Anscombe_rmd_code

Name: Rachakonda Hrithik Sagar

Roll No: 202390021

```
data <- read.csv("Anscombe_dataset.csv")
str(data)
```

```
## 'data.frame':    200 obs. of  3 variables:
##  $ Group: chr  "I" "I" "I" "I" ...
##  $ x    : num  8.17 14.85 12.32 10.77 5.64 ...
##  $ y    : num  8.02 10.8 9.09 8.04 3.83 ...
```

#descriptive stats

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

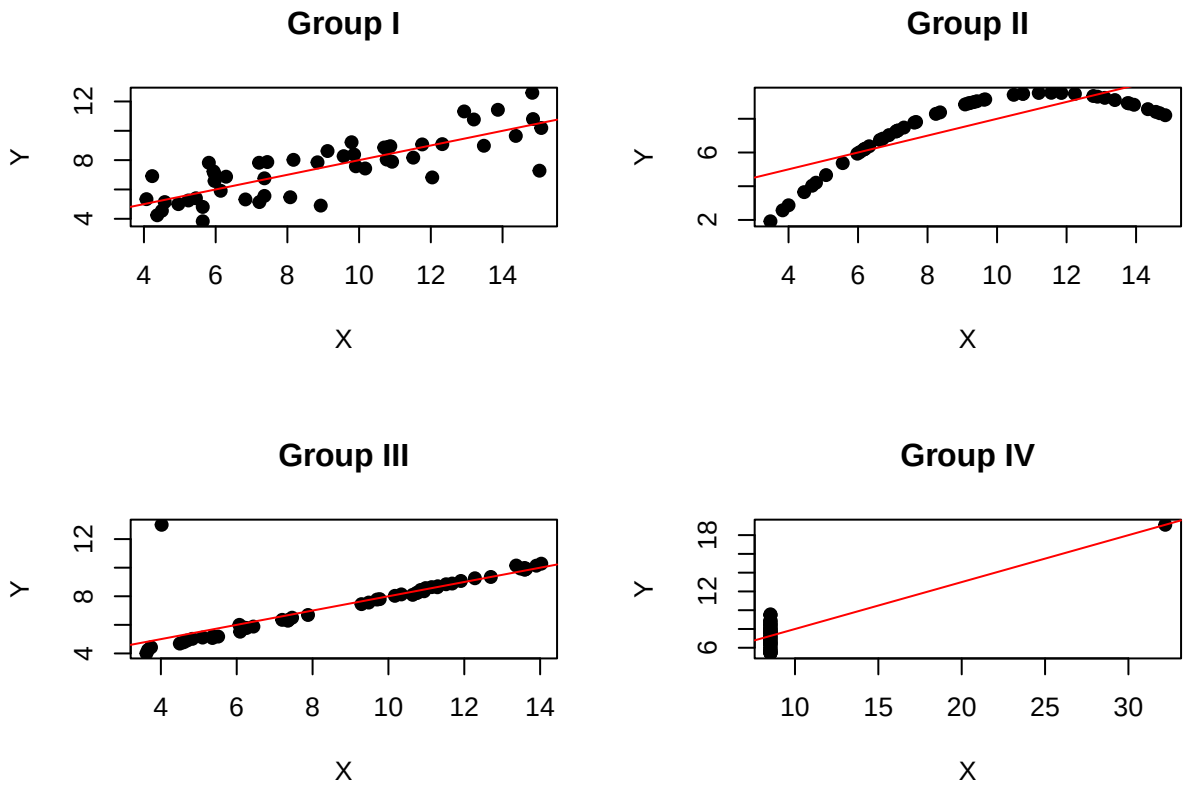
```
data %>%
  group_by(Group) %>%
  summarise(
    mean_x = mean(x),
    mean_y = mean(y),
    sd_x = sd(x),
    sd_y = sd(y),
    cor_xy = cor(x, y)
  )
```

```
## # A tibble: 4 x 6
##   Group mean_x mean_y  sd_x  sd_y cor_xy
##   <chr>  <dbl>  <dbl> <dbl> <dbl>  <dbl>
## 1 I       9.00   7.50  3.35  2.05  0.816
## 2 II      9.00   7.50  3.35  2.05  0.816
## 3 III     9.00   7.5   3.35  2.05  0.816
## 4 IV      9.00   7.50  3.35  2.05  0.816
```

#plots

```
par(mfrow = c(2,2))
groups <- unique(data$Group)

for(g in groups){
  subset_data <- subset(data, Group == g)
  plot(subset_data$x, subset_data$y,
       main = paste("Group", g),
       xlab = "X", ylab = "Y",
       pch = 19)
  abline(lm(y ~ x, subset_data), col = "red")
}
```



Observations

- Overall, the mean, standard deviation, and correlation are approximately the same across all four g
- Group 1: A clear linear relationship is observed between the variables X and Y.
- Group 2: A relationship between X and Y exists, but it is non-linear, which is not captured well by
- Group 3: A generally linear relationship is present, but it is strongly influenced by a single outl
- Group 4: There is no meaningful relationship between X and Y; the mean, standard deviation, and cor