```r
# Statistical Methods 2022

library(ggplot2)

data("women") # Loading dataset

head(women) # Seeing what is in this data, two variables' Heights and Weights.

View(women) # Viewing women dataset. Observation of 15 and two variable Height and Weight.
# (it's about individuals from a sample not whole population but some women in a study and
# Variable that characteristic has individual been measured from statistician's perspective)

help(women)  # For more information about women data.
# It's about Average Heights and Weights for American Women.
# Heights "in" and Weight "lbs" and it's 15 women from age 30 to 39.

df <- data.frame(women) # Making women dataset to a dataframe so i can use function

df # showing the data frame in console, 1-15 women with 15 different heights and weights.

# plotting women dataframe x with height and y with weight, geom_smooth lm making line with formula 'y ~ x'

gg <- ggplot(df, aes(x=height, y=weight)) +
  geom_point() +
  labs( x = "Height (in)", y = "Weight (lbs)", title = "Average Heights and Weights for American Women from age 30-39") +
  geom_smooth(method = "lm", se = T ) # using formula = 'y ~ x' also know as least-squares regression line

# showing the plot
```

print(gg)

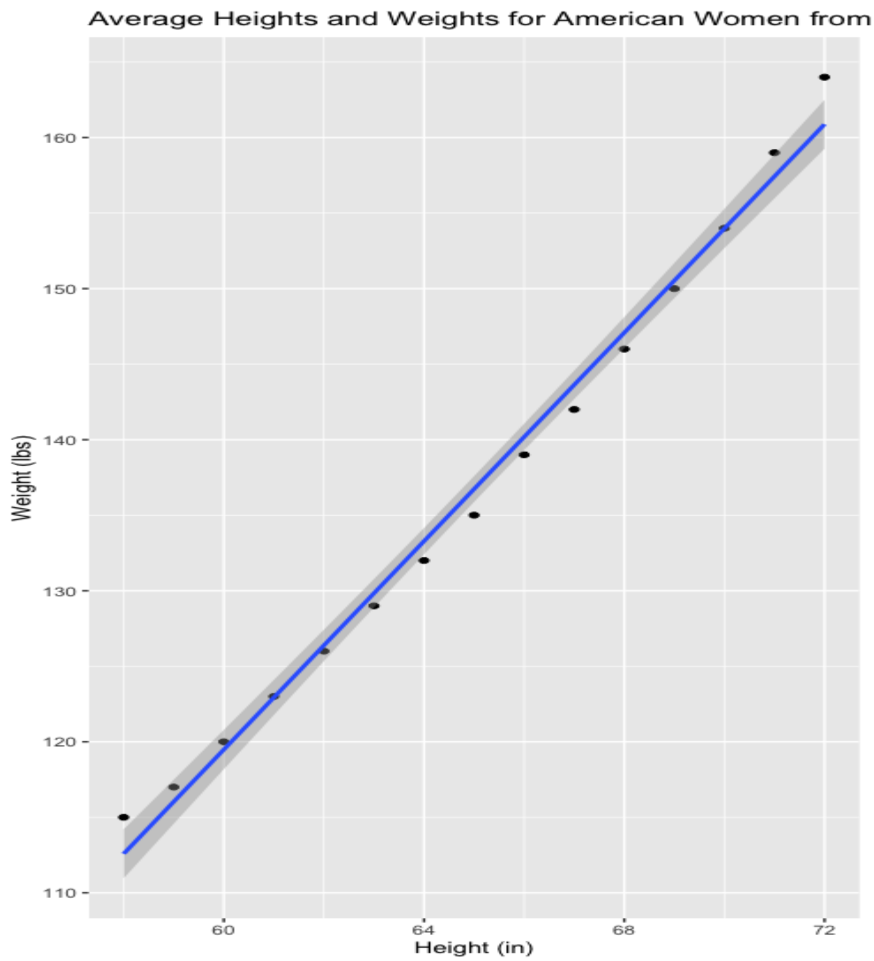


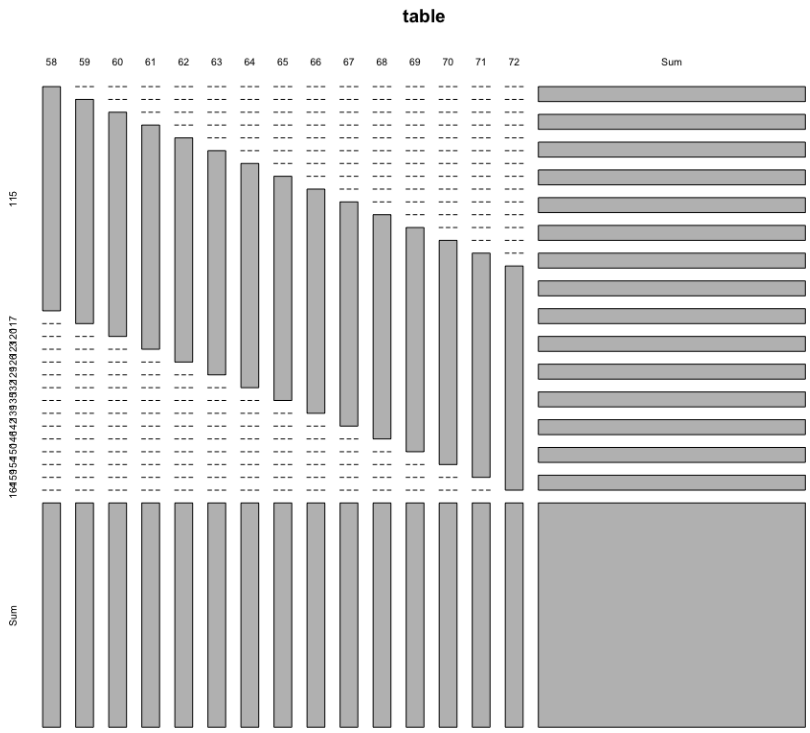
table <- addmargins(table(df$height, df$weight)) #table #two way table from this data frame.

print(table) #view table.

plot(table) # visualization of the table.

With table i can see categorical variable and its frequency as as output.

1 count for each variable.



Question,

What is the prediction that a women weight more, when she is taller?

What is Correlation and coefficient of determination with those two variables and what is the min mean and max value of each variable?

It's a Linear correlation, the more x gets lager (height) then y also increases (weight)

It's very close to the point so a perfect linear correlation with visualization.

The correlation line goes up so it's a positive correlation, also strong!

With this visualization i can confirm its strong positive correlation.

cor(df$height, df$weight) #the correlation between two variables.

cor(df$height, df$weight) # correlation quantitative analyze use the $ to pick what variable I want to analyze but this data has only two so the cor(df) works fine!

The correlation is 0.995 so very strong! # its very strong correlation close to 1.

0 mean there is no linear correlation and with 1 it's very strong the max value.

The formula for correlation coefficient

```
# r = (n * (sumX * sumY) - (sumX) * (sumY) ) / sqrt( [n * sumX^2 - (sumX)^2 ] * [n * sumY^2 -
(sumY)^2¨ ] )

# OUTCOME : 0.9954948 = 99,5%


#  Least squares line equation

model <- lm(formula = height~weight, data=df)


# Least squares line equation.

# where a = ŷ - b*x(hat) b i slope and a is intercept.

# ŷ = b*x + a, where b = (n * sumX * sumY) - (sumX) * (sumY) ) / (n * sumX^2 - (sumX)^2)

# a = ¯y- b*¯x

# ŷ = 25.7234557 + 0.2872492x that is the equation for the least-squares regression.


model # showing the model in the console. # Call: lm(formula = height ~ weight, data = df) y =
25.7235 + 0.2872..x


summary(model) #summary of the model


# The multiple R-squared r^2 is very close to 1

 #  Coefficient of Determination

summary(model)$r.squared # 0.991

# r-squared or r^2 measure a value from 0 to 1, where a value of 1 shows a perfect fit.

# 99,1 % dependent variable is predicted by the independent variable.

# The intercept and slope by using coef() function.

# The r^2 or called coefficient of determination show what is the predicted outcome it will happen in
the future with those variables.

# With 0.991 or 99,1 % is very more likely it would happen when a woman is taller the more, she
would weight more are very likely the outcome!

# formula # r = (n * (sumX * sumY) - (sumX) * (sumY) ) / sqrt( [n * sumX^2 - (sumX)^2 ] * [n * sumY^2
- (sumY)^2¨ ] )

# When r^2 = 1 - (the sum regression of squared / total sum of squares)


coef(model)
```

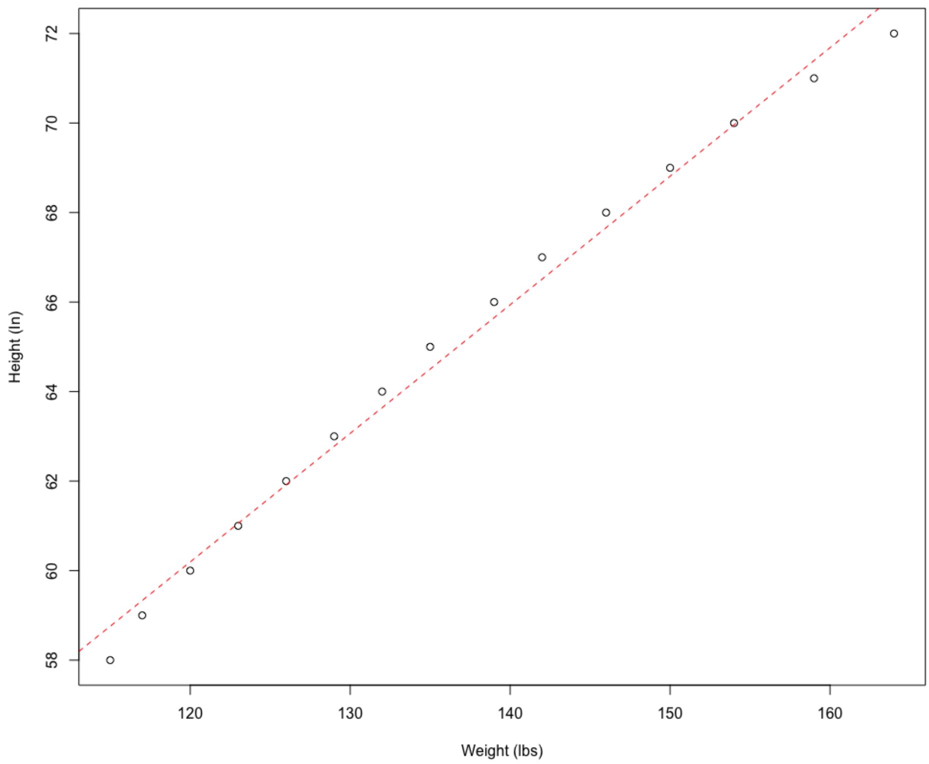
the intercept 25.7235 and slope 0.2872..x

a <- coef(model)[1] # picking 1 variable

b <- coef(model)[2] # picking the 2 variable

ploting women and adding least-squares regression line to this plot.

```
p <- plot(df$weight, df$height,
       xlab = "Weight (lbs)",
       ylab = "Height (In)")
abline(a = coef(model)[1],
     b = coef(model)[2],
     lty = 2,
     col = "red")
```



residual is the point from the model line show how far away,

its very small from each other it shows much strong east-squares line equation.

Residuals:

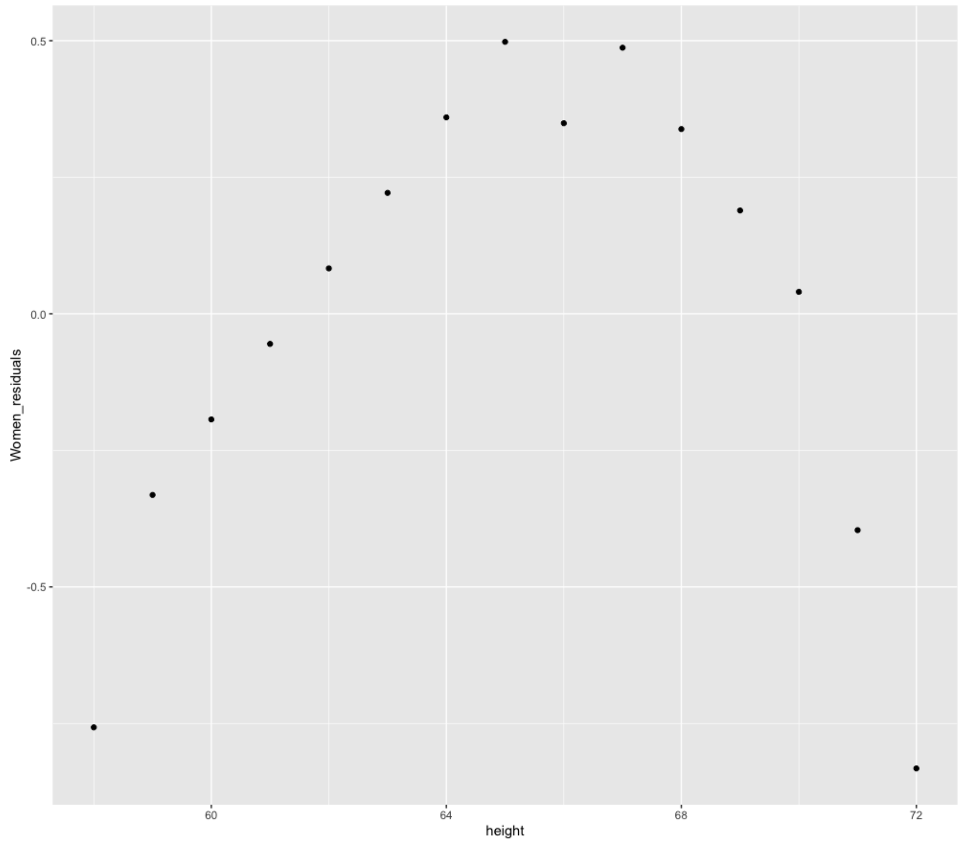
Min   1Q   Median   3Q   Max

-0.83233 -0.26249 0.08314 0.34353 0.49790

Women_residuals <- model$residuals

plotres <- ggplot(df, aes(x=height, y=Women_residuals)) + geom_point()

plotres



min(df$height) # getting the minimum of height

min(df$weight) # getting the minimum of weight

median(df$height) # getting the median of height

median(df$weight) # getting the median of weight

median is the middle of a sorted list of numbers which is 65 and 135 for those 2 variables.

(n + 1) ÷ 2th n = number of data the middle value in an ordered data set.

Median is the 50 % of the data.

max(df$height) # getting the maximum of height

max(df$weight) # getting the maximum of weight

mean(df$height) # getting the average of height

mean(df$weight) # getting the average of weight

Mean = (the sum of all values / divided by n = number of data)

mode(df$height) # getting the most repeated value

mode(df$weight) # getting the most repeated value

Outcome numeric cause it only shows 1 per variable so no mode in this dataset.

mode function will pick the most repeated value.

Making weight and height to own function also unlisting those variables so it works for function as "quantile".

weight_unlist <- unlist(df$weight)

height_unlist <- unlist(df$height)

This will show me first value and middle value and max on the whole dataset with height and weight together

So same as min, median and max. because 0% is the first element and 50% is the middle of the element and 100% is the max of the element.

quantile(weight_unlist, probs = data.frame(0,.5,1)) # The 0 percentile is 115 lbs in the 1st quartile percentile and 50th percentile the half is 135 lbs in the 2rd quartile/ 3rd quantile the median and 100 th the max percentile are 164 lbs 4th quantile

quantile(height_unlist, probs = data.frame(0,.5,1)) # The 0 percentile is 58 in is the 1 st quartile and 50th percentile the half is 65 in the 2rd quartile/ 3rd quantile the median and 100 th the max percentile is 72 in 4th quantile.

summary(df) # summary of the data set in height minimum are 58 median 65 and maximum value 72 lbs.

For weight minimum are 115 median 135 and maximum 164 in.

With this summary function it also shows min 1 st quartile, median, mena 3rd quartile, max

With the final analyze we can say the more a women weight the taller she is with this prediction of 99.1% of the times. It's a Linear correlation, the more x gets lager (height) then y also increases (weight). It's very close to the point so a perfect linear correlation with visualization and with the output in R. The correlation line goes up so it's a positive correlation, also strong! With this visualization and calculation, I can confirm its strong positive correlation and coefficient of determination.