Data visualisation

Exercise 1

First create the list, then plot the pie chart.

## creating the list

shopping <- c(12,2,1,11,20,2,2,1)

names(shopping) <- c("apples", "buns","loaf", "carrots","cereal bars","peppers", "oranges", "lemon")

# displaying the pie chart using the terrain colour palette and a suitable title.

pie(shopping, col=terrain.colors(length(shopping)), main = "Shopping list - items needed")

This pie chart is inadequate for our purposes. Its is difficult to distinguish proportions in the chart (e.g. buns, oranges and peppers - unclear if same amount needed because of the use of different colours). Some of the labels overlap each other (e.g. loaf and buns)).

Exercise 2

library("ggplot2")

## Warning: package 'ggplot2' was built under R version 3.5.1

ProvenOilReservesZerosT <-read.csv("ProvenOilReservesZerosT.csv", header=T)

## set x and y axis

p <- ggplot(ProvenOilReservesZerosT, aes(MTBarrels, Denmark))

## point geom - a square (22), with a dark blue line and a blue fill. The size of the square depends on the amount of proven oil reserves for Denkmark.

p <- p + geom\_point(shape=22, colour="darkblue", fill= "blue", aes(size=factor(Denmark)))

## suitable labels, title and legend.

p <- p + labs( x="year", y= "proven oil reserves", title = "Proven Oil Reserves for Denmark (million ton barrels)", size="Million Ton Barrels")

p

## Warning: Using size for a discrete variable is not advised.

The use of squares as the point is not very useful and clutters the point. The plot does show an increase in proven oil reserves till around 2005. After that time, it shows a decline. If data was only available till 2005, the decline would not be visible from the plot.The use of point size to emphasize the amount of reserves is not successful as it clutters the plot.

Exercise 3

## x axis is n.

channels <- read.csv("channels.csv", header=T)

p <- ggplot(channels, aes(n))

## one line for electric. Its colour label is specified as "electric current"

p <- p+geom\_line(aes(y = electric, colour = "electric current"))

## one line for length. Its colour label is specified as "object's length"

p <- p+geom\_line(aes(y = length, colour = "objects length"))

## one line for area. Its colour label is specified as "object's area"

p <- p+geom\_line(aes(y = area, colour = "objects area"))

## x and y labels plus legend for colours

p<-p+labs(x= "size of stimuli", y= "size of sensation", colour="Channels")

p

It is easy to see that electric currrent produces a much bigger sensation. However, it is very difficult to appreciate differences between length and area. Changing the y axis limits will help with this.

Limitting the y axis to 50 to appreciate differences between length and area

p <- p+ylim(0,50)

p

## Warning: Removed 4 rows containing missing values (geom\_path).

The plot clearly shows the differences between the 3 channels.

Exercise 4

ProvenOilReserveWEurope <- read.csv("ProvenOilReserveWEurope.csv", header=T)

## Present "year" against amount (MT barrels), grouping the data according to the country.

p <- ggplot(ProvenOilReserveWEurope, aes(Year, MT.Barrels, group=Country))

## Use coloured line - a colour per country

p <- p + geom\_line(aes(colour=factor(Country)))

## add appropriate labels, title and legend

p <- p+labs(x="year", y = "proven oil reserves", title="Proven oil reserves in West Europe (Million ton barrels)", colour="Country")

p

This is a good visualisation which clearly shows proven oil reserves, the dominance of Norway and the UK. It is clear that countries follow different trends. For example, the UK and Norway clearly showing a different trend during the first few years, with Norway’s proven reserves increasing and the UK’s decreasing.

Exercise 5

P <- ggplot(ProvenOilReserveWEurope, aes(Year))

## colour is coutry-dependent. The Y axis is MT.Barrels. the colour of the tiles is Country-dependent.

p <- p+geom\_tile(aes(y = MT.Barrels, colour=Country, fill=Country))

## add appropriate labels, title and legend

p <- p+labs(x="year", y = "proven oil reserves", title="Proven oil reserves in West Europe (Million ton barrels)", colour="Country")

## Facet specification - present them vertically.

p <- p + facet\_grid(Country ~ .)

p

All subplots have the same y axis scale While this makes it easy to compare large amounts of proven oil reservers vs smaller ones (e.g. Norway vs Italy), it makes it very difficult to compare similar-size data (e.g. Denmark vs Italy).The legend is not really needed. Using a free y axis would help to overcome this problem, but it will make it more difficult to compare large vs. small amounts.

Exercise 6

OilQuality <- read.csv("OilQuality.csv")

Correlation and covariance

cor.API.Price <- cor(OilQuality$API,OilQuality$Price)

cov.API.Price <- cov(OilQuality$API,OilQuality$Price)

cor.API.Price

## [1] 0.9799122

Exercise 7

Producing a scatterplot

p <- ggplot(OilQuality,aes(x = API,y = Price))

p <- p + geom\_point()

p <- p+labs(x="API gravity", y = "Price ($/barrel)", title="Oil Price vs. API Gravity")

p <- p+geom\_smooth()

p

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

## Exercise 8

lm.output <- lm(Price ~ API,data= OilQuality)

summary(lm.output)

##

## Call:

## lm(formula = Price ~ API, data = OilQuality)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.129624 -0.081163 -0.006995 0.076414 0.141965

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 8.772568 0.323412 27.12 1.66e-07 \*\*\*

## API 0.117737 0.009782 12.04 2.00e-05 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.1062 on 6 degrees of freedom

## Multiple R-squared: 0.9602, Adjusted R-squared: 0.9536

## F-statistic: 144.9 on 1 and 6 DF, p-value: 1.996e-05

The answers can be read from the above summary. Alternatively, we can obtain the coefficients by:

lm.output

##

## Call:

## lm(formula = Price ~ API, data = OilQuality)

##

## Coefficients:

## (Intercept) API

## 8.7726 0.1177

Exercise 9

The equation is: Price = 8.7726 + 0.1177\*API

The coefficient of determination (“Multiple R-squared”) is also the square of the correlation coefficient:

cor.API.Price\*cor.API.Price

## [1] 0.9602279

Exercise 10

Predictions

predict(lm.output,newdata = data.frame(API = c(35.5,45.1)))

## 1 2

## 12.95222 14.08250

As in the tutorial, you should note that the second prediction is a dubious extrapolation.

Exercise 11

Manipulating data

UKData <- read.csv("UKData.csv", header=T)

# make a copy - this is not strictly necessary

UKNewData <- UKData

# obtain month

UKNewData$month <- substring(UKNewData$Month,0,2)

# aggregate Oil production

aggdata <-aggregate(UKNewData$OPUKShare,

by=list(UKNewData$month),

FUN=sum, na.rm=TRUE)

# aggregate associated gas production

aggdata2 <-aggregate(UKNewData$AGPUKS,

by=list(UKNewData$month),

FUN=sum, na.rm=TRUE)

# Create new dataframe with the right columns from the 2 aggregations above

UKNewData <-data.frame(aggdata[1], aggdata[2], aggdata2[2])

#rename new columns

colnames(UKNewData) <- c( "month", "OPUKShare", "AGPSUKS")

UKNewData now contains the data needed for the t-test.

Exercise 12

MData <- subset(UKData, Operator =="MAERSK")

TData <- subset(UKData, Operator != "TOTAL")

Exercise 13

# First calculate the mean

meanOPUKShare <- mean(UKData$OPUKShare)

# Then add the new attribute (column) with appropriate values

UKData$bigProducerOil[UKData$OPUKShare <= meanOPUKShare] = FALSE

UKData$bigProducerOil[UKData$OPUKShare > meanOPUKShare] = TRUE

# Produce a summary to see how many have value true for big|ProducerOil

summary(UKData$bigProducerOil)

## Mode FALSE TRUE

## logical 31 17

There are 17 instances where the operator produced more than the average.

Exercise 14

Undertaking a t-test with true mean = 8000

t.test(UKData$AGPUKS, conf.level=0.95, mu=8000, alternative="greater")

##

## One Sample t-test

##

## data: UKData$AGPUKS

## t = 1.8215, df = 47, p-value = 0.03745

## alternative hypothesis: true mean is greater than 8000

## 95 percent confidence interval:

## 8379.614 Inf

## sample estimates:

## mean of x

## 12816.75

The p value is less than 0.05 so the null hypothesis (true mean is 8000) is rejected and the alternative hypothesis (true mean greater than 8000) is supported.

Undertaking a t-test with true mean = 10000

t.test(UKData$AGPUKS, conf.level=0.95, mu=10000, alternative="greater")

##

## One Sample t-test

##

## data: UKData$AGPUKS

## t = 1.0652, df = 47, p-value = 0.1461

## alternative hypothesis: true mean is greater than 10000

## 95 percent confidence interval:

## 8379.614 Inf

## sample estimates:

## mean of x

## 12816.75

The p value is greater than 0.05 and, therefore the null hypothesis (true mean is 10000 cannot be rejected).

Undertaking a t-test with true mean = 12000

t.test(UKData$AGPUKS, conf.level=0.95, mu=12000, alternative="greater")

##

## One Sample t-test

##

## data: UKData$AGPUKS

## t = 0.30886, df = 47, p-value = 0.3794

## alternative hypothesis: true mean is greater than 12000

## 95 percent confidence interval:

## 8379.614 Inf

## sample estimates:

## mean of x

## 12816.75

The p value is a lot larger than 0.05 and, therefore, the null hypothesis (true mean is 12000 cannot be rejected).

Exercise 15

Oil production analysis based on operator

anova <- aov(OPUKShare ~ Operator, data = UKData)

summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)

## Operator 3 4.473e+10 1.491e+10 297.1 <2e-16 \*\*\*

## Residuals 44 2.209e+09 5.019e+07

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pr(>F) is very, very small and, therefore, the null hypothesis (means of OPUKShare are the same regardless of the operator) is rejected. There is a difference in the means depending on the operator.