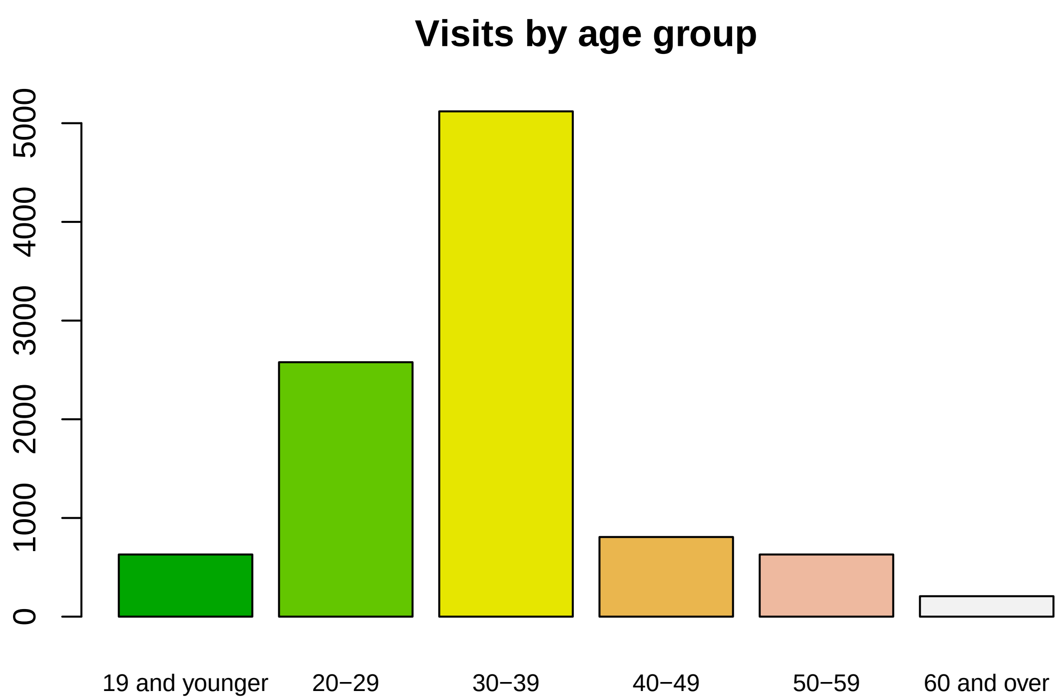
MATH40005 Coursework Spring 2022

Leo JONES, CID: 02036984

Question 1

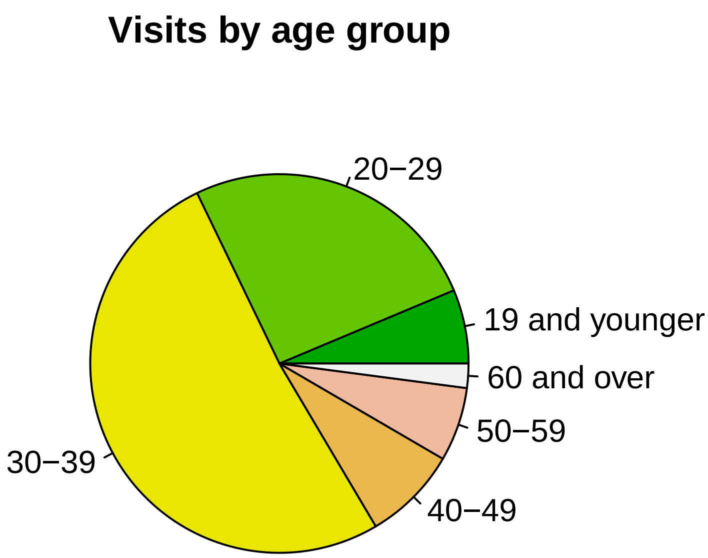
Question 1(a)

```
#reading data and creating breaks
d=read.table("spendingsurvey.txt",header=TRUE,sep=",")
labels= list("19 and younger","20-29","30-39","40-49","50-59","60 and over")
cat<-c(0,19,29,39,49,59,100)
#sorting entries into age ranges
age=as.numeric(t(d)[2,])
d$ageranges <- cut(age, breaks = cat)
#summing the visits over the different ranges and then plotting
ag=aggregate(d$visits,by =list(d$ageranges),FUN=sum)
barplot(as.vector(ag$x), names= labels, cex.names = 0.75, col=terrain.colors(6),main="Visits by age grou
```



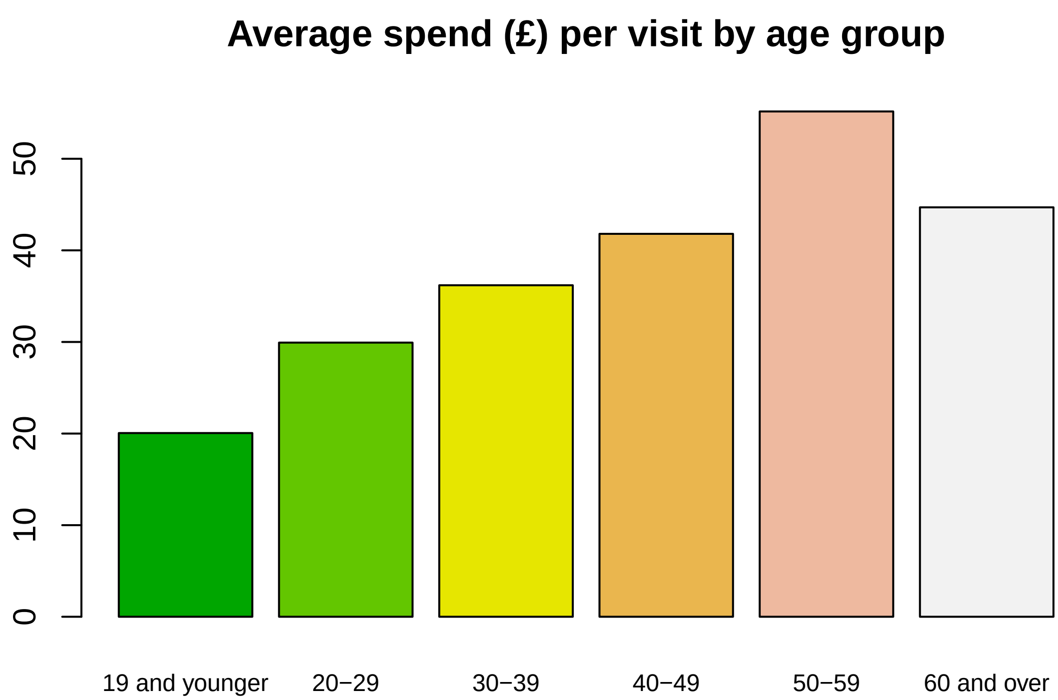
Question 1(b)

```
#same data as 1(a) but as a pie chart
pie(as.vector(ag$x), labels= labels, col=terrain.colors(6),main="Visits by age group")
```



Question 1(c)

```
#taking the mean of the total spend per visit for each age range
q1c=aggregate(d$totalspend/d$visits,by= list(d$ageranges), FUN=mean)
#plotting
barplot(as.vector(q1c$x), names =labels,cex.names=0.75,col=terrain.colors(6),main="Average spend (£) pe
```



Question 1(d)

The 30-39 group has the largest proportion of visits however the 50-59 group has the highest average spend.

```
#taking the mean of the total spend for each age group
q1d=aggregate(d$totalspend,by= list(d$ageranges), FUN=mean)
print(q1d)
```

```
##    Group.1         x
## 1   (0,19]  60.19871
## 2  (19,29] 146.89594
## 3  (29,39] 221.32500
## 4  (39,49] 160.72390
## 5  (49,59] 165.57357
## 6 (59,100]  88.34010
```

If you look at the total spend by age group, however, 30-39 has the highest spend, and therefore I would say they are most valuable to the restaurant chain.

Question 1(e)

```r
#creating lists of the spends per visit for each age range
h=aggregate(d$totalspend/d$visits, by =list(d$ageranges), FUN=list)
#t-test without equality of variances
t.test(unlist(h$x[2]),unlist(h$x[4]))
```

```
##
##  Welch Two Sample t-test
##
## data:  unlist(h$x[2]) and unlist(h$x[4])
## t = -49.329, df = 398.23, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.34921 -11.40262
## sample estimates:
## mean of x mean of y
##  29.92789  41.80380
```

Assuming normality but not equality of variances, the p-value is less than 0.05, which indicates a significant difference.

We can use the F test to test for equality in the variances:

```r
#F-test to compare two variances
var.test(unlist(h$x[2]),unlist(h$x[4]))
```

```
##
##  F test to compare two variances
##
## data:  unlist(h$x[2]) and unlist(h$x[4])
## F = 1.0744, num df = 523, denom df = 209, p-value = 0.5484
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.851174 1.341111
## sample estimates:
## ratio of variances
##           1.074431
```

This shows insufficient evidence of a significant difference in the variances as p-value>0.05. We can then use the classic t-test:

```r
#Classic t-test
t.test(unlist(h$x[2]),unlist(h$x[4]),var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  unlist(h$x[2]) and unlist(h$x[4])
## t = -48.577, df = 732, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
```

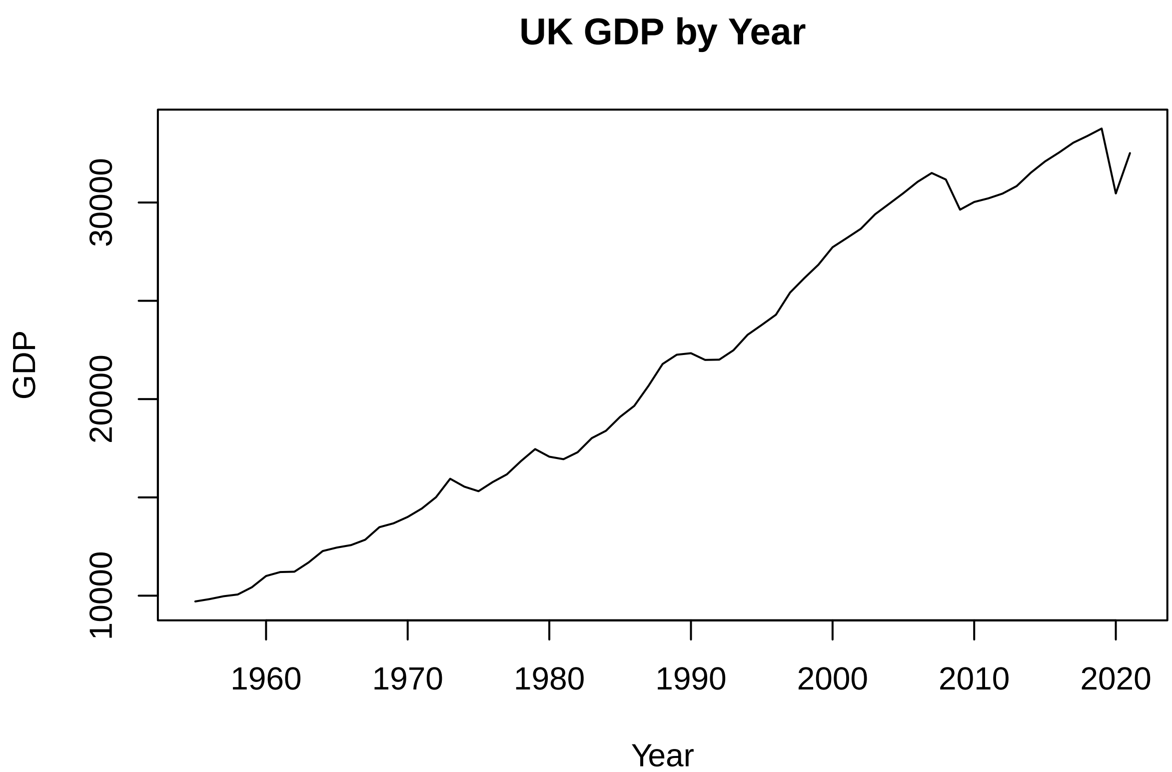
```
## 95 percent confidence interval:
##  -12.35587 -11.39595
## sample estimates:
## mean of x mean of y
##  29.92789  41.80380
```

This indicates a significant difference in the means of the two groups, 20-29 and 40-49.

Question 2

Question 2(a)

```
#reading data
g=read.table("ukgdp.txt",header=TRUE,sep=",")
#converting year and GDP to vectors
year=as.numeric(t(g)[1,])
gdp=as.numeric(t(g)[2,])
#plotting
plot(year,gdp,xlab="Year",ylab="GDP", main="UK GDP by Year",type='l')
```



Question 2(b)

```
#saving values for GDP in 2020 and 2021
g20=gdp[66]
g21=gdp[67]
#calculating their % difference
print((g21-g20)*100/g20)
```

```
## [1] 6.739766
```

The GDP increase between 2020 and 2021 is 6.74%. This is not equal to 7.5% therefore the claim is incorrect.

```
#same method as above
g19=gdp[65]
print((g21-g19)*100/g19)
```

```
## [1] -3.699316
```

The GDP decrease between 2019 and 2021 is 3.70%.

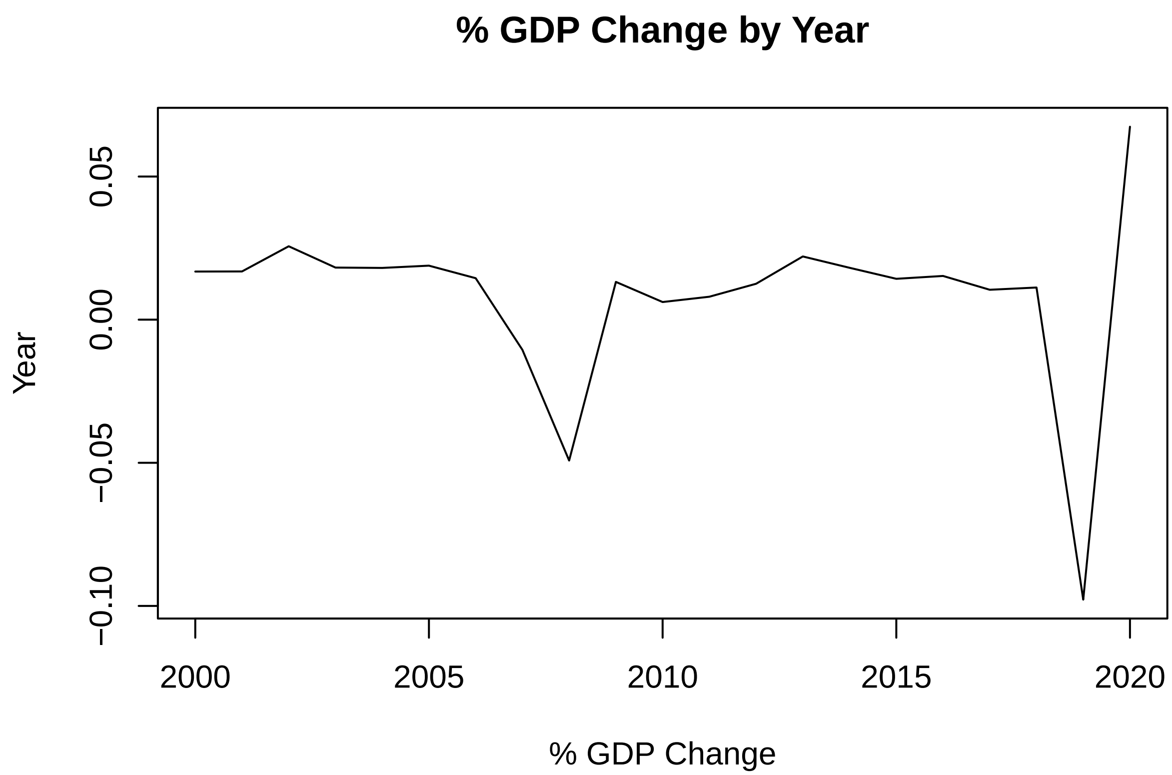
Question 2(c)

```r
gdpchange=NULL
vy=NULL
#using a for loop to iterate through GDP by year, and calculating % change using the next years GDP
#also creating a vector of the years
for (i in 46:66){
vy=append(vy,year[i])
change= (gdp[i+1]-gdp[i])/gdp[i]
gdpchange=append(gdpchange,change)
}
print(gdpchange)
```

```
##  [1]  0.016807935  0.016849349  0.025640131  0.018196660  0.018071887
##  [6]  0.018866686  0.014491820 -0.010539013 -0.049213988  0.013193413
## [11]  0.006161122  0.008010062  0.012543508  0.022084576  0.018117207
## [16]  0.014273249  0.015270694  0.010440941  0.011231580 -0.097799366
## [21]  0.067397656
```

Question 2(d)

```
#plotting % change
plot(vy,gdpchange, type='l',xlab="% GDP Change", ylab="Year", main="% GDP Change by Year")
```




```
#finding the GDPs of the anomalous years
anoms<- c(2008,2009,2019,2020,2021)
for (i in match(anoms, unlist(year))){
  print(gdp[i])}
```

```
## [1] 31170
## [1] 29636
## [1] 33763
## [1] 30461
## [1] 32514
```

The anomalies for the GDP change are the years 2008-2009, 2019-2020, 2020-2021. The 2008-2009 anomaly is due to the stock market crash in 2008. The anomalies in the past two years are likely due to COVID-19.