Sample practical examination - suggested solutions

Cleaning up the environment and loading required library.

rm(list=ls())

**library**(tidyverse)

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

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

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

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

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

Data visualisation

Exercise 1

*## set x and y axis*

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

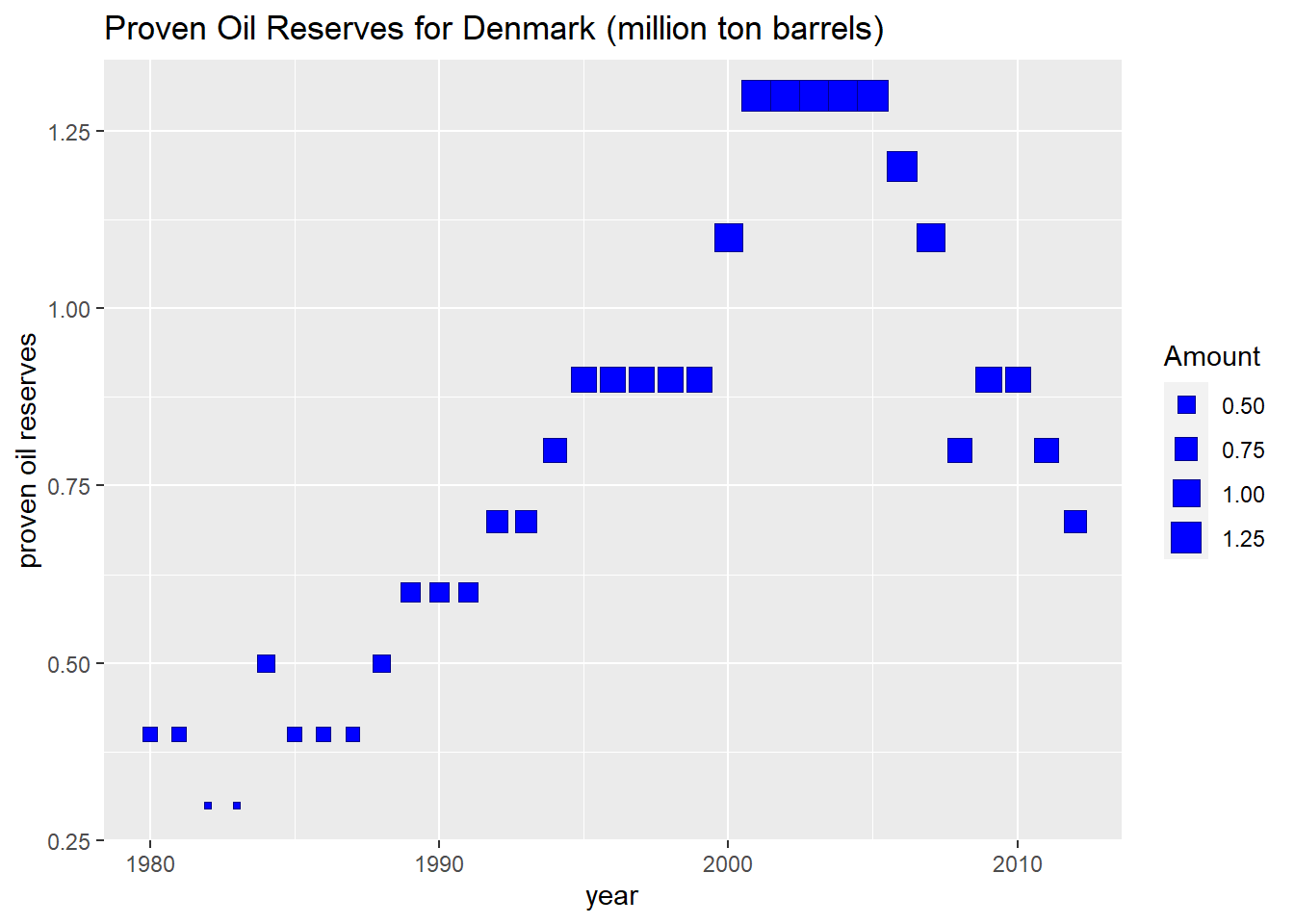
*## point geom - a square (22), with a dark blue line and a blue fill. The size of the square depends on*

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

p



The use of squares as the point is not very useful and clutters the graph. 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 required as the amount is easy to understand from the height of the point.

Exercise 2

*## x axis is n.*

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 length"))

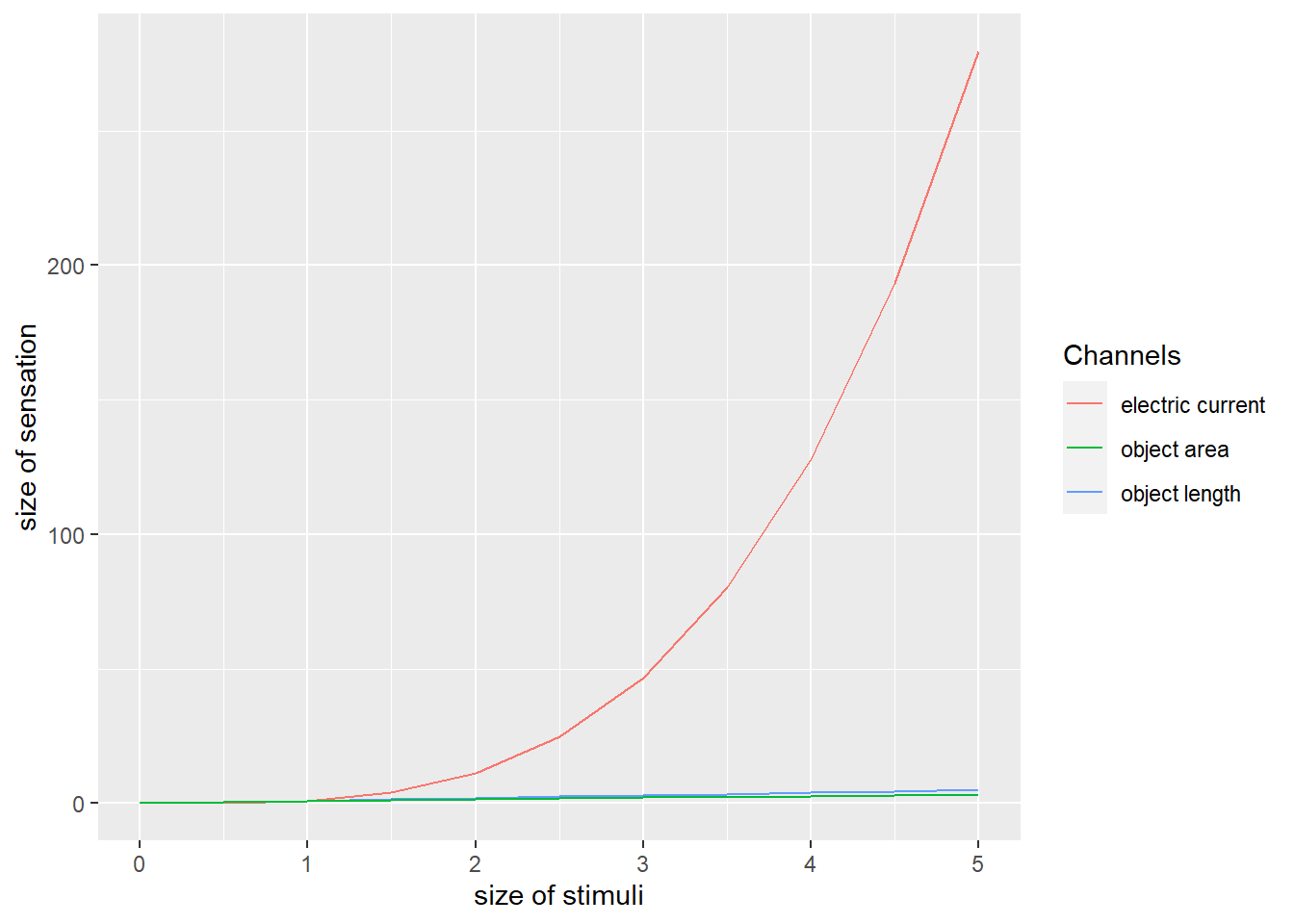
*## one line for area. Its colour label is specified as "object's area"*

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

*## x and y labels plus legend for colours*

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

p



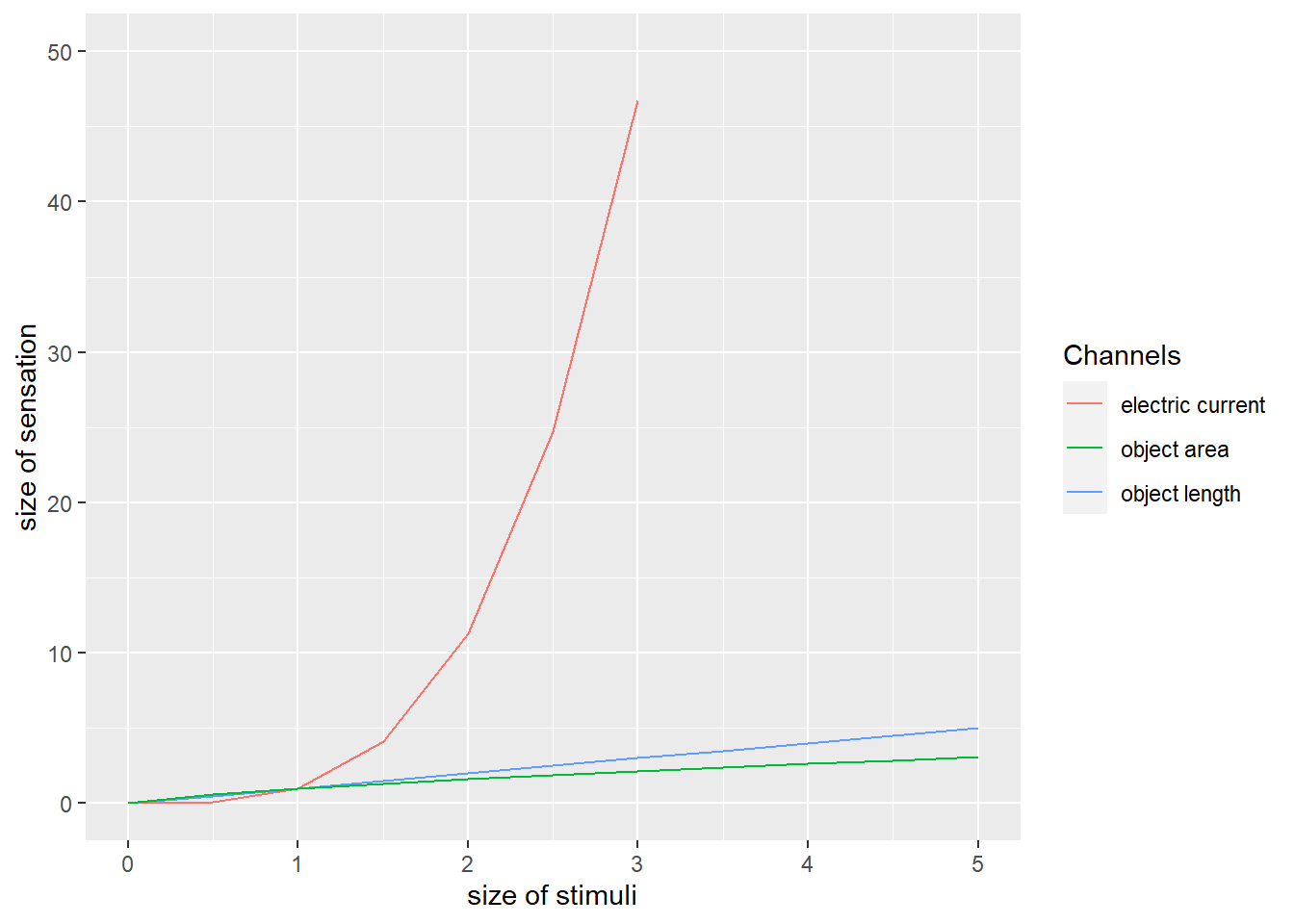
It is easy to see that electric currrent produces a much bigger sensation. However, it is very difficult to appreciate differences between length and area. Changing the y axis limits will help with this.

Limiting the y axis to 50 to appreciate differences between length and area. Note that this will result in a warning as some data will not be shown. The alternative will be to first remove the data and then display.

p <- p+ylim(0,50)

p

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



The plot clearly shows the differences between the 3 channels.

Exercise 3

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

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

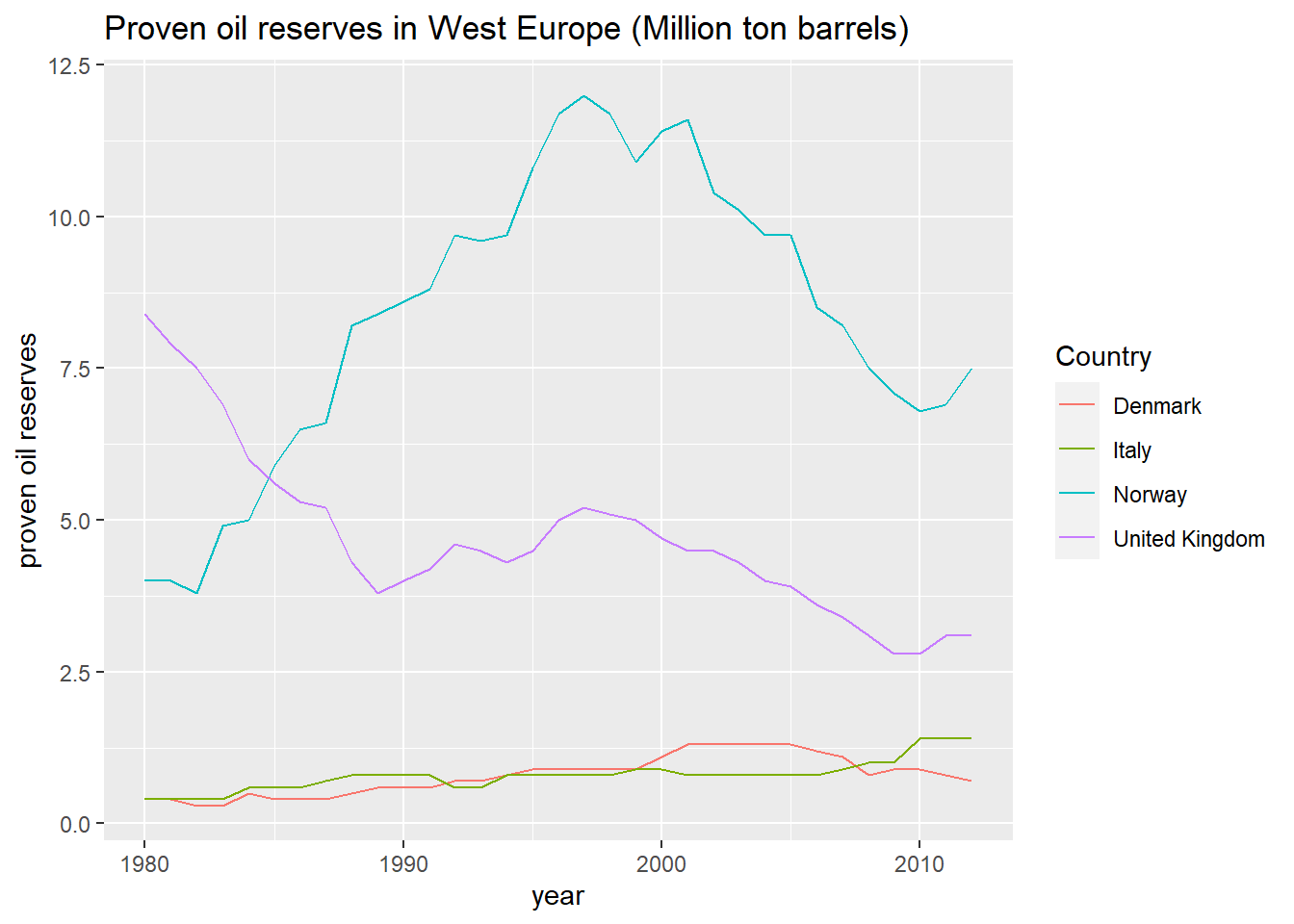
*## Use coloured line - a colour per country*

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

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 4

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))

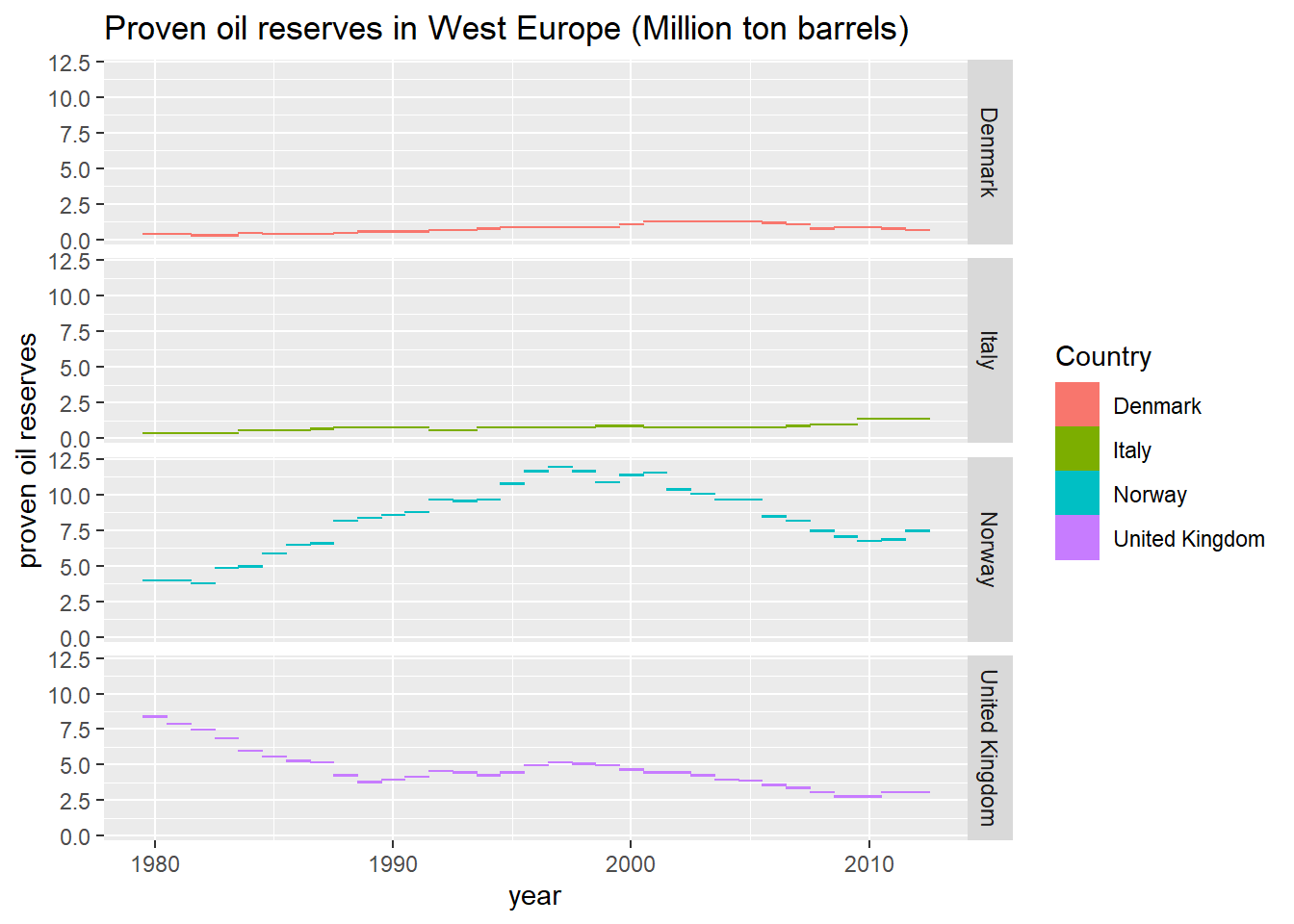
*## add appropriate labels, title and legend*

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

*## 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 reserves 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.

Data analysis

Exercise 5

Correlation and covariance

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

cor.API.Price

## [1] 0.9799122

The correlation coefficient is very close to 1, suggesting a very strong positive association.

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

cov.API.Price

## [1] 1.983339

The covariance is not as useful. It is positive, indicating a positive association, but the strength of the assoication is unclear.

Exercise 6

Producing a scatterplot

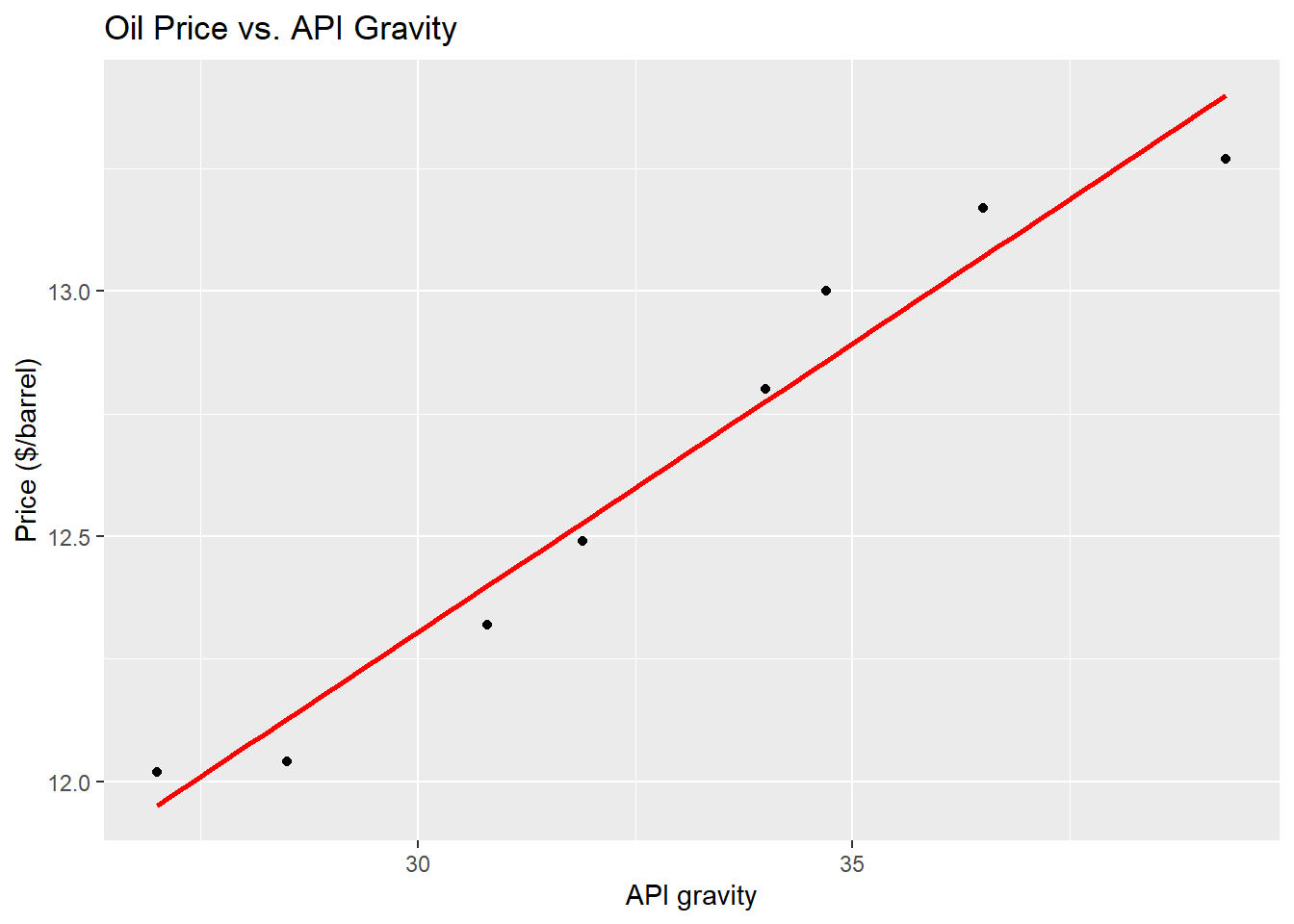
p <- ggplot(OilQuality,aes(x = API,y = Price))

p <- p + geom\_point()

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

p<- p+stat\_smooth(method="lm",se=F,col="red", formula=y~x)

p



The line fits the data quite well. However, there appears to be some curve in the shape of the data (so) ‘s’ shape, so a different kind of fitting may be better than a linear one.

Exercise 7

Linear regression

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 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

Predictions

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

## 1 2

## 12.95222 14.08250

Note that the second prediction is a dubious extrapolation.

Data manipulation and statistical inference

Manipulating data

Exercise 8

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")

theMean <- mean(UKNewData$OPUKShare)

UKNewData <- UKNewData |> mutate(bigProducerOil = (OPUKShare > theMean))

summary(UKNewData$bigProducerOil)

## Mode FALSE TRUE

## logical 7 5

There are 5 producers who produce more than average.

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

Exercise 9

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.

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 = 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.

Exercise 10

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.

The validity of the test would need to be explored.