2022 - Data Analytics for Immersive Environments - CA4 - RDBMS & Linear Regression Project

CA4 Part B - Linear Regression Analysis

Joe O'Regan

2023-01-16

Repo Link

https://github.com/joeaoregan/2022_DAIE_CA4_JOR1

Assumptions (Linear Regression)

1. **Homogeneity of variance (homoscedasticity):** The size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. **Independence of observations:** the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. **Normality:** The data follows a normal distribution.
4. **The relationship between the independent and dependent variable is linear:** the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Read data from CSV file

```
data <- read_csv("amalgamated_game_survey_250_2022.csv") # read data from csv
```

```
## Rows: 250 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): gender, top_reason_gaming, gaming_platform, favourite_game, ethnici...
## dbl (4): age, avg_monthly_hrs_gaming, avg_years_playing_games, avg_monthly_e...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Find usable columns

```
# assumption here is it would be very hard to plot a graph on anything else
# colnames(data) # list of column names
#sapply(data, class)
# str(data) # show column properties, find numeric columns
numeric_cols <- unlist(lapply(data, is.numeric))
numeric_data <- data[ , numeric_cols]
colnames(numeric_data)
```

```
## [1] "age"                          "avg_monthly_hrs_gaming"
## [3] "avg_years_playing_games"      "avg_monthly_expenditure_dlc"
```

Variables

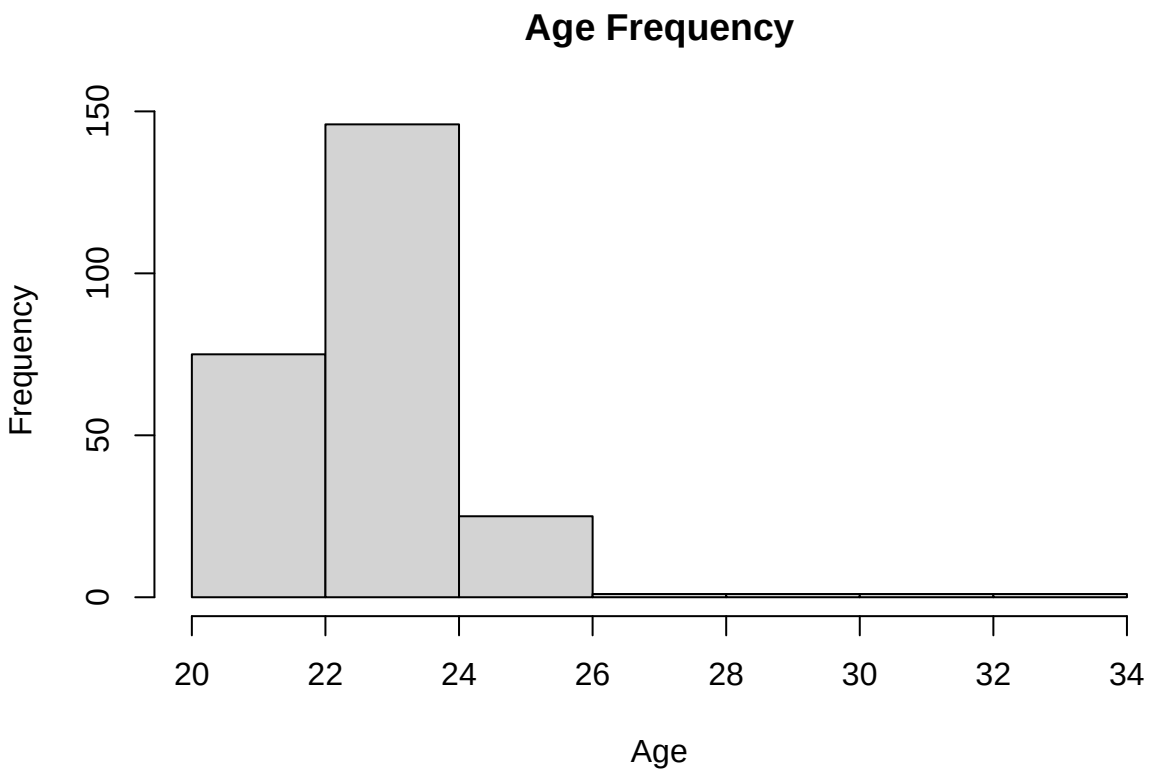
age, avg_monthly_hrs_gaming, avg_years_playing_games, avg_monthly_expenditure_dlc are all numeric fields.

Normality

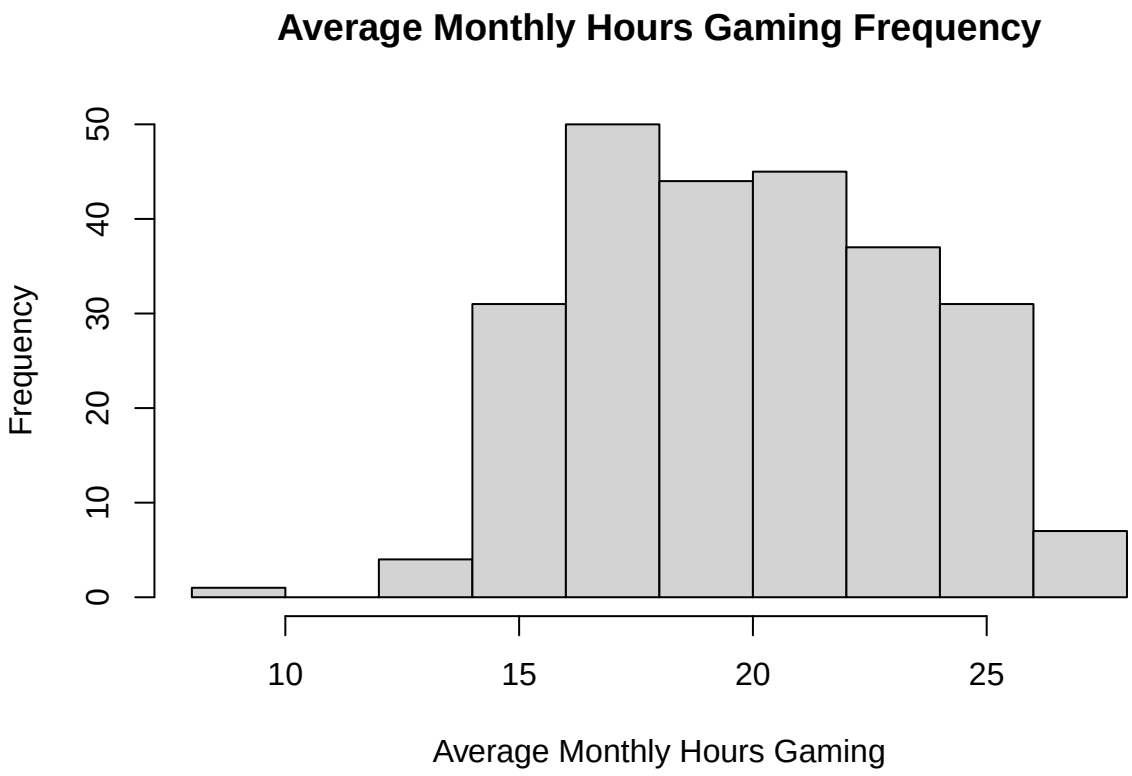
Histogram (Visual check)

Use histograms to visually check for normality. If the histogram is symmetrical/unimodal, then the data is assumed to be normally distributed.

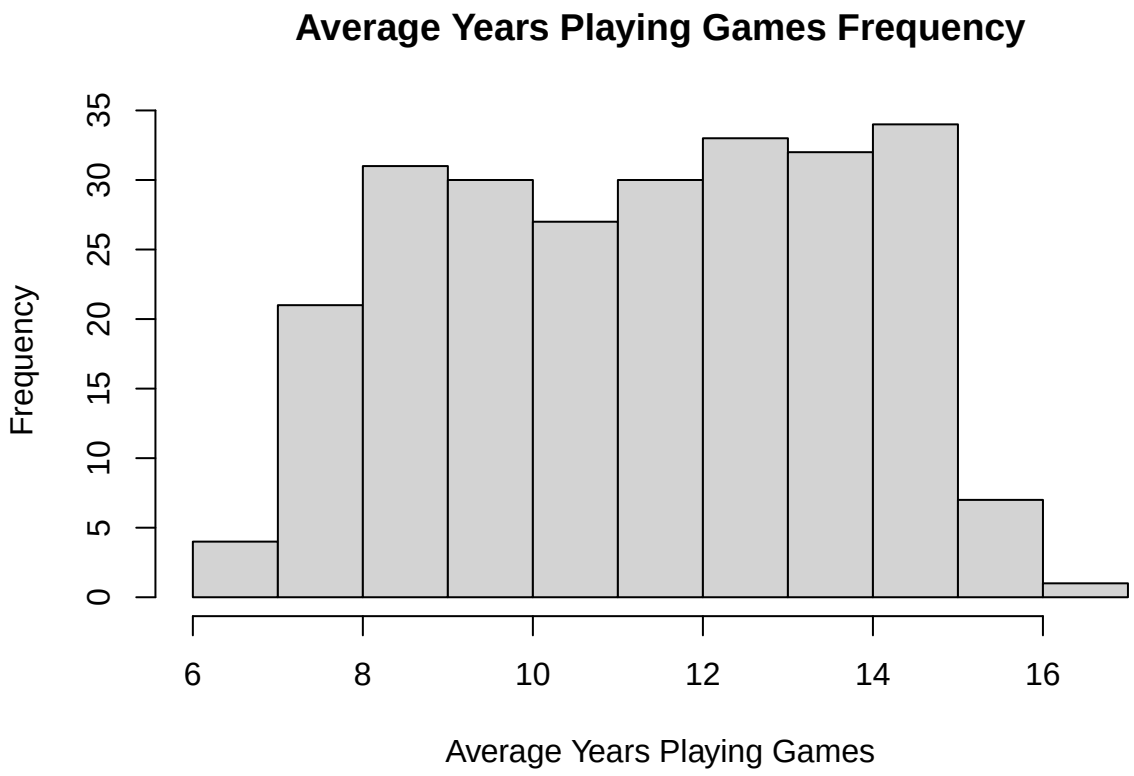
```
hist(data$age,
     main="Age Frequency",
     xlab = "Age")
```



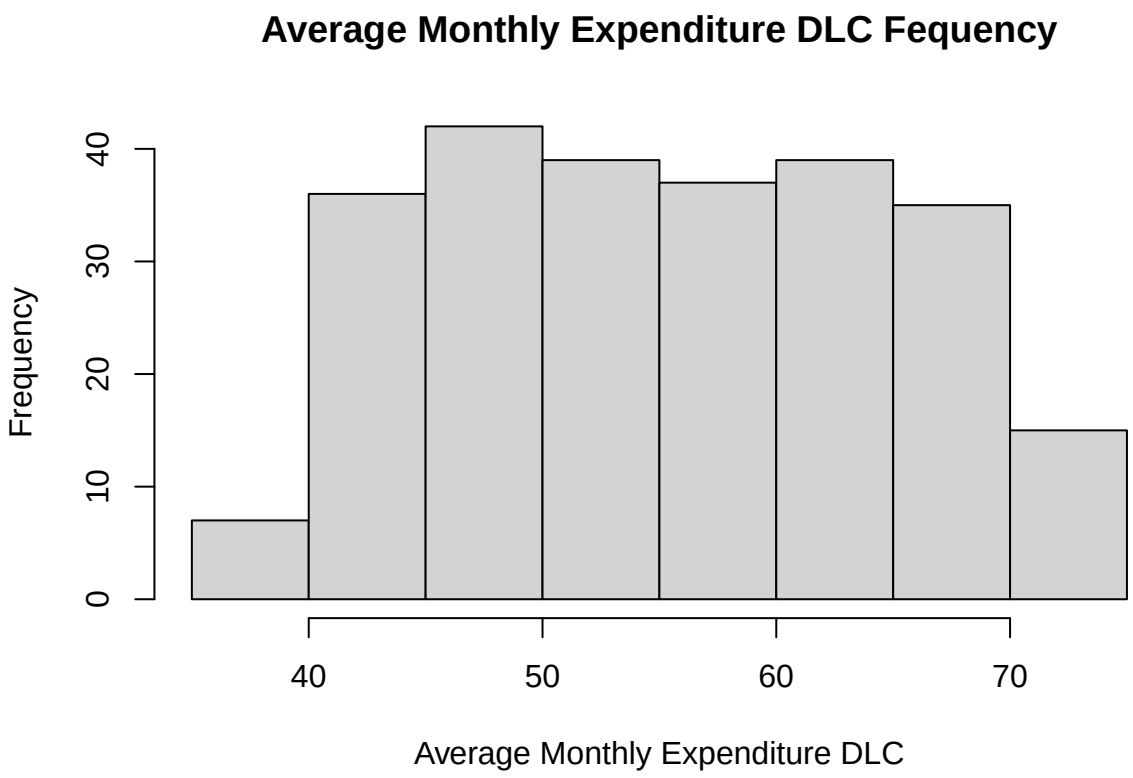

```
hist(data$avg_monthly_hrs_gaming,
     main="Average Monthly Hours Gaming Frequency",
     xlab="Average Monthly Hours Gaming")
```




```
hist(data$avg_years_playing_games,
     main="Average Years Playing Games Frequency",
     xlab = "Average Years Playing Games")
```




```
hist(data$avg_monthly_expenditure_dlc,
     main="Average Monthly Expenditure DLC Fequency",
     xlab = "Average Monthly Expenditure DLC")
```



Shapiro-Wilk's test

null hypothesis: the data are sampled from a Gaussian distribution.

If the P value is greater than 0.05 the answer is yes.

If the P value is less than or equal to 0.05 the answer is no.

```
st_age <- shapiro.test(data$age)
if(st_age$p.value < 0.05) print("nope") else print("yep")
```

```
## [1] "nope"
```

```
st_hours <- shapiro.test(data$avg_monthly_hrs_gaming)
if(st_hours$p.value < 0.05) print("nope") else print("yep")
```

```
## [1] "nope"
```

```
st_years <- shapiro.test(data$avg_years_playing_games)
if(st_years$p.value < 0.05) print("nope") else print("yep")
```

```
## [1] "nope"
```

```r
st_bucks <- shapiro.test(data$avg_monthly_expenditure_dlc)
if(st_bucks$p.value < 0.05) print("nope") else print("yep")
```

```
## [1] "nope"
```

Dependent Variable: avg_monthly_hrs_gaming

Independent Variable: avg_monthly_expenditure_dlc

```r
set.seed(321) # reproduce random values

sample_data <- sample_n(data, 200) # tibble 200 x 11

# lm() -
# dependent var. ~ independent var.
mod <- lm(avg_monthly_expenditure_dlc ~ avg_monthly_hrs_gaming,
          data = sample_data)
summary(mod)
```

```
##
## Call:
## lm(formula = avg_monthly_expenditure_dlc ~ avg_monthly_hrs_gaming,
##     data = sample_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8048  -8.1380   0.2678   8.2022  16.7890
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              59.4541     3.8204  15.562   <2e-16 ***
## avg_monthly_hrs_gaming  -0.1845     0.1873  -0.985    0.326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.212 on 198 degrees of freedom
## Multiple R-squared:  0.00488,    Adjusted R-squared:  -0.0001458
## F-statistic: 0.971 on 1 and 198 DF,  p-value: 0.3256
```

```r
#attributes(mod)
#mod$residuals
# hist(mod$residuals)

plot <- ggplot(data = mod, mapping = aes(x = avg_monthly_hrs_gaming,
                                         y = avg_monthly_expenditure_dlc)) +
  # geom_point(alpha = 0.1, color = "blue") # add colours for points
  geom_point() + # plot dataset in a scatter plot
  labs(title = "Relationship between games monthly hours played + DLC expenditure",
       x = "Average Monthly Hours Gaming",
       y = "Average Monthly Expenditure DLC")

# plot + geom_smooth(method = lm, se = FALSE, formula=y~x) # probably this one
```

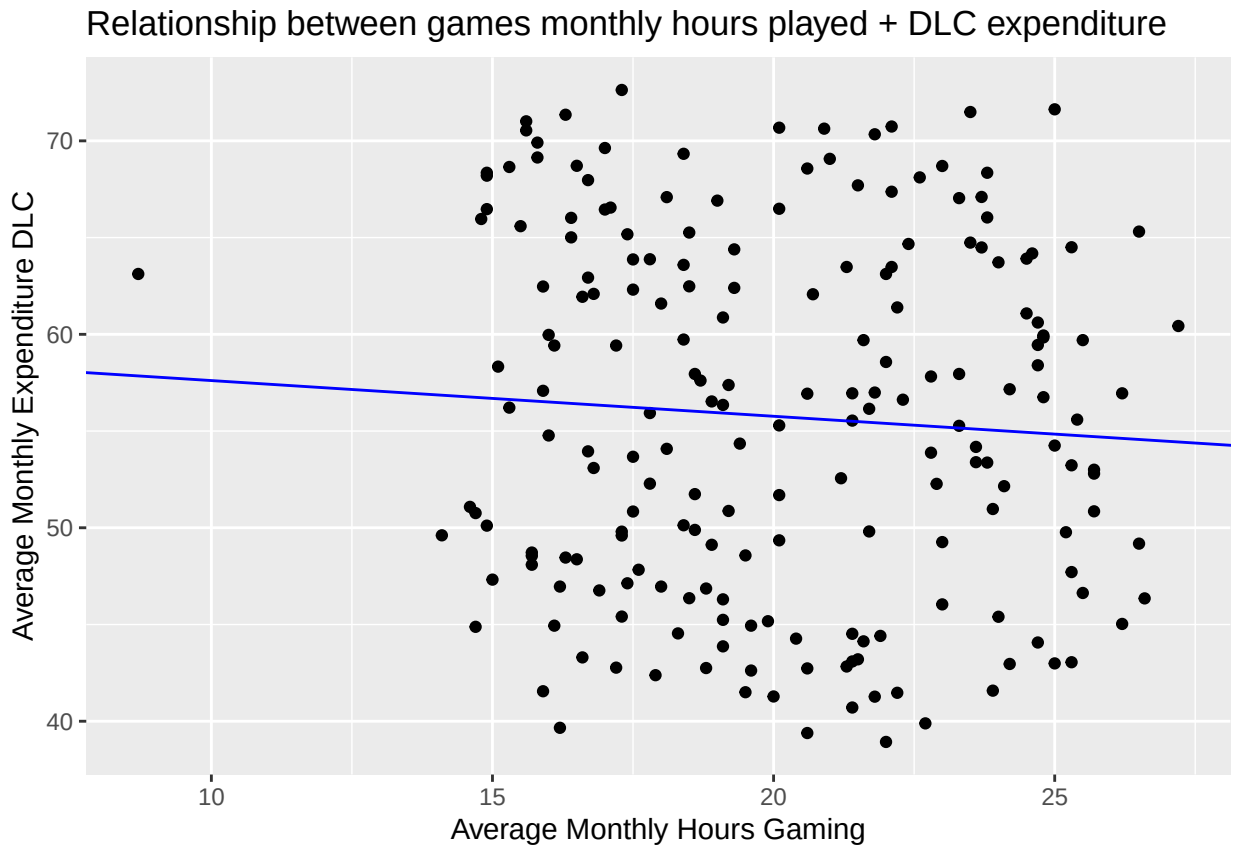
```
# plot + stat_smooth(method = lm, formula = y ~ x, geom = "smooth") # ok
# plot + geom_smooth(method = "loess", se = FALSE, formula=y~x) # curved line

coeff <- coefficients(mod)
coeff
```

```
##            (Intercept) avg_monthly_hrs_gaming
##              59.454087              -0.184524
```

```
intercept <- coeff[1]
slope <- coeff[2]
slope
```

```
## avg_monthly_hrs_gaming
##              -0.184524
```

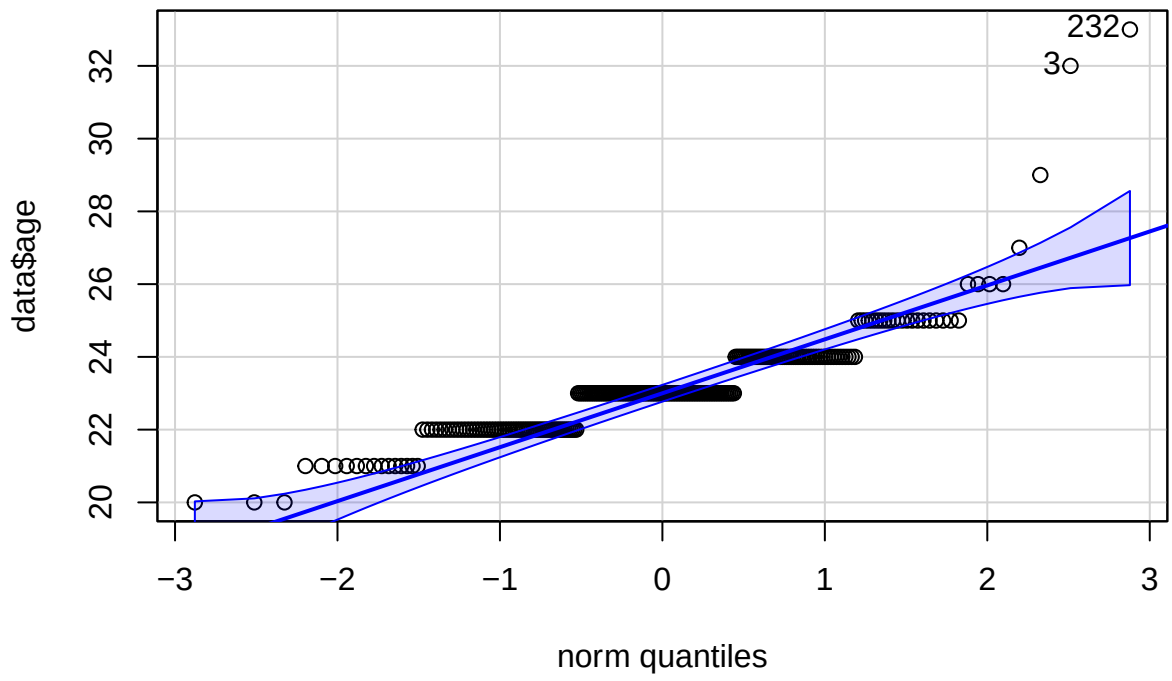
```
plot +
  geom_abline(intercept = intercept, slope = slope, color="blue") # regression line
```




```
# + geom_abline(mapping = aes(x = avg_monthly_hrs_gaming, y = avg_monthly_expenditure_dlc), data = mod)
```

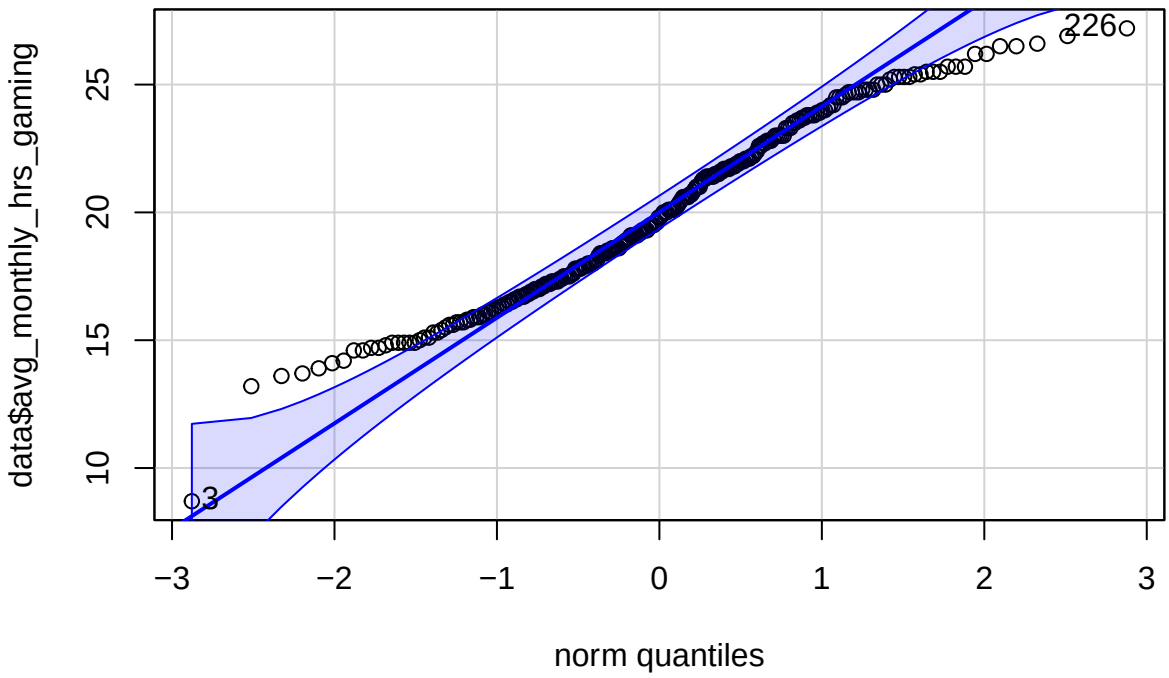
Visually Check Normality QQ Plots

```
car::qqPlot(data$age)
```



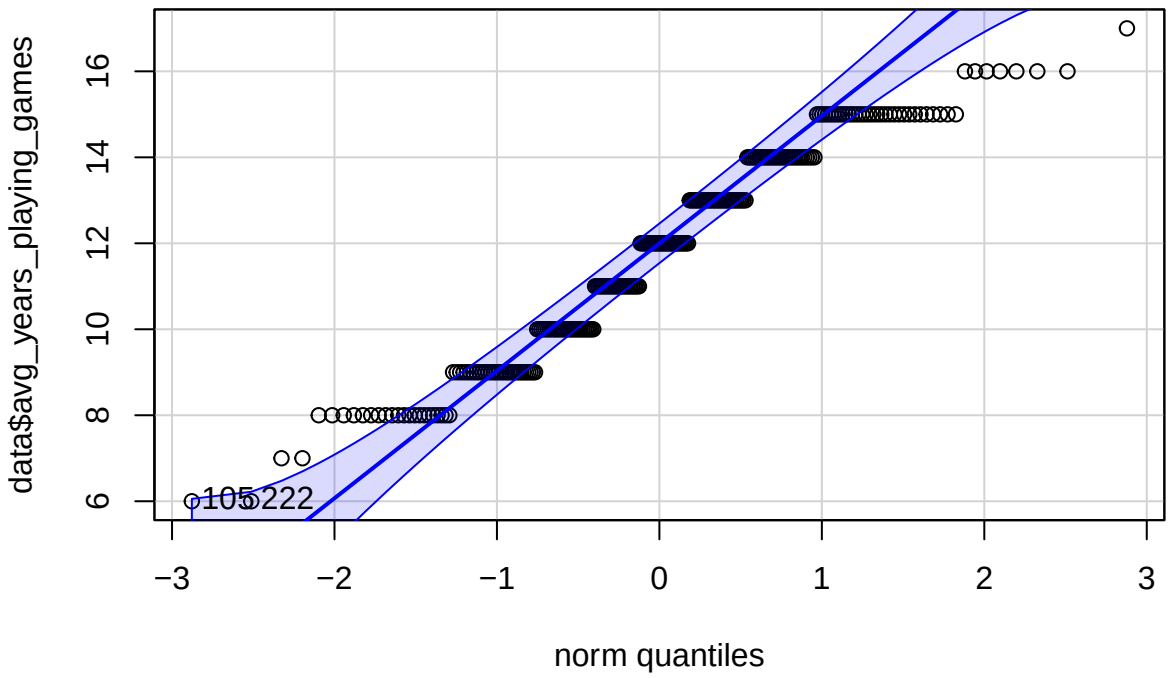

```
## [1] 232   3
```

```
car::qqPlot(data$avg_monthly_hrs_gaming)
```



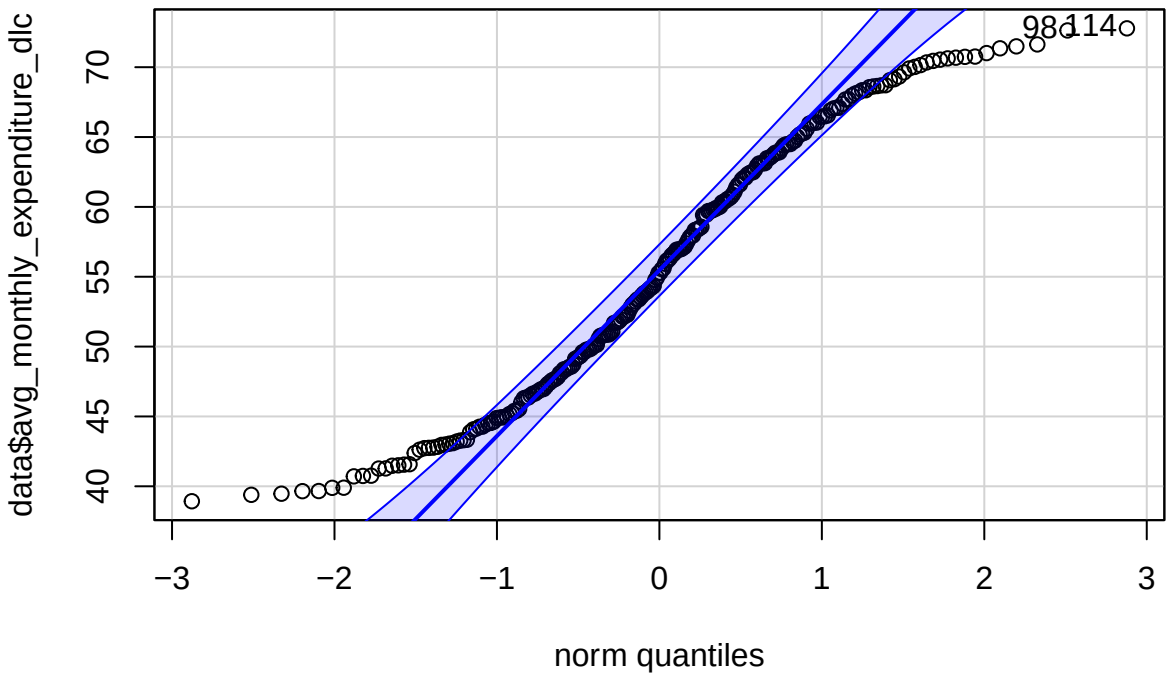

```
## [1]   3 226
```

```
car::qqPlot(data$avg_years_playing_games)
```



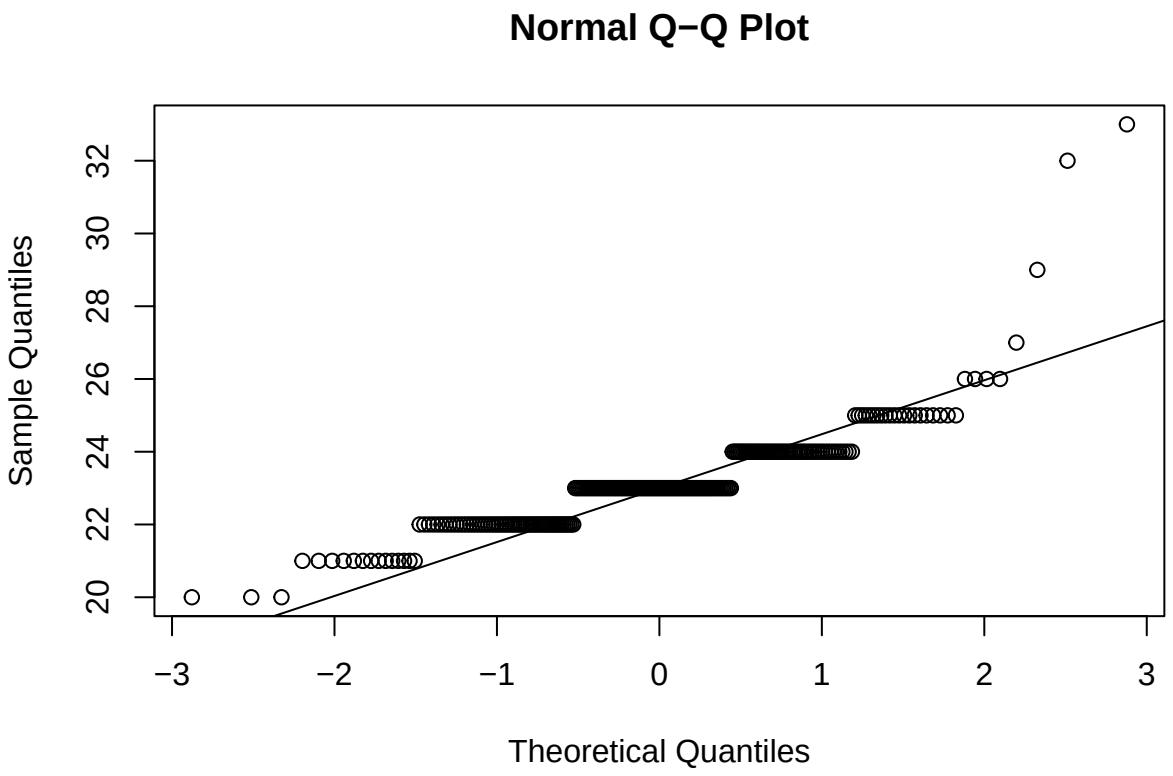

```
## [1] 105 222
```

```
car::qqPlot(data$avg_monthly_expenditure_dlc)
```

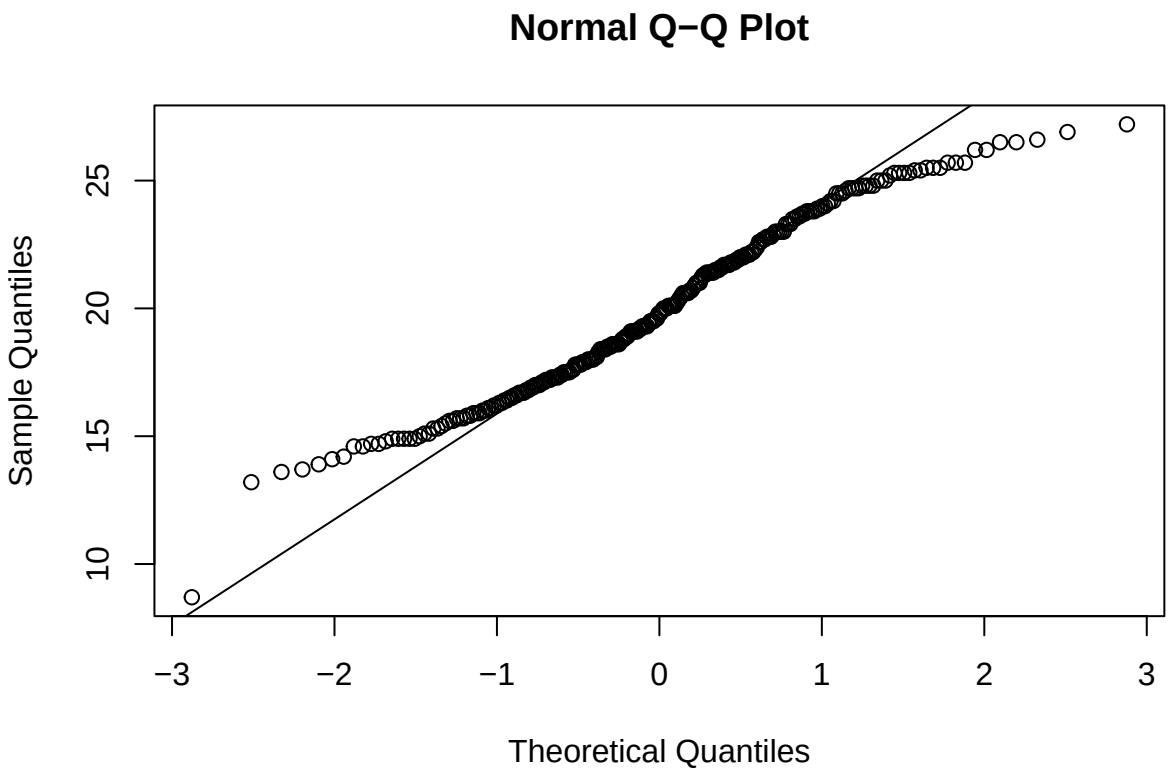



```
## [1] 114  98
```

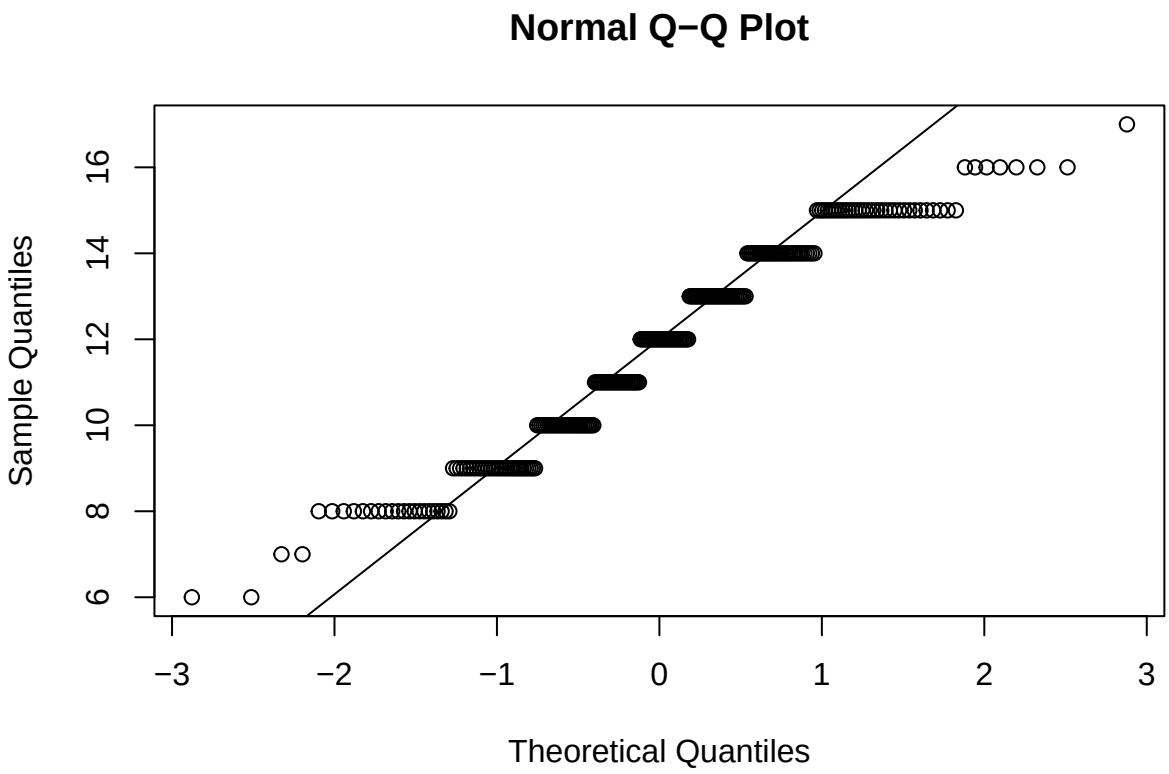
```
qqnorm(data$age)
qqline(data$age)
```



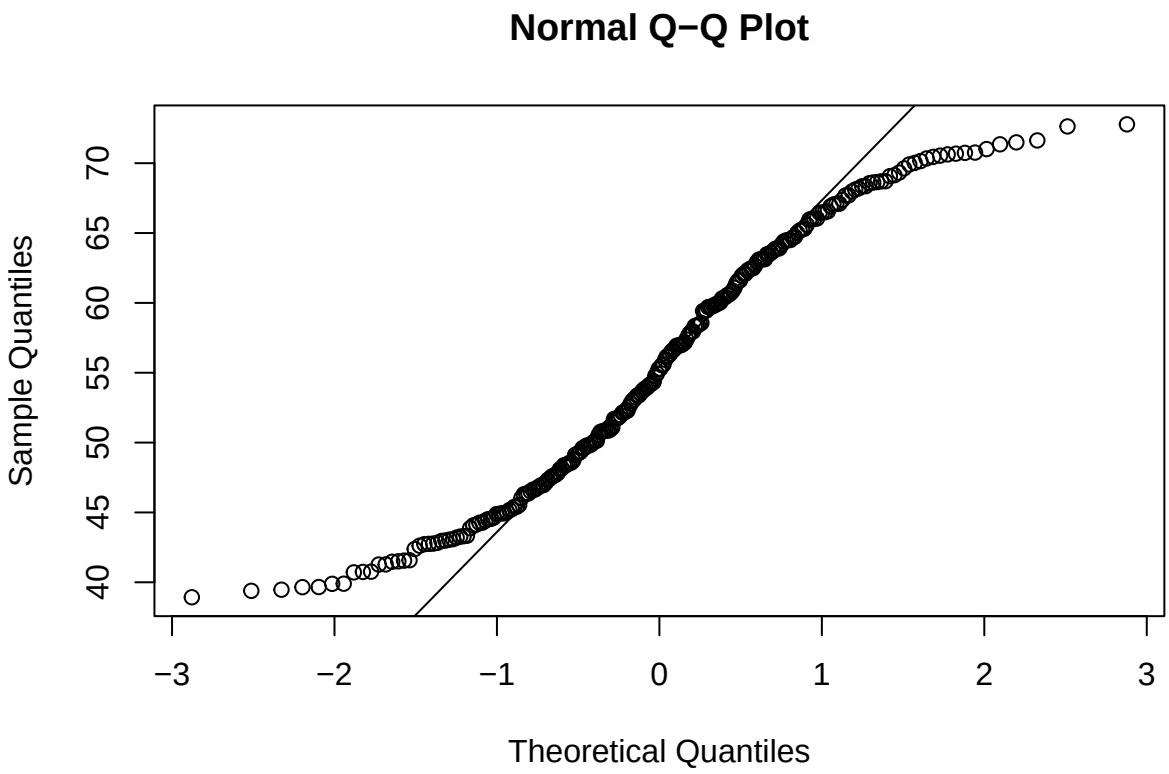

```
qqnorm(data$avg_monthly_hrs_gaming)
qqline(data$avg_monthly_hrs_gaming)
```




```
qqnorm(data$avg_years_playing_games)
qqline(data$avg_years_playing_games)
```




```
qqnorm(data$avg_monthly_expenditure_dlc)
qqline(data$avg_monthly_expenditure_dlc)
```

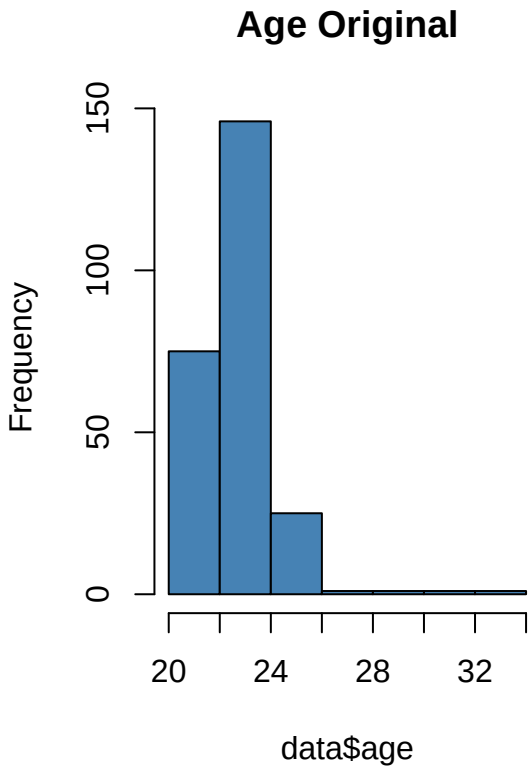


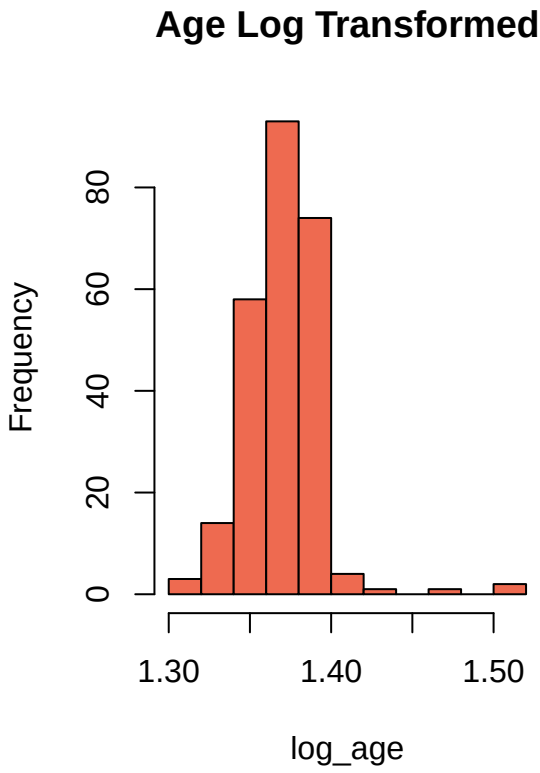
Age - transformations

```
significance <- 0.05

par(mfrow=c(1,2)) # define plotting region
shapiro.test(data$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$age
## W = 0.80693, p-value < 2.2e-16
```

```
# log transformation
log_age <- log10(data$age)
# histogram original distribution
hist(data$age, col='steelblue', main='Age Original')
# histogram log-transformed distribution
hist(log_age, col='coral2', main='Age Log Transformed')
```

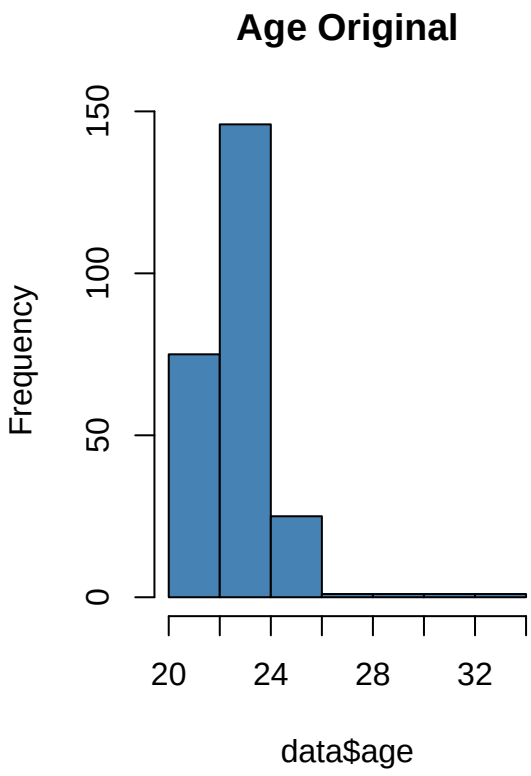


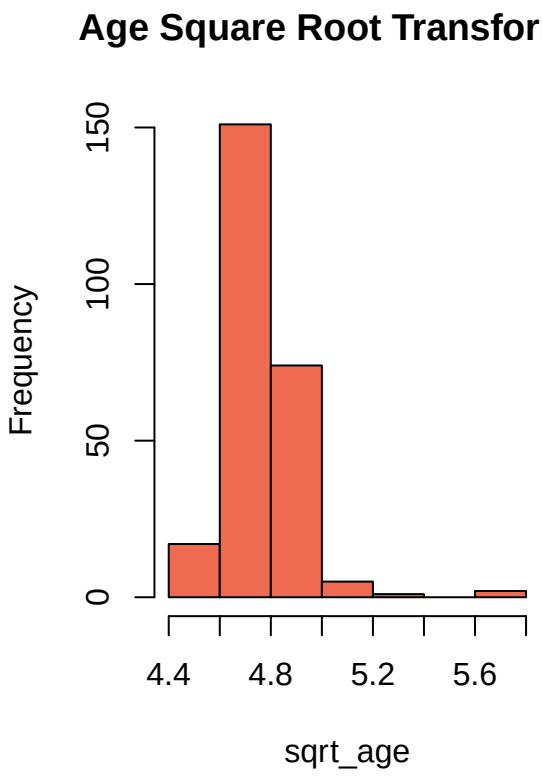



```
shapiro.test(log_age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log_age
## W = 0.85511, p-value = 1.423e-14
```

```
# square root transformation
sqrt_age <- sqrt(data$age)
# histogram original distribution
hist(data$age, col='steelblue', main='Age Original')
# histogram square root-transformed distribution
hist(sqrt_age, col='coral2', main='Age Square Root Transformed')
```

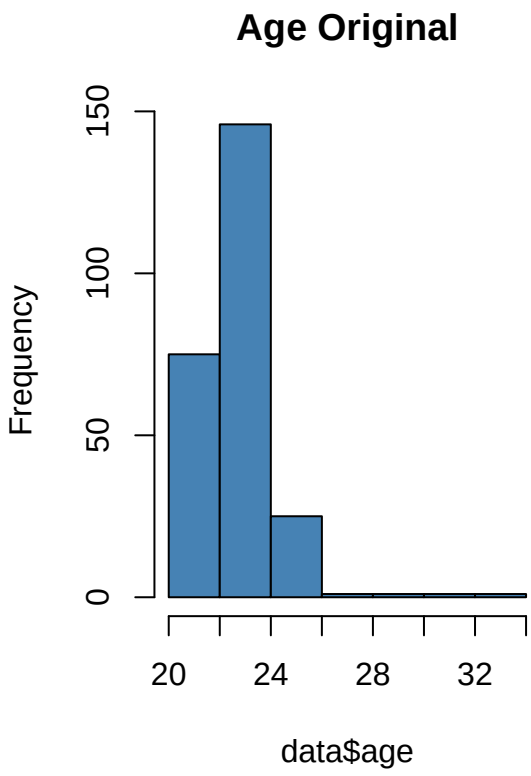


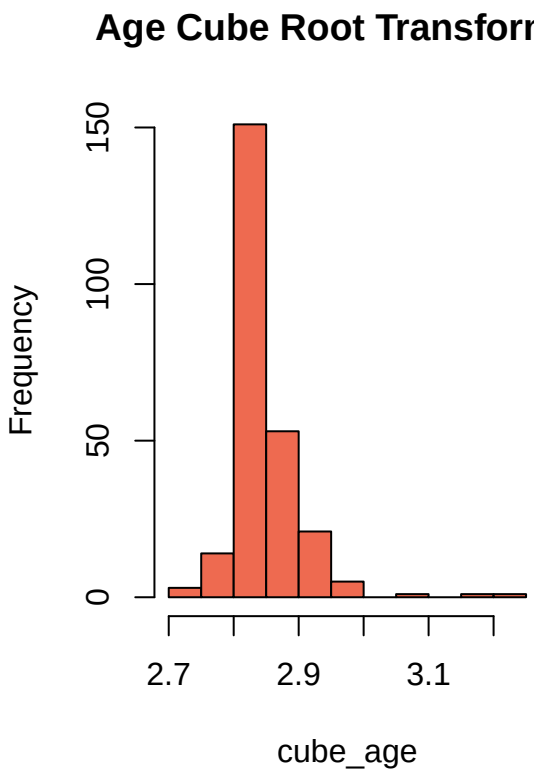



```
shapiro.test(sqrt_age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt_age
## W = 0.83284, p-value = 1.03e-15
```

```
# cube root transformation
cube_age <- data$age^(1/3)
# histogram original distribution
hist(data$age, col='steelblue', main='Age Original')
# histogram cube root-transformed
hist(cube_age, col='coral2', main='Age Cube Root Transformed')
```





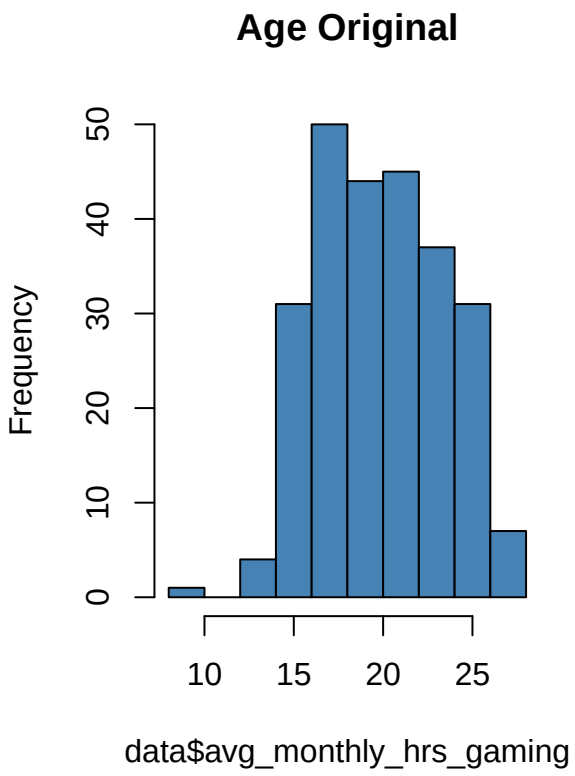

```
shapiro.test(cube_age)
```

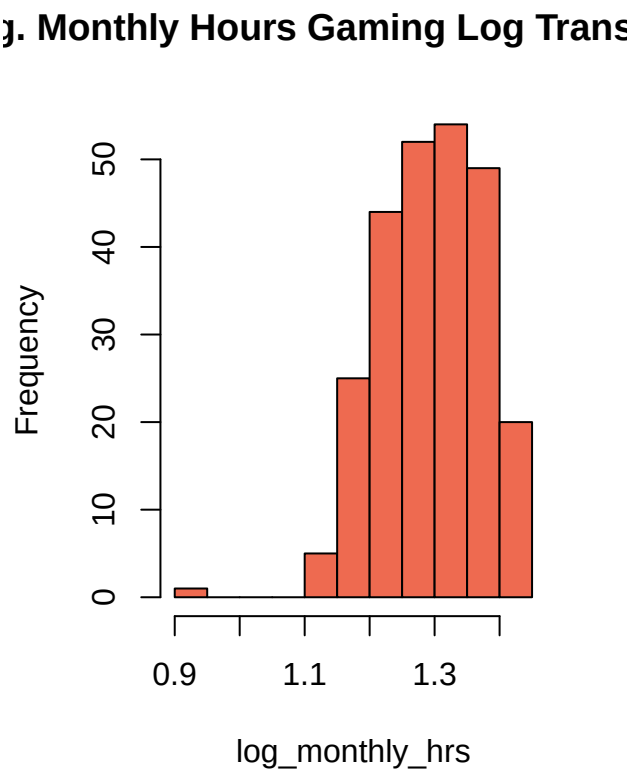
```
##
##  Shapiro-Wilk normality test
##
## data:  cube_age
## W = 0.84067, p-value = 2.516e-15
```

```
cube_age_p_value <- shapiro.test(cube_age)$p.value
if (cube_age_p_value < significance) {
  print(paste("Cube Root Transform of Age is less than ", significance))
} else {
  print(paste("Cube Root Transform of Age is less than ", significance))
}
```

```
## [1] "Cube Root Transform of Age is less than  0.05"
```

```
significance <- 0.05

par(mfrow=c(1,2)) # define plotting region

log_monthly_hrs <- log10(data$avg_monthly_hrs_gaming)
hist(data$avg_monthly_hrs_gaming, col='steelblue', main='Age Original')
hist(log_monthly_hrs, col='coral2', main='Avg. Monthly Hours Gaming Log Transformed')
```

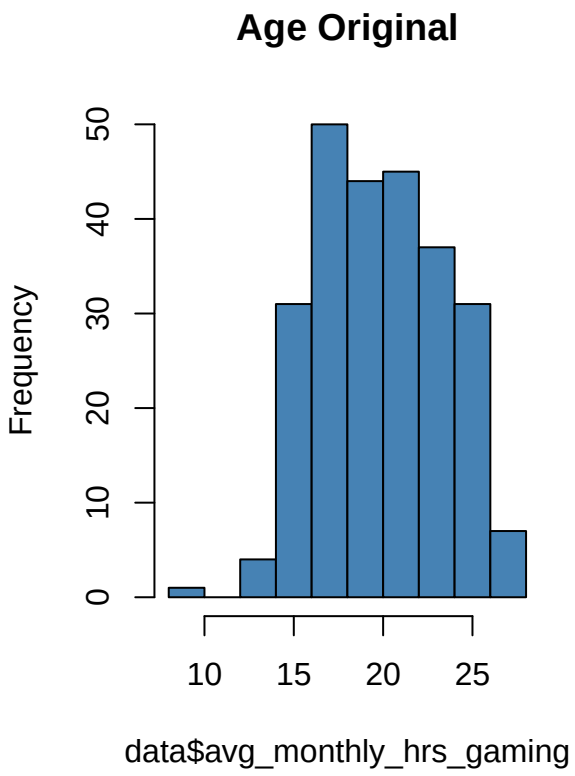


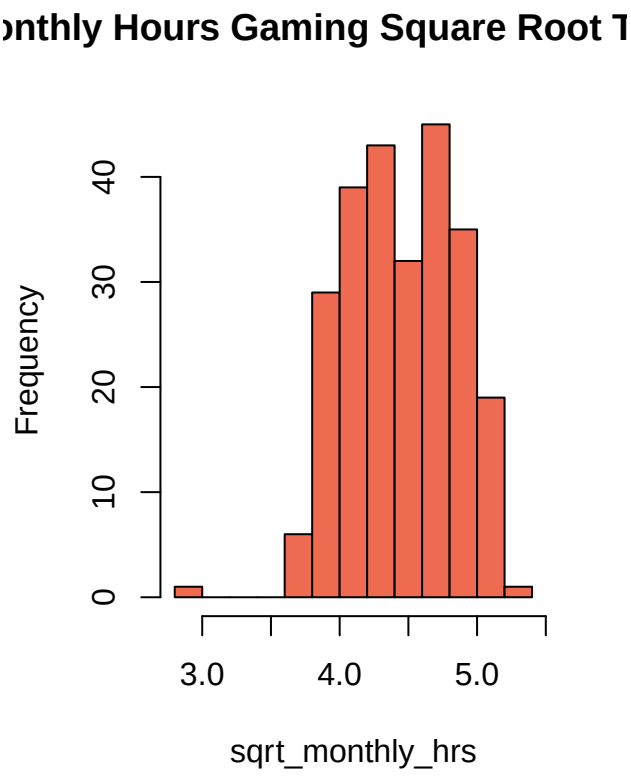



```
shapiro.test(log_monthly_hrs)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log_monthly_hrs
## W = 0.96709, p-value = 1.616e-05
```

```
sqrt_monthly_hrs <- sqrt(data$avg_monthly_hrs_gaming)
hist(data$avg_monthly_hrs_gaming, col='steelblue', main='Age Original')
hist(sqrt_monthly_hrs, col='coral2', main='Avg. Monthly Hours Gaming Square Root Transformed')
```

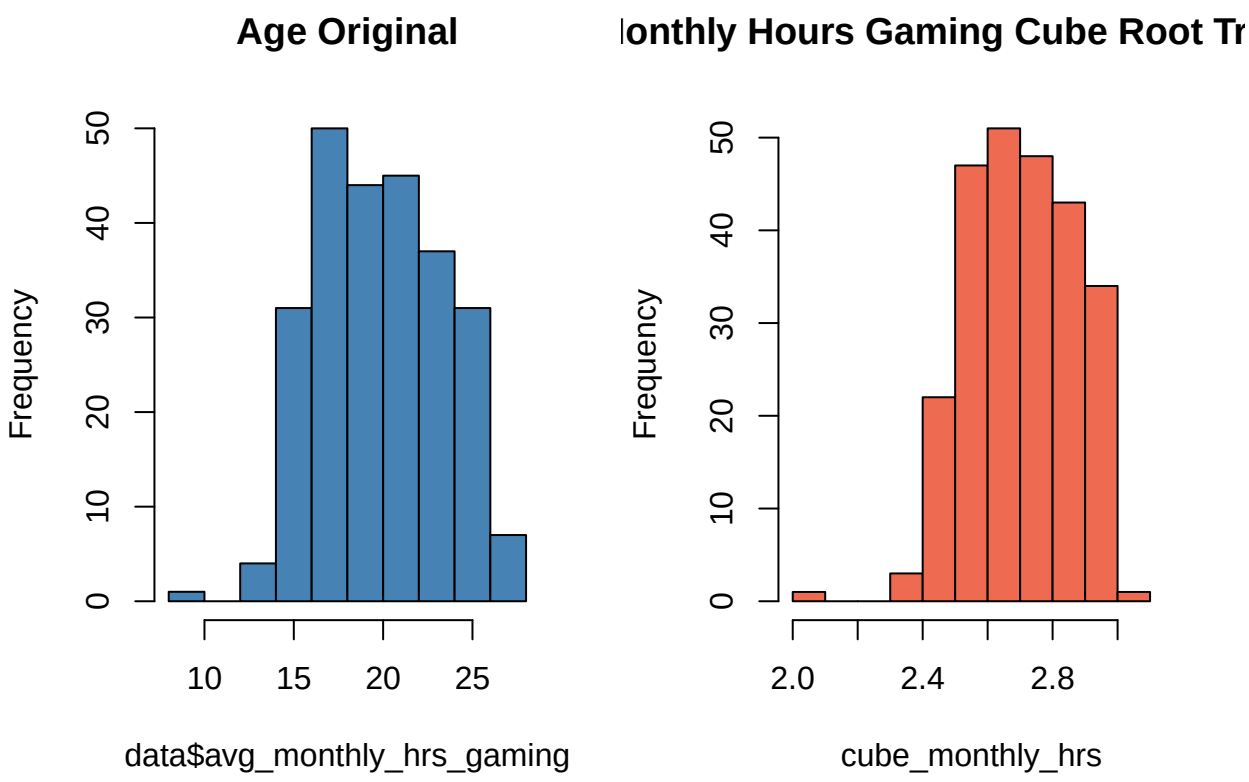





```
shapiro.test(sqrt_monthly_hrs)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt_monthly_hrs
## W = 0.97577, p-value = 0.0002864
```

```
cube_monthly_hrs <- data$avg_monthly_hrs_gaming^(1/3)
hist(data$avg_monthly_hrs_gaming, col='steelblue', main='Age Original')
hist(cube_monthly_hrs, col='coral2', main='Avg. Monthly Hours Gaming Cube Root Transformed')
```




```
shapiro.test(cube_monthly_hrs)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cube_monthly_hrs
## W = 0.9738, p-value = 0.0001439
```

```
cube_monthly_hrs_p_value <- shapiro.test(cube_monthly_hrs)$p.value
if (cube_monthly_hrs_p_value < significance) {
  print(paste("Cube Root Transform of Avg. Monthly Hours Gaming is less than ", significance))
} else {
  print(paste("Cube Root Transform of Avg. Monthly Hours Gaming is less than ", significance))
}
```

```
## [1] "Cube Root Transform of Avg. Monthly Hours Gaming is less than  0.05"
```