Stats101A, Spring 2023 - Homework 3

Luke Villanueva - 206039397

04/21/23

```
# libraries used
library(tidyverse)
```

Problem 1

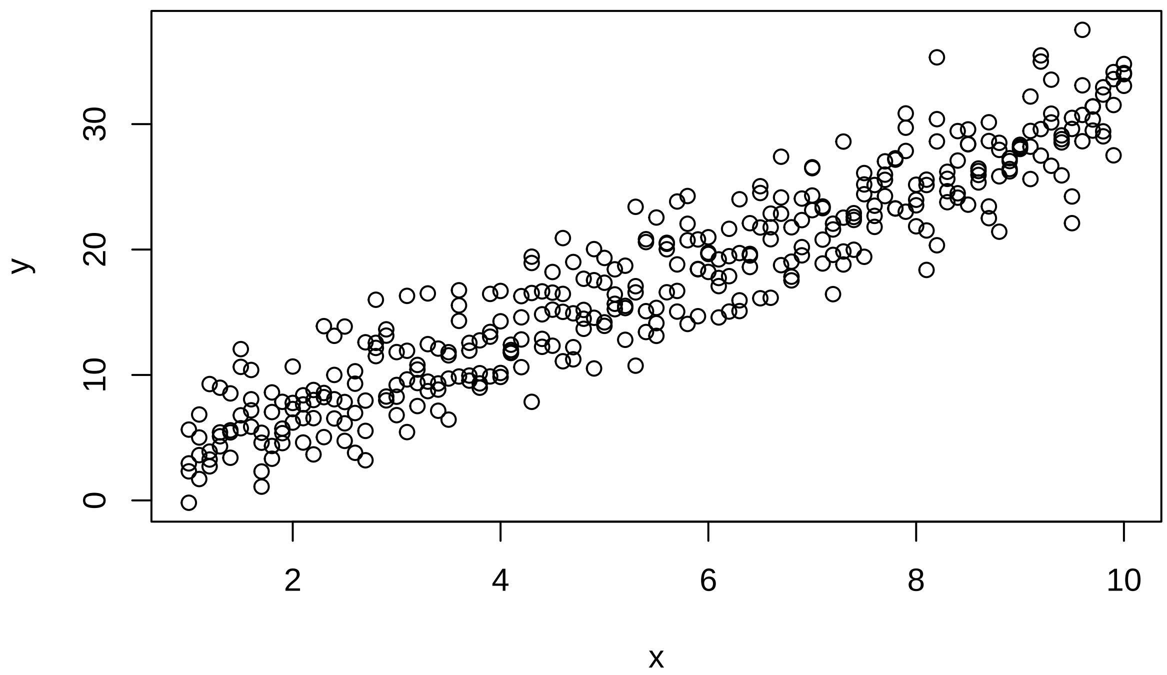
A.

a.

```
linear <- function(beta_0, beta_1, sigma, x, random.seed)
{
  # set seed
  set.seed(random.seed)

  # make result
  res <- numeric(0)

  # append result
  for(val in x)
  {
    res <- c(res, (beta_0 + beta_1 * val + rnorm(1, 0, sigma)))
  }

  # output
  res
}
```

b.

```
x <- rep(seq(1,10, by = 0.1), 4)

y <- linear(1, 3, 3, x, 123)

plot(x, y)
```



B.

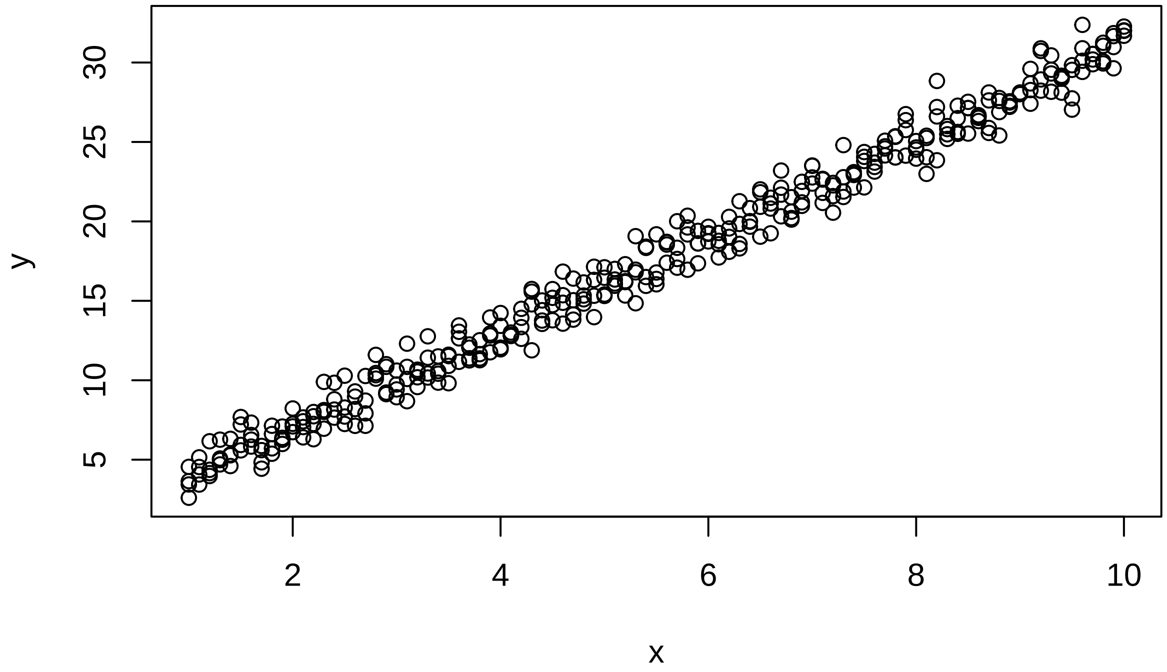
```
cor(x,y)
```

```
## [1] 0.9397388
```

The correlation coefficient is 0.9397.

C.

```
y <- linear(1, 3, 1, x, 123)

plot(x, y)
```



```
cor(x,y)
```

```
## [1] 0.9926201
```

The correlation coefficient is 0.99.

Problem 2

```
# initial loading of data
arms <- read.csv(file = "armspans2022_gender.csv", header = TRUE)

# view csv
glimpse(arms)
```

```
## Rows: 46
## Columns: 5
## $ height     <dbl> 74.00, 65.00, 60.00, 69.75, 70.00, 68.00, 64.00, 68.00, 68.~
## $ armspan    <dbl> 76.0, 65.0, 53.0, 69.0, 72.0, 70.5, 60.0, 67.0, 67.0, 60.0,~
## $ is.female  <int> 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1,~
## $ compmother <chr> "Taller", "Taller", "Shorter", "Taller", "Taller", "Taller"~
## $ compfather <chr> "Taller", "About the same", "Shorter", "About the same", "A~
```

```
# na check
any(is.na(arms$armspan))
```

```
## [1] TRUE
```

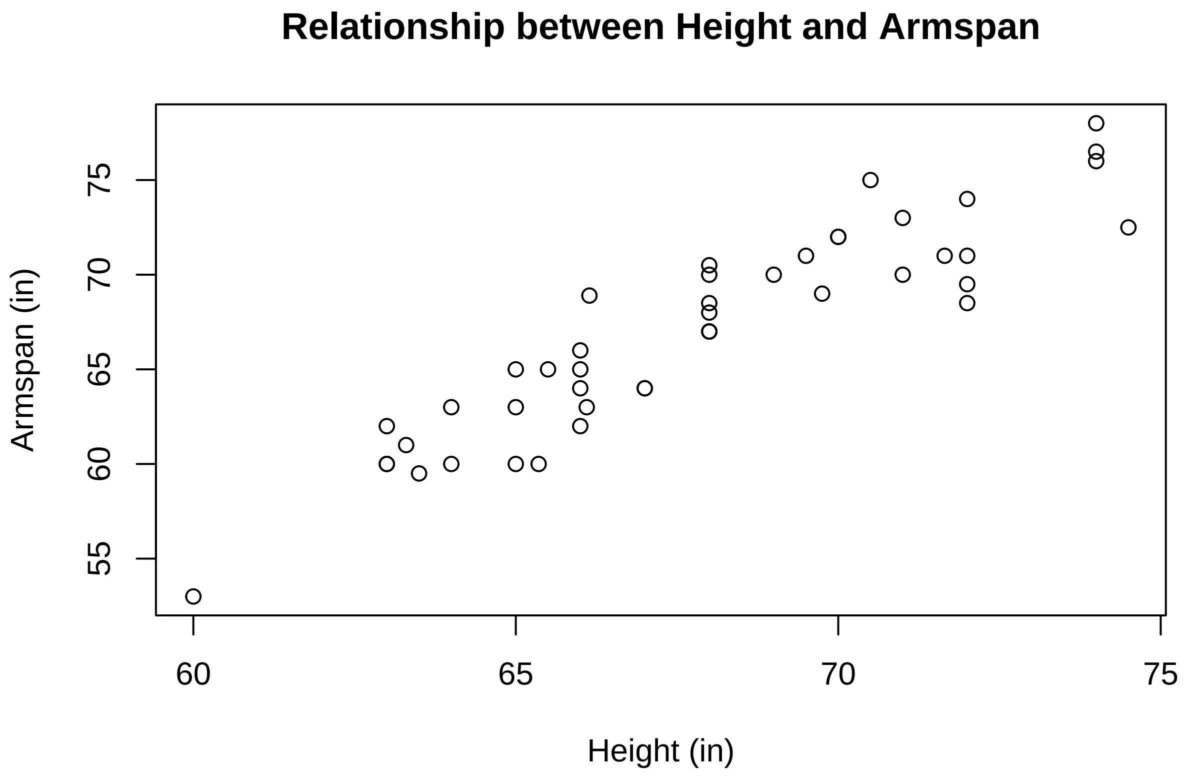
```
# index to those missing armspan measurement
arms[is.na(arms$armspan),]
```

```
##    height armspan is.female compmother compfather
## 33     72      NA         0     Taller     Taller
```

```
# update arms to exclude NA observation
arms <- arms[!is.na(arms$armspan),]
```

a.

```
plot(arms$height, arms$armspan, ylab = "Armspan (in)", xlab = "Height (in)", main = "Relationship betwe
```



```
# get correlation
cor(arms$height, arms$armspan)
```

```
## [1] 0.92147
```

There seems to only be one outlier, which is near the 60 inch mark in the height data. However, the relationship seems to be following the trending pattern of the rest of the data. And so, this may simply be an outlier in terms of height, but it is not abnormal behavior to the rest of the data in terms of the relationship between armspan and height.

In addition, the correlation coefficient is 0.92, which implies there is a strong positive linear relationship between armspan and height.

b.

```
lin <- lm(arms$armspan ~ arms$height)

coeff <- lin$coefficients

# coefficients
coeff[1]
```

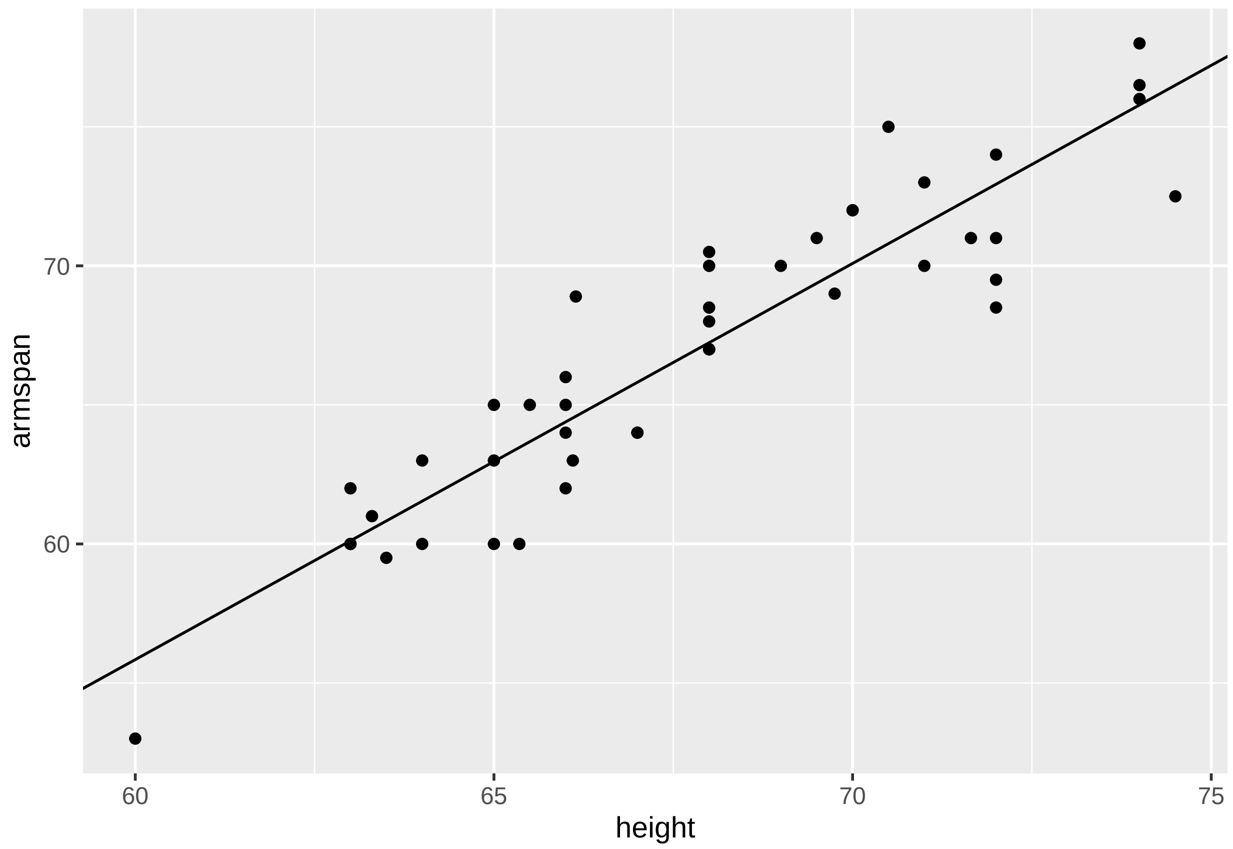
```
## (Intercept)
##    -29.6353
```

```
coeff[2]
```

```
## arms$height
##    1.424591
```

The equation for the line of best fit is: $-29.6353 + 1.424591x$

```
ggplot(data = arms) + geom_point(mapping = aes(x = height, y = armspan)) + geom_abline(intercept = coef
```



c.

$-29.6353 + 1.424591(64) \implies -29.6353 + 91.173824 = 61.5385$

My height is 64 inches, so the predicted armspan based on the linear model will be 61.5385 inches.

Since my armspan is about 63.5 inches, the residual is 1.9615.

d.

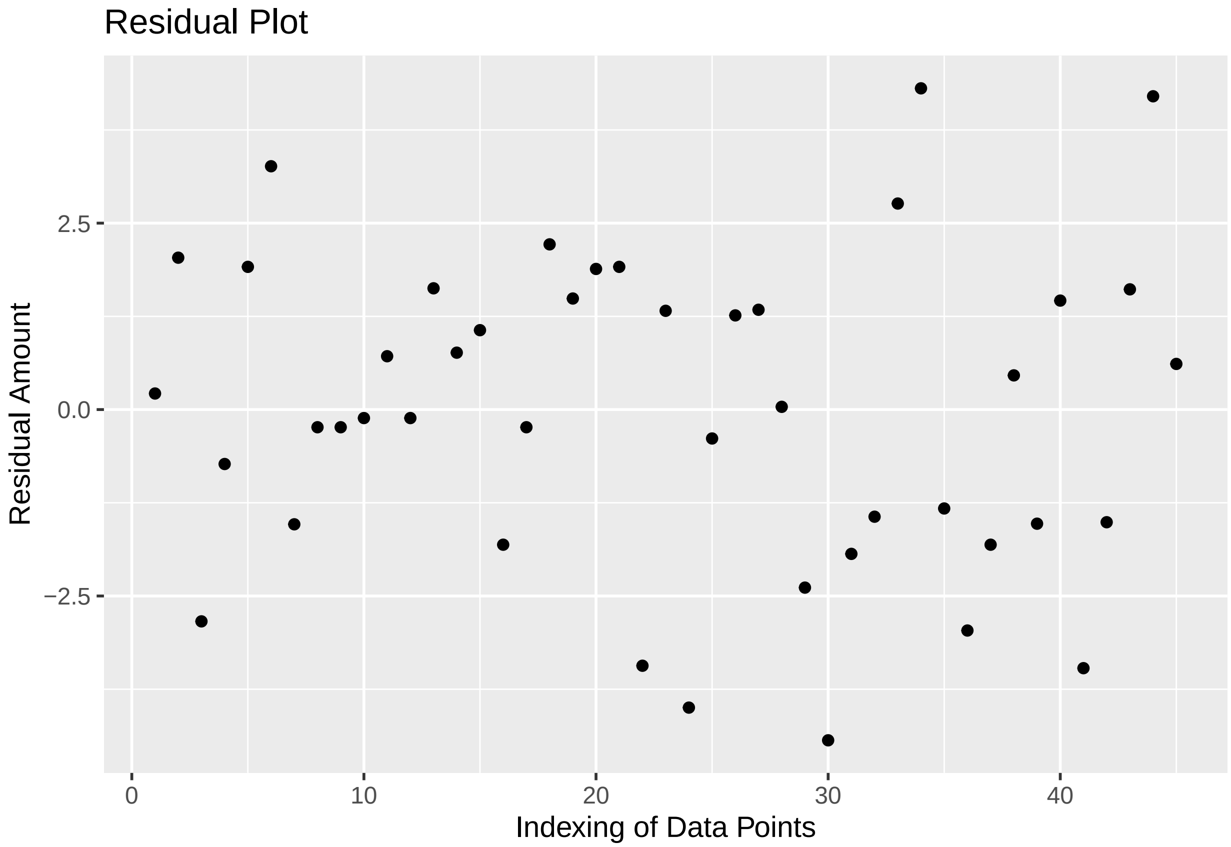
Phelps's armspan is not unusual if we base our conclusion off of the data we are given. This is because if we predict Phelps's armspan using the model we made in part c, then we get that Phelps's predicted armspan is around 78.633.

$-29.6353 + 1.424591(76) \implies -29.6353 + 108.268916 = 78.6336$

And since his actual armspan is reported to be around 79 inches, the residual is not abnormally big. And so, based off our data, Phelps's armspan is not unusual.

e.

```
ggplot() + geom_point(mapping = aes(x = as.numeric(names(lin$residuals)), y = lin$residuals)) + labs(ti
```



Since the residual plot is sporadic and seems to hold no pattern, this implies that the linear model placed onto the data is a valid model for the data.

Problem 3

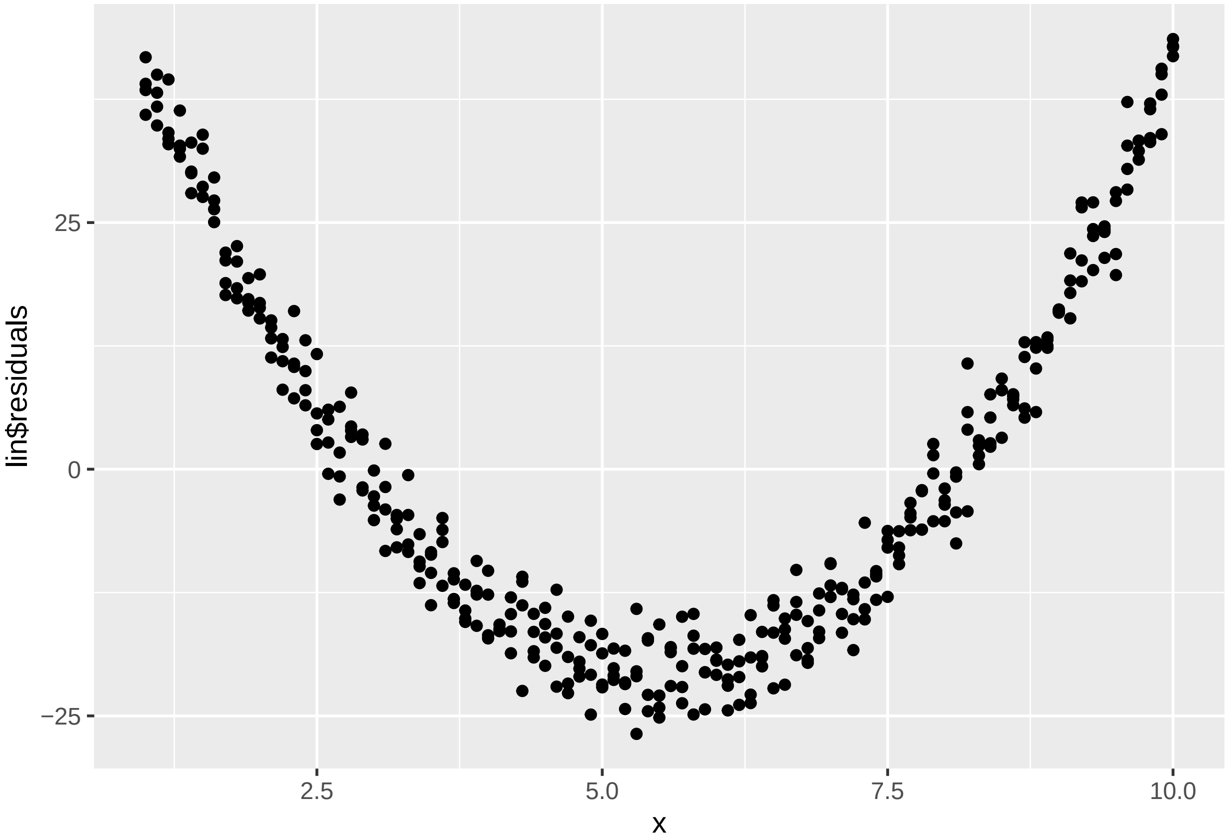
a.

```
# quadratic data
quadratic <- function(a, b, c, sigma, x, random.seed)
{
  # set seed
  set.seed(random.seed)

  # make result
  res <- numeric(0)

  # append result
  for(val in x)
  {
    res <- c(res, (a + (b * val) + (c * (val**2)) + rnorm(1, 0, sigma)))
  }

  # output
  res
}

# recreate x
x <- rep(seq(1,10, by = 0.1), 4)

# choose a = 1, b = 3, c = 3, sigma = 3, seed = 123
y <- quadratic(1, 3, 3, 3, x, 123)
```

The inputs for the data are: a = 1, b = 3, c = 3, sigma = 3, x = the last x input, random.seed = 123

b.

```
# get linear model
lin <- lm(y~x)

# make residual plot
ggplot() + geom_point(mapping = aes(x = x, y = lin$residuals))
```



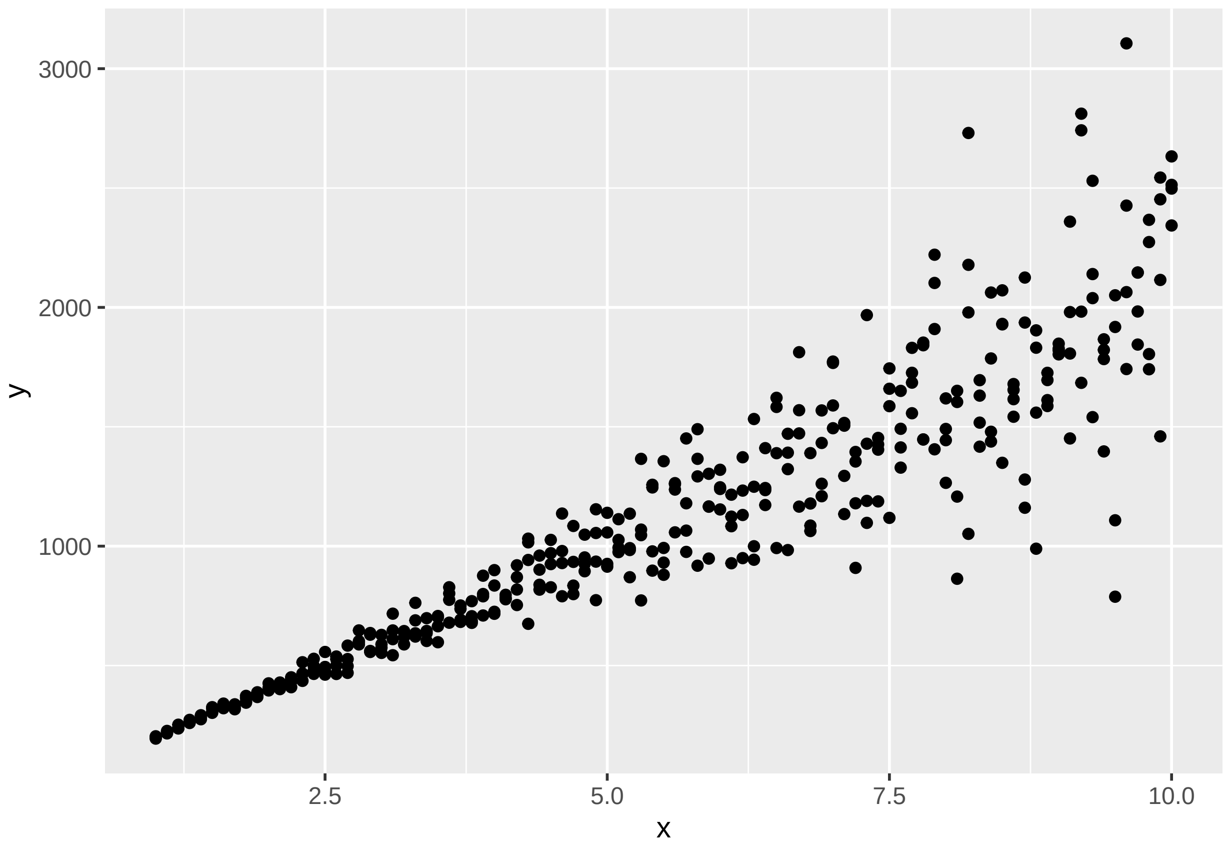
The residuals are plotted in a quadratic pattern.

c.

If the residual plot show cases a non-linear pattern, then that implies there is a nonlinear pattern in the data as well. This is because there is a non-linear pattern in the differences between the points and the fitted line. So, the non-linear residual plot tells there is a non-linear trend in the data.

d.

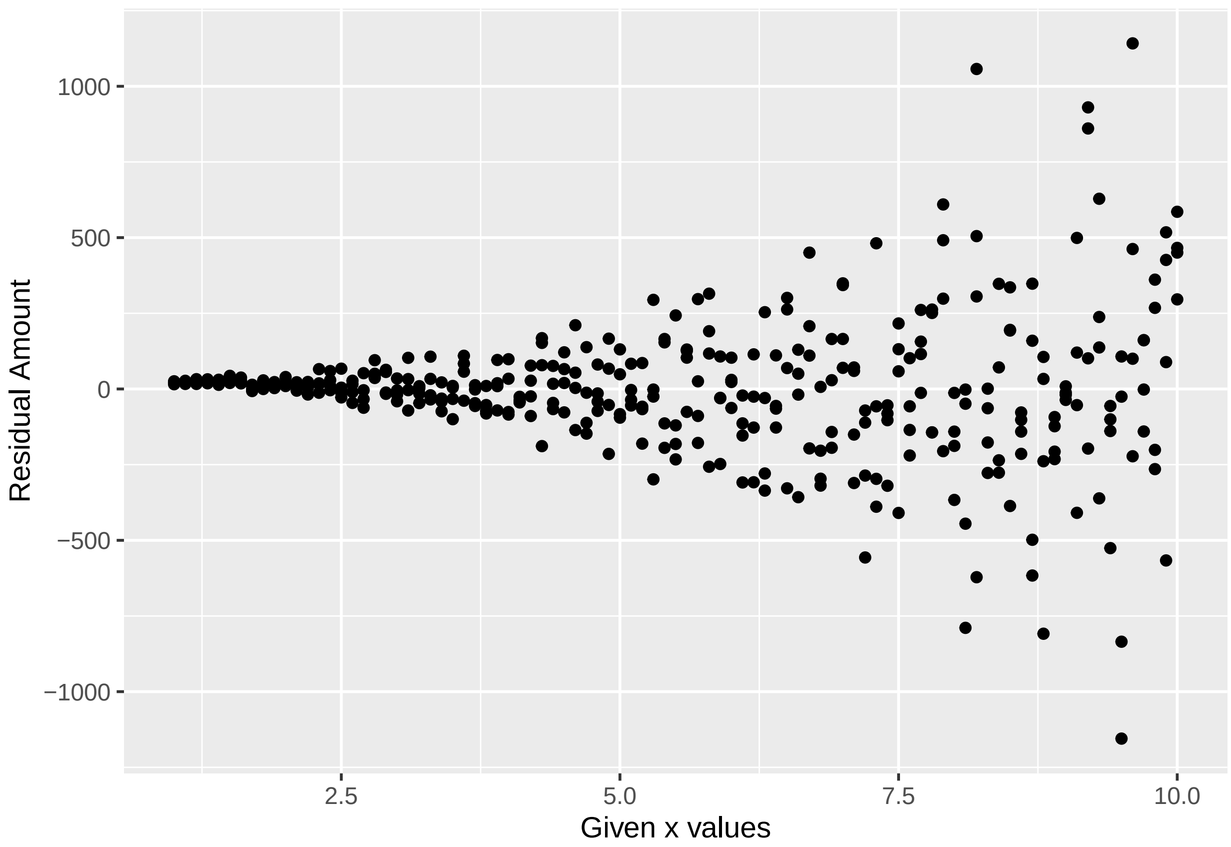
```r
notConstLinear <- function(a, b, sigma, random.seed)
{
  # set seed
  set.seed(random.seed)

  # make result
  res <- numeric(0)

  # append result
  for(val in x)
  {
    res <- c(res, (a + b * val + rnorm(1, 0, (sigma * val^2))))
  }

  # output
  res
}

x <- rep(seq(1,10, by = 0.1), 4)

y <- notConstLinear(1, 200, 5, 123)

ggplot() + geom_point(mapping = aes(x = x, y = y))
```



The plot starts of as linear, but as the predictor variable grows, the variability grows quickly. And so, this influences the data in the latter part of the graph to have high variance compared to the beginning part of the graph.

e.

```
# linear model of fanning graph
lin <- lm(y~x)

# residual plot of fanning graph
ggplot() + geom_point(mapping = aes(x = x, y = lin$residuals)) + labs(x = "Given x values", y = "Residu
```



The residual plot tells that the variance between the values is not majorly constant. This means there is a non-linear factor affecting the error amount in the data, implying that a linear model will not fit the data.

Problem 4

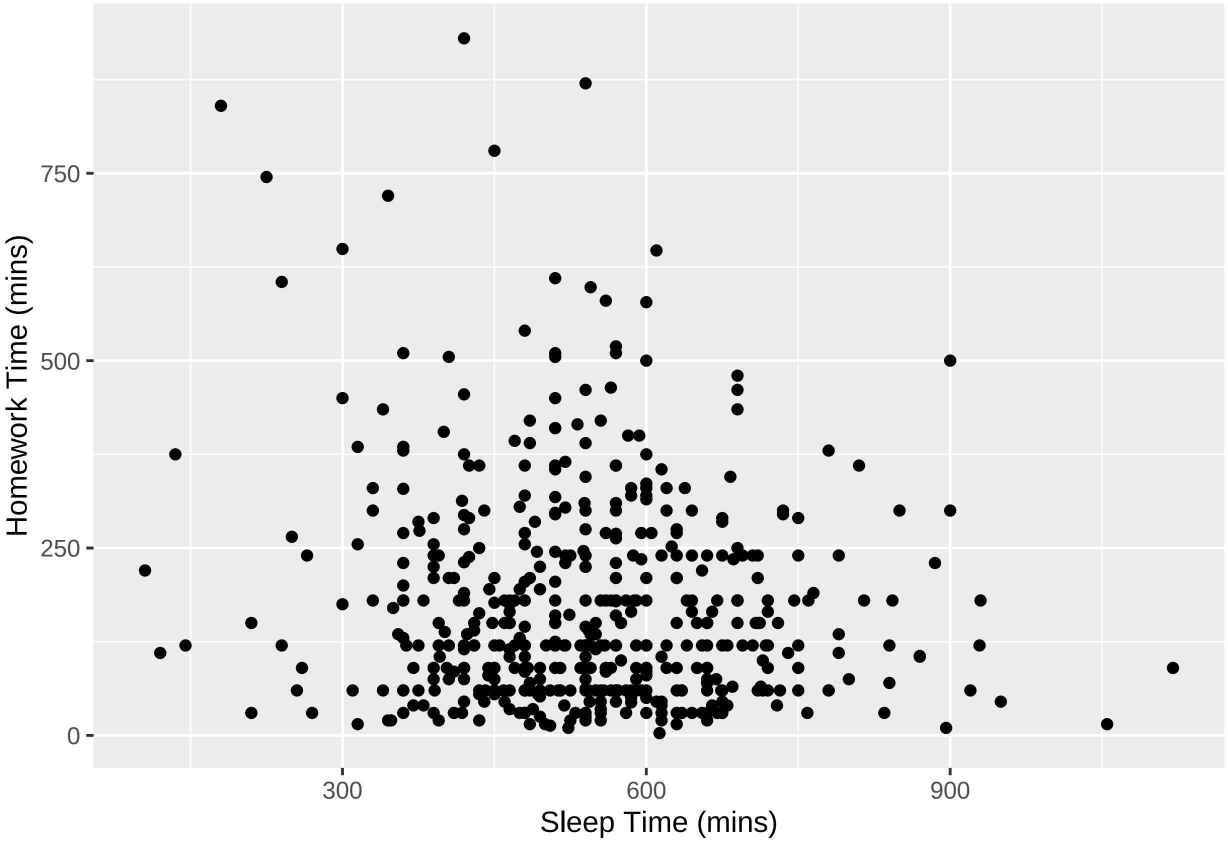
```
atus <- read.csv(file = "C:/Users/lavil/Downloads/atus.csv")

glimpse(atus)
```

```
## Rows: 12,443
## Columns: 48
## $ caseid              <dbl> 2.01201e+13, 2.01201e+13, 2.01201e+13, 2.01201e+13,~
## $ state               <chr> "Florida", "Virginia", "Georgia", "Florida", "South~
## $ age                 <int> 38, 17, 20, 58, 65, 19, 46, 24, 34, 40, 53, 54, 24,~
## $ gender              <chr> "Female", "Female", "Female", "Male", "Female", "Ma~
## $ citizen             <chr> "Native, Born in USA", "Native, Born in USA", "Nati~
## $ marital_stat        <chr> "Married", "Never married", "Never married", "Marri~
## $ veteran             <chr> "Non-Veteran", "Non-Veteran", "Non-Veteran", "Veter~
## $ active_armedforces <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ emp_status          <chr> "Unemployed, Looking", "Unemployed, Looking", "Empl~
## $ multi_jobs          <chr> "No answer", "No answer", "No", "No", "No", "No", "~
## $ work_class          <chr> "No answer", "No answer", "Private, for profit", "S~
## $ retired             <chr> "No answer", "No answer", "No answer", "No answer",~
## $ fulltime_emp       <chr> "No answer", "No answer", "Full time", "Full time",~
## $ hours_worked       <int> NA, NA, NA, 40, 16, 32, NA, 70, NA, NA, 40, NA, 20,~
## $ fam_income         <chr> "$40,000 to $49,999", "Less than $5,000", "Less tha~
## $ household_size     <int> 4, 7, 2, 2, 1, 2, 1, 2, 1, 4, 2, 1, 1, 3, 3, 3, 4, ~
## $ household_kids     <int> 2, 4, 1, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 1, 1, 2, 2, ~
## $ household_child    <chr> "Yes", "Yes", "Yes", "No", "No", "No", "No", "No", ~
## $ phys_challenge     <chr> "No difficulty", "No difficulty", "No difficulty", ~
## $ travel             <int> 2, 350, 4, 46, 120, 65, 0, 105, 119, 65, 82, 16, 24~
## $ phone              <int> 0, 0, 0, 0, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ volunteer          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ religion           <int> 0, 0, 0, 0, 240, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ sports             <int> 0, 0, 0, 0, 60, 60, 0, 0, 0, 0, 0, 0, 60, 0, 0, 0, ~
## $ social             <int> 528, 120, 13, 588, 415, 120, 835, 283, 452, 165, 22~
## $ food               <int> 60, 60, 55, 160, 40, 10, 0, 40, 30, 50, 50, 122, 5,~
## $ gov_civic          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ household          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 30, 5, 0, 0,~
## $ pro_services       <int> 0, 0, 0, 0, 195, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ purchasing         <int> 5, 0, 0, 10, 120, 5, 0, 160, 35, 45, 62, 0, 30, 110~
## $ education          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ work               <int> 0, 0, 518, 10, 0, 340, 0, 0, 3, 0, 475, 0, 353, 250~
## $ care_nonhousehold  <int> 0, 0, 425, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ care_household     <int> 0, 210, 15, 0, 0, 0, 0, 0, 0, 275, 0, 0, 0, 76, 45,~
## $ household_chores   <int> 95, 25, 15, 6, 15, 30, 65, 70, 26, 345, 85, 115, 55~
## $ personal_care      <int> 750, 675, 395, 620, 195, 810, 540, 735, 765, 495, 4~
## $ sleep              <int> 750, 649, 330, 620, 150, 720, 540, 690, 720, 495, 4~
## $ groom              <int> 0, 26, 65, 0, 40, 90, 0, 45, 45, 0, 20, 40, 30, 40,~
## $ health_related     <int> 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ eating             <int> 60, 60, 55, 160, 40, 10, 0, 40, 30, 50, 50, 122, 5,~
## $ class              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ homework           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ socializing        <int> 0, 0, 0, 114, 355, 0, 0, 110, 0, 45, 0, 292, 0, 0, ~
## $ holiday            <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ day                <chr> "Sunday", "Sunday", "Saturday", "Saturday", "Thursd~
## $ year               <int> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 201~
## $ month              <chr> "January", "January", "January", "January", "Januar~
## $ date               <chr> "2012-01-29", "2012-01-29", "2012-01-28", "2012-01-~
```

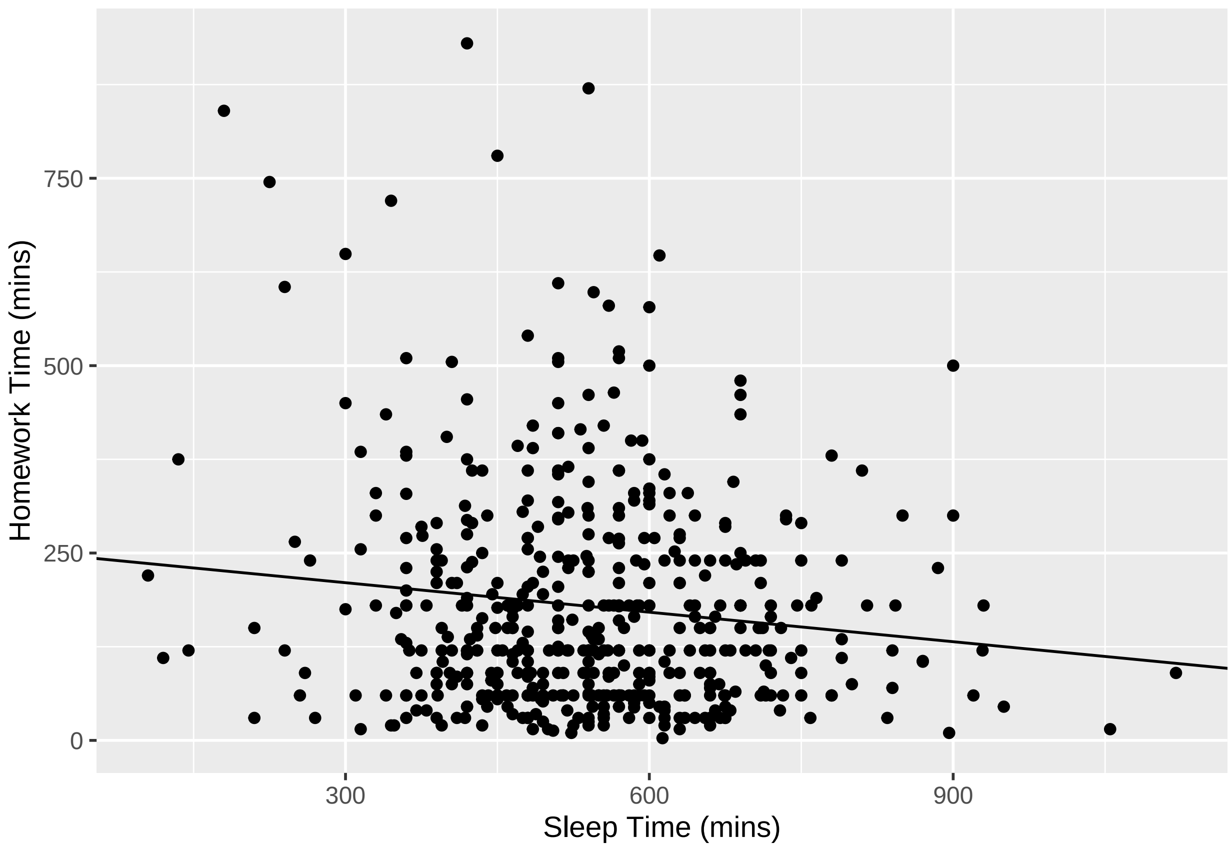
```r
# subset to those who did homework
atus1 <- subset(atus, homework > 0)
```

a.

```r
# make scatter plot of sleep to homework time
ggplot(atus1) + geom_point(mapping = aes(x = sleep, y = homework)) + labs(x = "Sleep Time (mins)", y =
```



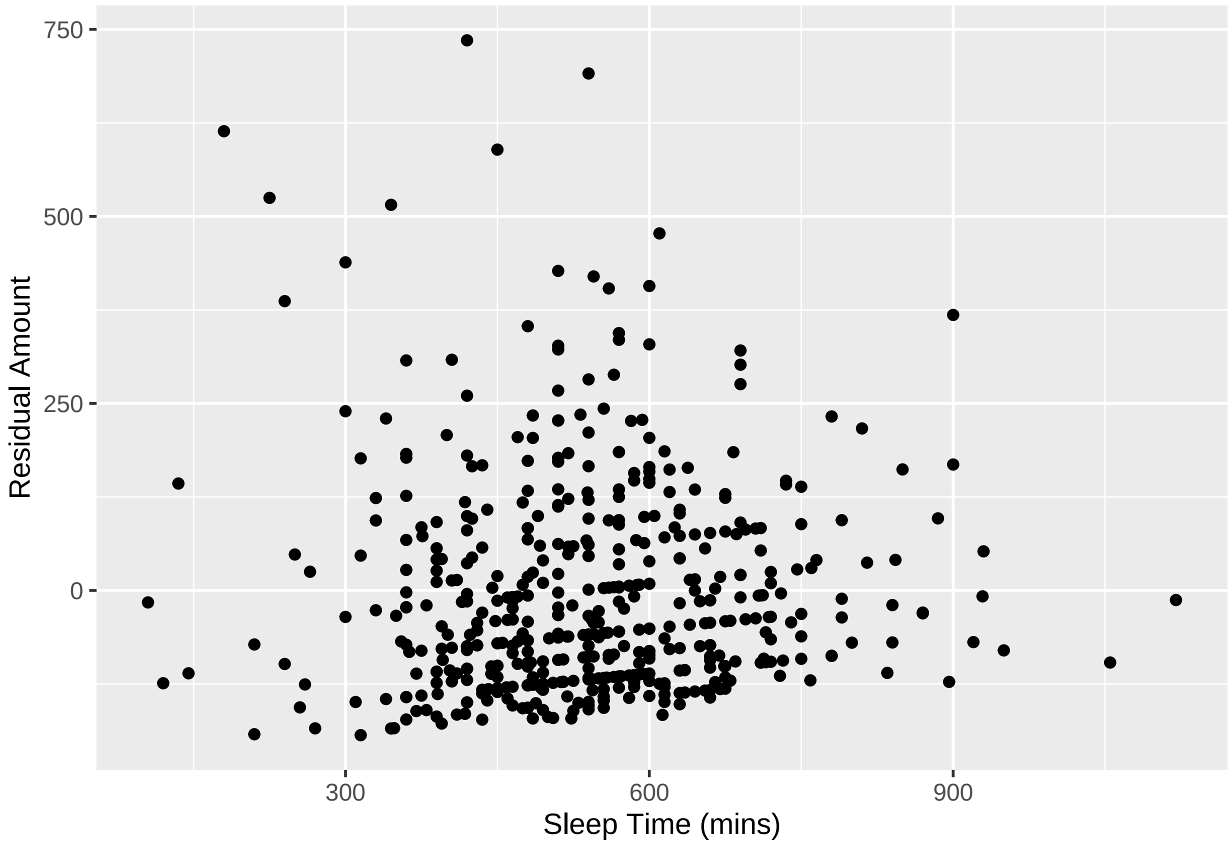
```
# linear model the two variables
lin <- lm(atus1$homework ~ atus1$sleep)

# plot the linear model onto the scatter plot
ggplot(atus1) + geom_point(mapping = aes(x = sleep, y = homework)) + labs(x = "Sleep Time (mins)", y =
```



Even if the graph looks to have no linear pattern, there seems to be a general descending pattern.

b.

```
ggplot(atus1) + geom_point(mapping = aes(x = sleep, y = lin$residuals)) + labs(x = "Sleep Time (mins)",
```



The residual plot is not randomly scattered because there is a closely clustered group towards the bottom of the graph. And so, this implies that possibly a linear model is not an appropriate model for the given data.

Problem 5

a.

Null hypothesis: The female average chore time is less than or equal to the male average chore time.

Alternative hypothesis: The female average chore time is greater than the male average chore time.

```
t.test(x = subset(atus, gender == "Female")$household_chores, y = subset(atus, gender == "Male")$househo
```

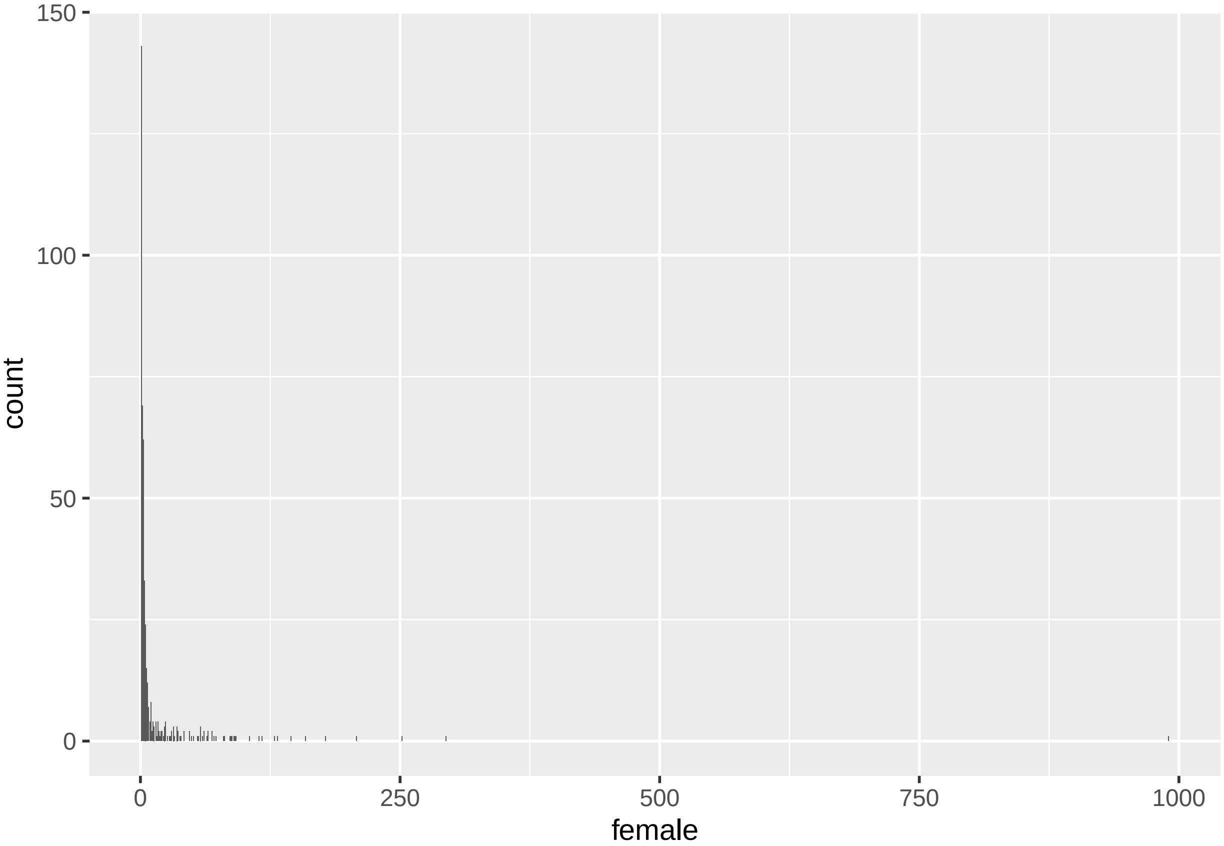
```
##
##  Welch Two Sample t-test
##
## data:  subset(atus, gender == "Female")$household_chores and subset(atus, gender == "Male")$househol
## t = 20.381, df = 12311, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  46.19276      Inf
## sample estimates:
## mean of x mean of y
##  140.2893   90.0410
```

Based on the results of the t test, the t statistic is 20.381, which means that the test places the difference between the means to be 20 standard deviations away from the usual output. Since the p-value is 0.0175 and the significance level needed is 0.05, we can reject the null hypothesis and say that the true mean of the amount of chore time females spend is greater than the amount of chore time males spend.

b.

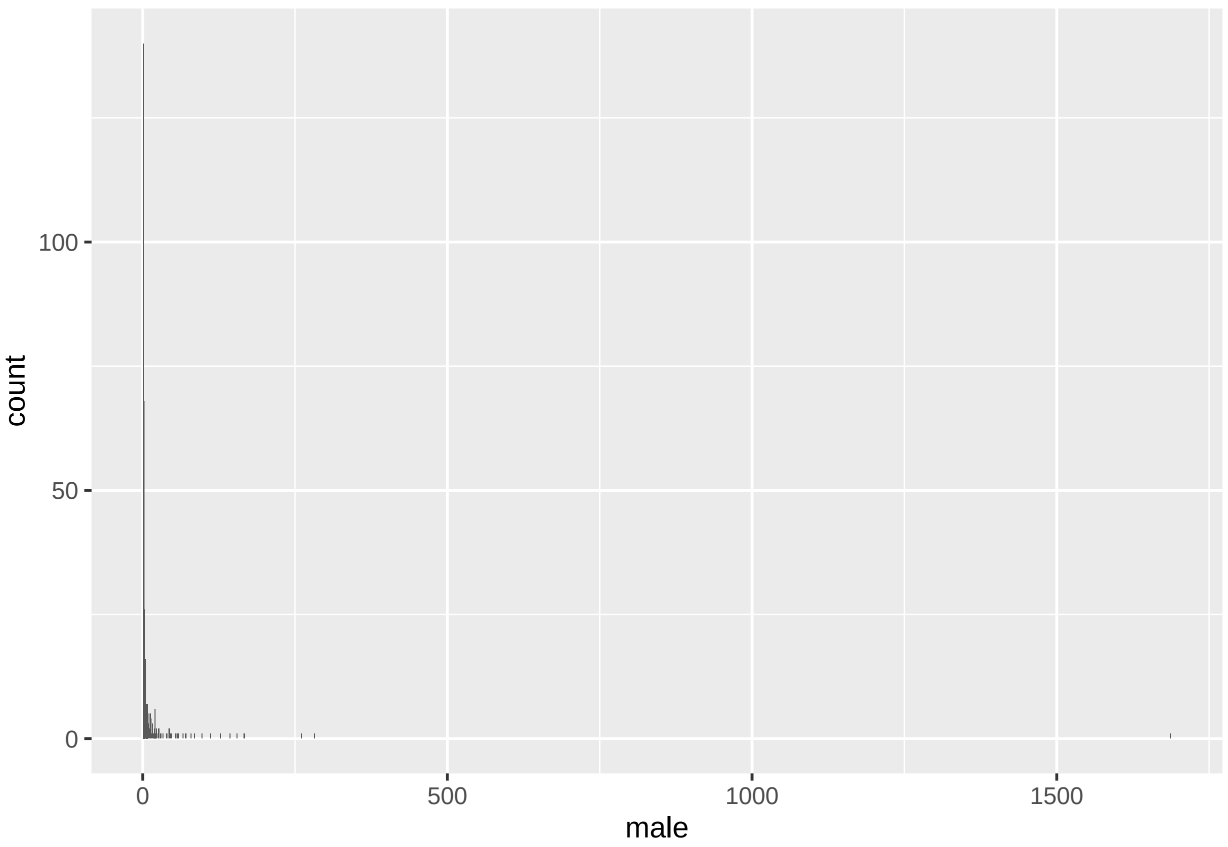
```
female <- table(subset(atus, gender == "Female")$household_chores)
male <- table(subset(atus, gender == "Male")$household_chores)

ggplot() + geom_bar(aes(female))
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```



```
ggplot() + geom_bar(aes(male))
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```



For the p-value to provide meaningful conclusions, the data has to be somewhat normal shape. As we can see from the graphs of the frequency of female and male chore times, both sets of data are highly skewed. In addition, there are highly extreme outliers, so the t-test is not as accurate.