Homework 1  
Statistical Inference

**PREPARED BY**

Mohsen Fayyaz - 810100524

Fall 2021

**فهرست مطالب**

[**Q1**](#_tbfpyld5fvy5) **5**

[Answer the following questions for each section:](#_7dcxtpqck16) 5

[a.](#_qj0wlgw9ybue) 5

[b.](#_zf56r6g554j) 5

[c.](#_iwojybqdyb96) 6

[**Q2**](#_r6tf1svio9tc) **7**

[In each of the following results, is there a potential confounding variable? If the answer is yes, find the confounding factor and explain.](#_rzgg0i52f2ki) 7

[a.](#_kckhtdmojkc6) 7

[b.](#_6mr9h7vjery1) 7

[**Q3**](#_9l83x94f1bhm) **8**

[Define the sampling method and explain your response:](#_muakmnirxgss) 8

[a.](#_sruq7taqbbj9) 8

[b.](#_boyzmljmd0ev) 8

[c.](#_wcu5dlfpa79y) 8

[d.](#_ub8tddt85x59) 8

[**Q4**](#_icq8bv5xtax5) **10**

[a.](#_dy65vs7v0qes) 10

[b.](#_z24ges7o5tki) 10

[**Q5**](#_5x3ec0qe3ulr) **11**

[For each of the following parts, explain (use more than three words) the most concerning potential source of bias (If any).](#_wabis0af1vhk) 11

[a.](#_w4g8cygqckqw) 11

[b.](#_qqruefs712kz) 11

[c.](#_r7qtiensm0ag) 11

[d.](#_ifgcqxai1jsj) 12

[e.](#_vecvp72c46) 12

[**Q6**](#_mgeokmetsuhb) **14**

[Determine if the following statements are true or false based on the corresponding diagram. If false, explain your reasoning.](#_5dknwvt3x7en) 14

[a.](#_lfu9ncu4k9jq) 14

[An employee’s daily income increases linearly over time in 2021 (from Jun 2 to Dec 8).](#_vp8ljz6iwde3) 14

[b.](#_qb3ky8itfmhy) 15

[The following map is the electoral vote map of America. According to this, most of the Americans vote for republicans.](#_9memo31xyp7p) 15

[c.](#_bo5ngb4flywm) 16

[d.](#_8x3u0qwbrzrb) 17

[**Q7**](#_e6mksrd1gi70) **18**

[a.](#_5mhv5jqf0zsy) 18

[b.](#_2hipaz3tyezx) 18

[c.](#_x6x8l0tgcyt6) 19

[**Q8 (R)**](#_ozgkbbvq3ee3) **20**

[Below are the final exam scores of twenty introductory statistics students.](#_kv6p5vqxob6i) 20

[a.](#_kqhaqq8kr9pm) 20

[b.](#_fgtsdyy0gg1i) 20

[c.](#_y1te2y3cs8uy) 20

[d.](#_agy80avjscsx) 21

[e.](#_eecj1d2hqudo) 21

[**Q9 (R)**](#_wm80p2elr69o) **23**

[In this part, you are going to study and analyze the IMDB movies dataset. Your plots must have a proper title, x-label and y-label. Also, do not use any non-built-in R packages, e.g., ggplot2, etc. Follow the instructions:](#_5ajb0xehgryq) 23

[a.](#_5uxrrlxv3z4h) 23

[b.](#_hw6n4ahzz38i) 23

[c.](#_hec488niuzs) 24

[d.](#_4s6zxxxfloio) 26

[e.](#_d3q0riymyog0) 28

[f.](#_j6b3d511qmzm) 29

# Q1

## Answer the following questions for each section:

* Explain the type of study (Observational or Experimental).
* Identify explanatory variables.
* Identify response variable.
* Comment on whether or not the findings of the study can be used to establish causal relationships.

### a.

In a study published in the July issue of Psychosomatic Medicine, Joseph Boscarino examined the prevalence of heart disease and PTSD in more than 4,000 Vietnam veterans. The more severe the PTSD diagnosis, the greater the likelihood of death from heart disease, the study showed.

* Type of study: Observational: since there is no random assignment
* Explanatory Variables: Severity of the PTSD diagnosis
* Response Variables: the likelihood of death from heart disease
* causal relationships: As this is an observational study, we can only establish an association, not causal connections, between explanatory and response variables.

### b.

In order to assess the effectiveness of 3 different types of diets on weight loss, in a 2-year trial, researchers randomly assigned 322 moderately obese subjects to one of three diets: low-fat, Mediterranean, low-carbohydrate. Results show that the Mediterranean and the low-carbohydrate diets may be effective alternatives to low-fat diets.

* Type of study: Experimental: since there is random assignment of subjects to treatments
* Explanatory Variables: 3 different types of diets (low-fat, Mediterranean, low-carbohydrate)
* Response Variables: Weight loss
* causal relationships: As this is an experimental study, we can establish causal connections between explanatory and response variables.

### c.

Educational psychologists investigate the impact of different types of instruction on learning. In one study, researchers taught a math lesson to 9th-graders using the “Inventing to Prepare for Learning (IPL)” instructional cycle. The second group of students received traditional “tell and practice” instruction. After the lessons, both groups studied a worked example of a math problem on their own. Then they took a test that included problems like the worked example. The journal Cognition and Instruction published the results in 2004.

* Type of study: Experimental: since there is random assignment of subjects to treatments
* Explanatory Variables: Different types of instruction (“Inventing to Prepare for Learning (IPL)”, “tell and practice”)
* Response Variables: Learning
* causal relationships: As this is an experimental study, we can establish causal connections between explanatory and response variables.

# 

# Q2

## In each of the following results, is there a potential confounding variable? If the answer is yes, find the confounding factor and explain.

### a.

It is known that murder rates and ice cream sales are highly positively correlated throughout the year, That is, as murder rates rise, so does the sale of ice cream.

Potential confounding variable:   
Weather -> In warm weather, when people eat more ice cream, there are more people on the streets and therefore more potential targets for murder which may affect the murder rate.

### b.

A study claims that living next to high-voltage transmission lines causes cancer.

Potential confounding variable:   
Poor Neighborhood -> It might be the case that most low-income families live in places where a high-voltage line is near; thereforeother factors such as less healthy diets and fewer doctor checkups can cause more cancer.

However, selection bias might have a larger effect on the results of this study than these variables.

# Q3

## Define the sampling method and explain your response:

### a.

A hospital wants to survey religious participants in their city about what they seek from a hospital chaplain, so they randomly select 5 religious meetings in the city and survey every participant in those meetings.

Cluster Sampling:  
Divide the population into clusters (religious meetings), randomly sample a few clusters (5), then sample all observations within these clusters

### b.

A factory manager takes an alphabetized list of workers’ names and picks a random starting point. Every 20th worker is selected to take a survey.

Systematic Sampling: (This was not in the course)  
This is a selection of elements (workers) from an ordered sampling frame (alphabetized list of workers’ names) where we first pick a random starting point and select every 20th element thereafter.

### c.

A research team is seeking opinions about the Covid-19 vaccine amongst various age groups. Instead of collecting feedback from 326,044,985 U.S citizens, random samples of around 10000 can be selected for research. These 10000 citizens can be divided into age groups of 18-29, 30-39, 40-49, 50-59, and 60 and above.

Stratified Sampling:  
Divide the population into homogenous strata (different age groups), then randomly sample from within each stratum (10000 in total).

### d.

A school principal tends to conduct a survey concerning personal hygiene. He has the list of all students in a table in Excel. He generates random numbers using Excel and asks corresponding students to answer the survey.

Simple Random Sampling (SRS): Numbers are generated randomly from all the students’ numbers

# 

# Q4

Many people believe that Telegram Messenger is superior to WhatsApp because it is much more user-friendly. Experimenters decided to test this claim. They take a sample of 52 social media users, each claiming to have experience (each person has a different level of expertise) with both applications. They give them a series of tasks to do on each application and record the amount of time it takes to complete the tasks in total.

### a.

Describe how you might design an experiment for this purpose.

During random sampling, I would choose a larger sample of people and use a stratified sampling based on people’s level of experience, age, and based on which app they use more. After the random assignment as described in b, I would give each group a series of tasks to do on the app they liked to use and record the amount of time. The rest is explained in b.

### b.

Does your experimental design use blocking? Explain why you did or did not include blocking in your design.

It is suspected that age and experience can affect the pace at which a person works with electronic devices. Therefore, during the random assignment, we need to use the blocking variable of age and level of expertise. So, we divide the samples into age groups and expertise groups (expertise self-report can be a source of bias). Also, which app the participant likes more can affect their performance, e.g., somebody who hates Whatsapp may intentionally work slowly on this platform. Consequently, we should have age groups, expertise, and preferred messenger equally represented in the resulting treatment (Telegram) and control (Whatsapp) groups.

# 

# Q5

## For each of the following parts, explain (use more than three words) the most concerning potential source of bias (If any).

### a.

Rasul wants to use Mahan Airlines to travel around the world, but he is worried about his life, so he decides to interview passengers who have used Mahan in the last month about the safety of Mahan planes. According to another survey, 90% of Mahan customers have a cell phone and 91% of them answer the unknown phone number. So, Rasul finds their cell phone number and asks them whether their travel with Mahan was safe or not. Finally, he finds that all the answers are ‘Yes, my travel was safe!’.

Anecdotal evidence:   
Those who had not a safe flight are likely to be dead :\ Rasul is naive only to ask those who have survived! He should compare the accidents rate to other airlines for a sound conclusion.

Non-response bias:   
9% don’t answer phone

Convenience sample bias:   
Rasul finds those who have a cellphone which is 90% (apart from just those who are alive).

### b.

A psychology professor wants to study the popularity of meat-based foods amongst undergraduate students at her university. She sends out a survey to everyone enrolled in Introduction to Psychology courses at her university. They all complete it in exchange for course credits

Convenience sample bias:  
She only asks Psychology students who may not be representative of the whole population (undergraduate students).

### c.

Coca-Cola company wants to make an advertisement video based on a real survey. So, the company employs a marketing person and asks him to do a survey about Coca-Cola’s popularity. The marketing person prepares two groups of cola cups. One is filled with Coca-Cola and the other with Pepsi-Cola. He labels Coca-Cola and Pepsi-Cola cups with ‘A’ and ‘B’, respectively. Then he takes a random sample of people and gives them a cola cup with the label ‘A’ and a cola cup with the label ‘B’ and asks them which cola cup they prefer.

With the given information, the sample is randomly sampled, and it can represent the population. The only problem is that it is better to have the Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group. Becase the company for sure will rig the experiment in its own favor as they may have done with other studies in the past.

### d.

A social scientist selects a random sample of girls walking in Naser-Khosro street in the morning, asking their name, last name, and if they have experienced sexual abuse. Of the 20 samples surveyed, 2 say ‘yes’ and the others say ‘no’. The scientist uses data and concludes that 90% of girls walking in Naser-Khosro in the morning have not experienced sexual abuse.

Convenience sample bias:  
Like the airplane example, those who have had a negative experience in this street may not continue to walk in that street. Therefore, the sample may be biased.  
Moreover, those who have had the experience might not be comfortable reporting it and lie.

### e.

According to StatCounter, more than 72% of social media traffic in Iran is dedicated to the Instagram app. Snapp company wants to gauge its popularity amongst Iranian people. So, the company decides to send a survey as a direct message to every Iranian Instagram user. The company also gifts participants a free travel. Reports show that 90% of participants are satisfied with Snapp.

Voluntary response bias:   
Only those who have a strong opinion or really want free travel will answer, and probably in favor of Sanpp as they would think if they vote against them, they might not consider them for the gift.

Convenience sample bias:  
Those who use Instagram may have specific traits and personalities such as liking to travel more; therefore, their response may not represent the population of Iran.

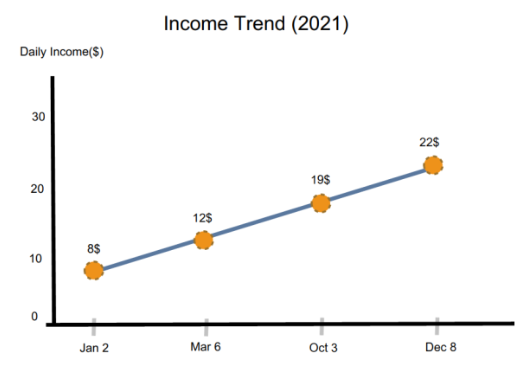
# 

# Q6

## Determine if the following statements are true or false based on the corresponding diagram. If false, explain your reasoning.

### a.

## An employee’s daily income increases linearly over time in 2021 (from Jun 2 to Dec 8).

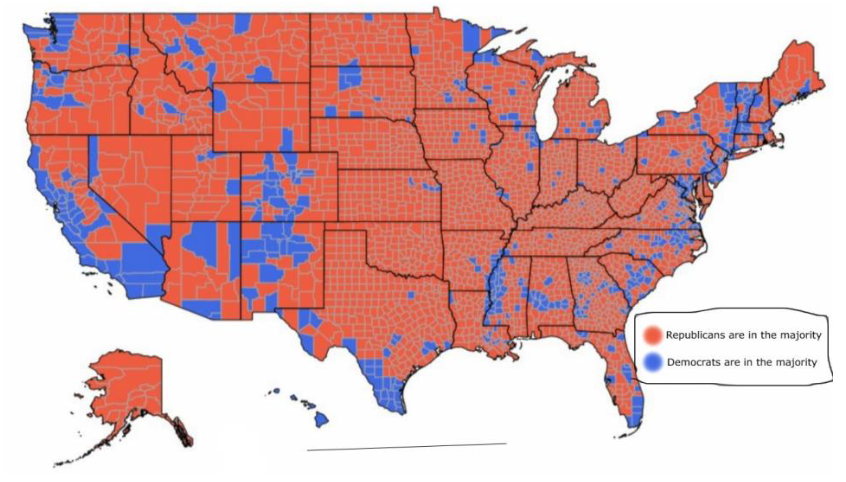


False:  
The x-axis where we have the time is not linearly increasing, rather increasing disproportionately.   
Jan2->Mar6 = **63 days**, Mar6->Oct3 = **211 days**, Oct3->Dec8 = **66 days**

Therefore, linear increase when the intervals of x-axis are not the same will not imply linear increase over time.

### b.

## The following map is the electoral vote map of America. According to this, most of the Americans vote for republicans.

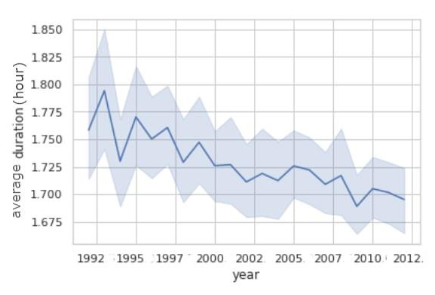


False:  
As we do not know the population of each region, we cannot conclude that most Americans vote for republicans. For instance, suppose that every place that Republicans are in the majority have only ten people, whereas Democrat regions have 1 million population each. Clearly, most Americans vote for Democrats in this case; nonetheless, the map will look the same.

### 

### c.

The average duration of IMDB movies decreases strongly over time.



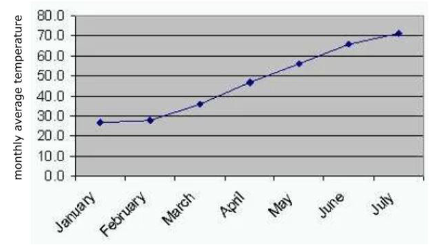
False:

Here we see plot abuse by using a non-zero baseline. The same plot, if had used zero as the baseline for the y-axis i.e., average duration, would show less of a decrease. This is a common way of persuading the audience and giving them the feeling that the change has been strong (Fox News); however, in a zero-baseline plot, the change is not as extreme.

### 

### d.

According to the monthly average temperature trend of the U.S., we must take the global warming problem seriously.

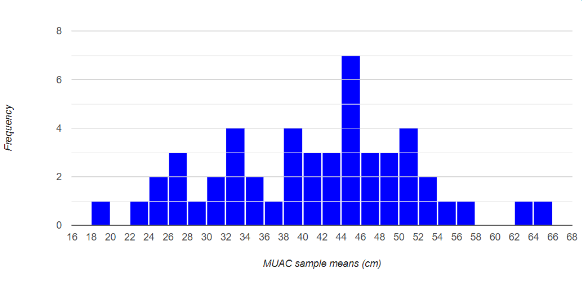


False,

The U.S. is in the northern hemisphere, and therefore, from December until March is winter and from March until June is spring, and afterward is summer. The warming trend is a yearly repetitive cycle that does not show the global warming problem by itself. We need the annualtemperature for at least the last decade to actually compare and see if the weather has got warmer in recent years or not.

# Q7

Researchers claim that the average mid-upper arm circumference (MUAC) of people who smoke cigarettes is 24cm. Rasul (that all of his friends smoke cigarettes) thinks researchers are wrong because his friends have much bigger arms. So, he selected a random sample of 15 persons from his friends and gauged their MUACs. The mean MUAC of his randomly selected friends was 37cm. He decided to perform a simulation to validate his findings. For this purpose, he repeatedly took a random sample of size 15 from smokers and gauged their MUACs. He did this 50 times and recorded the mean of the MUACs in each sample and prepared the following histogram:



### a.

Explain the appropriate null hypothesis and alternative hypothesis for his significance test in terms of words.

Null Hypothesis H0:  
MUAC is 24cm

Alternative Hypothesis:  
MUAC is larger than 24cm

Sample Result = 37cm

### b.

Based on the simulation results, what is the approximate p-value of the test?

P-value = P(p ≥ 37cm | H0 ture) = 33/50 = 0.66  
P-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. Here it means observing results as extreme as 37cm, which is MUAC≥37.

### c.

Interpret the p-value?

The p-value is 0.66 which is much larger than 0.05 (significance threshold). It means that 0.66 of test results are as extreme as the results given the null hypothesis. Since a p-value less than 0.05 (typically ≤ 0.05) is statistically significant, we cannot reject the null hypothesis.

# Q8 (R)

## Below are the final exam scores of twenty introductory statistics students.

57, 66, 72, 78, 79, 79, 81, 81, 82, 83, 84, 87, 88, 88, 89, 90, 91, 92, 94, 95

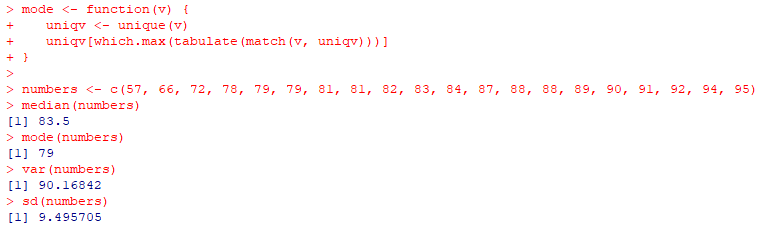
### a.

Create a vector of scores.

| numbers <- c(57, 66, 72, 78, 79, 79, 81, 81, 82, 83, 84, 87, 88, 88, 89, 90, 91, 92, 94, 95) |
| --- |

### b.

Calculate median, mode, variance, and standard deviation of scores.



### c.

Are there any outliers in any group? What are the exact values? Show the calculation of detecting outliers.

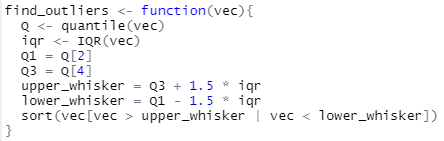
An outlier in a distribution is a number that **is more than 1.5 times the length of the box away from either the lower or upper quartiles**. Speciﬁcally, if a number is less than Q1 – 1.5×IQR or greater than Q3 + 1.5×IQR, then it is an outlier.



Low outliers = Q1 – 1.5(Q3 – Q1) = 79 - 1.5(89.25 - 79) = 63.625 15.375

High outliers = Q3 + 1.5(Q3 – Q1) = 89.25 + 1.5(89.25 - 79) = 104.625

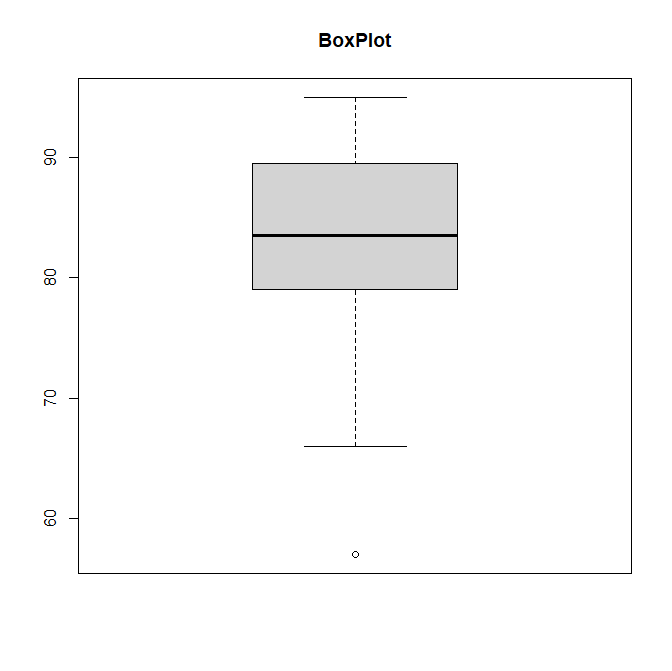
The only outlier out of 63.625 to 104.625 is 57.



### d.

Plot the boxplot.

| boxplot(numbers, main = "BoxPlot") |
| --- |



### e.

Plot the histogram and the density of scores in a single plot.

* Based on the plots, discuss the skewness of scores.

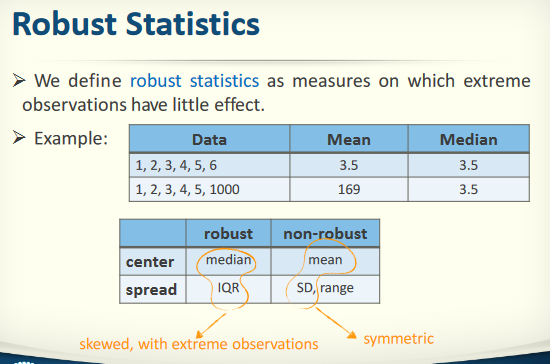
Left skewed (mean smaller than median)

* Based on the plots, would you expect the mean of this dataset to be smaller or larger than the median? Explain your reasoning.

As this is left-skewed, we expect the mean to be smaller than the median. We have a small portion of data with extreme values which affect the mean and makes it smaller, whereas the median is robust and stays larger than mean.

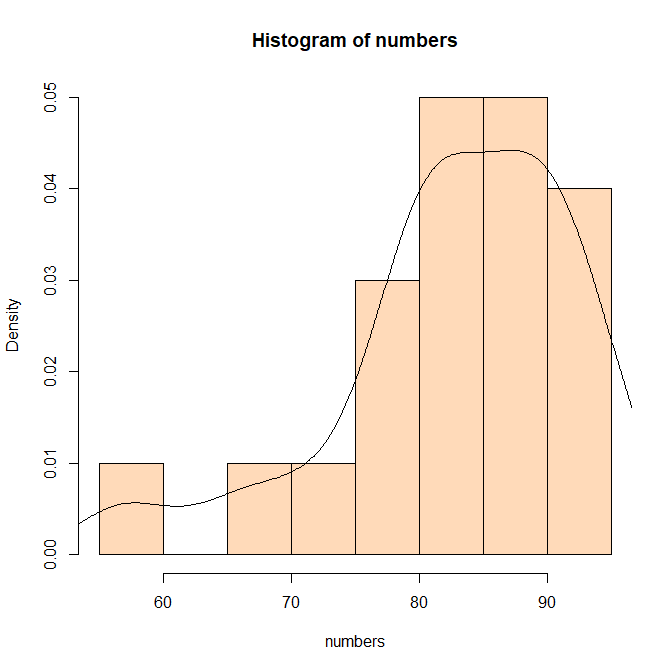
* What is the best measurement of the center for the scores? Why?

**The median** is the value in the center of the data. Half of the values are less than the median and half of the values are more than the median. It is probably the best measure of center to use in a skewed distribution because we have lots of data with small density that will change the mean.



As mentioned in the course, median for center and IQR for spread are robust in skewed with extreme observations, whereas mean and SD are better for symmetric distributions.

| hist(numbers, prob = TRUE, col="peachpuff") lines(density(numbers)) |
| --- |

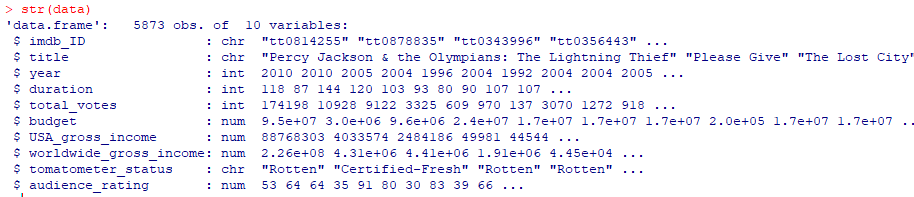


# Q9 (R)

## In this part, you are going to study and analyze the IMDB movies dataset. Your plots must have a proper title, x-label and y-label. Also, do not use any non-built-in R packages, e.g., ggplot2, etc. Follow the instructions:

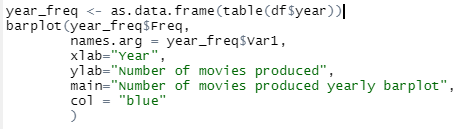
### a.

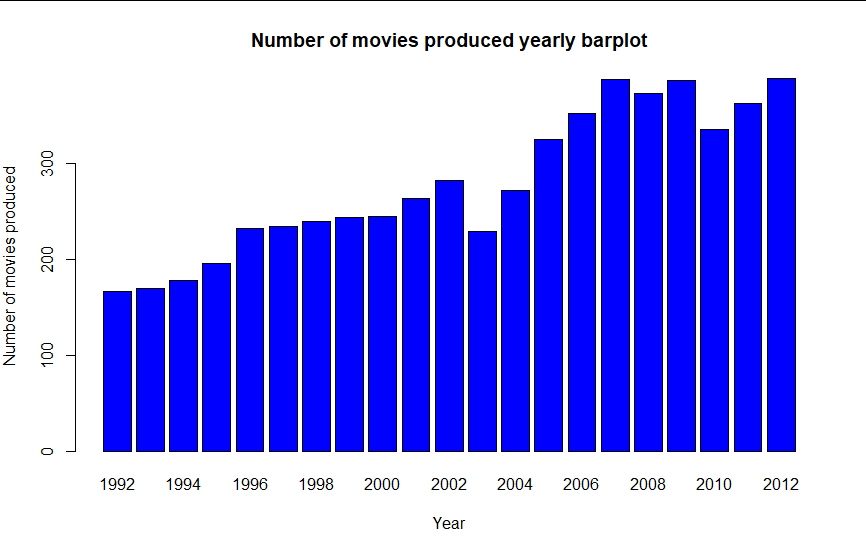
Identify the variables and their types.



### b.

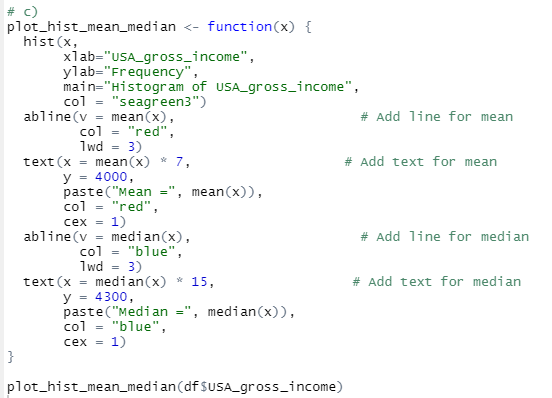
Use an appropriate diagram to visualize the number of movies produced yearly.

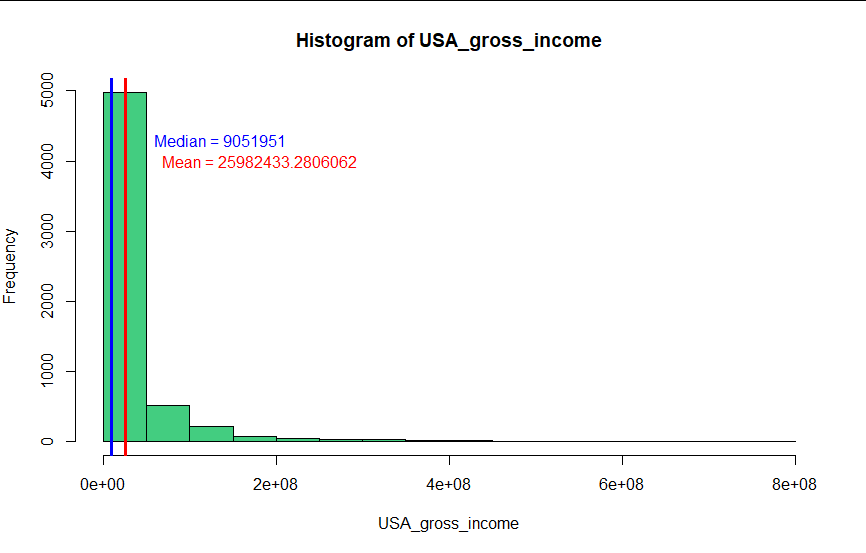




### c.

Plot a histogram of the distribution of ‘USA\_gross\_income’ and discuss its skewness.





As Mean is larger than Median, it is right skewed.

### 

### 

### 

### 

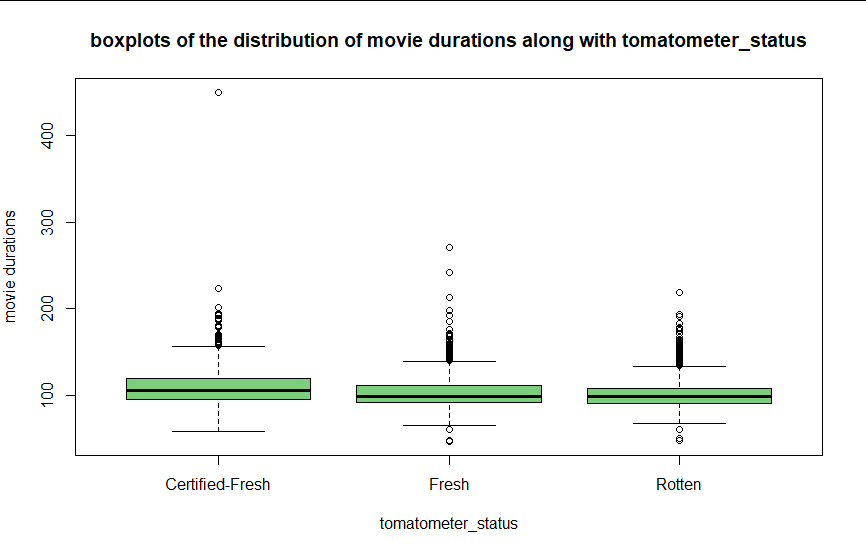
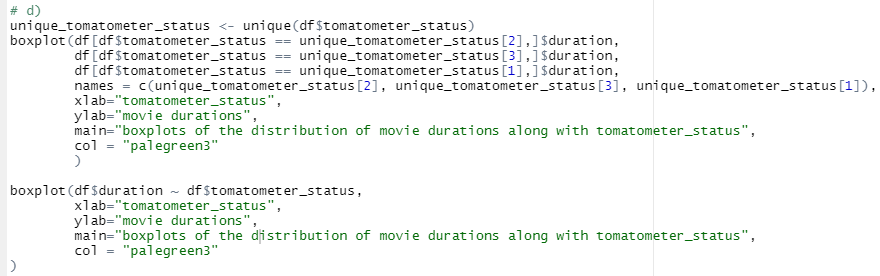
### 

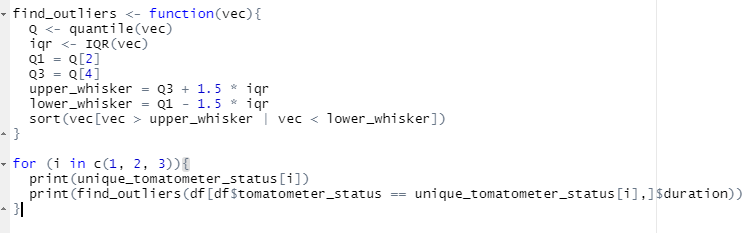
### 

### 

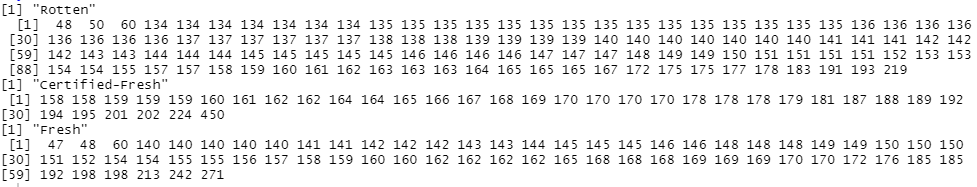
### d.

Use side-by-side boxplots to display the distribution of movie durations along with ‘tomatometer\_status’. Then identify outliers for each group.





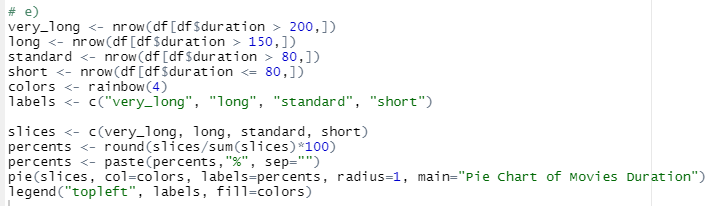
Outliers out of the lower and upper whiskers drawn with circles in boxplot:

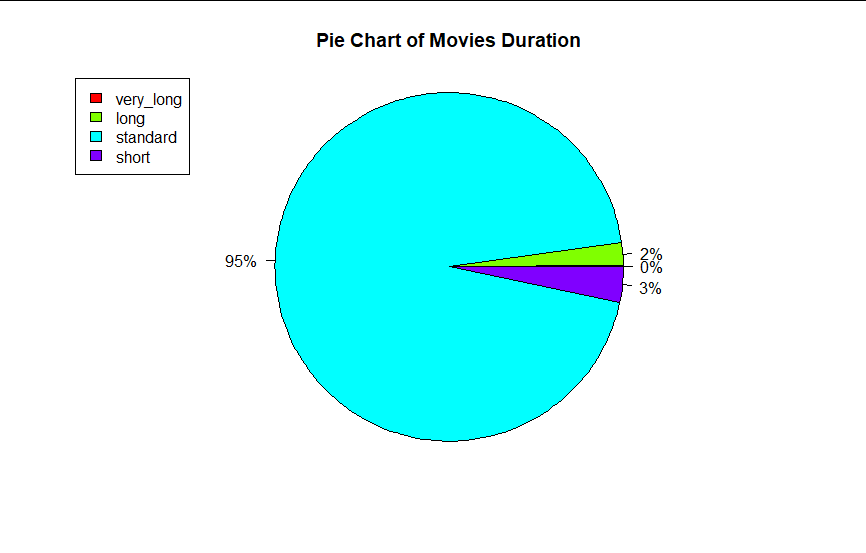


### 

### e.

Categorize all the movies based on their durations into 4 groups: “very long” (>200), “long” (>150), “standard” (>80), and “short” (<=80). Plot a pie chart that visualizes the frequency of these five categories. Each category must have a percentage and should have a unique color. Draw a legend for your pie chart.

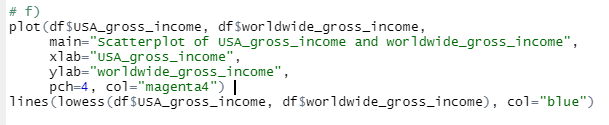


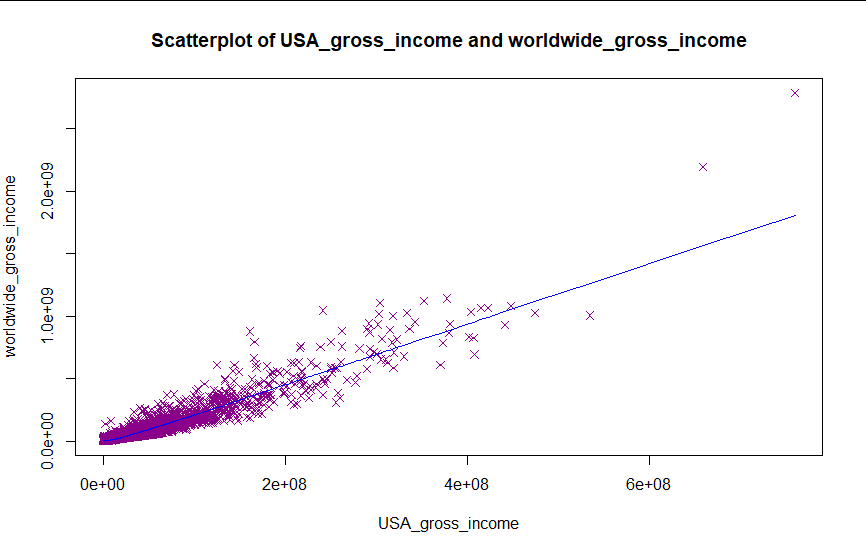


### 

### f.

Use a scatter plot to determine the relationship between ‘USA\_gross\_income’ and ‘worldwide\_gross\_income’. Interpret your plot.





It seems these two have a positive and high correlation (cor = 0.94) meaning that when one increases the other follows. However, we should consider that correlating does not imply causation.