DATA 606 Fall 2016 - Final Exam

Shyam BV

December 9, 2016

Contents

Part I

Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively.

a. Describe the two distributions (2 pts).

Figure A:

Given figure is a histogram chart. The data might be an output from Observational or experimental study.

Also it is an population distribution(Observations) which is of unimodel. Right skewness mentions that it has longer tails with outliers.

As the histogram bins are close, the variable might be an continuous numerical value. The mean is called as (μ)

Figure B:

This is the sampled distribution from the population distribution. The samples are simple random sampels which is of size 30. The chart seems to be a normal distribution which is of unimodel.

The chart shows that 30 samples taken from the population distribution and it is performed for 500 iterations. Mean is calculated for these 500 iterations and then this chart is formed. So it is the sampling distribution.

Also as the sample size is 30, we can apply central limit theorem(CLT). The mean from this sample will be the point estimate of the population distribution($\bar{x}$).

b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts).

Population Mean(μ): Population distribution has the mean of 5.05. This is the only mean which we get from the overall population. It is calcuated from the sum of all values divided by the total population size.

Population SD(σ): Population standard deviation is the deviation of all the values from the mean.

$\sigma = \sqrt{(\sum_{i=1}^{N}(X_i - \mu)^2/N)}$

Sample Distribution($\bar{x}$): Sampling distribution mean is 5.04. This approximatly equal the population mean. This is calculated by getting the mean of all the sample means. As per CLT, this mean will be approximatly equal to population mean.

Sample SD(s): Sample standard deviation is the deviation from the sample mean to all sample means(here 500 sample means). Generally the deviation of sampling distribution is called as Standard error.

Each random samples is an estimate of true population. So the sample SD will be always smaller than true standard deviation.

c. What is the statistical principal that describes this phenomenon (2 pts)?

The statistical principal that describes this phenomenon is called as Central Limit Theorem (CLT). It describes if the population distribution is normal distribution or skewed, if we have enough samples from the population distribution, then the mean of the sampling distribution will be be equal to the population mean.

CLT also states that if the sample size is atleast 30 independent random observations, then the sampling distribution will be a normal model given that the data is not strongly skewed.

Part II

Consider the four datasets, each with twocolumns (x and y), provided below.

```
options(digits=2)

data1 <-data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5), y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4

data2 <-data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5), y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26

data3 <-data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5), y=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6

data4 <-data.frame(x=c(8,8,8,8,8,8,8,19,8,8,8), y=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6
```

a. The mean (for x and y separately; 1 pt).

```
data_list <- list(data1,data2,data3,data4)

data <- rbind(data1,data2,data3,data4)

#For x in all individual dataframe
for(i in 1:4){

print(paste0("Mean of data",i,"$","x"," is " ,round(mean(data_list[[i]]$x),2)))
}
```

```
## [1] "Mean of data1$x is 9"
## [1] "Mean of data2$x is 9"
## [1] "Mean of data3$x is 9"
## [1] "Mean of data4$x is 9"
```

```
#For y in all individual dataframe
for(i in 1:4){

print(paste0("Mean of data",i,"$","y"," is " ,round(mean(data_list[[i]]$y),2)))
}
```

```
## [1] "Mean of data1$y is 7.5"
## [1] "Mean of data2$y is 7.5"
## [1] "Mean of data3$y is 7.5"
## [1] "Mean of data4$y is 7.5"
```

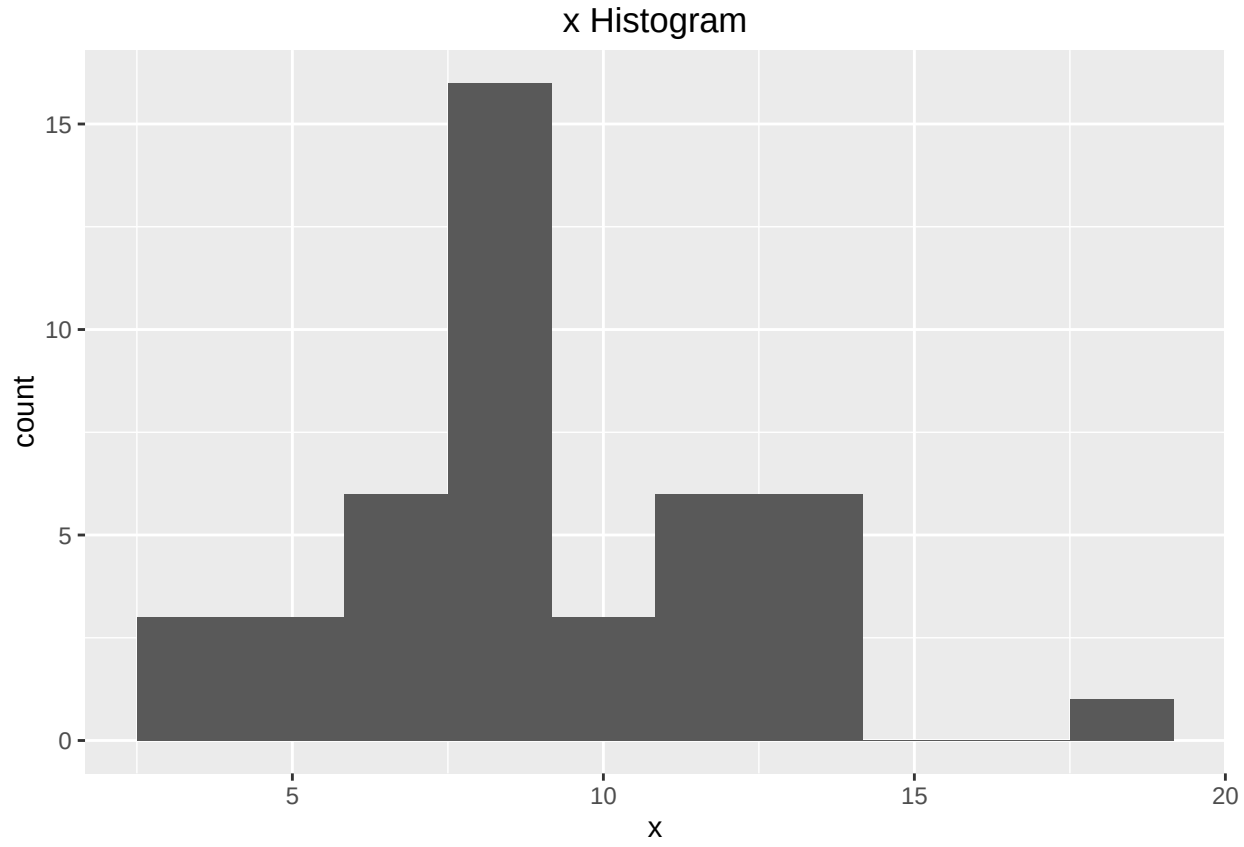
```
print(paste0("Mean of data"," x"," is " ,round(mean(data$x),2)))
```

```
## [1] "Mean of data x is 9"
```

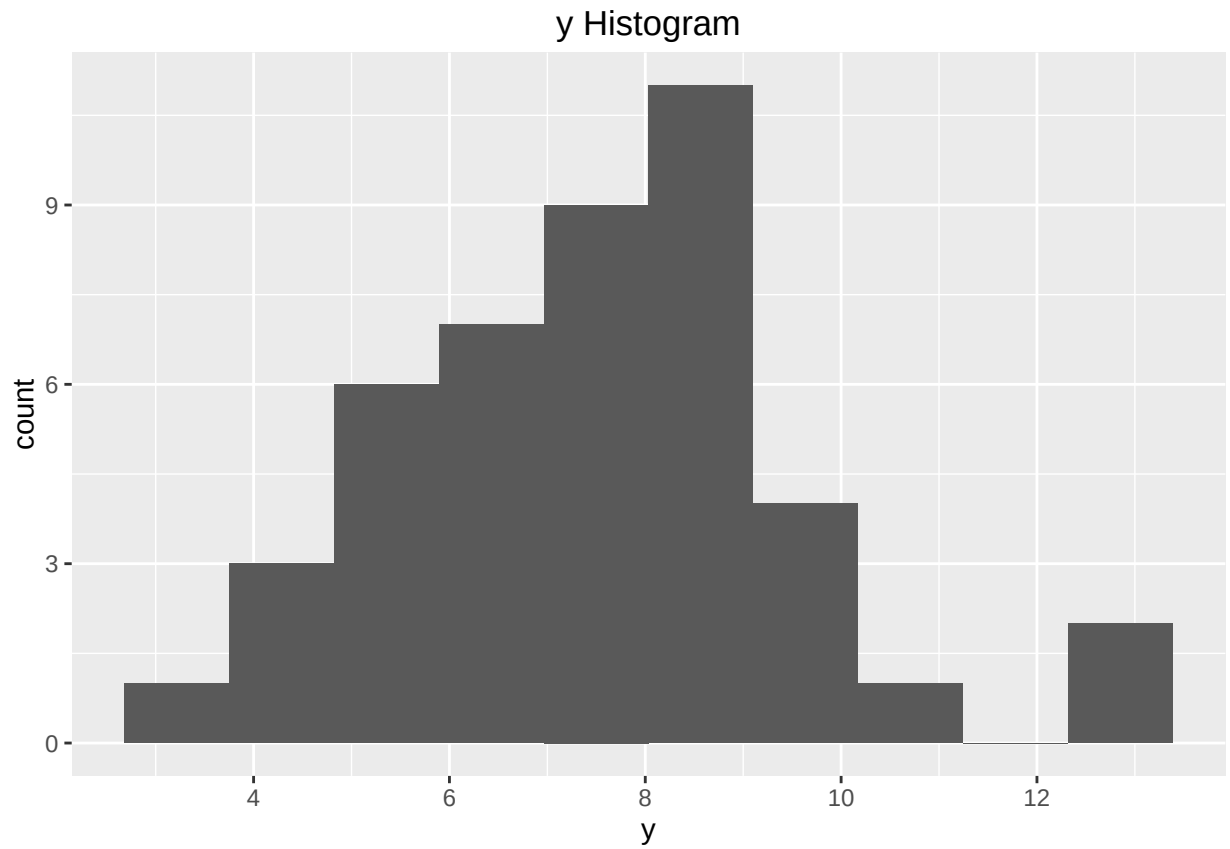
```
print(paste0("Mean of data"," y"," is " ,round(mean(data$y),2)))
```

```
## [1] "Mean of data y is 7.5"
```

```
ggplot(data,aes(x)) + geom_histogram(bins = 10) + ggtitle("x Histogram")
```



```
ggplot(data,aes(y)) + geom_histogram(bins = 10)  + ggtitle("y Histogram")
```



b. The median (for x and y separately; 1 pt).

```
#For x in all individual dataframe
for(i in 1:4){

print(paste0("Median of data",i,"$","x"," is " ,round(median(data_list[[i]]$x),2)))
}
```

```
## [1] "Median of data1$x is 9"
## [1] "Median of data2$x is 9"
## [1] "Median of data3$x is 9"
## [1] "Median of data4$x is 8"
```

```
#For y in all individual dataframe
for(i in 1:4){

print(paste0("Median of data",i,"$","y"," is " ,round(median(data_list[[i]]$y),2)))
}
```

```
## [1] "Median of data1$y is 7.58"
## [1] "Median of data2$y is 8.14"
## [1] "Median of data3$y is 7.11"
## [1] "Median of data4$y is 7.04"
```

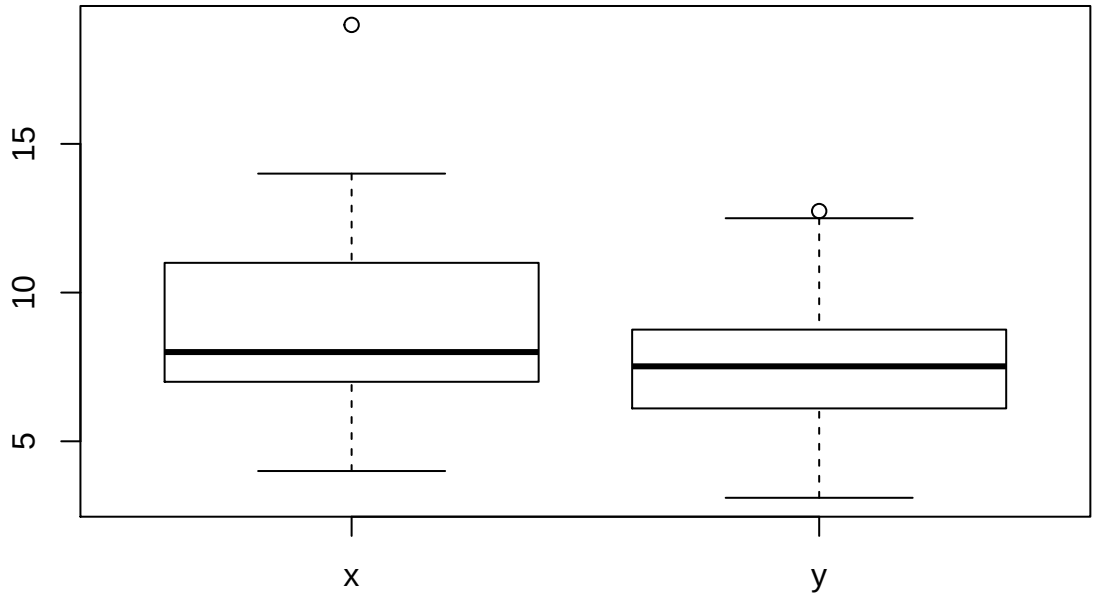
```r
#total median

paste0("Median of data"," is " ,median(data$x))
```

```
## [1] "Median of data is 8"
```

```r
paste0("Median of data"," is " ,median(data$y))
```

```
## [1] "Median of data is 7.52"
```

```r
boxplot(data)
```



c. The standard deviation (for x and y separately; 1 pt).

```r
#For x in all individual dataframe
for(i in 1:4){

print(paste0("SD of data",i,"$","x"," is " ,round(sd(data_list[[i]]$x),2)))
}
```

```
## [1] "SD of data1$x is 3.32"
## [1] "SD of data2$x is 3.32"
## [1] "SD of data3$x is 3.32"
## [1] "SD of data4$x is 3.32"
```

```r
#For y in all individual dataframe
for(i in 1:4){
```

```
print(paste0("SD of data",i,"$","y"," is " ,round(sd(data_list[[i]]$y),2)))
}
```

```
## [1] "SD of data1$y is 2.03"
## [1] "SD of data2$y is 2.03"
## [1] "SD of data3$y is 2.03"
## [1] "SD of data4$y is 2.03"
```

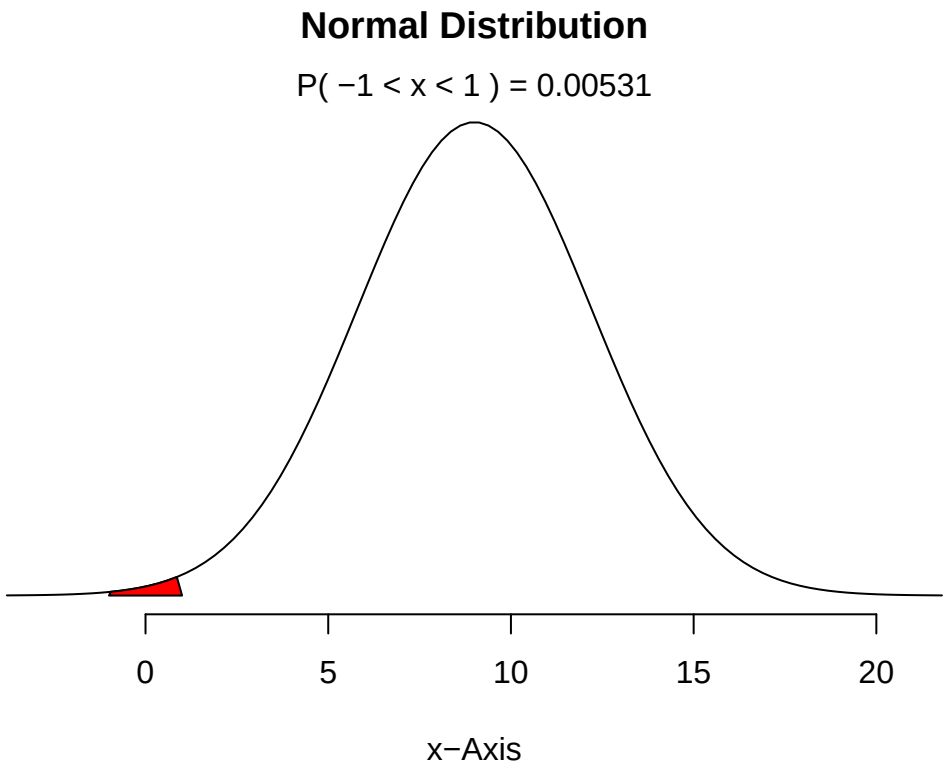
```
#total x

paste0("Standard Deviation of data"," is " ,round(sd(data$x),2))
```

```
## [1] "Standard Deviation of data is 3.2"
```

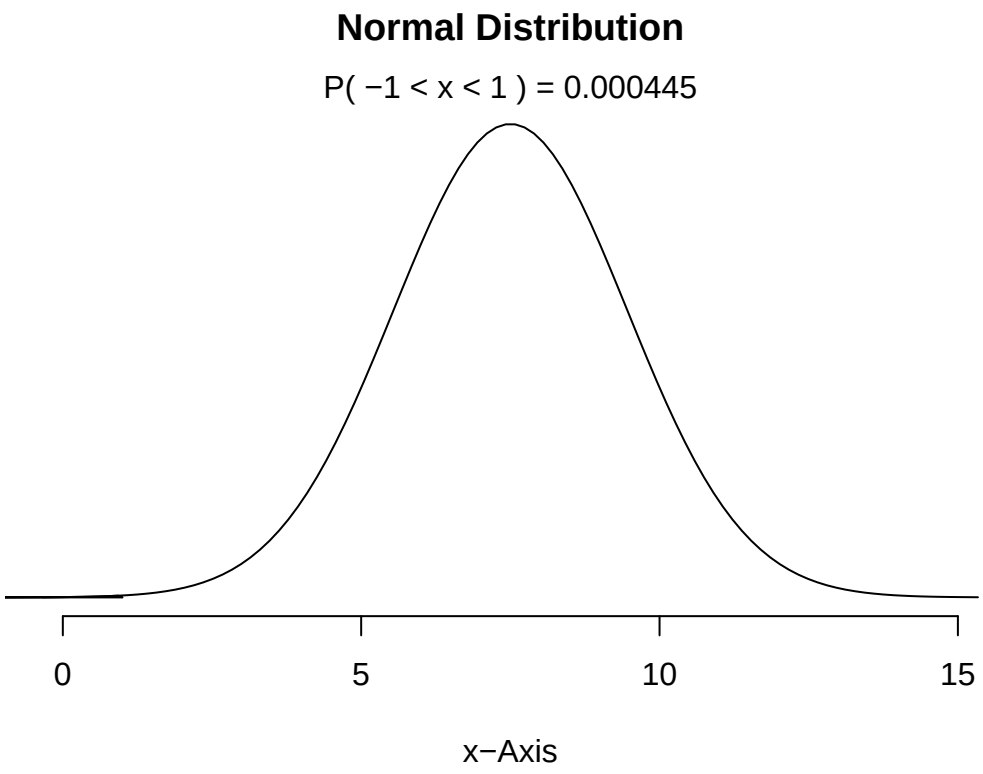
```
paste0("Standard Deviation of data"," is " ,round(sd(data$y),2))
```

```
## [1] "Standard Deviation of data is 1.96"
```

```
#Normal distribution for x
normalPlot(mean(data$x),sd(data$x))
```




```
#Normal distribution for y

normalPlot(mean(data$y),sd(data$y))
```



d. The correlation (1 pt).

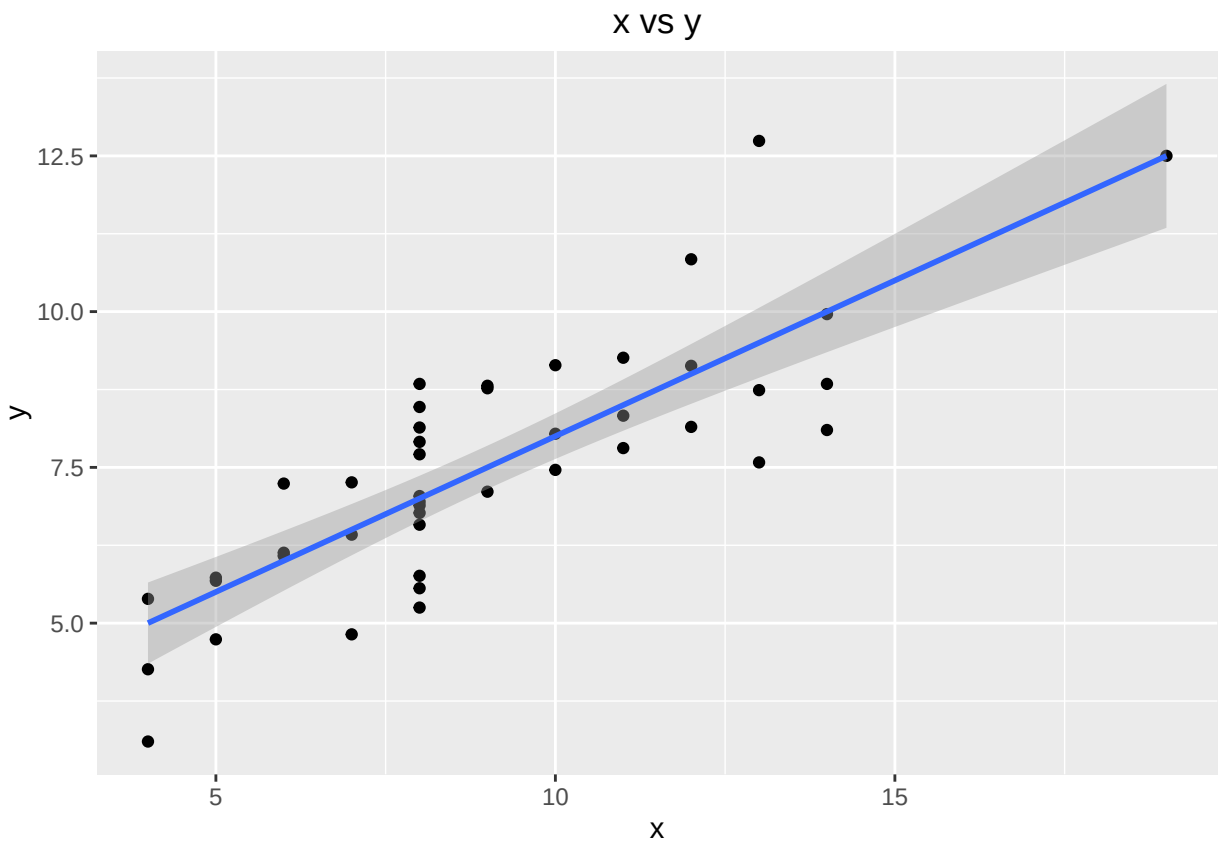
```
#For x an y in all individual dataframes
for(i in 1:4){

print(paste("Correlation within x and y in data",i, "is",cor(data_list[[i]]$x,data_list[[i]]$y)))
}
```

```
## [1] "Correlation within x and y in data 1 is 0.81642051634484"
## [1] "Correlation within x and y in data 2 is 0.816236506000243"
## [1] "Correlation within x and y in data 3 is 0.816286739489598"
## [1] "Correlation within x and y in data 4 is 0.816521436888503"
```

```
#Correlation between x and y
print(paste("Correlation within x and y in combined data is",cor(data$x,data$y)))
```

```
## [1] "Correlation within x and y in combined data is 0.81636624276147"
```

```
ggplot(data,aes(x,y)) + geom_point() +geom_smooth(method="lm")  + ggtitle("x vs y")
```



The correlation between x and y is and figure, it shows there is a strong correlation between x and y.

e. Linear regression equation (2 pts).

```
#For x an y in all individual dataframes
for(i in 1:4){

print(paste("Linear equation for x and y in data",i, "is"))

print(summary(lm(y ~ x,data = data_list[[i]])))

}
```

```
## [1] "Linear equation for x and y in data 1 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9213 -0.4558 -0.0414  0.7094  1.8388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.000      1.125    2.67   0.0257 *
## x              0.500      0.118    4.24   0.0022 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00217
##
## [1] "Linear equation for x and y in data 2 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.901 -0.761  0.129  0.949  1.269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125    2.67   0.0258 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00218
##
## [1] "Linear equation for x and y in data 3 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.159 -0.615 -0.230  0.154  3.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00218
##
## [1] "Linear equation for x and y in data 4 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.63
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00216
```

```
data_lm <- lm(y ~ x,data = data)

print("Linear equation for x and y from all data is")
```

```
## [1] "Linear equation for x and y from all data is"
```

```
print(summary(data_lm))
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.920 -0.746 -0.020  0.759  3.240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0013     0.5206    5.77  8.6e-07 ***
## x             0.4999     0.0546    9.16  1.4e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1 on 42 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.659
## F-statistic: 83.9 on 1 and 42 DF,  p-value: 1.44e-11
```

```
names(data_lm)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

General equation of linear equation is

$y = \beta_0 + \beta_1 * x$

Linear equation of the comined data frame is

y = 3.00130 + 0.49993 * x

Intercept: If x is zero, then the value of y is `3.00130`.

Slope: For each unit of increase in x will increase `0.49993` of y.

f. R-Squared (2 pts)

The strength of the fit of a linear model is most commonly evaluated using R-squared. R-squared can be calculated in different ways. For single linear regression, square of correlation coefficient is called as R-squared.

Below are the conditions for R-squared:

1. Linearity
2. Nearly normal residuals
3. Constant Variability
4. Linearity

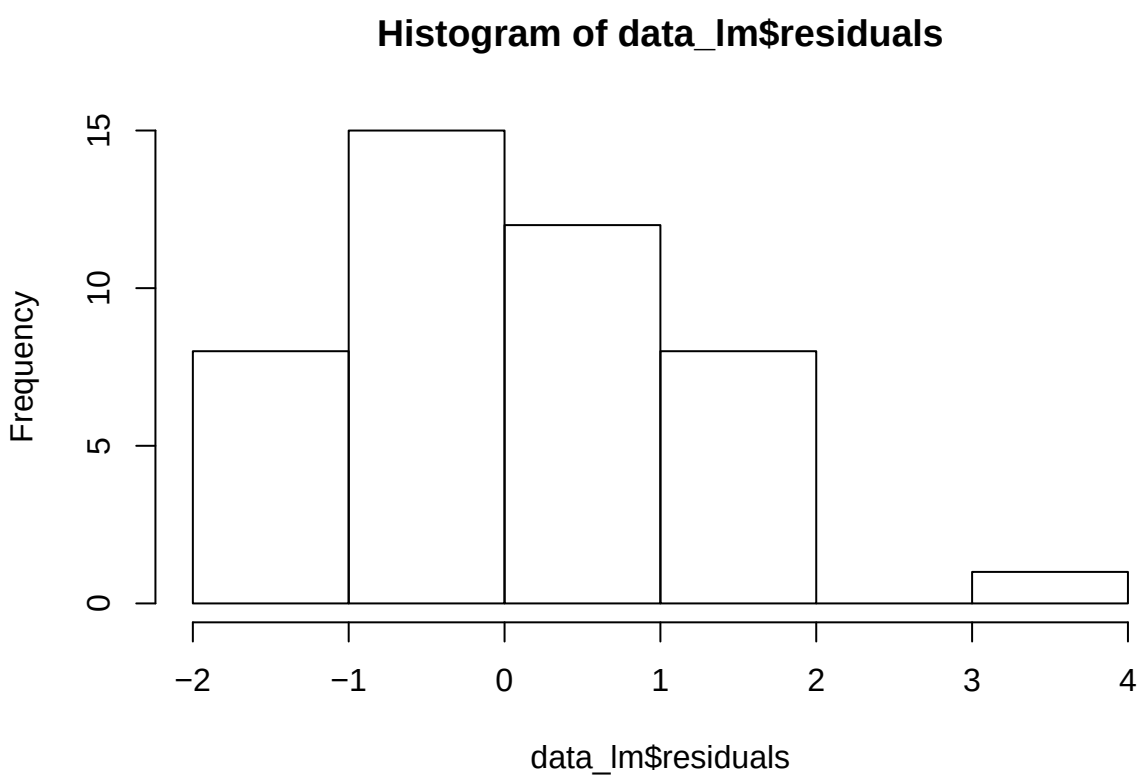
Relationship between the explanatory variable (x) and the response variable (y) should be linear.

From the plot in section d, shows that the relationship is linear.

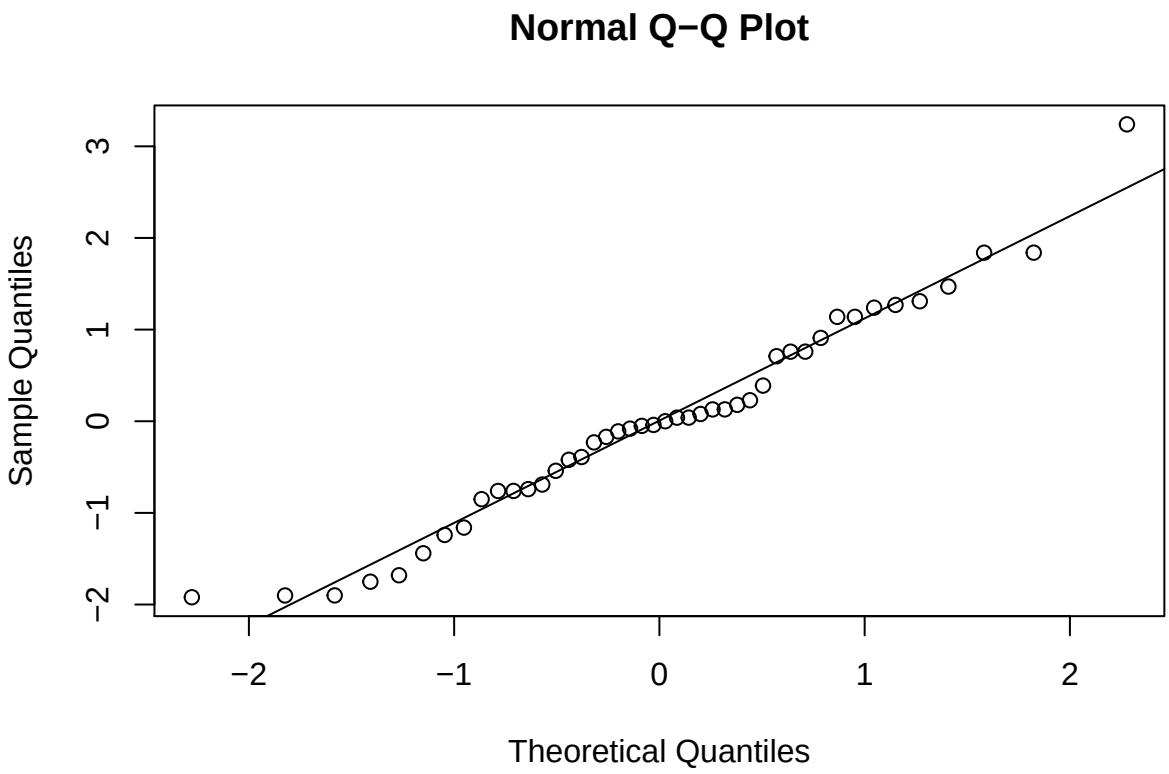
2. Nearly normal Residuals

Residuals should be normally distributed. Below are the plots which shows normal distribution

```
hist(data_lm$residuals)
```




```
qqnorm(data_lm$residuals)
qqline(data_lm$residuals)
```

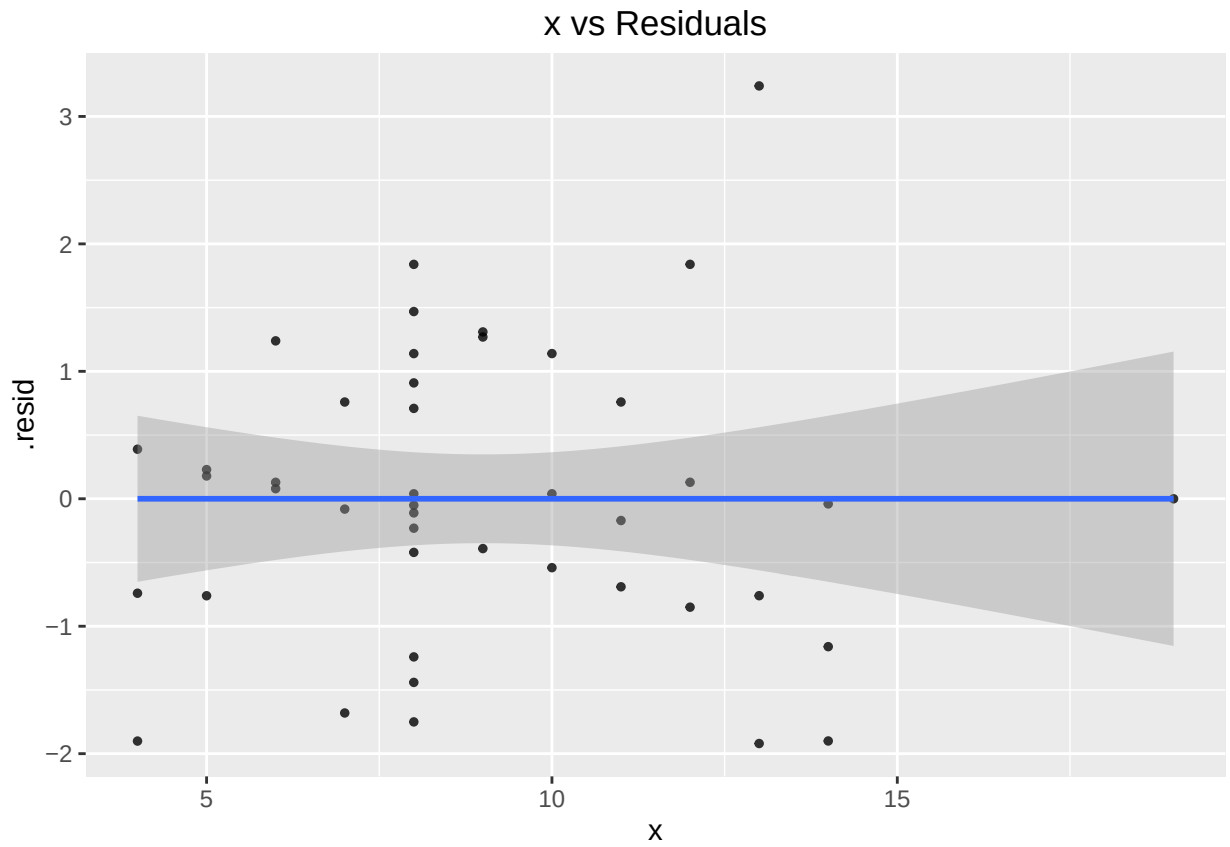


Above plot shows that the residuals are normally distributed. There are some outliners present in it.

3. Constant Variability

Variablity of these resiuals should be constant. Below plot shows the variability of residuals.

```
data_residuals <- augment(data_lm)

ggplot(data_lm,aes(x=x,y=.resid)) + geom_point(size=1,alpha=0.8) + geom_smooth(method = "lm")  + ggtitl
```



Above chart shows the constant variability of x and residuals. Although the curve shows that it is not constant, it is because of outliers at the end. So linear regression can be used. If a we require a perfect model, we can use logarithmic regression.

R-squared value of the combined dataframe is `0.6665`. It means the 66.65% of variability of y by x is explained by this model.

For each pair

Is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

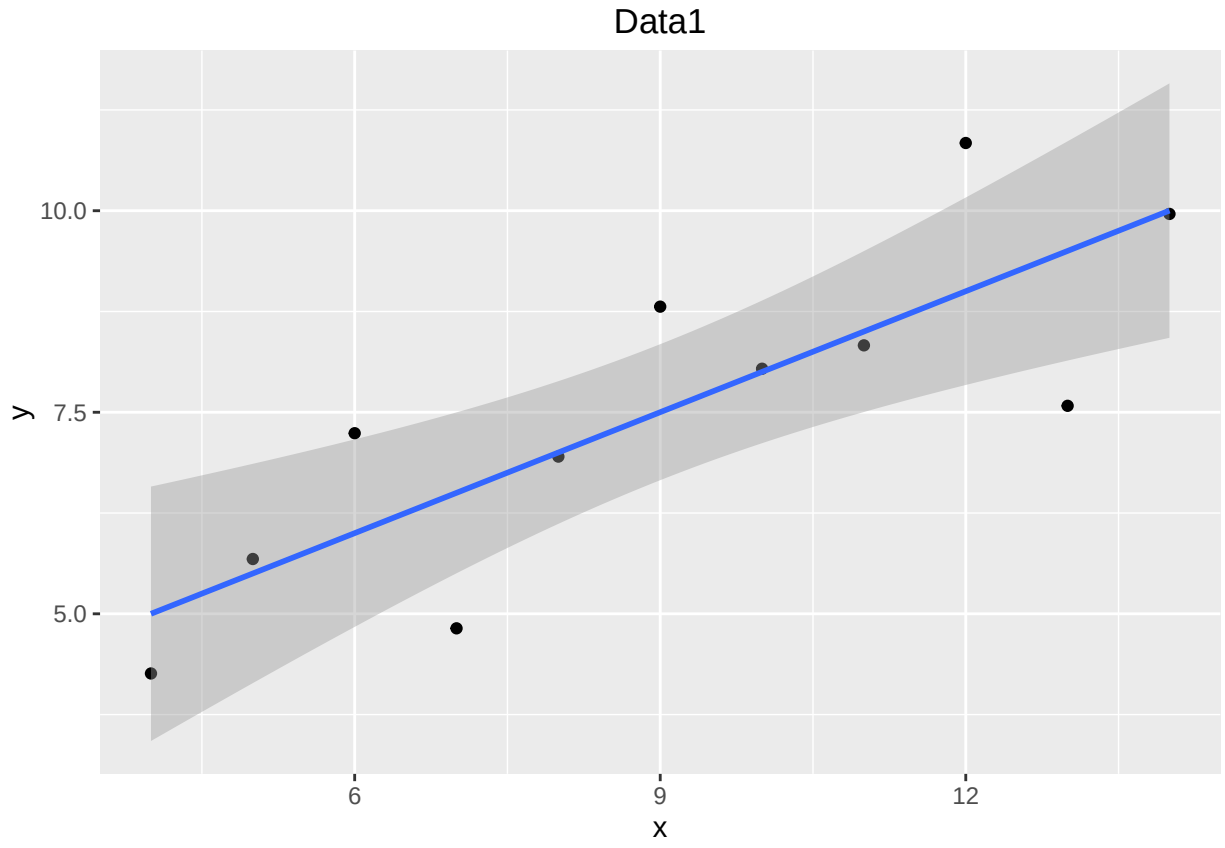
```
#For x an y in all individual dataframes
for(i in 1:4){

print(paste("Linear equation for x and y in data",i, "is"))

print(summary(lm(y ~ x,data = data_list[[i]])))

}
```

```
## [1] "Linear equation for x and y in data 1 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
```

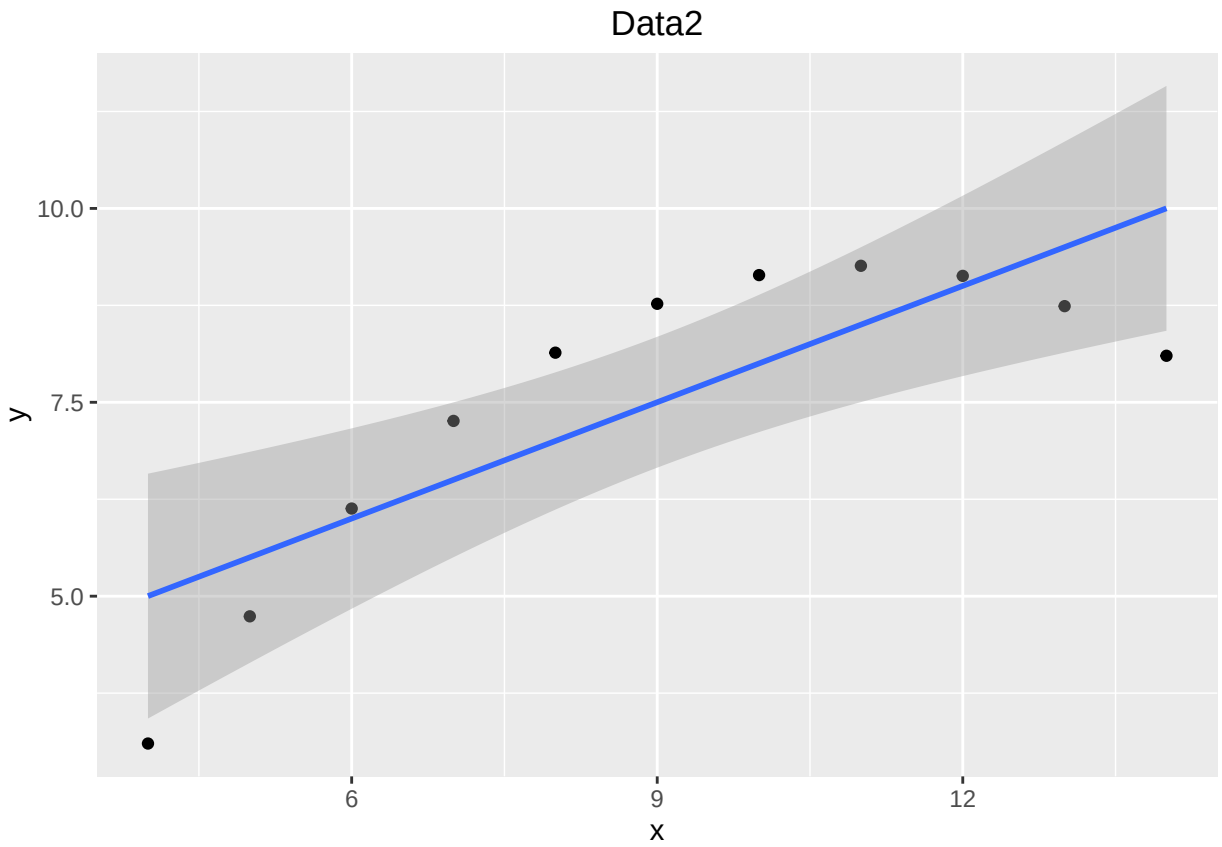
```
##     Min      1Q  Median      3Q     Max
## -1.9213 -0.4558 -0.0414  0.7094  1.8388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.000      1.125    2.67   0.0257 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00217
##
## [1] "Linear equation for x and y in data 2 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.901 -0.761  0.129  0.949  1.269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125    2.67   0.0258 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00218
##
## [1] "Linear equation for x and y in data 3 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.159 -0.615 -0.230  0.154  3.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00218
##
```

```
## [1] "Linear equation for x and y in data 4 is"
##
## Call:
## lm(formula = y ~ x, data = data_list[[i]])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.63
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00216
```

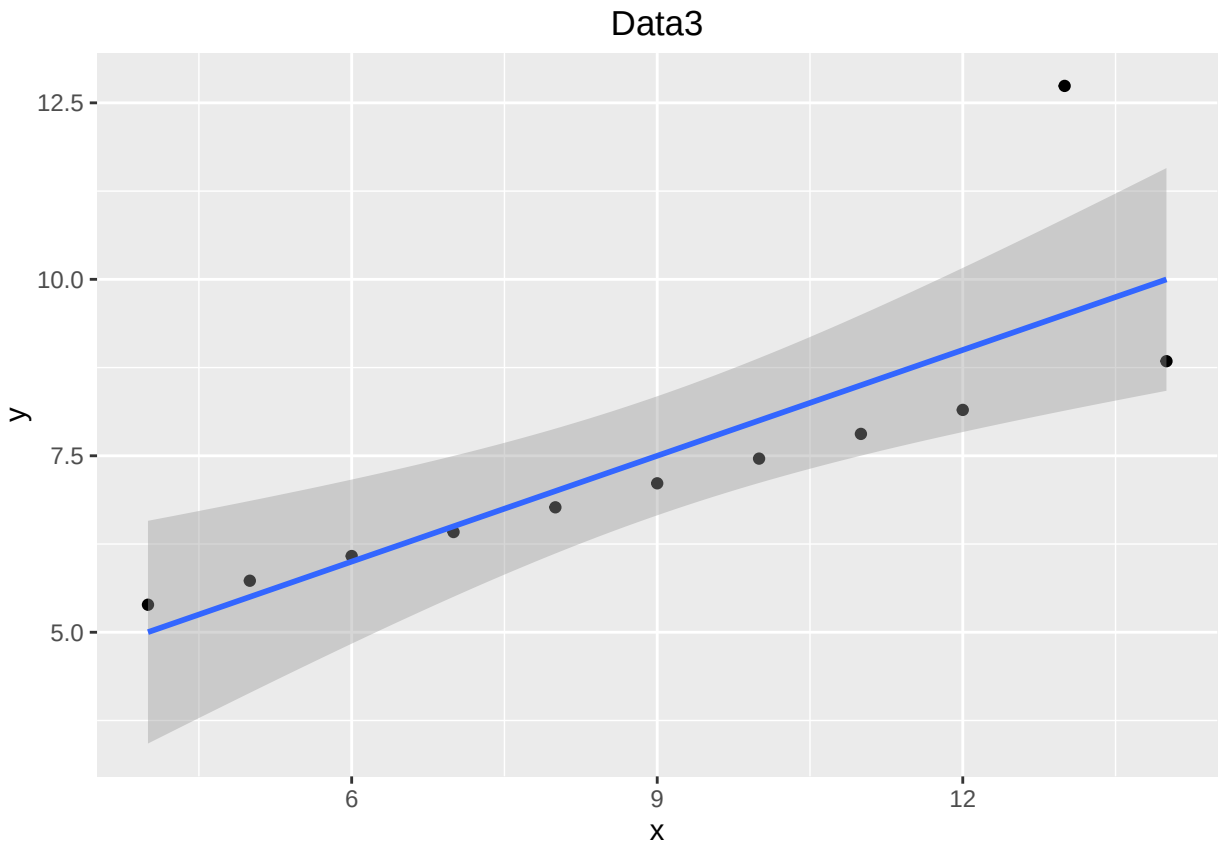
```
#For x an y in all individual dataframes

  ggplot(data_list[[1]],aes(x,y)) + geom_point() +geom_smooth(method="lm") + ggtitle("Data1")
```



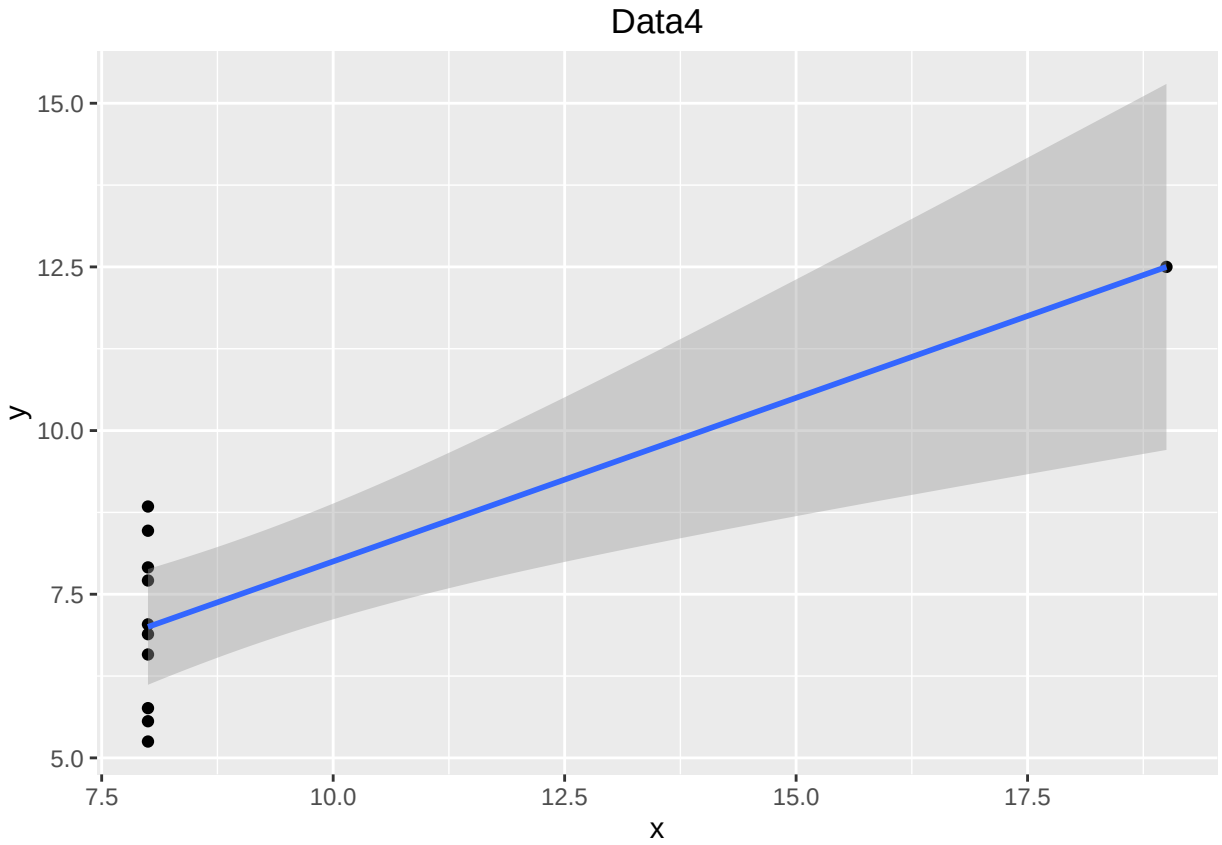

```
  ggplot(data_list[[2]],aes(x,y)) + geom_point() +geom_smooth(method="lm")  + ggtitle("Data2")
```




```
ggplot(data_list[[3]],aes(x,y)) + geom_point() +geom_smooth(method="lm")  + ggtitle("Data3")
```




```
ggplot(data_list[[4]],aes(x,y)) + geom_point() +geom_smooth(method="lm")  + ggtitle("Data4")
```



For each pair we can create a linear regression model. Above output and charts show for each pair.

Data1: The variability is constant. x and y has linear relationship. So linear relationship can be used.

Data2: The variability is not constant in the beginning. x vs y follows a non-linear patternt. It would be more approporiate to use other models(like logarithmic) than linear regression.

Data3: This has a constant linear relationship, but it has an outlier. The outlier is high leverage point and it is influential. Would be more approporiate to use other models(like logarithmic) than linear regression.

Data4: This does not have any constant linear relationship. There is one influential point which decides the complete line. It would be more approporiate to use other models(like logarithmic) than linear regression.

Analyze Visuals

Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts)

Visualizations are very important in statistical studies. It visually shows the relationship between two variables (In this example x and y). Below are the points where it is particulary useful

1. Outliers - Without visuals, it is difficult to figure out the outliers in the distribution.
2. Variablity - It is offen required to validate this condition. Depending on this condition the model will be decided. If it has constant variability we can use linear regression model else we have to use non-linear regression models.
3. Independence - For almost all stat problems it is important to find out independence between two values in a distribution. If it is independent, it will follow a trend. So we need to model appropriately.

4. Fitting Regression Line - Once we plot the points it is required to plot a regression line and check. It is very hard to plot a regression line without an visualizations.

I have added visualizations when required in all the problems.