۱-

a)

- مطالعه از نوع Observational است. (اگر آن ۴۰۰۰ سرباز بازنشسته به صورت رندوم انتخاب شده باشند می‌توان نتیجه مطالعه را به کل سربازان بازنشسته درجنگ ویتنام نسبت داد. اگر به صورت رندوم انتخاب نشده باشند این مطالعه ی observational خوبی نبوده و نتیجه مطالعه را تنها می‌توان به همان نمونه نسبت داد)
- Explanatory variable: شدت PTSD (the severity of PTSD)
- Response variable: مرگ بر اثر بیماری قلبی (Death from heart disease)
- خیر – نمیشه نتیجه گیری علّی کرد و رابطه علّی(causal) مشخص کرد. چرا که مطالعه از نوع مشاهده ای است. تنها می‌توان یک همبستگی(correlation) بین متغیر های Explanatory و Response نشان داد.

---

b)

- مطالعه از نوع Experimental است.
- Explanatory variable: نوع رژیم غذایی (diet type: low-fat, Mediterranean, low-carbohydrate)
- Response variable: کاهش وزن (Weight Loss)
- بله – می‌توان نتیجه‌گیری علّی کرد و رابطه‌ی causal مشخص کرد چون random assignment انجام شده. (البته در صورتی می‌توان نتیجه مطالعه را به همه افراد جامعه نسبت داد که نمونه‌گیری نیز به صورت رندوم صورت گرفته باشد(random sampling). در غیر این صورت نتیجه مطالعه را فقط می‌توان به اعضای نمونه نسبت داد.)

---

c)

- مطالعه از نوع Experimental است. (البته در صورتی که تقسیم دانش‌آموزان کلاس نهم به دو گروه به صورت رندوم صورت گرفته باشد. یعنی random assignment داشته باشیم)
- Explanatory variable: نوع شیوه آموزشی (type of Instruction: IPL , tell and practice)
- Response variable: میزان یادگیری (Learning)
- می‌توان نتیجه‌گیری علّی کرد و رابطه‌ی causal مشخص کرد در صورتیکه random assignment انجام شده باشد.

-۲

a) بله وجود دارد.

Confounding variable: temperature

با گرم شدن هوا، اجتماع افراد در مکان‌های عمومی بیشتر شده و در نتیجه احتمال وقوع دعوا و مشاجره و از کوره در رفتن افراد هم بیشتر می‌شود. که این امر می‌تواند منجر به افزایش آمار وقوع جرم از جمله قتل شود.

همچنین با گرم تر شدن هوا، تمایل افراد به خوردن بستنی بیشتر می‌شود. پس همبستگی بین میزان فروش بستنی و میزان قتل، ناشی از یک متغیر سوم است که آن دمای هوا می‌باشد.

---

b) خیر- وجود ندارد. تنها یک correlation است که به طور اتفاقی بین دو متغیر مربوطه وجود دارد.

-۳

a) Cluster sampling (نمونه برداری خوشه‌ای)
هر دیدار مذهبی شهر به یک خوشه نسبت داده شده. از بین این خوشه ها، ۵ تا به صورت تصادفی انتخاب شده و سپس تمام اعضای این خوشه به عنوان نمونه انتخاب شدند.

b) Simple Random Sampling
انتخاب اولین فرد به صورت رندوم صورت گرفته و در نتیجه انتخاب افراد بعدی که با فاصله ۲۰ انتخاب می‌شوند هم وابسته به این انتخاب رندوم اولیه خواهد بود.

c) stratified sampling (نمونه برداری لایه لایه)
اگر قبل از نمونه برداری، افراد جامعه هدف رو با توجه به گروه های سنی گروهبندی شده و سپس تعداد ثابتی از افراد رو به طور رندوم از هر گروه انتخاب شوند، در اون صورت روش نمونه برداری stratified sampling می‌باشد(در اینجا این روش به طور مستقیم نگفته شده ولی روشی که برای این نوع کار آماری مناسب است نمونه برداری لایه لایه می‌باشد)

d) Simple Random Sampling
چون احتمال انتخاب هر کدام از دانش آموزان یکسان بوده و تصادفی است.

-۴

a)۵۲ نفر از کاربران رسانه های اجتماعی رو به صورت رندوم انتخاب می‌کنیم. سپس آنها را بر اساس سطح مهارتشون در کار کردن با واتسپ و تلگرام در سه گروه مبتدی، متوسط و حرفه ای قرار می‌دهیم. بعد به طور رندوم به تعداد نصف افراد مبتدی، نصف افراد متوسط و نصف افراد حرفه‌ای از هر گروه می‌گیریم و آنها را به تلگرام assign می‌کنیم. بقیه افراد باقیمانده از هر گروه را به واتسپ

assign می‌کنیم. سپس از افراد گروه تلگرام و واتسپ میخواهیم که taskهای مشخص شده را انجام دهند و مدت زمان لازم برای اتمام هر task را می‌نویسیم. در آخر، مدت زمان گروه تلگرام و واتسپ را میانگین می‌گیریم و با هم مقایسه می‌کنیم.

b) بله - <<میزان مهارت>> را به عنوان متغیر blocking در نظر گرفته و در مرحله random assignmentاستفاده کردم. چرا که میزان مهارت افراد در سرعت اتمام taskها توسط آنها موثر است. پس با اینکار نسبت افراد مبتدی، متوسط و حرفه‌ای در گروه تلگرام با نسبت آنها در گروه واتسپ یکسان می‌شود.

۵-

a) Convenience sample: رسول فقط از افرادی که تلفن همراه داشتند سوال پرسید و ۱۰٪ دیگر ازافراد جامعه هدف را درنظر نگرفت. پس نمونه ممکن است به خوبی representative جامعه نباشد.
Non-response bias: از افرادی که تلفن همراه دارن، ۹۱٪ آنها پاسخ دادند و ۹٪ آنها هیچ پاسخی ندادند که این باز هم یک منبع بایاس می‌باشد.

b) Advertising or Pre-screening Bias: چون استاد درس با وعده دادن نمره در ازای پاسخگویی به سوالات، دانشجو ها را به شرکت در این survey تشویق کرده و باعث ایجاد bias در نمونه شده.

c) عامل sampling bias وجود ندارد.

d) منبع بایاس وجود ندارد. البته برای اینکه بتوان نتیجه گیری مذکور را انجام داد بهتر بود فرد محقق این سوال را در روزهای مختلف سال از افراد می‌پرسید. چون در این حالت فعلی، نتیجه رو فقط می‌توان به صبح آن روز خاص نسبت داد. اگر اینگونه در نظر بگیریم شاید بتوان گفت که convenience bias دارد.

e)
Non- response bias: اسنپ فقط افرادی که به سوالات پاسخ دادند را در نظر گرفت و در واقع افرادی که به نظرسنجی پاسخ ندادند را از نمونه آماری دور ریخت.
Advertising or Pre-screening Bias: چون اسنپ با وعده ی سفر مجانی، افراد را به شرکت در نظرسنجی تشویق کرده، با اینکار یک منبع دیگر بایاس ایجاد کرده است و باعث شده افرادی که شرکت می‌کنند احتمالا نظر مثبت دهند.

۶-

a) False – چون شیب خط بین نقاط نارنجی مجاور هم که روی نمودار مشخص شده، یکسان نیست. (چون فاصله بین ماه های مشخص شده روی بردار xها یکسان نیست)

b) False – این جمله لزوما صحیح نمی‌باشد. باید جمعیت افراد هر ناحیه رنگ شده را بدانیم و همچنین میزان اختلاف بین رای های دموکرات ها و جمهوری‌خواه ها در هر ناحیه. چرا که ممکن است جمعیت نقاط آبی بیشتر باشه و اکثر افراد دموکرات باشند در حالیکه در نقاط قرمز، علاوه بر کمتر بودن جمعیت، اختلاف ناچیزی باعث برتری جمهوری‌خواه ها شده باشد.

c) True

d) False – چون همانطور که در نمودار مشاهده می‌شود ماه ها به ترتیب از زمستان تا تابستان نوشته شدند. که از زمستان تا تابستان دمای هوا به تدریج افزایش می‌یابد. پس ماه های سال یک متغیر blocking محسوب می‌شود. که برای چنین نتیجه گیری ای باید کنترل شود. یعنی برای بررسی چنین نتیجه ای بهتر بود که میانگین دمای سال های متوالی را در نمودار رسم می‌کردیم.

۷-

a)

H0(Null Hypothesis): میانگین دور وسط بالای بازو افراد سیگاری برابر با 24 cm می‌باشد. (MUAC = 24 cm)

HA(Alternative Hypothesis): میانگین دور وسط بالای بازو افراد سیگاری بیشتر از 24 cm می‌باشد. (MUAC > 24 cm)

---

b) با مشاهده‌ی هیستوگرام می‌بینیم که احتمال رخ دادن میانگین MUAC برابر با 24 cm یا کمتر از آن(extremeتر) برابر با 2/50 می‌باشد. (از این ۵۰ تا simulationی که رسول run کرد، فقط دوبار میانگین 24 cm یا کمتر مشاهده شده.)

$$p - value = \frac{2}{50} = 4\%$$

---

c) از این ۵۰ تا simulationی که رسول run کرد، فقط دوبار میانگین 24 cm یا کمتر مشاهده شده. پس یعنی احتمال خیلی کمی(4%) وجود دارد که با فرض صحیح بودن H0 ، چنین دیتایی اتفاق بیفتد. پس رسول می‌تواند نتیجه بگیرد که H0 صحیح نیست و H0 را reject کند. و می‌توان گفت: شواهد کافی برای اینکه میانگیم MUAC افراد سیگاری کمتر مساوی 24 cm باشد، وجود ندارد.

۸-

a)

```
#creating a vector
nums <- c(57,66,72,78,79,79,81,81,82,83,84,87,88,88,89,90,91,92,94,95)
```

---

b)

```r
#Calculating the Median
med <- median(nums)

#Calculating the Variance
vari <- var(nums)

#Calculating the Standard Deviation
sdev <- sd(nums)

#printing median, variance and sd
sprintf("median: %s", med)
sprintf("variance: %s", vari)
sprintf("standard deviation: %s", sdev)
```

The result:

```
> sprintf("median: %s", med)
[1] "median: 83.5"
> sprintf("variance: %s", vari)
[1] "variance: 90.1684210526316"
> sprintf("standard deviation: %s", sdev)
[1] "standard deviation: 9.49570540047613"
>
```

```r
#Calculating the Mode(s)
y <- table(nums)
df <- data.frame(y)
modes <- df$values[df$Freq == max(df$Freq)]

print(modes)
```

The result:

```
[1] 79 81 88
Levels: 57 66 72 78 79 81 82 83 84 87 88 89 90 91 92 94 95
>
```

---

c) بله – یک outlier وجود دارد که مقدار کمتری از lower whisker دارد. (outlier = 57)

```r
#plotting the boxplot of nums
boxnums = boxplot(nums)

#calculating the outliers
#getting the values of upper whisker and lower whisker in the boxnums boxplot
lower_whisker = boxnums$stats[1]
upper_whisker = boxnums$stats[5]

upper_outliers <- nums[nums < lower_whisker]
lower_outliers <- nums[nums > upper_whisker]

print(upper_outliers)
print(lower_outliers)
```

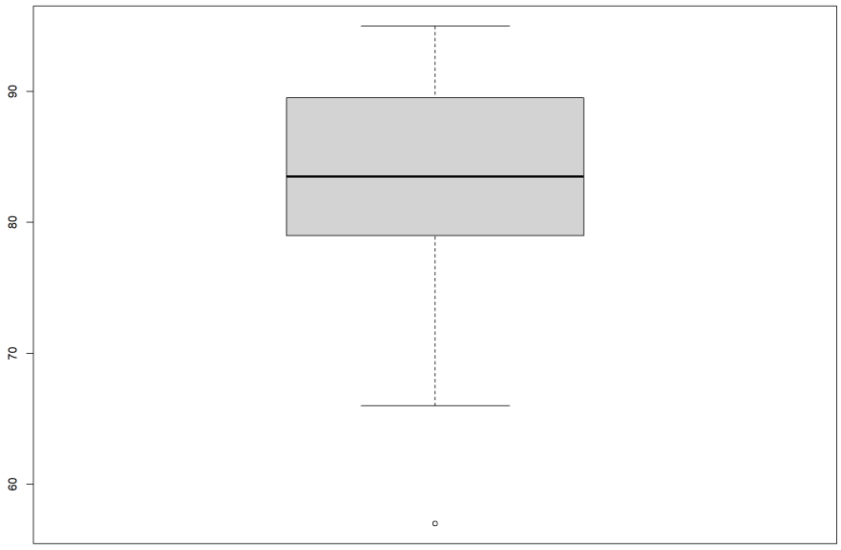
The result:

```
> print(upper_outliers)
[1] 57
> print(lower_outliers)
numeric(0)
> 
```

(d

```
#plotting the box plot of nums
boxnums = boxplot(nums)
```

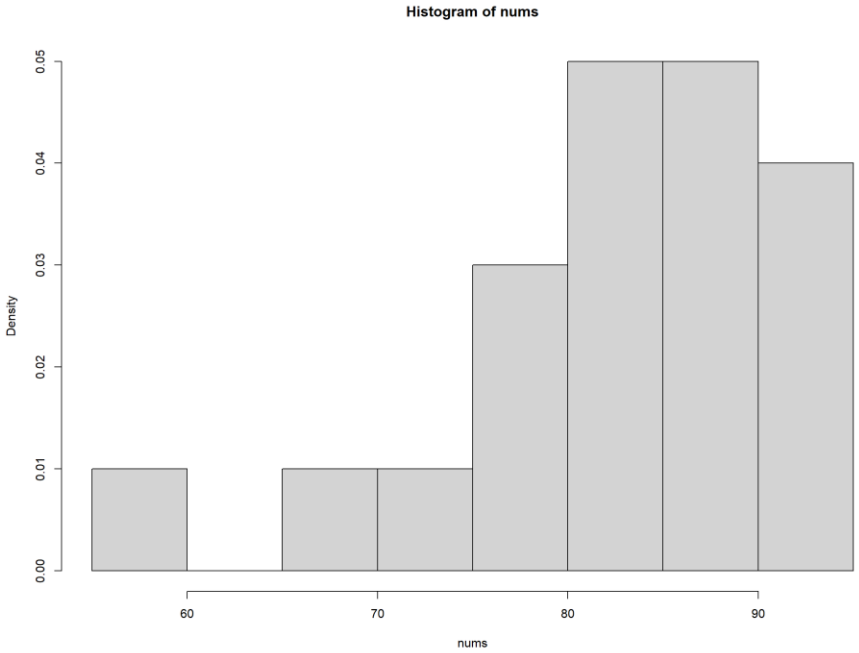
The result:



(e

```
#plotting the histogram
hist(nums, freq = FALSE)
```

The result:



I. نمرات چولگی به چپ دارند(Left - skewed). چرا که با توجه به هیستوگرام برای رفتن از نقطه ماکسیمم نمودار به سمت مینیمم باید به سمت چپ حرکت کنیم. در واقع outlier ها (البته در اینجا یک outlier داریم) بیشتر در مقادیر پایین قرار گرفته اند. (در box plot قابل مشاهده می‌باشد)

II. انتظار می‌رود که مقدار میانگین کمتر از مقدار میانه باشد. زیرا طبق تعریف ساده‌ی چولگی داریم:

$$\frac{mean-median}{SD} < 0 \longrightarrow \text{چولگی به چپ}$$

III. میانگین نمرات برابر با 82.8 بوده و میانه نمرات برابر با 83.5 می‌باشد. اختلاف زیادی بین این دو مقدار وجود ندارد. ولی با توجه به اینکه یک outlier با مقدار کمتر از lower whisker داریم و میانگین یک measure of center حساس به outlier می‌باشد، بهتر است که میانه را به عنوان measure of center انتخاب کنیم.

۹-

(a

```
# Identifying the variables and their types
sapply(imdb_df, typeof)
```

The result:

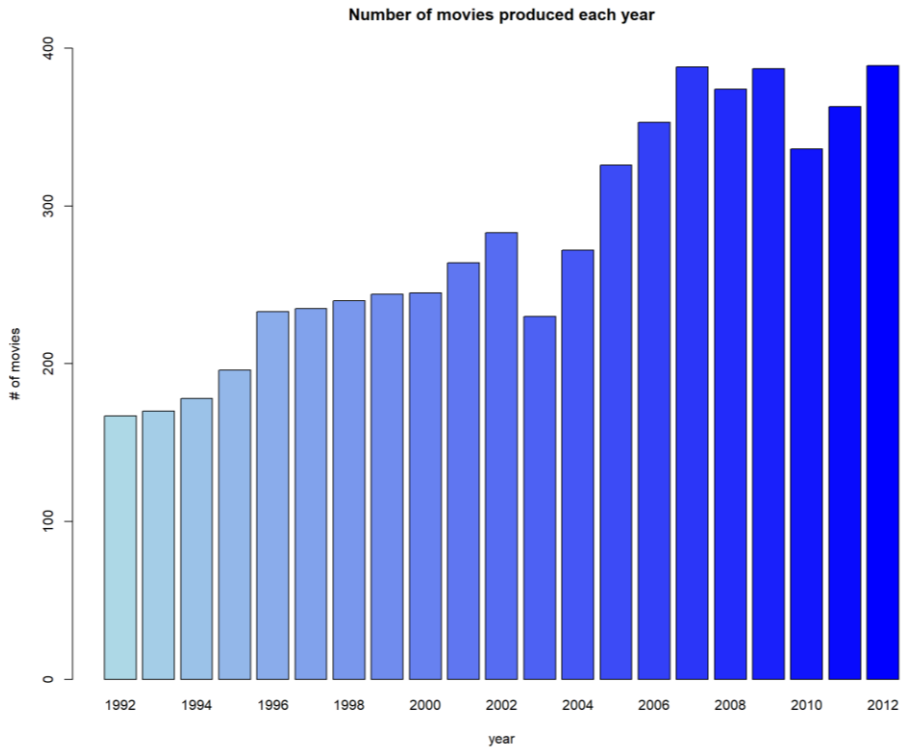
```
> sapply(imdb_df, typeof)
             imdb_ID                 title                  year              duration
         "character"           "character"             "integer"             "integer"
       total_votes                budget       USA_gross_income worldwide_gross_income
         "integer"              "double"              "double"              "double"
 tomatometer_status      audience_rating
         "character"              "double"
>
```

---

b) از آنجایی که <<سال تولید>> یک متغیر categorical می‌باشد، می‌بایست از یکی از روشهای visualization برای این نوع متغیرها استفاده کنیم. برای اینکار از bar plot استفاده می‌کنیم.

```
#plotting Bar plot for the number of movies produced each year
barplot(table(imdb$year), main = "Number of movies produced each year",
        xlab = "year", ylab = "# of movies", ylim = c(0,400),
        col = colorRampPalette(colors = c("lightblue", "blue"))(21))
```

The result:

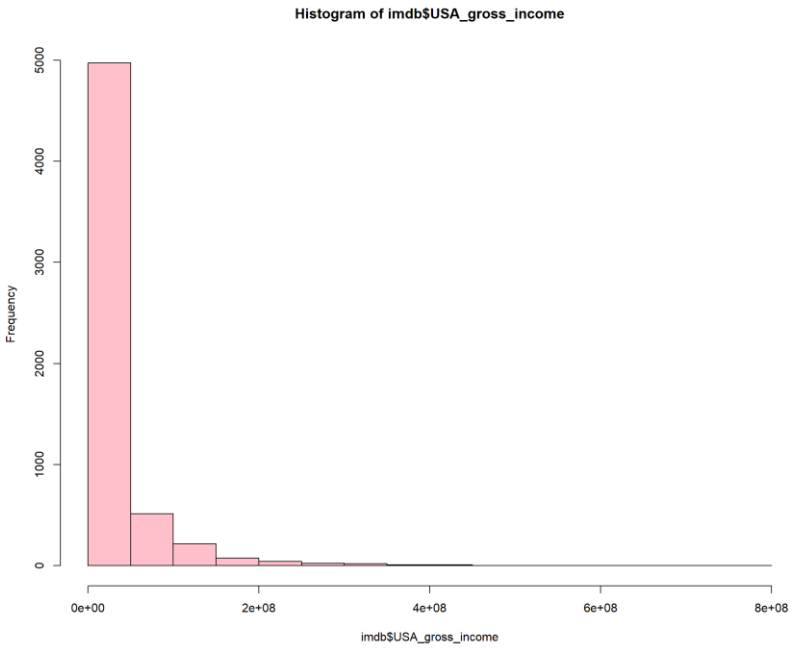


---

c)

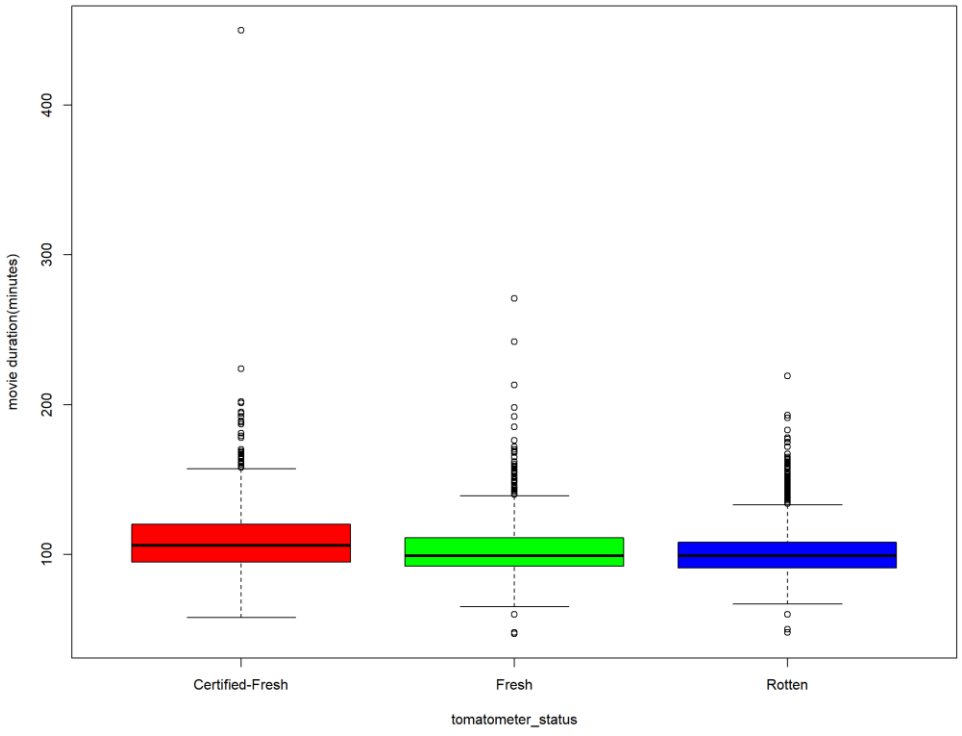
```
#plotting the Histogram of USA_gross_income
hist(imdb$USA_gross_income, col = "pink")
```

The result:



همانطور که در هیستوگرام مشاهده می‌کنید میزان درآمد ناخالص فیلم ها در آمریکا به شدت چولگی به راست(right-skewed) دارد. (outlierها در نقاط بالا قرار دارند) یعنی سود حاصل از بیشتر فیلم ها در آمریکا پایین بوده و هرچه میزان درآمد ناخالص فیلم ها به سمت بالا می‌رود، تعداد فیلم ها کمتر می‌شود که البته این نوع توزیع درآمد مورد انتظار بود چرا که معمولا تعداد اندکی فیلم درآمد خیلی بالا دارند و اکثریت فیلم ها سود کمی داشته در حالیکه بخش قابل قبولی از فیلم ها درآمد متوسط دارند.

(d



حال برای محاسبه‌ی outlierهای هرکدام از box ها به صورت زیر عمل می‌کنیم(ابتدا whiskerهای بالا و پایین هر کدام از box ها را بدست می‌آوریم):

```
#finding the outliers for each tomatometer status

#finding the upper and lower whiskers for each of the boxes
dur_tomat_df <- data.frame(tomatometer_status = c("Certified-Fresh","Fresh","Rotten"),
                           lower_whisker = c(side_box$stats[1,1:3]),
                           upper_whisker = c(side_box$stats[5,1:3]))
```

سپس outlier ها را محاسبه و چاپ می‌کنیم:

Certified-Fresh outliers:

```
##finding the outliers for "Certified_Fresh"
outliers_Certified_Fresh <- imdb$duration[imdb$tomatometer_status == "Certified-Fresh"]
upper_outliers_Certified_Fresh <-
  outliers_Certified_Fresh[outliers_Certified_Fresh > dur_tomat_df$upper_whisker[dur_tomat_df$tomatometer_status == "Certified-Fresh"]]
lower_outliers_Certified_Fresh <-
  outliers_Certified_Fresh[outliers_Certified_Fresh < dur_tomat_df$lower_whisker[dur_tomat_df$tomatometer_status == "Certified-Fresh"]]
print(lower_outliers_Certified_Fresh)
print(upper_outliers_Certified_Fresh)
```

The result:

```
> print(lower_outliers_Certified_Fresh)
integer(0)
> print(upper_outliers_Certified_Fresh)
 [1] 202 178 178 170 170 160 162 170 159 166 165 162 159 167 189 161 187 224 168 188 450 164 192 169 195 181 164 158 178 201 179 158 194 159 170
> |
```

همانگونه که در box plot نیز قابل مشاهده می‌باشد، می بینیم که دسته‌ی Certified-Fresh فقط outlierهایی بزرگتر از upper whisker دارد(35 تا) و outlierی با مقدار کمتر از lower whisker ندارد.

Fresh outliers:

```
##finding the outliers for "Fresh"
outliers_Fresh <- imdb$duration[imdb$tomatometer_status == "Fresh"]
upper_outliers_Fresh <-
  outliers_Fresh[outliers_Fresh > dur_tomat_df$upper_whisker[dur_tomat_df$tomatometer_status == "Fresh"]]
lower_outliers_Fresh <-
  outliers_Fresh[outliers_Fresh < dur_tomat_df$lower_whisker[dur_tomat_df$tomatometer_status == "Fresh"]]
print(lower_outliers_Fresh)
print(upper_outliers_Fresh)
```

The result:

```
> print(lower_outliers_Fresh)
[1] 48 60 47
> print(upper_outliers_Fresh)
 [1] 140 145 140 271 145 149 150 170 185 141 168 158 157 142 176 152 151 150 148 143 155 170 150 172 154 140 144 148 142 146 160 242 169 169 160 198 168 213 146 140 185
[42] 154 148 156 143 141 165 162 149 168 159 162 142 140 169 198 145 192 162 162 155
> |
```

می بینیم که دسته‌ی Fresh سه outlier با مقدار کمتر از lower whisker داشته و 61تا outlier بزرگتر از upper whisker دارد.

Rotten outliers:

```
##finding the outliers for "Rotten"
outliers_Rotten <- imdb$duration[imdb$tomatometer_status == "Rotten"]
print(outliers_Rotten)
upper_outliers_Rotten <-
  outliers_Rotten[outliers_Rotten > dur_tomat_df$upper_whisker[dur_tomat_df$tomatometer_status == "Rotten"]]
lower_outliers_Rotten <-
  outliers_Rotten[outliers_Rotten < dur_tomat_df$lower_whisker[dur_tomat_df$tomatometer_status == "Rotten"]]
print(lower_outliers_Rotten)
print(upper_outliers_Rotten)
```

The result:

```
> print(lower_outliers_Rotten)
[1] 48 60 50
> print(upper_outliers_Rotten)
  [1] 144 165 136 140 191 135 148 183 137 135 138 135 153 140 154 158 164 157 137 175 145 162 151 165 163 147 172 159 151 138 137 143 135 135 149 142 147 136 140 134 140
 [42] 137 136 175 219 135 167 136 140 135 145 137 150 139 138 193 157 135 144 139 152 146 139 161 178 145 155 146 160 141 147 143 151 137 177 134 135 163 165 140 142 136
 [83] 145 146 145 135 134 134 136 146 142 151 153 135 135 141 139 136 140 134 134 154 149 144 163 136 141 135 135 135 134
>
```

می بینیم که دسته‌ی rotten سه outlier با مقدار کمتر از lower whisker داشته و 111تا outlier بزرگتر از upper whisker دارد.

(e

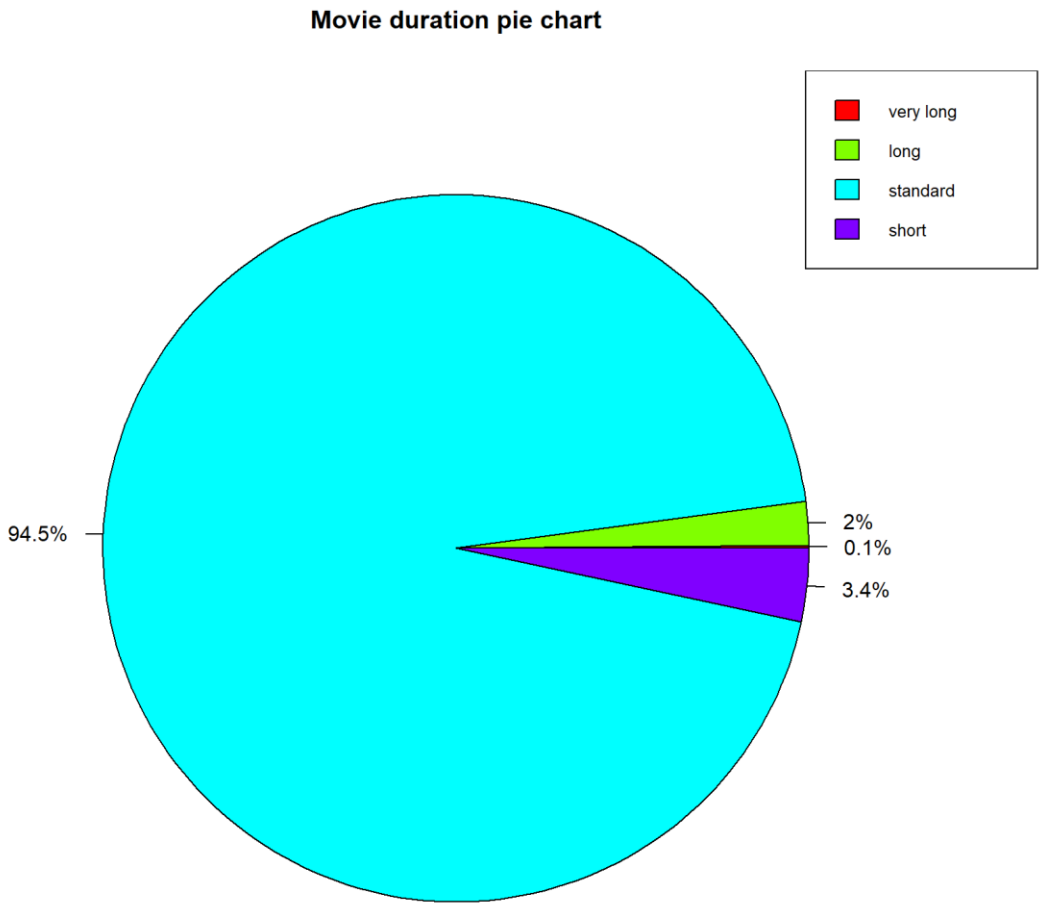
```
#e) Categorizing the movies based on their durations into 4 groups and plotting pie chart
imdb$duration_category <-
    ifelse(imdb$duration > 200, "very long",
           ifelse(imdb$duration > 150, "long",
                  ifelse(imdb$duration > 80, "standard","short")))
#getting the frequency of movies in each category
very_long_freq <- length(imdb$imdb_ID[imdb$duration_category == "very long"])
long_freq <- length(imdb$imdb_ID[imdb$duration_category == "long"])
standard_freq <- length(imdb$imdb_ID[imdb$duration_category == "standard"])
short_freq <- length(imdb$imdb_ID[imdb$duration_category == "short"])

#setting the arguments of the chart
x <- c(very_long_freq,long_freq,standard_freq,short_freq)
labels <- c("very long","long","standard","short")

piepercent<- round(100*x/sum(x), 1)
piepercent <- as.character(piepercent)
piepercent <- paste(piepercent, "%", sep="")

# Plotting the pie chart
pie(x, labels = piepercent, main = "Movie duration pie chart",col = rainbow(length(x)))
legend("topright", c("very long","long","standard","short"), cex = 0.8,
       fill = rainbow(length(x)))
```

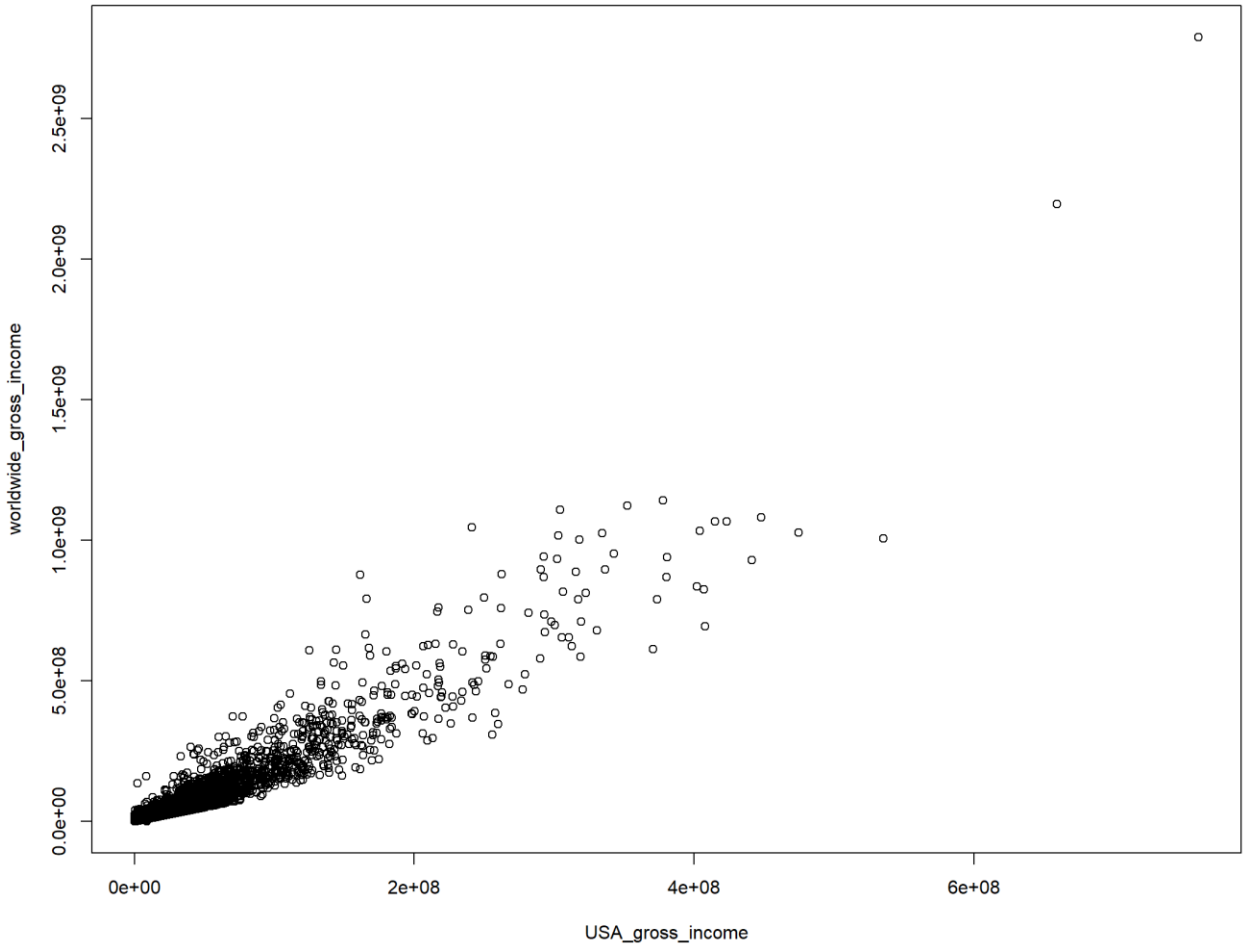
The result:



(f

```
#f) scatter plot for determining the relationship "USA_gross_income" and "worldwide_gross_income"
plot(imdb$USA_gross_income,imdb$worldwide_gross_income, xlab = "USA_gross_income",
     ylab = "worldwide_gross_income")
```

The result:



با توجه به scatter plot به نظر می‌رسد که correlation مثبتی بین دو متغیر مذکور وجود دارد. یعنی با افزایش درآمد ناخالص فیلم ها در آمریکا(میزان سود فیلم از فروش درآمریکا)، میزان درآمد ناخالص فیلم ها در جهان(میزان سود فیلم از فروش جهانی) نیز افزایش می‌یابد. این correlation را می‌توان این گونه توضیح داد که معمولا فیلمی که در آمریکا مورد توجه مردم قرار می‌گیرد و زیاد فروش می‌رود، به طور کلی در سایر کشور ها هم محبوب شده و زیاد به فروش می‌رود.