Homework 2  
Statistical Inference

**PREPARED BY**

Mohsen Fayyaz - 810100524

Fall 2021

**فهرست مطالب**

[**Q1**](#_tbfpyld5fvy5) **3**

[**Q2**](#_qs3aqg4qlamu) **3**

[**Q3**](#_ubhlkrvsgyo6) **4**

[Benedict is a pizza salesperson. He sells on average 20 pizzas on a round of his route.](#_6glwsiwf9t12) 4

[**Q4**](#_xkg1vwut9iu1) **5**

[**Q5**](#_t70v3u4xcgr) **7**

[Answer the following questions.](#_qvas6oofstlw) 7

[**Q6**](#_3yqiguekvnfz) **9**

[A family has two children, each of the children is either a girl or boy. Find the probability of both children being boys in the following scenarios:](#_de7fjrurit7c) 9

[**Q7**](#_8g4pj24x3ruh) **11**

[Below is the structure of a 16-Team Tennis Tournament (Single Elimination) where “Federer” and “Nadal” both will participate.](#_2epnpfotptkv) 11

[**Q8**](#_qp5snrwvmhsd) **12**

[(R, ggplot2) In this question, you are going to use the “Foods” dataset. This dataset consists of some details about 1,700 foods. Note that you must use the ggplot2 library to draw the diagrams.](#_sm1be5p5uhir) 12

# Q1

Sixty students have lined up to enter the final exam of the Graph Theory course. Each of these 60 people has a card corresponding to a class chair. Unfortunately, the first person loses her card and then randomly chooses a chair and sits on it. Each of the subsequent people will sit on his/her seat if it is empty. Otherwise, they will select a vacant seat at random and sit on it. What is the probability that the last person to enter the class will sit in his/her chair?

As the individual persons are not of our interest here, we can think of when someone cannot sit in their designated seat and chooses another seat at random, as that person asks the first person to change her seat and she chooses a seat randomly. Therefore, in the end, everybody is sitting in their own seat, and the last time, the first person will have to choose between her own seat and the last person’s seat. So the probability that she chooses the correct seat is ½ which is the probability of the last person sitting in his/her chair.

# Q2

Gloria has become the project manager of a software company recently. She has decided to give each software developer a day off whenever it is the birthday of at least one of them. Except for these days, the developers work a 365-day year. Gloria wants to increase the productivity of her company. How many developers should she hire to reach this goal?

For any one person in a group of *n* people, the probability that he or she does NOT share his birthday with someone else is meaning others do not possess the day of his birthday.

We need to calculate the expectation of number of days that will not have any birthdays. The probability of the ith day of the year having no birthdays is that none of the n employees be born in that day which is .  
The number of “no birthday days” is , so the expected number of “no birthday days” is .  
Now the productivity for the company is the   
To maximize this value we take the derivative:  
  
  
Hence, for n about 364 and 365 she can have the most productive company.

# Q3

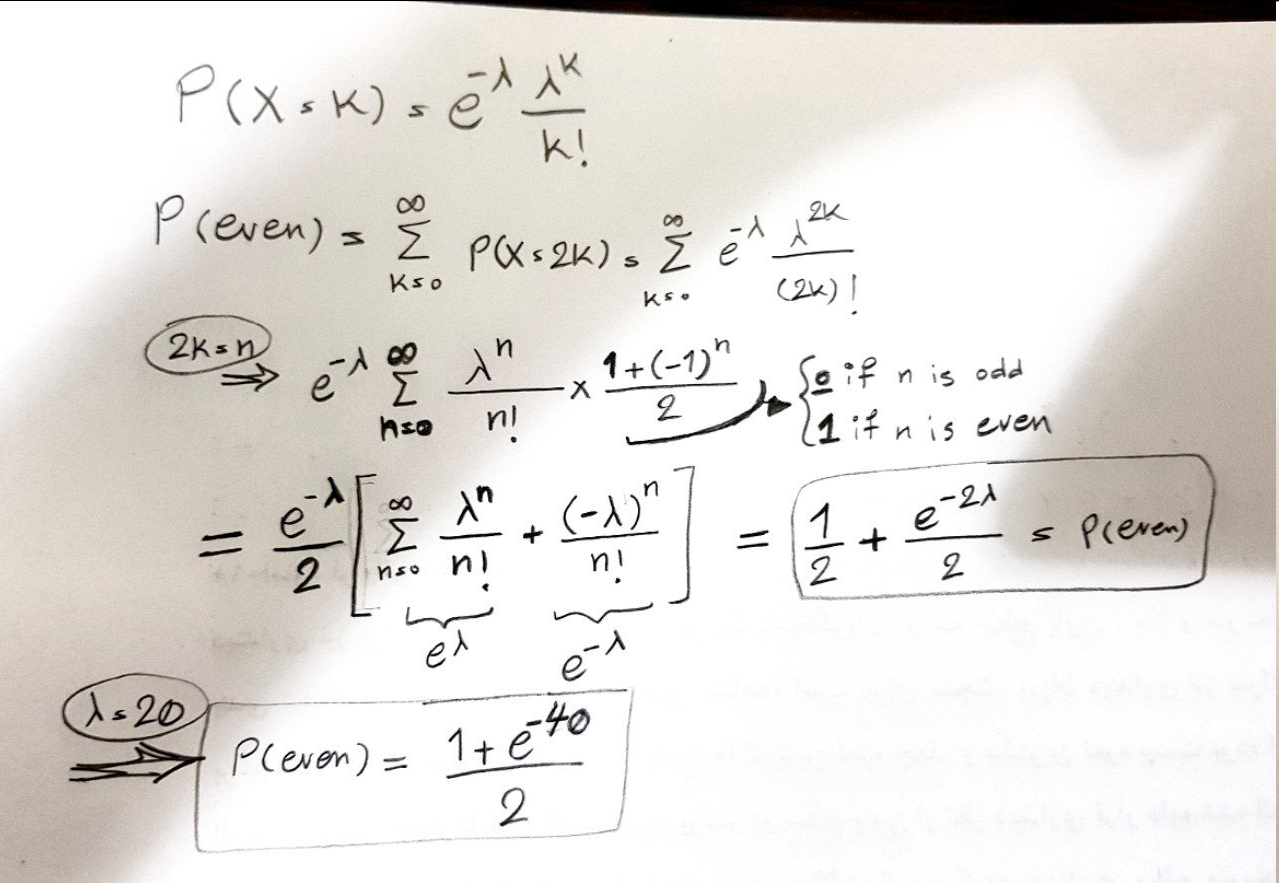
## Benedict is a pizza salesperson. He sells on average 20 pizzas on a round of his route.

* Explain why the distribution of the number of pizzas follows Poisson.

The Poisson distribution is used to model the number of events occurring within a given time interval. Here the number of sold pizzas is the event and the time interval is on a round of his route.

* What is the chance that he sells an even number of pizzas?

Now we consider 2k as n and 0 out the odd parts.



* (R) Do the calculations of part b in the form of an R script.



# Q4

There are N students currently enrolled in the Statistical Inference course. All students will vote on the Elearn website and they will either vote for date A or date B as potential exam dates. Ali has unofficially asked the students and realized that pN of the students support date A, and (1-p)N of the students will vote for date B, where N is known and 𝑝 ∈(0,1). Unfortunately, because the Elearn website has some technical problems (as always), each student will randomly and independently be kicked out of the website with a probability of 0.5, and therefore, will not vote. Let 𝑋𝐴 be the number of people who have successfully voted for date A and let 𝑋𝑏 be the number of people who have successfully voted for date B.

a. What are the **mean** and **standard deviation** of random variables 𝑋𝐴 and 𝑋𝐵 in terms of p and N?

Each student will vote for A with the Bernoulli distribution with a probability of and will vote for B or will not vote at all. The number of votes for a date is like the number of successes in N trials and is a binomial distribution that has a mean of np and variance of np(1-p).  
mean:   
  
Variance:

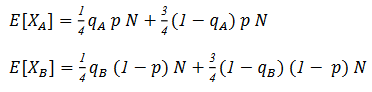
b. Show that when N is large enough, the fraction of students who voted for date A should be very close to p.

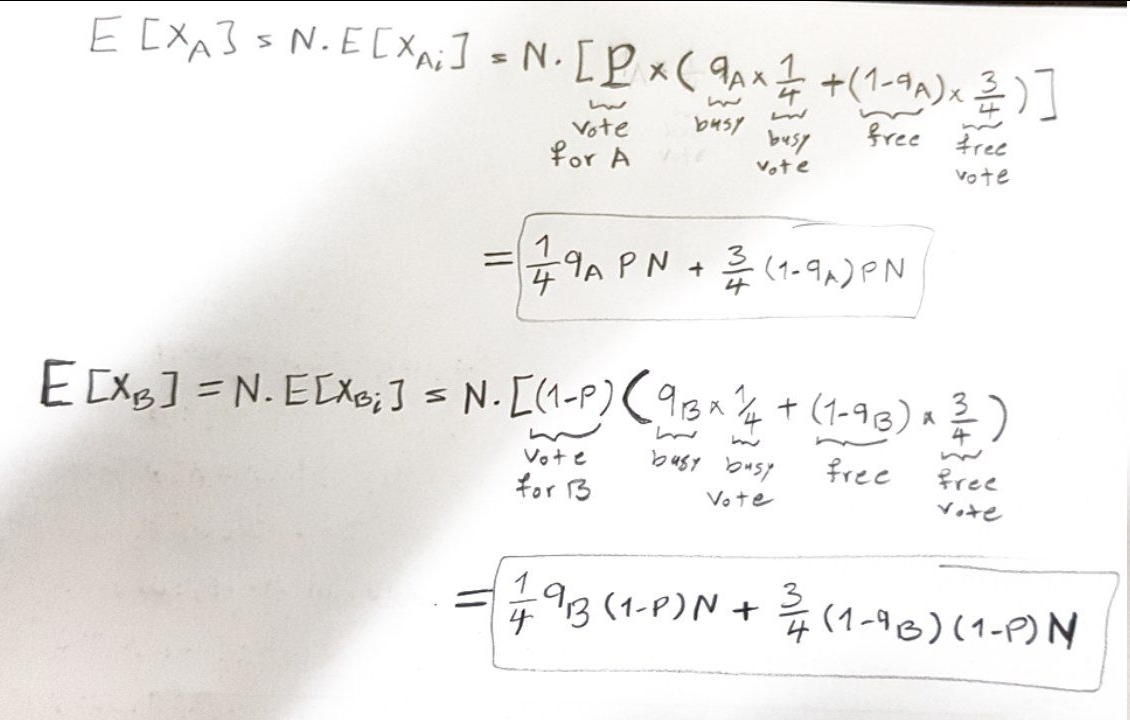
Fraction of students who voted for date A =

If N is large enough so that we get close to the expectation then the fraction becomes close to p.

Now suppose that we have fixed the website and want to repeat the vote. This time, some students are “busy” and will vote with probability 1/4 and will not participate with probability 3/4. Others are “free” and will vote with probability 3/4 and will not participate with probability 1/4. Suppose that a fraction 𝑞𝐴 of the supporters of date 𝐴 are busy and 1−𝑞𝐴 are free, and a fraction 𝑞𝐵 of the supporters of date 𝐵 are busy, and 1−𝑞𝐵 are free.

c. show that:





d. Does the statement in part b hold here as well?

Fraction of students who voted for date A =

No, the statement no longer holds here.

# 

# Q5

## Answer the following questions.

a. What is the difference between the “Independence” and “Disjointness” of random variables?

Two random processes are independent if knowing the outcome of one provides no useful information about the outcome of the other. Therefore if A and B are independent, then

Disjoint (mutually exclusive) events cannot happen at the same time. Therefore if A and B are disjoint then

b. Is it possible for two disjoint variables to be independent as well? If yes, explain when it happens.

Yes. For instance when the probability of one variable is zero (null) , then both and are true.

c. There are n children in a particular family. Find the n for which the events A and B are independent:  
A: “The family has children of both sexes.”  
B: “There is at most one boy.”

We should find the n for which .  
:

:

(BB, BG, GB, GG):

(BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG):  
  
  
 **n=3**

# Q6

## A family has two children, each of the children is either a girl or boy. Find the probability of both children being boys in the following scenarios:

a. We asked their mother, “Do you have at least one son?” and she answered, “Yes!”.

All possible outcomes are BB, BG, GB, GG.

b. We asked their mother, “Do you have at least one son named ‘Ali’ ?” and she answered “Yes!”. (Assume that if the family has a son, they name him “Ali” with probability 𝛼).

The question is not well-defined and we have to make assumptions such as✅the mother will not name both her sons Ali in BB case.  
Let “” be A.

For p(A|BB) We can consider   
(Here we assumed that the number of names is infinite and if the first B’s name is not Ali, will not change (increase) the probability of the second B’s name being Ali.)

✅If we consider 𝛼 as choosing Ali in a pool of n names as , then we have . Therefore

✅If we assume that the mother can name both her children Ali then we have:  
  
(0.5 when and when )

c. Are the answers for the two scenarios different? Explain why.

They are different. In the second case where we know that one of them is Ali, the two variables become conditionally dependant based on this information. The calculations are written in detail in b and we can see that the new information changes the conditional probabilities and therefore the final result.

# Q7

## Below is the structure of a 16-Team Tennis Tournament (Single Elimination) where “Federer” and “Nadal” both will participate.



a. What is the chance that Federer and Nadal will meet in a match during the tournament?

There are 15 matches in the tournament and all the possible matches between every pair of 16 players are .  
So the probability of any two players playing in the tournament games is

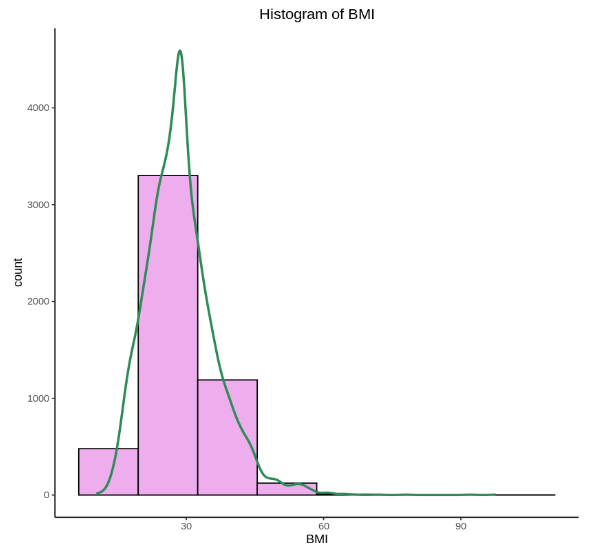
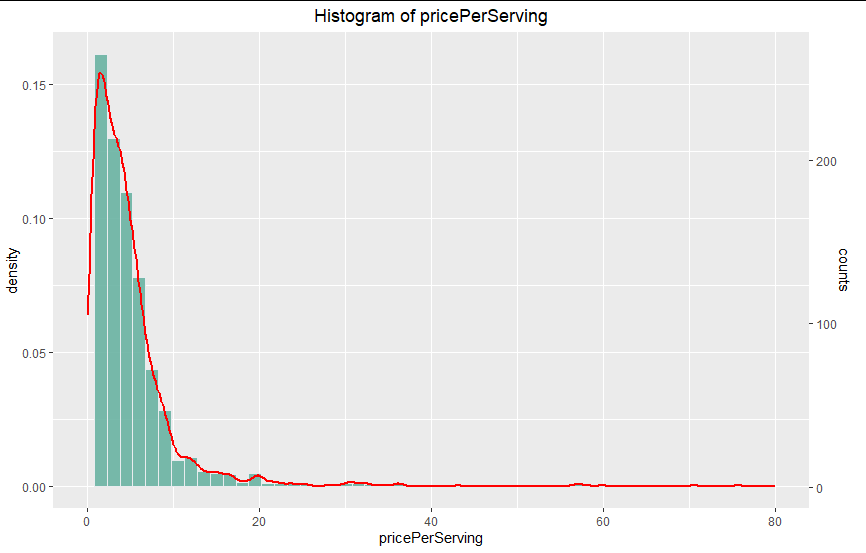
b. Solve the problem for a tournament with 2^𝑛 participants instead of 16 teams.

There are matches in the tournament and all the possible matches between every pair of players are .  
So the probability of any two players playing in the tournament games is

# Q8

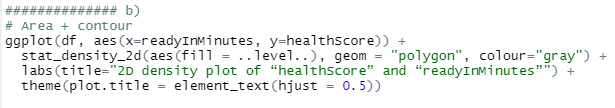
## (R, ggplot2) In this question, you are going to use the “Foods” dataset. This dataset consists of some details about 1,700 foods. Note that you must use the ggplot2 library to draw the diagrams.

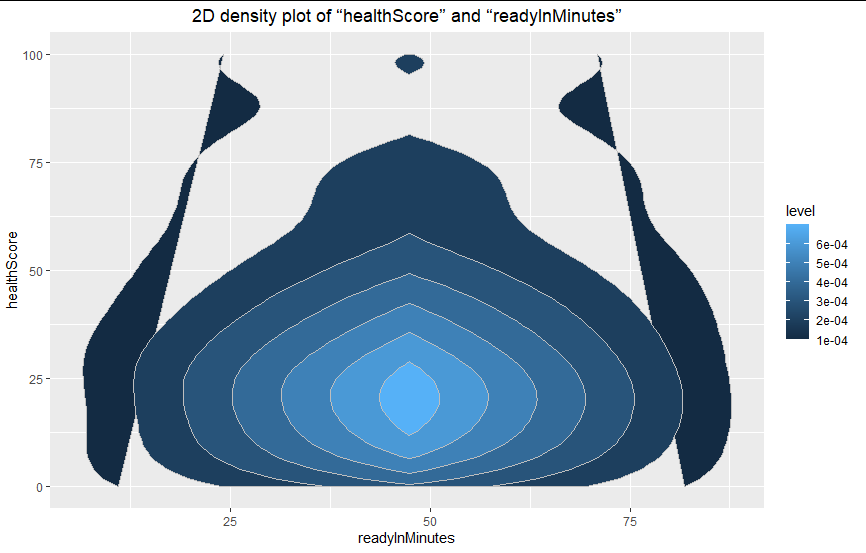
a. Plot the histogram for “pricePerServing” with an appropriate bin size, then overlay that with the density curve. Your diagram must be similar to the following figure.

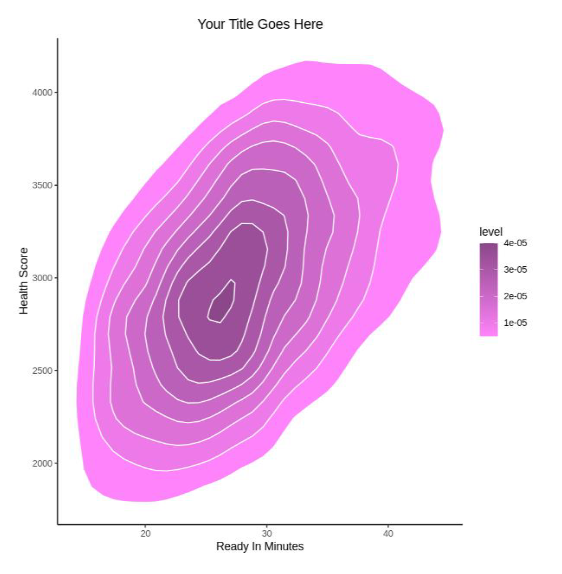


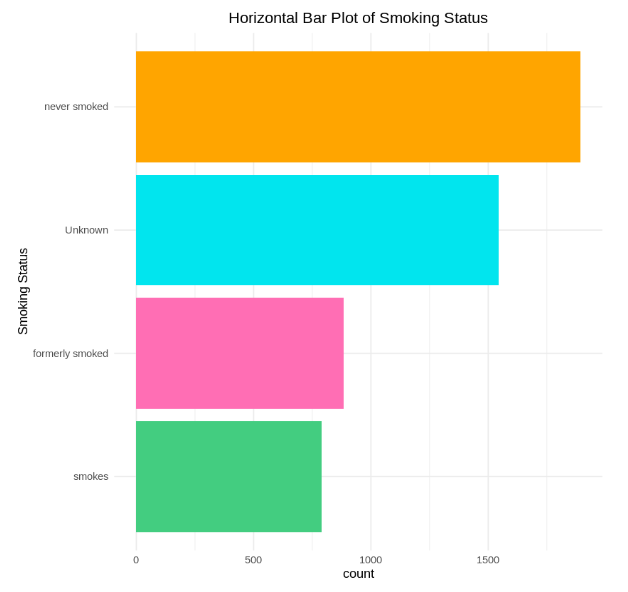


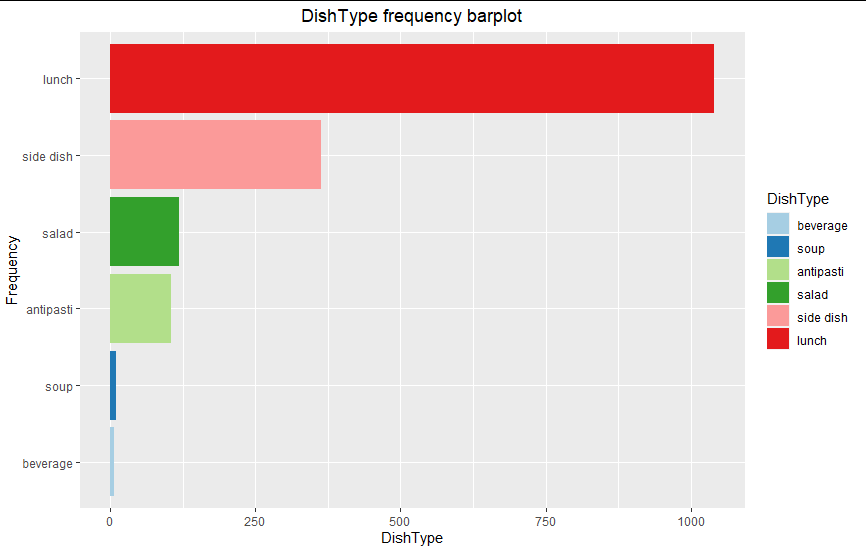
b. Draw the 2D density plot of “healthScore” and “readyInMinutes”. Your output must be similar to the following image.

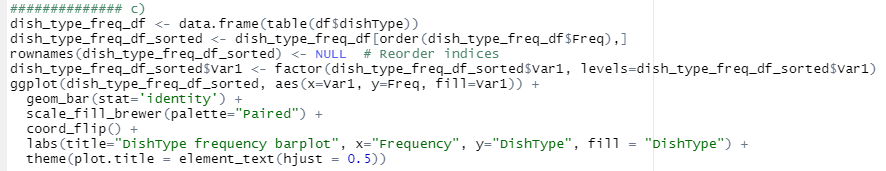


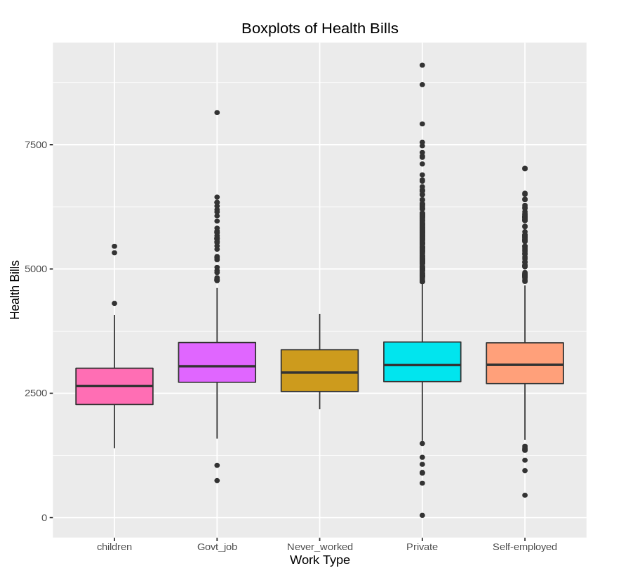


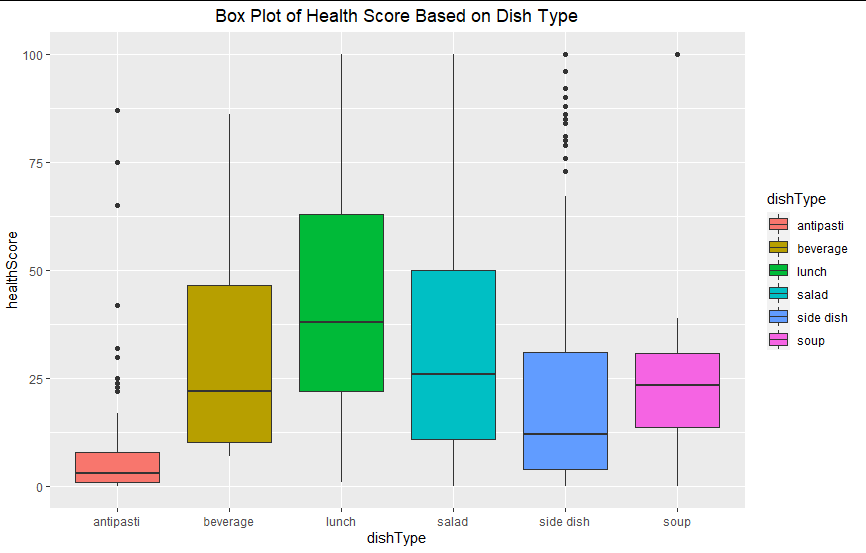


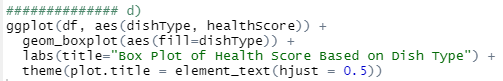
c. Sort the categories in “dishType” by their frequencies, then draw a horizontal barplot to show the result. Your output must be similar to the following image.





d. Draw the separate boxplots of the “healthScore” variable for each “dishType''. Your diagram must look like the following image.





e. Draw the mosaic plot of “veryHealthy” and “dairyFree”. Your output must be similar to the following image. Please pay attention to all of the details you can see in this figure.

