EDA LA-1 REPORT

AMOGH JAVALI USN : 1NT21IS026

2024-01-26

HOTEL BOOKINGS DEMAND DATASET

Load a dataset:

```
my_data <- read.csv("C:/Users/amogh/Desktop/Dataset/amogh.csv")
```

Packages:

```
install.packages("base")
```

```
## Warning: package 'base' is in use and will not be installed
```

```
library(base)
```

Help and Documentation:

```
?mean
```

```
## starting httpd help server ... done
```

View the first few rows of the dataset:

```
head(my_data)
```

```
##   IsCanceled LeadTime ArrivalDateYear ArrivalDateMonth ArrivalDateWeekNumber
## 1          0      342            2015             July                    27
## 2          0      737            2015             July                    27
## 3          0        7            2015             July                    27
## 4          0       13            2015             July                    27
## 5          0       14            2015             July                    27
## 6          0       14            2015             July                    27
##   ArrivalDateDayOfMonth StaysInWeekendNights StaysInWeekNights Adults Children
## 1                     1                    0                 0      2        0
## 2                     1                    0                 0      2        0
## 3                     1                    0                 1      1        0
## 4                     1                    0                 1      1        0
## 5                     1                    0                 2      2        0
## 6                     1                    0                 2      2        0
##   Babies      Meal Country MarketSegment DistributionChannel IsRepeatedGuest
## 1      0 BB            PRT        Direct              Direct               0
## 2      0 BB            PRT        Direct              Direct               0
## 3      0 BB            GBR        Direct              Direct               0
## 4      0 BB            GBR     Corporate           Corporate               0
## 5      0 BB            GBR     Online TA               TA/TO               0
## 6      0 BB            GBR     Online TA               TA/TO               0
##   PreviousCancellations PreviousBookingsNotCanceled ReservedRoomType
## 1                     0                           0 C
## 2                     0                           0 C
## 3                     0                           0 A
## 4                     0                           0 A
## 5                     0                           0 A
## 6                     0                           0 A
##   AssignedRoomType BookingChanges     DepositType       Agent     Company
## 1 C                             3 No Deposit             NULL        NULL
## 2 C                             4 No Deposit             NULL        NULL
## 3 C                             0 No Deposit             NULL        NULL
## 4 A                             0 No Deposit              304        NULL
## 5 A                             0 No Deposit              240        NULL
## 6 A                             0 No Deposit              240        NULL
##   DaysInWaitingList CustomerType ADR RequiredCarParkingSpaces
## 1                 0    Transient   0                        0
## 2                 0    Transient   0                        0
## 3                 0    Transient  75                        0
## 4                 0    Transient  75                        0
## 5                 0    Transient  98                        0
## 6                 0    Transient  98                        0
##   TotalOfSpecialRequests ReservationStatus ReservationStatusDate
## 1                      0         Check-Out            01-07-2015
## 2                      0         Check-Out            01-07-2015
## 3                      0         Check-Out            02-07-2015
## 4                      0         Check-Out            02-07-2015
## 5                      1         Check-Out            03-07-2015
## 6                      1         Check-Out            03-07-2015
```

View the structure of the dataset:

```
str(my_data)
```

```
## ’data.frame’:    40060 obs. of  31 variables:
##  $ IsCanceled                 : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ LeadTime                   : int  342 737 7 13 14 14 0 9 85 75 ...
##  $ ArrivalDateYear            : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ ArrivalDateMonth           : chr  "July" "July" "July" "July" ...
##  $ ArrivalDateWeekNumber      : int  27 27 27 27 27 27 27 27 27 27 ...
##  $ ArrivalDateDayOfMonth      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ StaysInWeekendNights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ StaysInWeekNights          : int  0 0 1 1 2 2 2 2 3 3 ...
##  $ Adults                     : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ Children                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Babies                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Meal                       : chr  "BB       " "BB       " "BB       " "BB       " ...
##  $ Country                    : chr  "PRT" "PRT" "GBR" "GBR" ...
##  $ MarketSegment              : chr  "Direct" "Direct" "Direct" "Corporate" ...
##  $ DistributionChannel        : chr  "Direct" "Direct" "Direct" "Corporate" ...
##  $ IsRepeatedGuest            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PreviousCancellations      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PreviousBookingsNotCanceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ReservedRoomType           : chr  "C               " "C               " "A               " "A
##  $ AssignedRoomType           : chr  "C               " "C               " "C               " "A
##  $ BookingChanges             : int  3 4 0 0 0 0 0 0 0 0 ...
##  $ DepositType                : chr  "No Deposit     " "No Deposit     " "No Deposit     " "No Deposi
##  $ Agent                      : chr  "       NULL" "       NULL" "       NULL" "304" ...
##  $ Company                    : chr  "       NULL" "       NULL" "       NULL" "       NULL" ...
##  $ DaysInWaitingList          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CustomerType               : chr  "Transient" "Transient" "Transient" "Transient" ...
##  $ ADR                        : num  0 0 75 75 98 ...
##  $ RequiredCarParkingSpaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ TotalOfSpecialRequests     : int  0 0 0 0 1 1 0 1 1 0 ...
##  $ ReservationStatus          : chr  "Check-Out" "Check-Out" "Check-Out" "Check-Out" ...
##  $ ReservationStatusDate      : chr  "01-07-2015" "01-07-2015" "02-07-2015" "02-07-2015" ...
```

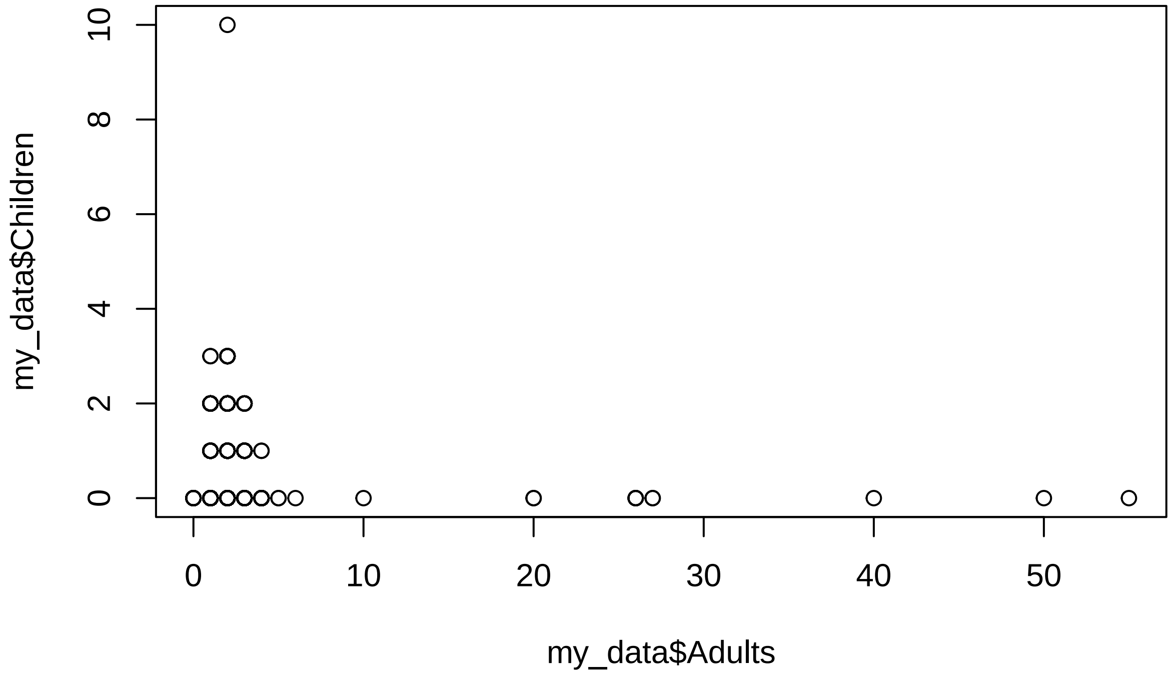
Summary statistics:

```
summary(my_data)
```

```
##    IsCanceled        LeadTime      ArrivalDateYear ArrivalDateMonth
##  Min.   :0.0000   Min.   :  0.00   Min.   :2015    Length:40060
##  1st Qu.:0.0000   1st Qu.: 10.00   1st Qu.:2016    Class :character
##  Median :0.0000   Median : 57.00   Median :2016    Mode  :character
##  Mean   :0.2776   Mean   : 92.68   Mean   :2016
##  3rd Qu.:1.0000   3rd Qu.:155.00   3rd Qu.:2017
##  Max.   :1.0000   Max.   :737.00   Max.   :2017
##  ArrivalDateWeekNumber ArrivalDateDayOfMonth StaysInWeekendNights
##  Min.   : 1.00         Min.   : 1.00         Min.   : 0.00
##  1st Qu.:16.00         1st Qu.: 8.00         1st Qu.: 0.00
##  Median :28.00         Median :16.00         Median : 1.00
##  Mean   :27.14         Mean   :15.82         Mean   : 1.19
##  3rd Qu.:38.00         3rd Qu.:24.00         3rd Qu.: 2.00
##  Max.   :53.00         Max.   :31.00         Max.   :19.00
##  StaysInWeekNights     Adults          Children           Babies
##  Min.   : 0.000    Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 1.000    1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 3.000    Median : 2.000   Median : 0.0000   Median :0.0000
##  Mean   : 3.129    Mean   : 1.867   Mean   : 0.1287   Mean   :0.0139
##  3rd Qu.: 5.000    3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :50.000    Max.   :55.000   Max.   :10.0000   Max.   :2.0000
##      Meal             Country          MarketSegment      DistributionChannel
##  Length:40060       Length:40060       Length:40060       Length:40060
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  IsRepeatedGuest   PreviousCancellations PreviousBookingsNotCanceled
##  Min.   :0.00000   Min.   : 0.0000       Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 0.0000       1st Qu.: 0.0000
##  Median :0.00000   Median : 0.0000       Median : 0.0000
##  Mean   :0.04438   Mean   : 0.1017       Mean   : 0.1465
##  3rd Qu.:0.00000   3rd Qu.: 0.0000       3rd Qu.: 0.0000
##  Max.   :1.00000   Max.   :26.0000       Max.   :30.0000
##  ReservedRoomType   AssignedRoomType   BookingChanges   DepositType
##  Length:40060       Length:40060       Min.   : 0.000   Length:40060
##  Class :character   Class :character   1st Qu.: 0.000   Class :character
##  Mode  :character   Mode  :character   Median : 0.000   Mode  :character
##                                        Mean   : 0.288
##                                        3rd Qu.: 0.000
##                                        Max.   :17.000
##     Agent             Company          DaysInWaitingList  CustomerType
##  Length:40060       Length:40060       Min.   :  0.0000   Length:40060
##  Class :character   Class :character   1st Qu.:  0.0000   Class :character
##  Mode  :character   Mode  :character   Median :  0.0000   Mode  :character
##                                        Mean   :  0.5278
##                                        3rd Qu.:  0.0000
##                                        Max.   :185.0000
##       ADR         RequiredCarParkingSpaces TotalOfSpecialRequests
##  Min.   : -6.38   Min.   :0.0000           Min.   :0.0000
##  1st Qu.: 50.00   1st Qu.:0.0000           1st Qu.:0.0000
##  Median : 75.00   Median :0.0000           Median :0.0000
##  Mean   : 94.95   Mean   :0.1381           Mean   :0.6198
##  3rd Qu.:125.00   3rd Qu.:0.0000           3rd Qu.:1.0000
##  Max.   :508.00   Max.   :8.0000           Max.   :5.0000
##  ReservationStatus  ReservationStatusDate
##  Length:40060       Length:40060
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

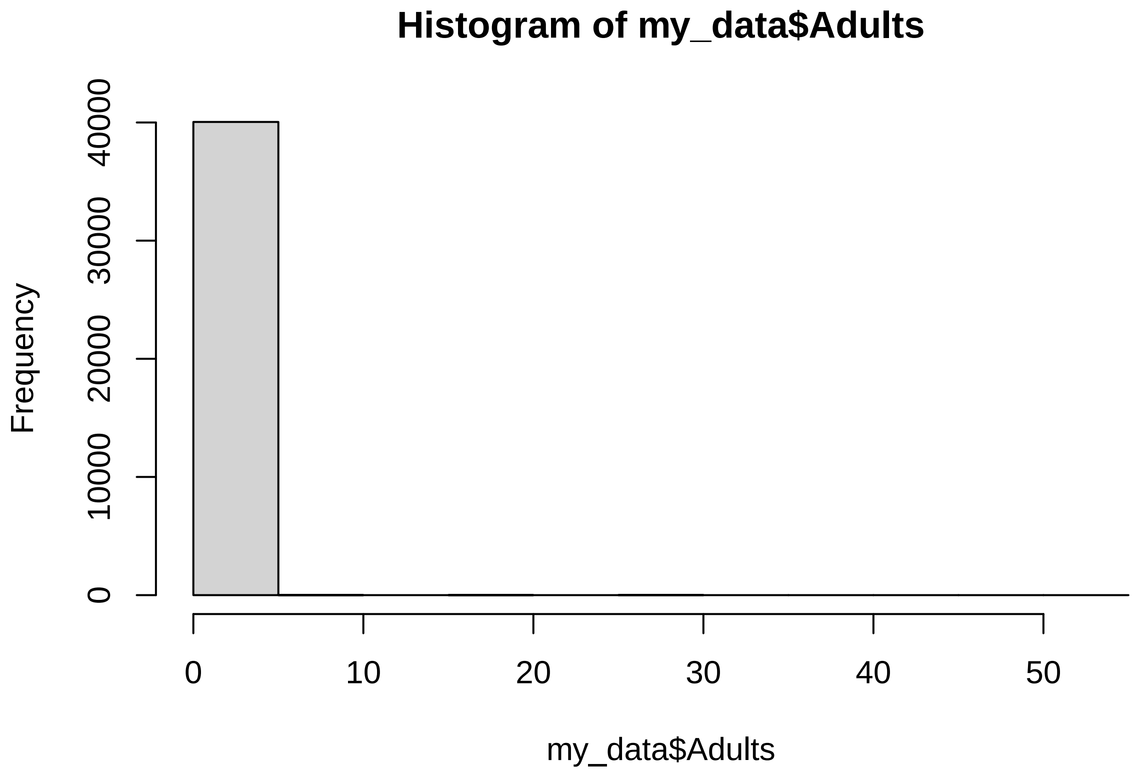
Basic plot:

```
plot(my_data$Adults, my_data$Children)
```



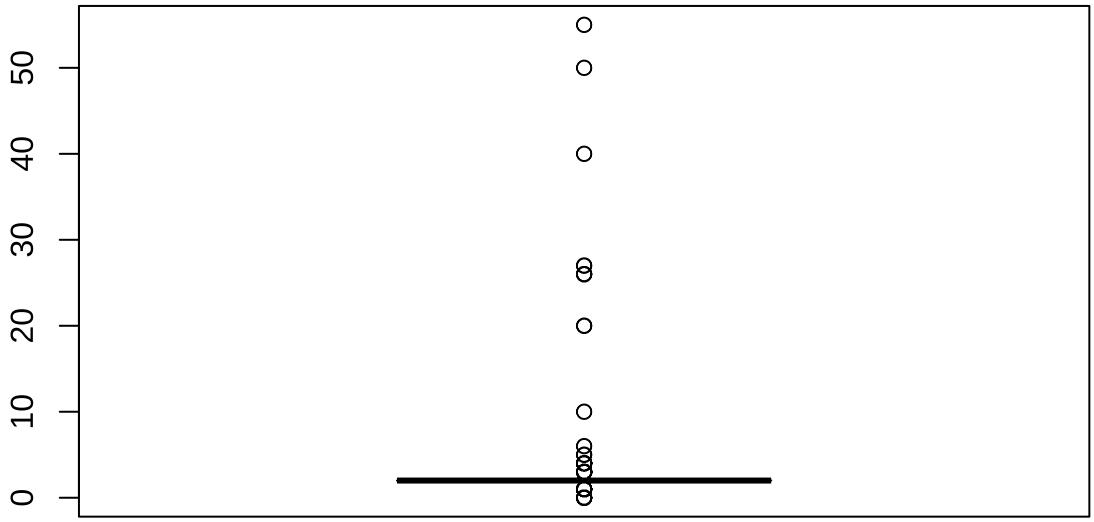
Histogram:

```
hist(my_data$Adults)
```



Boxplot:

```
boxplot(my_data$Adults)
```



Calculate mean:

```
mean(my_data$Adults)
```

```
## [1] 1.867149
```

Calculate median:

```
median(my_data$Adults)
```

```
## [1] 2
```

Data Manipulation:

Filter rows based on a condition:

```
subset_data <- my_data[my_data$Adults > 100, ]
head(subset_data, n = 10)
```

```
##  [1] IsCanceled                  LeadTime
##  [3] ArrivalDateYear             ArrivalDateMonth
##  [5] ArrivalDateWeekNumber       ArrivalDateDayOfMonth
##  [7] StaysInWeekendNights        StaysInWeekNights
##  [9] Adults                      Children
## [11] Babies                      Meal
## [13] Country                     MarketSegment
## [15] DistributionChannel         IsRepeatedGuest
## [17] PreviousCancellations       PreviousBookingsNotCanceled
## [19] ReservedRoomType            AssignedRoomType
## [21] BookingChanges              DepositType
## [23] Agent                       Company
## [25] DaysInWaitingList           CustomerType
## [27] ADR                         RequiredCarParkingSpaces
## [29] TotalOfSpecialRequests      ReservationStatus
## [31] ReservationStatusDate
## <0 rows> (or 0-length row.names)
```

Create a new variable:

```
my_data$new_variable <- my_data$Adults + my_data$Children
head(my_data$new_variable, n = 10)
```

```
##  [1] 2 2 1 1 2 2 2 2 2 2
```

Rename a variable:

```
names(my_data)[2] <- "Hotel"
```

Remove a variable:

```
my_data$Babies <- NULL
head(my_data, n = 10)
```

```
##    IsCanceled Hotel ArrivalDateYear ArrivalDateMonth ArrivalDateWeekNumber
## 1           0   342            2015             July                    27
## 2           0   737            2015             July                    27
## 3           0     7            2015             July                    27
## 4           0    13            2015             July                    27
## 5           0    14            2015             July                    27
## 6           0    14            2015             July                    27
## 7           0     0            2015             July                    27
## 8           0     9            2015             July                    27
## 9           1    85            2015             July                    27
## 10          1    75            2015             July                    27
##    ArrivalDateDayOfMonth StaysInWeekendNights StaysInWeekNights Adults Children
## 1                      1                    0                 0      2        0
## 2                      1                    0                 0      2        0
## 3                      1                    0                 1      1        0
## 4                      1                    0                 1      1        0
## 5                      1                    0                 2      2        0
## 6                      1                    0                 2      2        0
## 7                      1                    0                 2      2        0
## 8                      1                    0                 2      2        0
## 9                      1                    0                 3      2        0
## 10                     1                    0                 3      2        0
##         Meal Country MarketSegment DistributionChannel IsRepeatedGuest
## 1  BB            PRT        Direct              Direct               0
## 2  BB            PRT        Direct              Direct               0
## 3  BB            GBR        Direct              Direct               0
## 4  BB            GBR     Corporate           Corporate               0
## 5  BB            GBR     Online TA               TA/TO               0
## 6  BB            GBR     Online TA               TA/TO               0
## 7  BB            PRT        Direct              Direct               0
## 8  FB            PRT        Direct              Direct               0
## 9  BB            PRT     Online TA               TA/TO               0
## 10 HB            PRT Offline TA/TO               TA/TO               0
##    PreviousCancellations PreviousBookingsNotCanceled ReservedRoomType
## 1                      0                           0 C
## 2                      0                           0 C
## 3                      0                           0 A
## 4                      0                           0 A
## 5                      0                           0 A
## 6                      0                           0 A
## 7                      0                           0 C
## 8                      0                           0 C
## 9                      0                           0 A
## 10                     0                           0 D
##    AssignedRoomType BookingChanges     DepositType       Agent     Company
## 1  C                             3 No Deposit             NULL        NULL
## 2  C                             4 No Deposit             NULL        NULL
## 3  C                             0 No Deposit             NULL        NULL
## 4  A                             0 No Deposit              304        NULL
## 5  A                             0 No Deposit              240        NULL
## 6  A                             0 No Deposit              240        NULL
## 7  C                             0 No Deposit             NULL        NULL
## 8  C                             0 No Deposit              303        NULL
## 9  A                             0 No Deposit              240        NULL
## 10 D                             0 No Deposit               15        NULL
##    DaysInWaitingList CustomerType   ADR RequiredCarParkingSpaces
## 1                  0    Transient   0.0                        0
## 2                  0    Transient   0.0                        0
## 3                  0    Transient  75.0                        0
## 4                  0    Transient  75.0                        0
## 5                  0    Transient  98.0                        0
## 6                  0    Transient  98.0                        0
## 7                  0    Transient 107.0                        0
## 8                  0    Transient 103.0                        0
## 9                  0    Transient  82.0                        0
## 10                 0    Transient 105.5                        0
##    TotalOfSpecialRequests ReservationStatus ReservationStatusDate new_variable
## 1                       0         Check-Out            01-07-2015            2
## 2                       0         Check-Out            01-07-2015            2
## 3                       0         Check-Out            02-07-2015            1
## 4                       0         Check-Out            02-07-2015            1
## 5                       1         Check-Out            03-07-2015            2
## 6                       1         Check-Out            03-07-2015            2
## 7                       0         Check-Out            03-07-2015            2
## 8                       1         Check-Out            03-07-2015            2
## 9                       1          Canceled            06-05-2015            2
## 10                      0          Canceled            22-04-2015            2
```

Sort data by a variable:

```r
my_data <- my_data[order(my_data$Adults), ]
```

Data Exploration:

Correlation matrix:

```r
cor(my_data$Adults, my_data$Children)
```

```
## [1] 0.07324593
```

Frequency table:

```r
table_freq <- table(my_data$Adults)
head(table_freq, n = 50)
```

```
## 
##      0      1      2      3      4      5      6     10     20     26     27     40     50 
##     13   7148  31425   1427     31      2      1      1      2      5      2      1      1 
##     55 
##      1 
```

Descriptive statistics by group:

```r
table_tap <- tapply(my_data$Adults, my_data$Children, summary)
head(table_tap, n = 10)
```

```
## $‘0‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   2.000   2.000   1.847   2.000  55.000 
## 
## $‘1‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.166   2.000   4.000
##
## $‘2‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.977   2.000   3.000
##
## $‘3‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.882   2.000   2.000
##
## $‘10‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2       2       2       2       2       2
```

Missing Data Handling:

Check for missing values:

```
any(is.na(my_data))
```

```
## [1] FALSE
```

Remove missing values:

```
my_data_no_na <- na.omit(my_data)
```

Impute missing values:

```
my_data$High[is.na(my_data$Adults)] <- mean(my_data$Adults, na.rm = TRUE)
```

Statistical Analysis:

t-test:

```
result <- t.test(my_data$Adults)
result
```

```
##
##  One Sample t-test
##
## data:  my_data$Adults
## t = 535.95, df = 40059, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.860321 1.873978
## sample estimates:
## mean of x
##  1.867149
```

ANOVA:

```
anova_model <- aov(Adults ~ Children, data = my_data)
summary(anova_model)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Children         1    104  104.49   216.1 <2e-16 ***
## Residuals    40058  19372    0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Regression:

```
lm_model <- lm(Adults ~ Children, data = my_data)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Adults ~ Children, data = my_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.852  0.033  0.148  0.148 53.148
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.852387   0.003617   512.2   <2e-16 ***
## Children    0.114721   0.007805    14.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6954 on 40058 degrees of freedom
## Multiple R-squared:  0.005365,   Adjusted R-squared:  0.00534
## F-statistic: 216.1 on 1 and 40058 DF,  p-value: < 2.2e-16
```

Chi-square test:

```
chisq.test(table(my_data$Adults, my_data$Children))
```

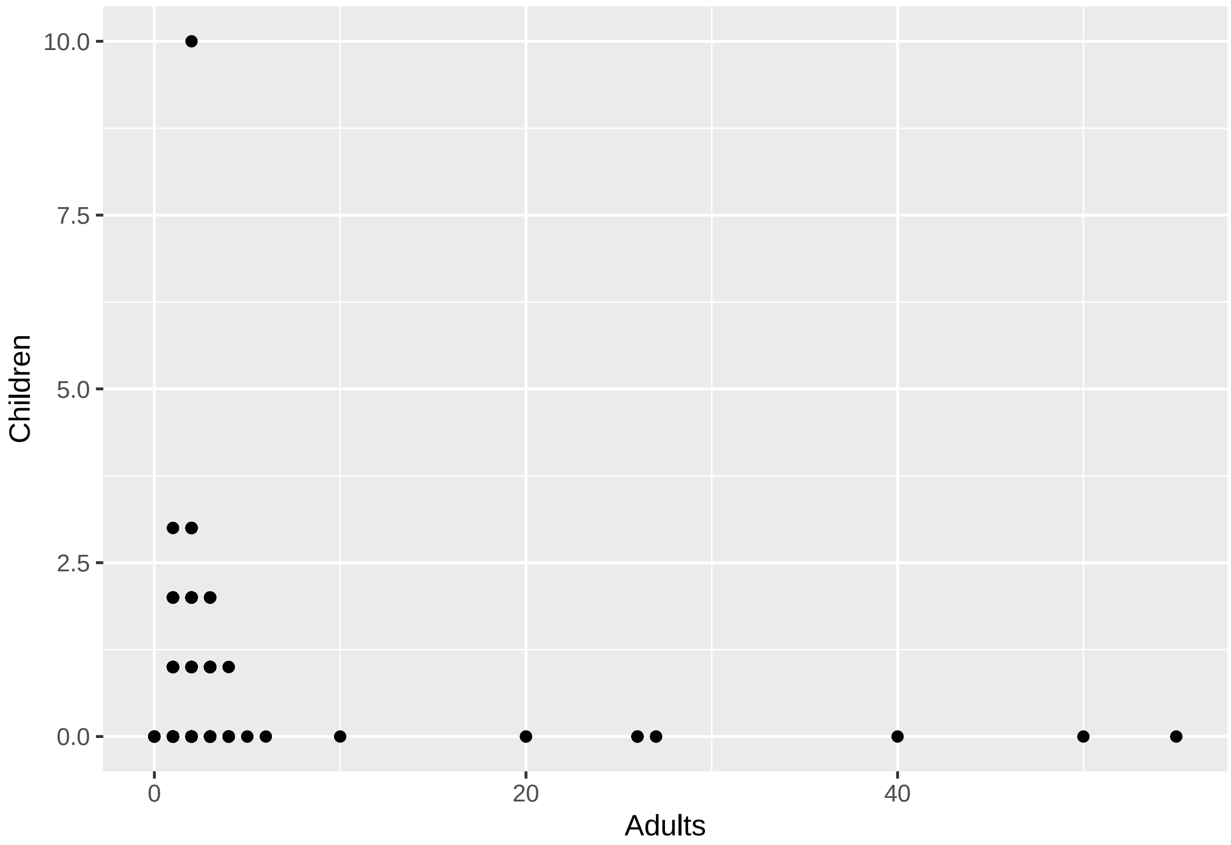
```
## Warning in chisq.test(table(my_data$Adults, my_data$Children)): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(my_data$Adults, my_data$Children)
## X-squared = 2124.2, df = 52, p-value < 2.2e-16
```

Advanced Visualization:

ggplot2 - Scatter plot:

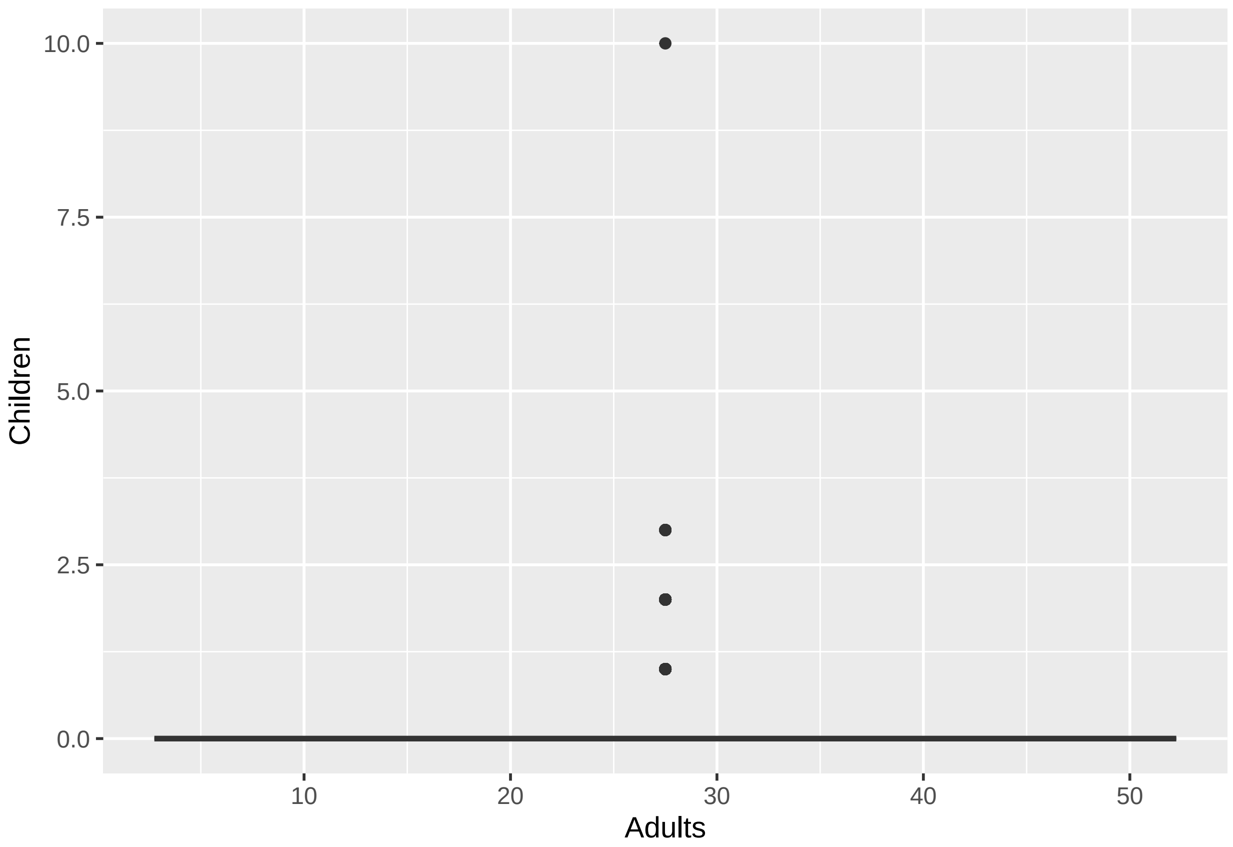
```
library(ggplot2)
ggplot(my_data, aes(x = Adults, y = Children)) +
  geom_point()
```



ggplot2 - Boxplot:

```
ggplot(my_data, aes(x = Adults, y = Children)) +
  geom_boxplot()
```

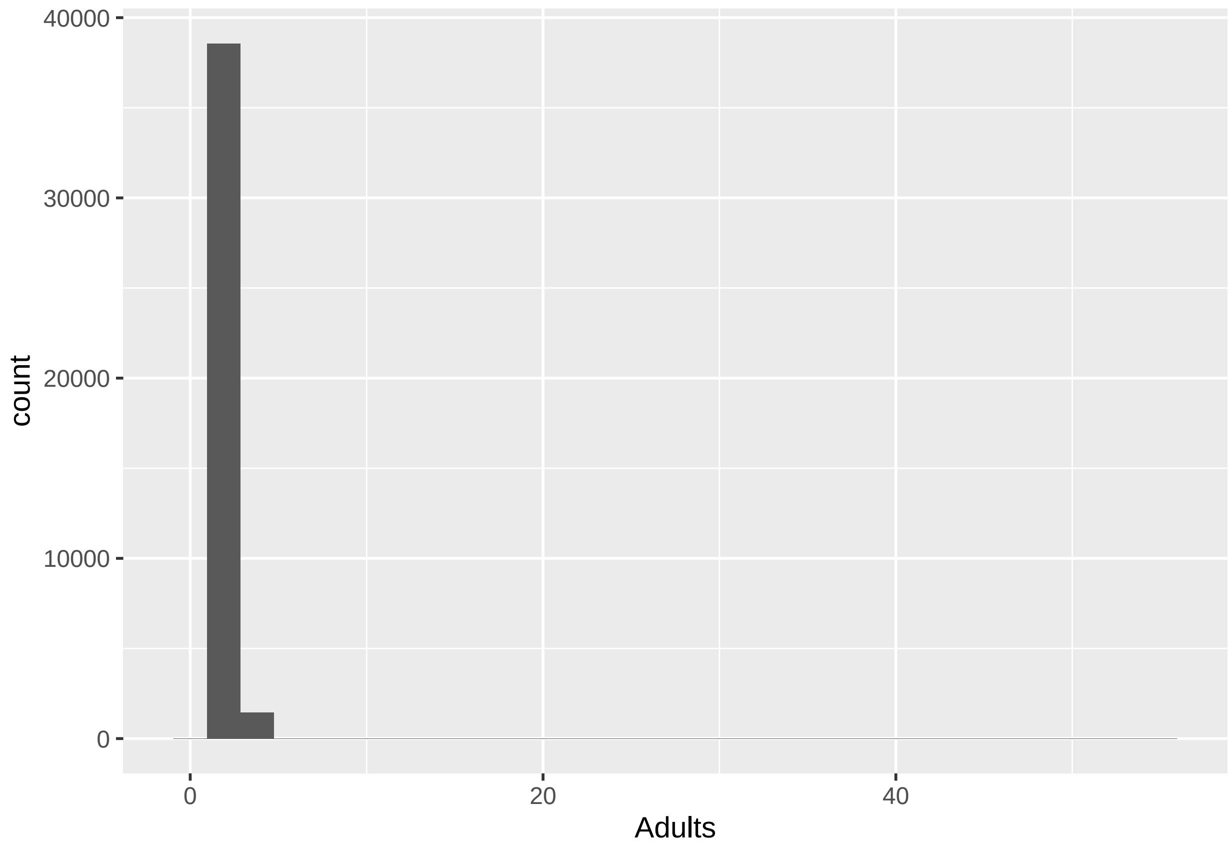
```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```



ggplot2 - Histogram:

```
ggplot(my_data, aes(x = Adults)) +
  geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Time Series Analysis:

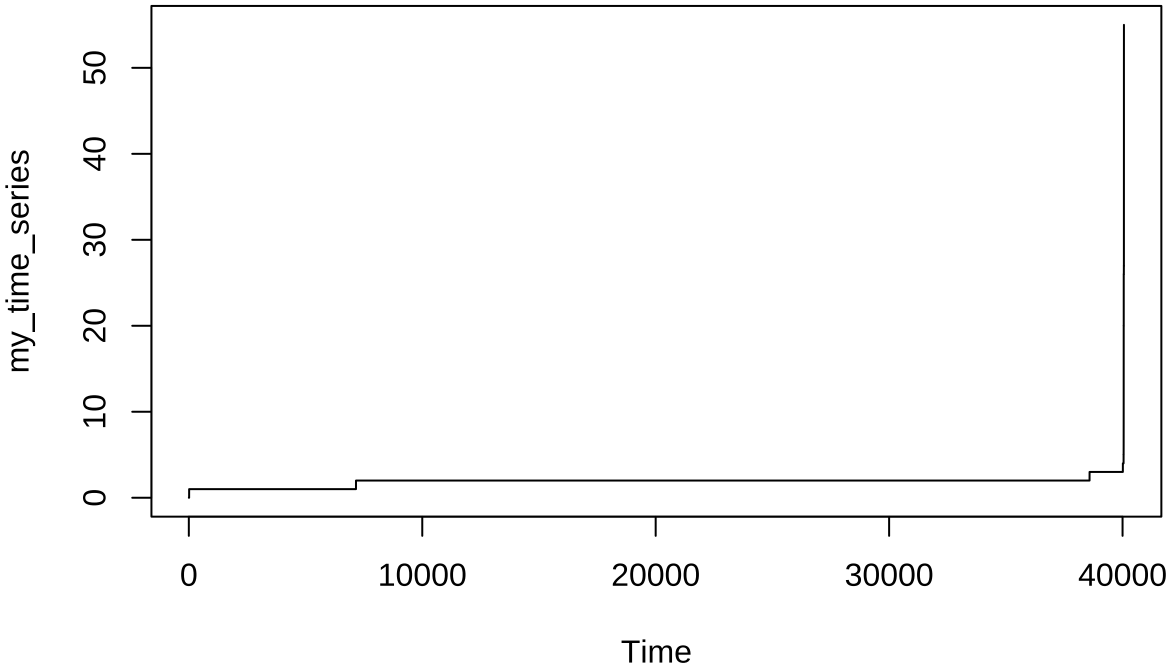
Convert to time series:

```
my_time_series <- ts(my_data$Adults, start = 1, end = length(my_data$Children), frequency = 1)
head(my_time_series, n = 10)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0
```

Time series plot:

```
plot(my_time_series)
```



ARIMA modeling:

```
arima_model <- arima(my_time_series, order = c(1,1,1))
arima_model
```

```
##
## Call:
## arima(x = my_time_series, order = c(1, 1, 1))
##
## Coefficients:
##          ar1      ma1
##       0.7940  -0.3393
## s.e.  0.0078   0.0100
##
## sigma^2 estimated as 0.008129:  log likelihood = 39546.55,  aic = -79087.11
```