RLab Assignment 1

Arun Kumar P

March 25, 2018

Statistical Analysis on Data sets

Reading the input data from CSV file and using ‘summary’ command to get the statistical summary of the data

```
dt1<-read.csv("r3.csv")
summary(dt1)
```

```
##        X1              Y1               X2              Y2        
##  Min.   : 4.0   Min.   : 4.260   Min.   : 4.0   Min.   :3.100  
##  1st Qu.: 6.5   1st Qu.: 6.315   1st Qu.: 6.5   1st Qu.:6.695  
##  Median : 9.0   Median : 7.580   Median : 9.0   Median :8.140  
##  Mean   : 9.0   Mean   : 7.501   Mean   : 9.0   Mean   :7.501  
##  3rd Qu.:11.5   3rd Qu.: 8.570   3rd Qu.:11.5   3rd Qu.:8.950  
##  Max.   :14.0   Max.   :10.840   Max.   :14.0   Max.   :9.260  
##        X3              Y3               X4           Y4        
##  Min.   : 4.0   Min.   : 5.39   Min.   : 8   Min.   : 5.250  
##  1st Qu.: 6.5   1st Qu.: 6.25   1st Qu.: 8   1st Qu.: 6.170  
##  Median : 9.0   Median : 7.11   Median : 8   Median : 7.040  
##  Mean   : 9.0   Mean   : 7.50   Mean   : 9   Mean   : 7.501  
##  3rd Qu.:11.5   3rd Qu.: 7.98   3rd Qu.: 8   3rd Qu.: 8.190  
##  Max.   :14.0   Max.   :12.74   Max.   :19   Max.   :12.500
```

Summary Statistics Insights:

1.**Correlation Coefficient**:
a.All data sets have identical correlation coefficient of 0.816.
b.This indicates the values are almost positively correlated with each other respectively.
2.**Mean**:The average or mean of all the data sets (Y) are identical (7.501).
3.**Outliers**:
a.Data set 1 does not have any supected outliers or outliers.
b.Data set 2 does not have any supected outliers or outliers.
c.Data set 3 contains one outlier value (12.74).
d.Data set 4 contains one outlier value (12.50).
4.**Skewness**:
a.Data set 1 is normally distributed.
b.Data set 2 is slightly left skewed (mean < median).
c.Data set 3 does not have any skewness (mean is almost equal to median).
d.Data set 3 does not have any skewness (mean is almost equal to median.

Plots

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

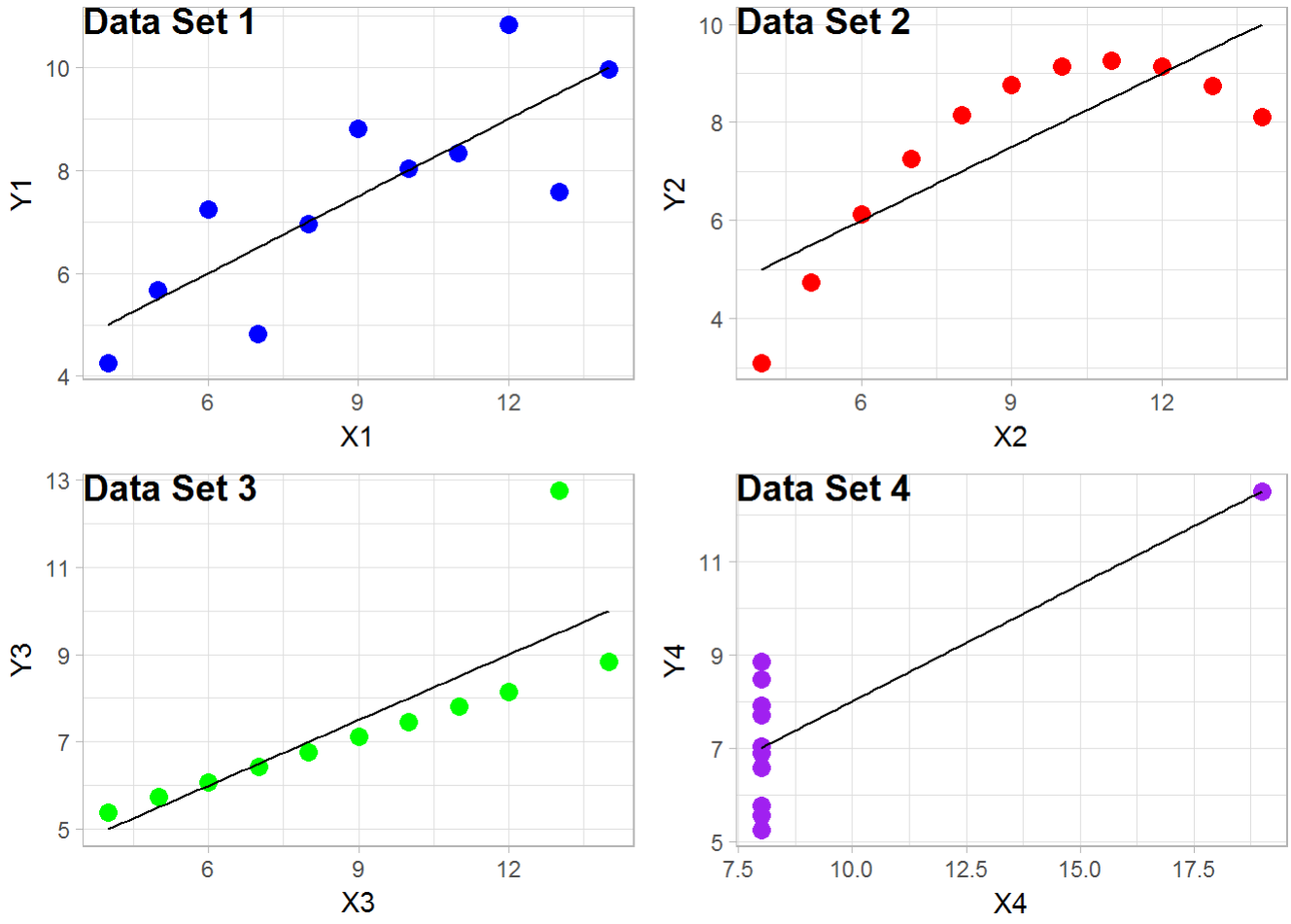
```
## Warning: package 'cowplot' was built under R version 3.4.4
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

Using the scatter plot for analyzing the data

```
plot1_1<-ggplot(dt1,aes(x=X1,y=Y1))+geom_point(shape=19,color="blue",size=3)
plot1_2<-plot1_1+geom_smooth(method="lm",se=FALSE,color="black",size=0.5)+theme_light()
plot2_1<-ggplot(dt1,aes(x=X2,y=Y2))+geom_point(shape=19,color="red",size=3)
plot2_2<-plot2_1+geom_smooth(method="lm",se=FALSE,color="black",size=0.5)+theme_light()
plot3_1<-ggplot(dt1,aes(x=X3,y=Y3))+geom_point(shape=19,color="green",size=3)
plot3_2<-plot3_1+geom_smooth(method="lm",se=FALSE,color="black",size=0.5)+theme_light()
plot4_1<-ggplot(dt1,aes(x=X4,y=Y4))+geom_point(shape=19,color="purple",size=3)
plot4_2<-plot4_1+geom_smooth(method="lm",se=FALSE,color="black",size=0.5)+theme_light()
plotnames<-c("Data Set 1","Data Set 2","Data Set 3","Data Set 4")
plot_grid(plot1_2,plot2_2,plot3_2,plot4_2,labels=plotnames,ncol=2,nrow=2)
```



Scatter Plot Insights:

1.**DataSet 1**:
a.The variables X1 and Y1 are correlated with each other.
b.The relationship is linear.
2.**DataSet 2**:
a.The variables X2 and Y2 are correlated with each other.
b.The relationship is non-linear.
3.**DataSet 3**:
a.The variables X3 and Y3 are correlated with each other.
b.The relationship is linear.
4.**DataSet 4**:
a.The variables X4 and Y4 are not correlated(not related) with each other as X4 as is almost constant.
b.The relationship is non-linear.

Summary:

1.The correlation coefficient value of DataSet 3 is decreased due to the presence of one outlier(13.00,12.74).
2.The correlation coefficient value of DataSet 4 is increased due to the presence of one outlier(19.00,12.50).
3.The statistical summary represents identical values across data sets though there are differences which can be infered through scatter plots.