = Count_unrelated ~ Age + Neighborhood, data = df_basic_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5974 -1.3147 -0.0098  1.0911  4.7195
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.35814    1.30588  -0.274  0.78457
## Age                                0.24673    0.07686   3.210  0.00189 **
## NeighborhoodCamponado/Campolancho  0.94571    0.95924   0.986  0.32705
## NeighborhoodCentro                -1.60589    1.29556  -1.240  0.21864
## NeighborhoodCola de Mica/Lozano   -0.07539    1.01258  -0.074  0.94083
## NeighborhoodCountryside            6.63103    2.15026   3.084  0.00277 **
## NeighborhoodDonkey City            2.65719    2.11033   1.259  0.21151
## NeighborhoodJericho               -1.36897    2.15026  -0.637  0.52610
## NeighborhoodLa Loma                0.38430    2.13565   0.180  0.85763
## NeighborhoodMonte Fresco          -0.35589    1.61255  -0.221  0.82587
## NeighborhoodPumpkin Hill          -3.09608    1.60925  -1.924  0.05779 .
## NeighborhoodRocky Hill             1.35466    1.14691   1.181  0.24092
## NeighborhoodSandy Bay              0.10606    0.97489   0.109  0.91363
```

```
## NeighborhoodThe Point              1.48823    1.24560   1.195  0.23557
## NeighborhoodTreasure Hill          1.88430    1.64842   1.143  0.25628
## NeighborhoodWestern Path           3.63103    2.15026   1.689  0.09504 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.92 on 83 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.3125, Adjusted R-squared:  0.1883
## F-statistic: 2.515 on 15 and 83 DF,  p-value: 0.004103
```

```r
# Fit a model with interaction between Age and Neighborhood
model_interaction <- lm(Count_unrelated ~ Age * Neighborhood, data = df_basic_clean)

# Summarize the model with interaction
summary(model_interaction)
```

```
##
## Call:
## lm(formula = Count_unrelated ~ Age * Neighborhood, data = df_basic_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5000 -1.2442  0.0000  0.9684  4.7558
##
## Coefficients: (5 not defined because of singularities)
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           1.78947    5.28149   0.339   0.7357
## Age                                   0.07895    0.40699   0.194   0.8467
## NeighborhoodCamponado/Campolancho    -1.46853    5.46418  -0.269   0.7889
## NeighborhoodCentro                   -4.70614    7.36382  -0.639   0.5247
## NeighborhoodCola de Mica/Lozano      -4.00798    5.64663  -0.710   0.4801
## NeighborhoodCountryside               5.65789    3.17871   1.780   0.0792 .
## NeighborhoodDonkey City               3.02632    2.30948   1.310   0.1941
## NeighborhoodJericho                  -2.34211    3.17871  -0.737   0.4636
## NeighborhoodLa Loma                  -0.42105    2.88936  -0.146   0.8845
## NeighborhoodMonte Fresco             11.21053    9.32225   1.203   0.2330
## NeighborhoodPumpkin Hill             -1.78947    8.42057  -0.213   0.8323
## NeighborhoodRocky Hill               -3.28947    5.90657  -0.557   0.5793
## NeighborhoodSandy Bay                -1.36640    5.50592  -0.248   0.8047
## NeighborhoodThe Point                 4.34566    7.08857   0.613   0.5417
## NeighborhoodTreasure Hill             0.37719    6.57348   0.057   0.9544
## NeighborhoodWestern Path              2.65789    3.17871   0.836   0.4058
## Age:NeighborhoodCamponado/Campolancho 0.19580    0.43125   0.454   0.6511
## Age:NeighborhoodCentro                0.25439    0.61274   0.415   0.6792
## Age:NeighborhoodCola de Mica/Lozano   0.34393    0.44950   0.765   0.4466
## Age:NeighborhoodCountryside               NA         NA      NA       NA
## Age:NeighborhoodDonkey City               NA         NA      NA       NA
## Age:NeighborhoodJericho                   NA         NA      NA       NA
## Age:NeighborhoodLa Loma                   NA         NA      NA       NA
## Age:NeighborhoodMonte Fresco         -1.07895    0.79858  -1.351   0.1808
## Age:NeighborhoodPumpkin Hill         -0.07895    0.61274  -0.129   0.8978
## Age:NeighborhoodRocky Hill            0.42105    0.48173   0.874   0.3849
## Age:NeighborhoodSandy Bay             0.10214    0.43223   0.236   0.8138
```

```
## Age:NeighborhoodThe Point             -0.37624    0.64871  -0.580   0.5637
## Age:NeighborhoodTreasure Hill          0.08772    0.61274   0.143   0.8866
## Age:NeighborhoodWestern Path                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.943 on 74 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.3722, Adjusted R-squared:  0.1686
## F-statistic: 1.828 on 24 and 74 DF,  p-value: 0.02575
```

It failed to improve the situation.