СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ”
Факултет по математика и информатика

Курсов проект на тема:
Здравословно хранене

Изготвила: Весела Илиянова Стоянова
Дисциплина: Статистика и емпирични методи
II курс, Информационни системи
Факултетен номер: 71949

София 2021

Съдържание:

1. Описание на данните:.....4

1.1. Основни въпроси, на които се отговаря чрез анализа на данни:.....4

1.2. Променливи, чрез които са предоставени данните.....5

1.3. Използвани статистически методи.....6

2. Едномерен анализ.....7

2.1. Засичане на outlier-и.....7

2.2. Определяне на локацията и разсейването на разпределението.....9

2.3. Графично представяне.....11

2.4. Определяне вида на разпределението.....20

3. Многомерен анализ.....22

3.1. Категорийна VS числова.....22

3.2. Числова VS числова.....25

3.2.1.Корелационен анализ.....25

3.2.2Линейна регресия.....29

4. Изводи.....32

Таблици и графики:

Фиг. 1: Таблица с потенциалния outlier QuantityFruits.....7

Фиг. 2: Таблица с потенциалния outlier QuantityVegetables.....8

Фиг. 3: Таблица с потенциалния outlier Calories.....8

Фиг. 4: Таблица с дескриптивните статистики за различните вектори за мъжете.....9

Фиг. 5: Таблица с дескриптивните статистики за различните вектори за жените.....9

Фиг. 6: Таблица с описателна статистика за центъра на разпределението за различните вектори за мъжете.....10

Фиг. 7: Таблица с описателна статистика за центъра на разпределението за различните вектори за жените.....10

Фиг. 8: Хистограма, показваща броя жени, които средно изяждат по съответната стойност калории на ден.....11

Фиг. 9: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност калории на ден.....11

Фиг. 10: Хистограма, показваща броя жени, които средно изпиват по съответната стойност вода на ден.....12

Фиг. 11: Хистограма, показваща броя мъже, които средно изпиват по съответната стойност вода на ден.....12

Фиг. 12: Хистограма, показваща броя жени, които средно изяждат по съответната стойност плодове на ден.....13

Фиг. 13: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност плодове на ден.....13

Фиг. 14: Хистограма, показваща броя жени, които средно изяждат по съответната стойност зеленчуци на ден 14
Фиг. 15: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност зеленчуци на ден 14
Фиг. 16: Хистограма, показваща броя жени, които средно изпиват по съответната стойност газирани напитки на ден 15
Фиг. 17: Хистограма, показваща броя мъже, които средно изпиват по съответната стойност газирани напитки на ден 15
Фиг. 18: Хистограма, показваща броя жени, които средно изяждат по съответната стойност сладкиши на ден 16
Фиг. 19: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност сладкиши на ден 16
Фиг. 20: Кръгова диаграма, показваща процентите от анкетираните хора, които обичат най-много съответния плод 17
Фиг. 21: Кръгова диаграма, показваща процентите от анкетираните хора, които обичат най-много съответния зеленчук 17
Фиг. 22: Хистограма, показваща най-често използваните храни от анкетираните хора 19
Фиг. 23: Кръгова диаграма, показваща процентите от анкетираните хора, които водят съответния тип начин на живот 20
Фиг. 24: Таблица с определяне на разпределенията на данните за мъжете 21
Фиг. 25: Таблица с определяне на разпределенията на данните за жените 21

1. Описание на данните:

За целите на настоящия проект са използвани данни от изследване, проведено чрез анкета. Анкетираните хора са 66 на брой като 38 са жени и 28 мъже. Изследването е проведено с цел да се разбере разликата между навиците в хранене на различни хора. Анализирани са резултатите от изследването.

1.1. Въвеждане на данните:

Използвам функцията:

```
healthyEatingData <- read.csv2(file = “HealthyEating1.csv”)
```

като чрез нея копирам цялата информация от анкетата.

Въведох всичките вектори чрез функциите:

```
Gender <- healthyEatingData[[“Gender”]]
Kcal <- healthyEatingData[[“Calories”]]
QuantityWater <- healthyEatingData[[QuantityWaterPerDay”]]
QuantityFruits <- healthyEatingData[[“QuantityFruitsPerDay”]]
QuantityVegetables <- healthyEatingData[[“QuantityVegetablesPerDay”]]
QuantitySweets <- healthyEatingData[[“QuantitySweetsPerDay”]]
QuantityCarbonatedDrinks <- healthyEatingData <-
healthyEatingData[[“QuantityCarbonatedDrinksPerDay”]]
Lifestyle <- healthyEatingData[[“Lifestyle”]]
FavFruit <- healthyEatingData[[“FavouriteFruit”]]
FavVegetable <- healthyEatingData[[“FavouriteVegetable”]]
UsedFood <- healthyEatingData[[“TheMostFood”]]
```

1.1. Основни въпроси, на които се отговаря чрез анализа на данни:

1. Съществува ли съществена статистическа разлика между количеството вода, което изпиват мъжете и жените?
2. Съществува ли съществена статистическа разлика между калориите, които изяждат мъжете и жените?
3. Съществува ли съществена статистическа разлика между количеството зеленчуци, които изяждат мъжете и жените?
4. Съществува ли съществена статистическа разлика между количеството плодове, които изяждат мъжете и жените?
5. Съществува ли съществена статистическа разлика между количеството газирани напитки, които изпиват мъжете и жените?
6. Съществува ли съществена статистическа разлика между количеството сладкиши, които изяждат мъжете и жените?

7. Каква е зависимостта между количеството плодове и количеството зеленчуци, които изяждат жените?
8. Каква е зависимостта между количеството плодове и количеството зеленчуци, които изяждат мъжете?
9. Каква е зависимостта между количеството вода, което изпиват жените и калориите, които изяждат жените?
10. Каква е зависимостта между количеството вода, което изпиват мъжете и калориите, които изяждат мъжете?
11. Каква е зависимостта между калориите и количеството плодове, които изяждат жените?
12. Каква е зависимостта между калориите и количеството плодове, които изяждат мъжете?
13. Каква е зависимостта между калориите и количеството зеленчуци, които изяждат жените?
14. Каква е зависимостта между калориите и количеството зеленчуци, които изяждат мъжете?
15. Каква е зависимостта между калориите и количеството сладкиши, които изяждат жените?
16. Каква е зависимостта между калориите и количеството сладкиши, които изяждат мъжете?
17. Каква е зависимостта между калориите и количеството газирани напитки, което изпиват жените?
18. Каква е зависимостта между калориите и количеството газирани напитки, което изпиват мъжете?
19. Каква е зависимостта между количеството газирани напитки, което изпиват жените и количеството сладкиши, което изяждат жените?
20. Каква е зависимостта между количеството газирани напитки, което изпиват мъжете и количеството сладкиши, което изяждат мъжете?
21. Каква е зависимостта между количеството плодове и количеството сладкиши, което изяждат жените?
22. Каква е зависимостта между количеството плодове и количеството сладкиши, което изяждат мъжете?

1.2. Променливи, чрез които са предоставени данните

- Gender – от категориен тип, обозначени с f(female) и m(male);
- QuantityWater – от числов тип;
- Kcal – от числов тип;
- QuantityVegetables – от числов тип;
- QuantityFruits – от числов тип;

- QuantityCarbonatedDrinks – от числов тип;
- QuantitySweets – от числов тип;
- Lifestyle – от тип стринг;
- FavFruit – от тип стринг;
- FavVegetable – от тип стринг;
- UsedFood – от тип стринг;

1.3. Използвани статистически методи

- Описателна (дескриптивна) статистика, включваща изчисляване на средна стойност, медиана, мода, стандартно отклонение и вариация(дисперсия);
- Имплементирана е допълнителна функция за изчисляване на модата:

```
getMode <- function(values){
    uniqValues <- unique(values)
    uniqValues[which.max(tablulate(match(values, uniqValues)))]
}
```

- Представяне на данните графично посредством хистограми;
- Определяне типа на всяко едно от изследваните разпределения на данните с цел последващ избор на статистически тест за сравняване на данните. Приложен е тест на Shapiro-Wilcoxon с нулева хипотеза H0 „Разпределението е нормално“ и алтернативна хипотеза „Разпределението не е нормално“ с равнище на значимост p = 0,05;
- Поставените задачи за изследване изискват сравнение на независими извадки. В случай на нормалност на двете сравнявани разпределения използваме едностранен t-test за сравнение на средните стойности. В случай, че поне едно от разпределенията на сравняваните извадки не е нормално, използваме непараметричния тест на Mann-Whitney-Wilcoxon;
- Приложен е корелационен анализ за установяване на зависимостта между различните данни при мъжете и при жените.

2. Едномерен анализ

За целите на изследването данните за разделени в няколко отделни dataframe-ове, както следва:

- allMales – данните за всички мъже;
- allFemales – данните за всички жени;

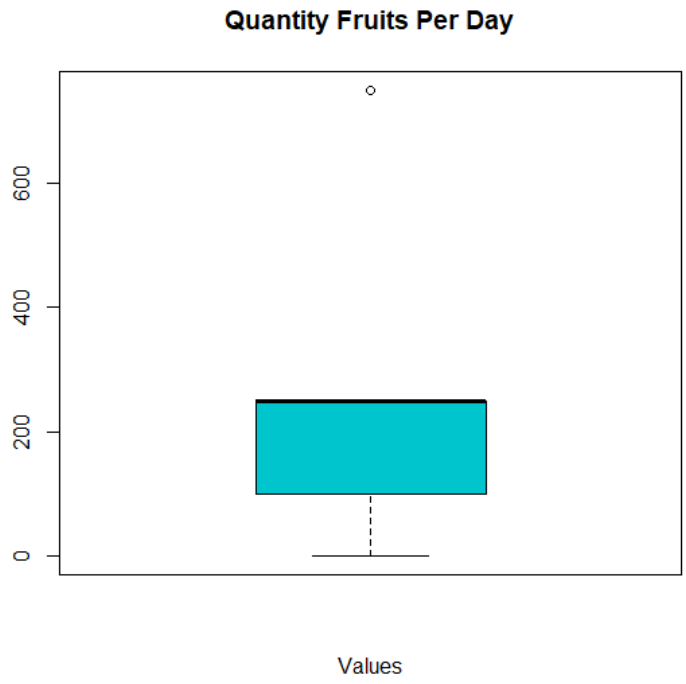
С цел оптимизация и избягване на повтаряне на код в скрипта е обособена функция, която пресмята и отпечатва петте характеристики

```
printDescriptiveStatistics <- function(values){
        print (mean (values))
        print (median (values))
        print (getMode (values))
        print (sd (values))
        print (var(values))
    }
```

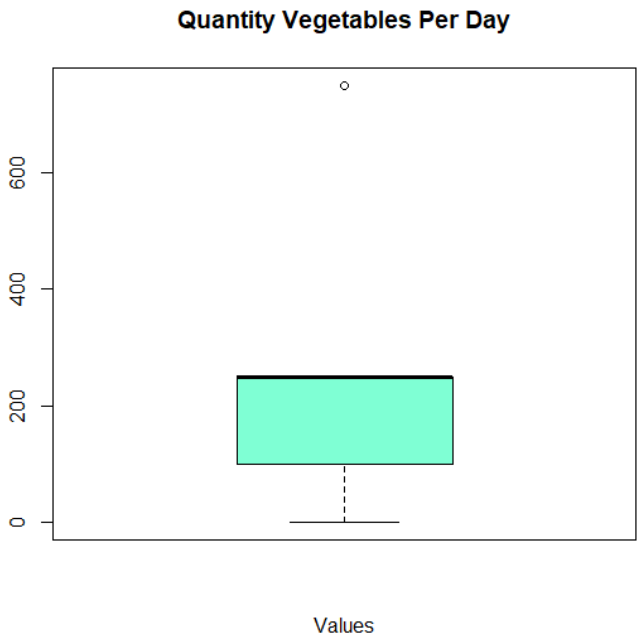
2.1. Засичане на outlier-и

Използвам boxplot за откриване на потенциални outlier-и.

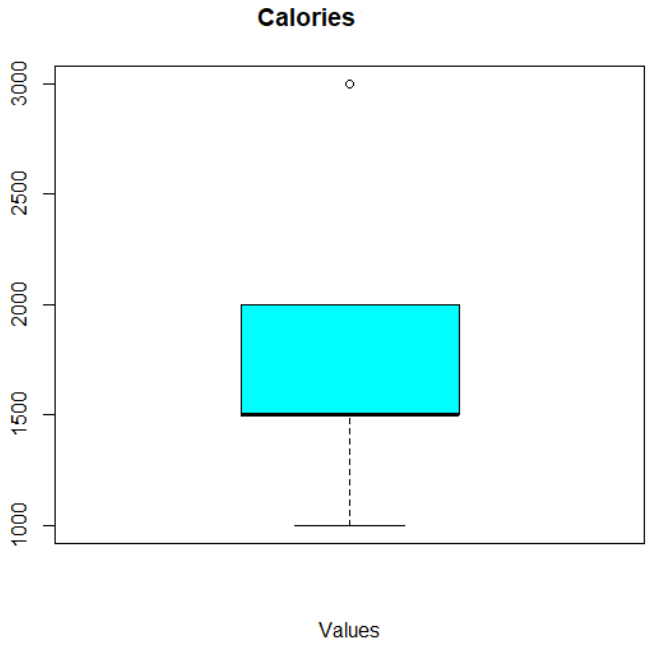
Открих 3 потенциални outlier-и:



Фиг. 1: Таблица с потенциалния outlier QuantityFruits



Фиг. 2: Таблица с потенциалния outlier QuantityVegetables



Фиг. 3: Таблица с потенциалния outlier Calories

2.2. Определяне на локацията и разсейването на разпределението

- В следващите таблици са показани резултатите от прилагането на функцията **printDescriptiveStatistics** за различните вектори:

Извадка	Средна стойност	Медиана	Мода	Стандартно отклонение	Дисперсия (вариация)
allMales$Kcal	2142.857	2000	2000	650.6	423280.4
allMales$QuantityWater	2250	2000	3000	799.3053	638888.9
allMales$QuantityCarbonated Drinks	678.5714	750	750	556.3486	309523.8
allMales$QuantityFruits	171.4286	100	100	143.6486	20634.92
allMales$QuantityVegetables	187.5	100	100	177.7561	31597.22
allMales$QuantitySweets	271.4286	300	300	207.0197	42857.14

Фиг. 4: Таблица с дескриптивните статистики за различните вектори за мъжете

Извадка	Средна стойност	Медиана	Мода	Стандартно отклонение	Дисперсия (вариация)
allFemales$Kcal	1684.211	1500	1500	409.6975	167852.1
allFemales$QuantityWater	1342.105	1000	1000	717.5904	514936
allFemales$Quantity CarbonatedDrinks	230.2632	0	0	355.3718	126289.1
allFemales$QuantityFruits	285.5263	250	250	214.321	45933.5
allFemales$Quantity Vegetables	253.9474	250	250	189.3736	35862.38
allFemales$QuantitySweets	173.6842	100	300	138.884	19288.76

Фиг. 5: Таблица с дескриптивните статистики за различните вектори за жените

Прилагам и **summary** – описателна статистика за центъра на разпределението. Тя дава информация за минимална стойност, 1-вия квартил, 2-рия квартил (медианата), 3-тия квартил и максималната стойност. В следващите таблици са показани резултатите от прилагането на тази функция:

Извадка	Минимална стойност	Стойност 1-ви квартил	Стойност 2-ри квартил	Стойност 3-ти квартил	Максимална стойност
allMales$Kcal	1000	1500	2000	3000	3000
allMales$Quantity Water	1000	2000	2000	3000	3000
allMales$Quantity CarbonatedDrinks	0	250	750	937.5	1500
allMales$Quantity Fruits	0	100	100	250	750
allMales$Quantity Vegetables	0	100	100	250	750
allMales$Quantity Sweets	0	100	300	300	600

Фиг. 6: Таблица с описателна статистика за центъра на разпределението за различните вектори за мъжете

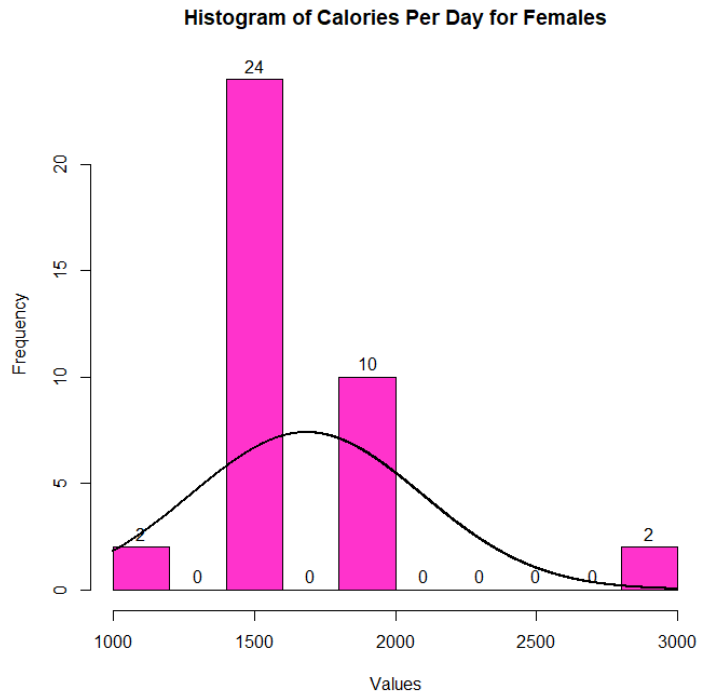
Извадка	Минимална стойност	Стойност 1-ви квартил	Стойност 2-ри квартил	Стойност 3-ти квартил	Максимална стойност
allFemales$Kcal	1000	1500	1500	2000	3000
allFemales$Quantity Water	500	1000	1000	2000	3000
allFemales$Quantity CarbonatedDrinks	0	0	0	250	1500
allFemales$Quantity Fruits	100	100	250	250	750
allFemales$Quantity Vegetables	0	100	250	250	750
allFemales$Quantity Sweets	0	100	100	300	600

Фиг. 7: Таблица с описателна статистика за центъра на разпределението за различните вектори за жените

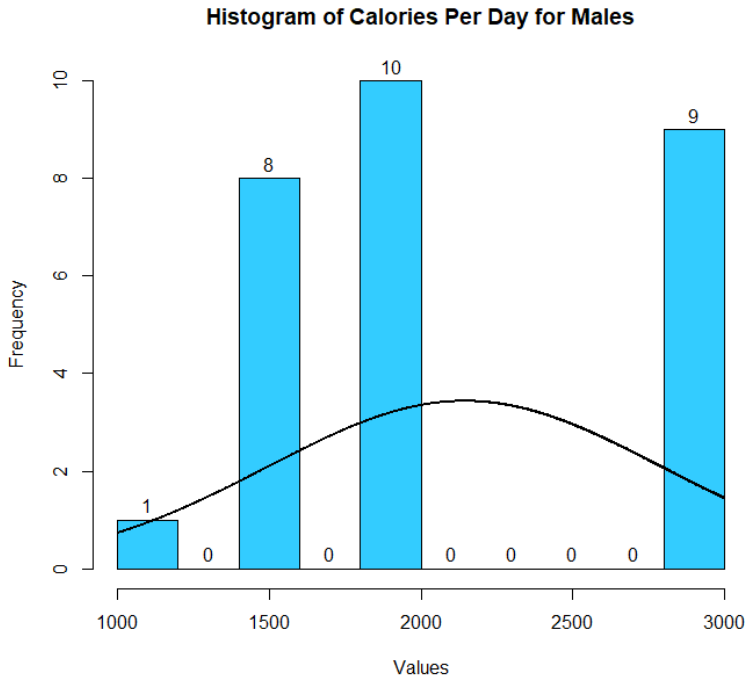
От фигура 1 и фигура 3 става ясно, че мъжете средно изяждат по 2142.857 калории на ден. Модата е 2000 – най-често срещаната стойност във вектора. Следователно най-много мъже изяждат по 2000 калории на ден. Стандартното отклонение е оценка на вариацията, която показва колко далече са наблюденията от очакването. За разлика от обхвата, взема под внимание всички наблюдения. Стандартното отклонение е производно на вариацията и представлява корен квадратен от дисперсията.

2.3. Графично представяне

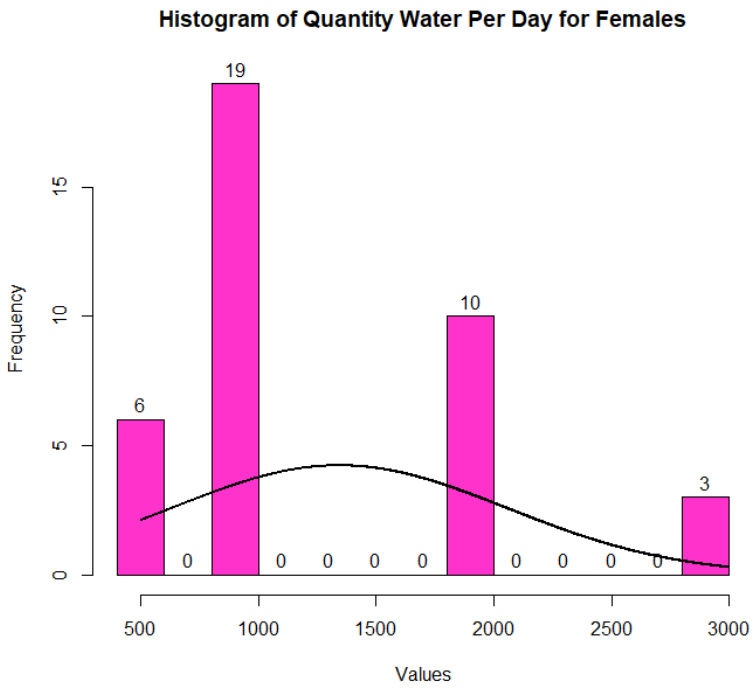
Представен е набор от хистограми и кръгови диаграми, които онагледяват данните от изследването.



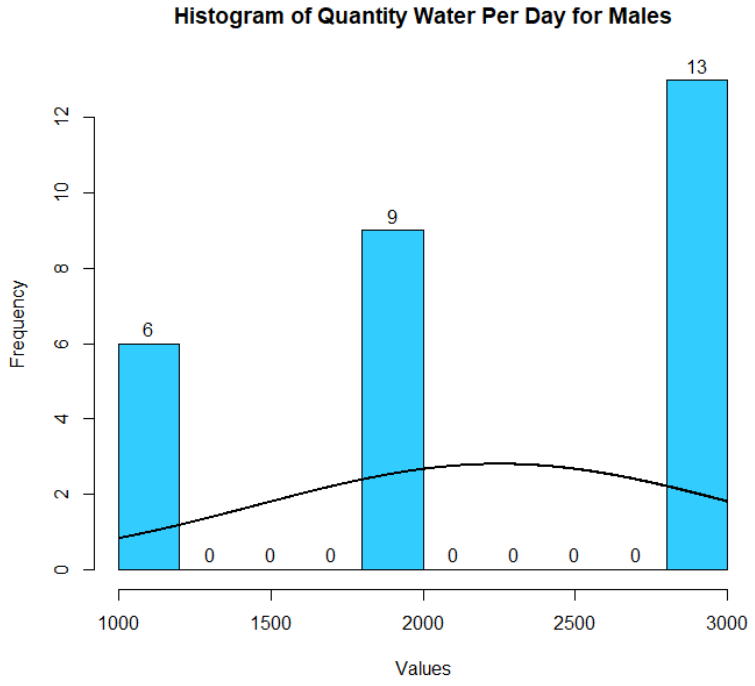
Фиг. 8: Хистограма, показваща броя жени, които средно изяждат по съответната стойност калории на ден



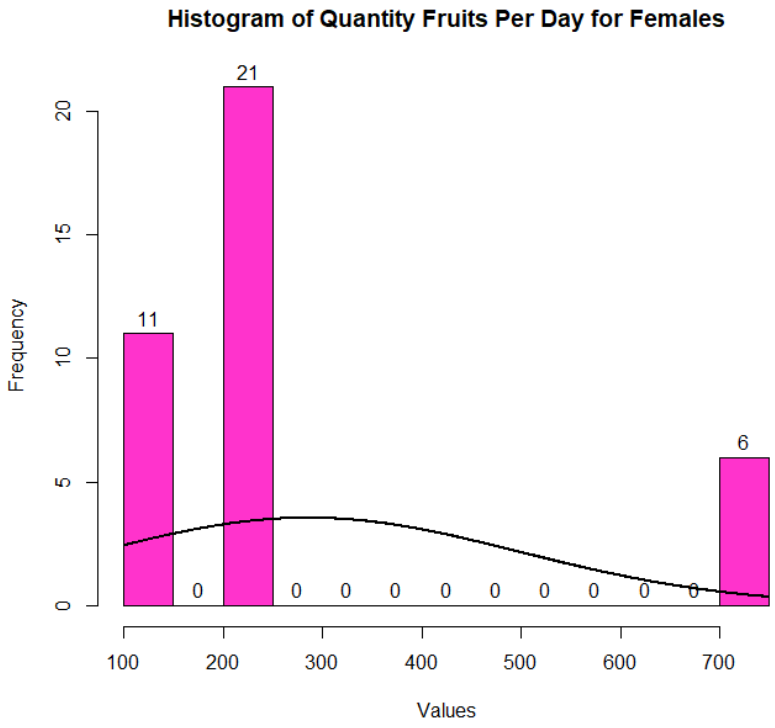
Фиг. 9: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност калории на ден



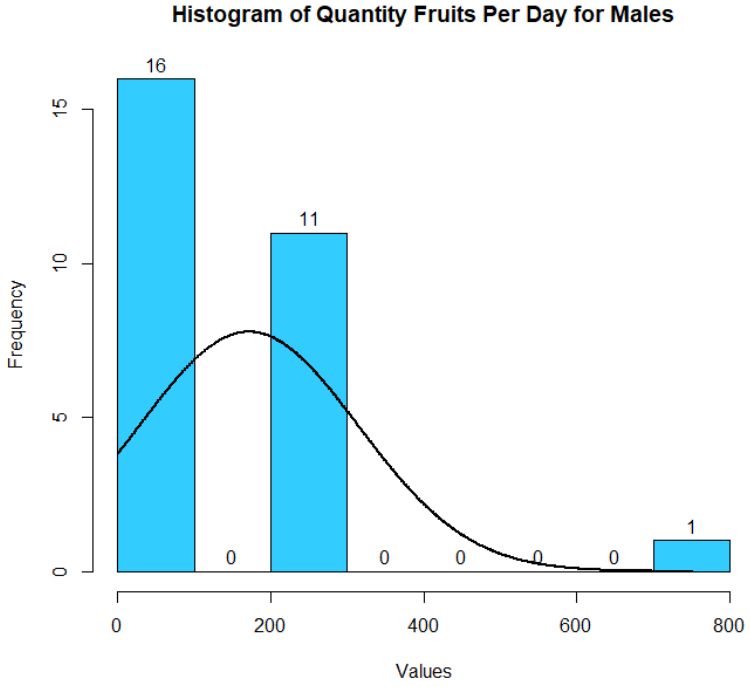
Фиг. 10: Хистограма, показваща броя жени, които средно изпиват по съответната стойност вода на ден



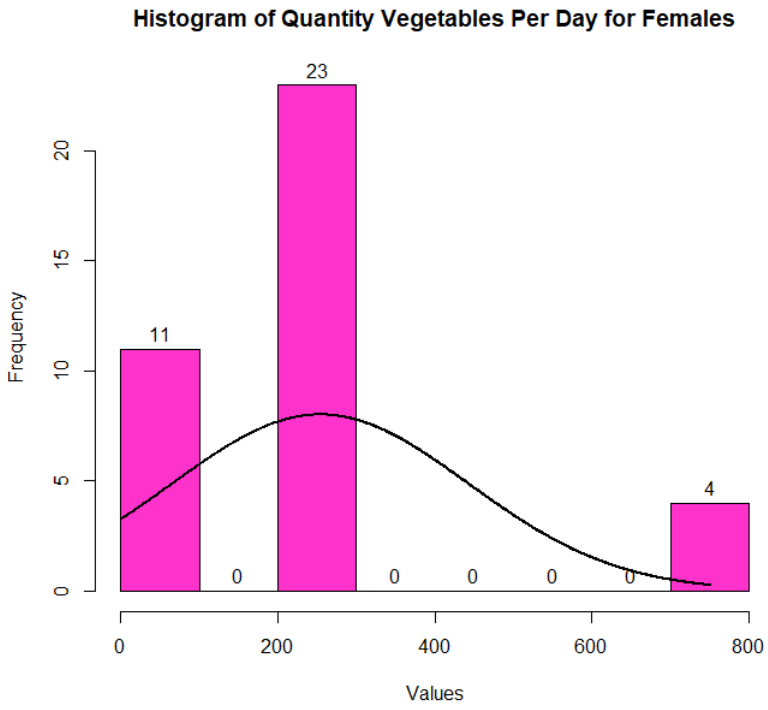
Фиг. 11: Хистограма, показваща броя мъже, които средно изпиват по съответната стойност вода на ден



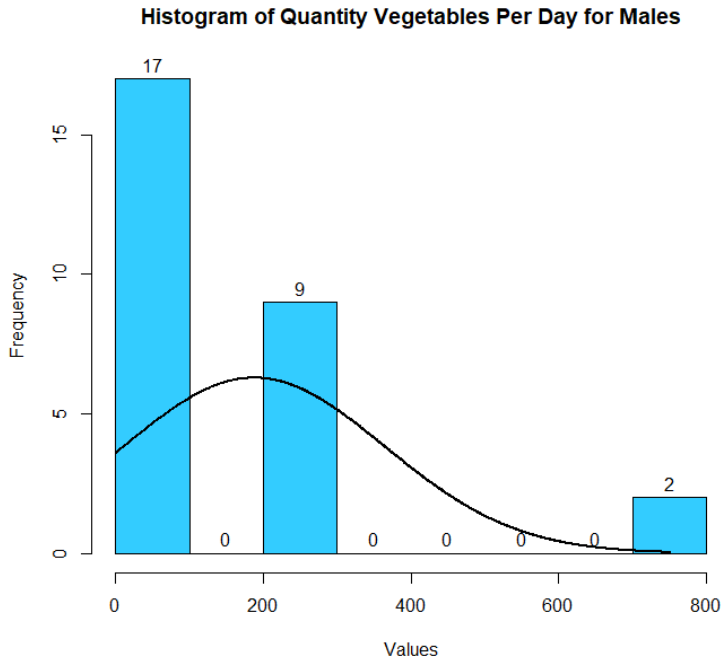
Фиг. 12: Хистограма, показваща броя жени, които средно изяждат по съответната стойност плодове на ден



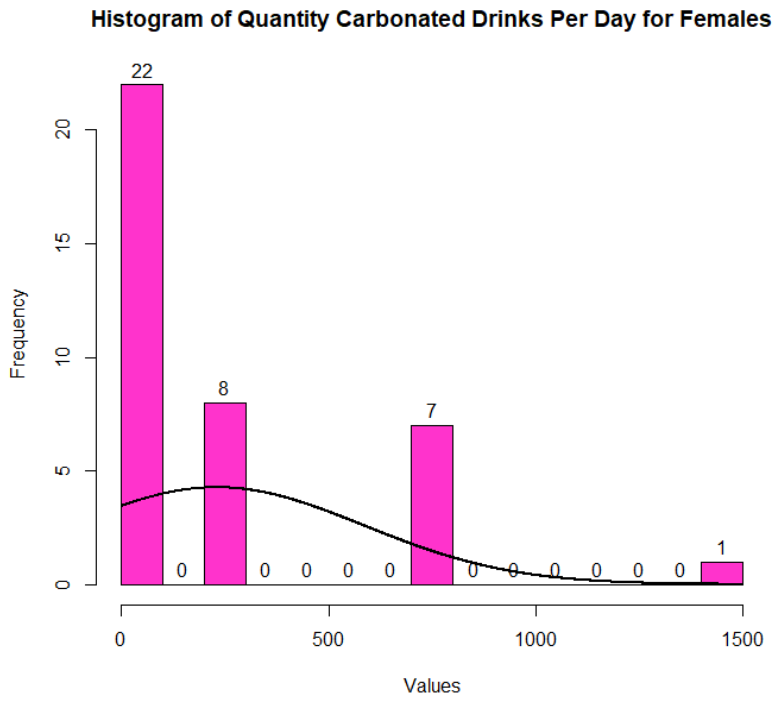
Фиг. 13: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност плодове на ден



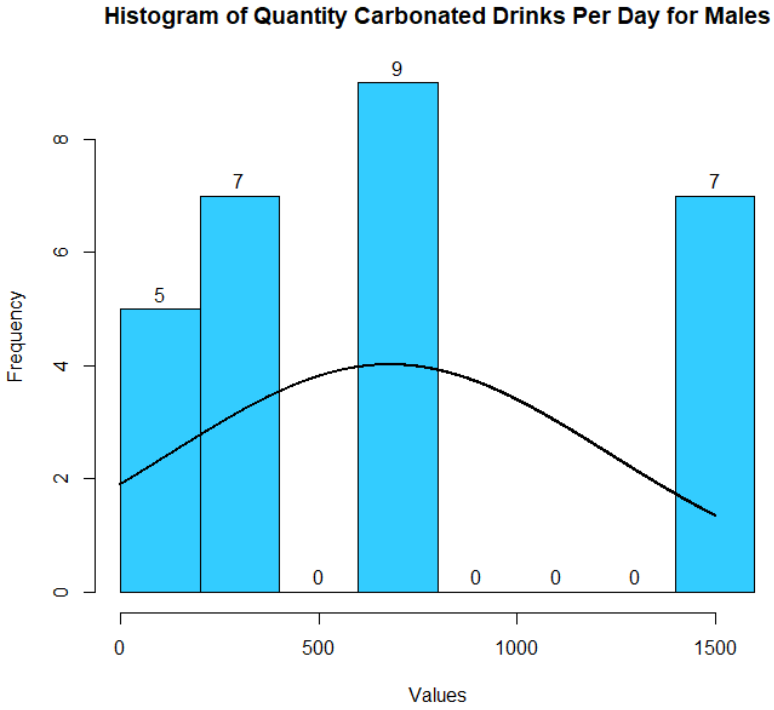
Фиг. 14: Хистограма, показваща броя жени, които средно изяждат по съответната стойност зеленчуци на ден



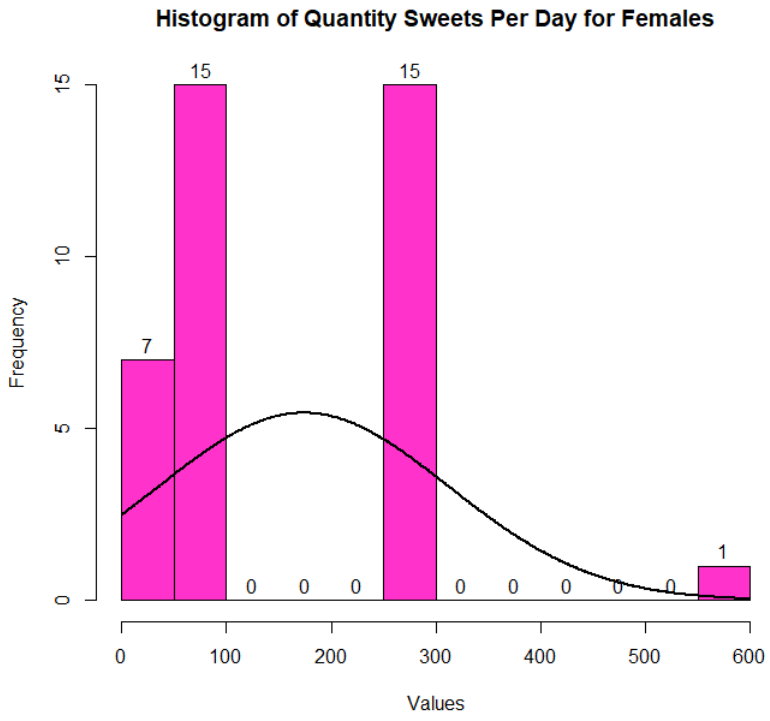
Фиг. 15: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност зеленчуци на ден



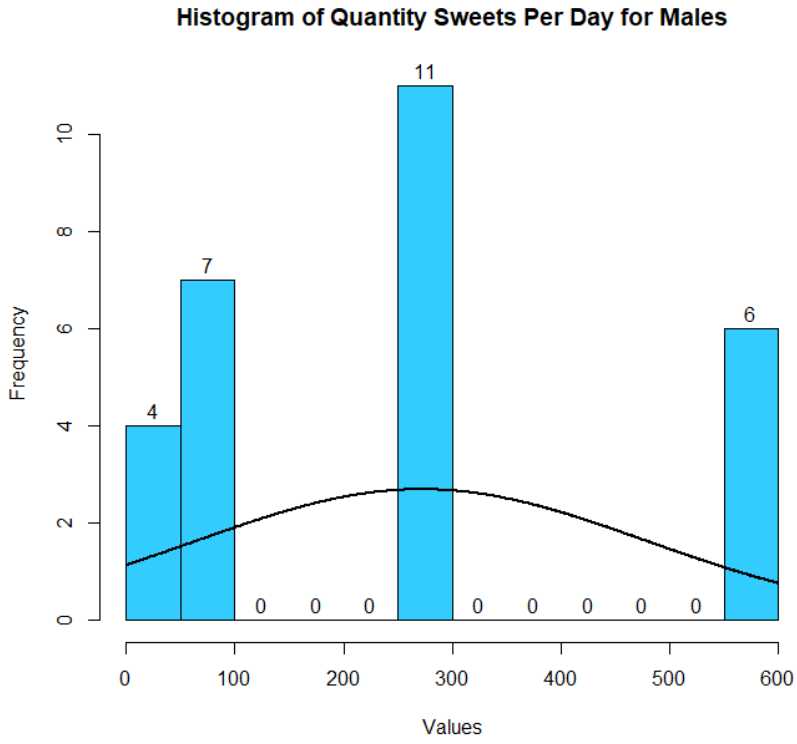
Фиг. 16: Хистограма, показваща броя жени, които средно изпиват по съответната стойност газирани напитки на ден



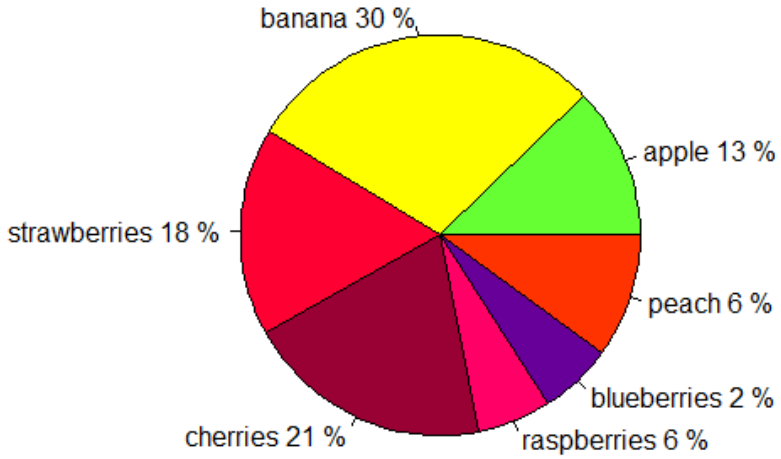
Фиг. 17: Хистограма, показваща броя мъже, които средно изпиват по съответната стойност газирани напитки на ден



Фиг. 18: Хистограма, показваща броя жени, които средно изяждат по съответната стойност сладкиши на ден

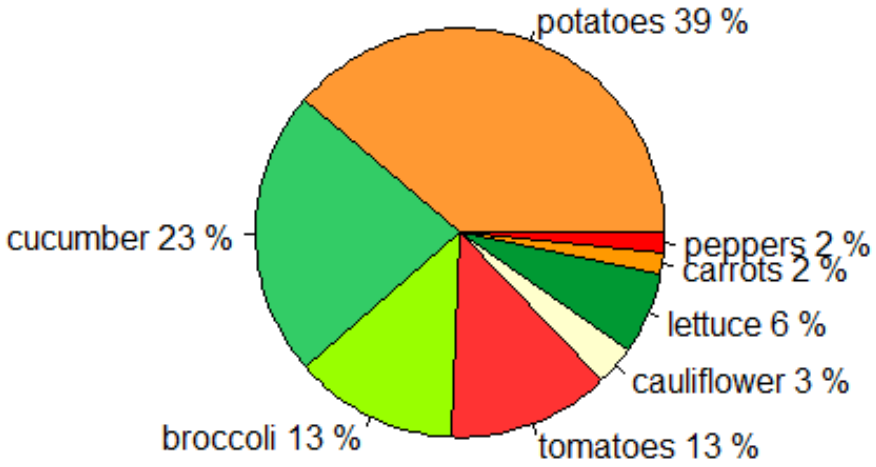


Фиг. 19: Хистограма, показваща броя мъже, които средно изяждат по съответната стойност сладкиши на ден



Фиг. 20: Кръгова диаграма, показваща процентите от анкетираните хора, които обичат най-много съответния плод

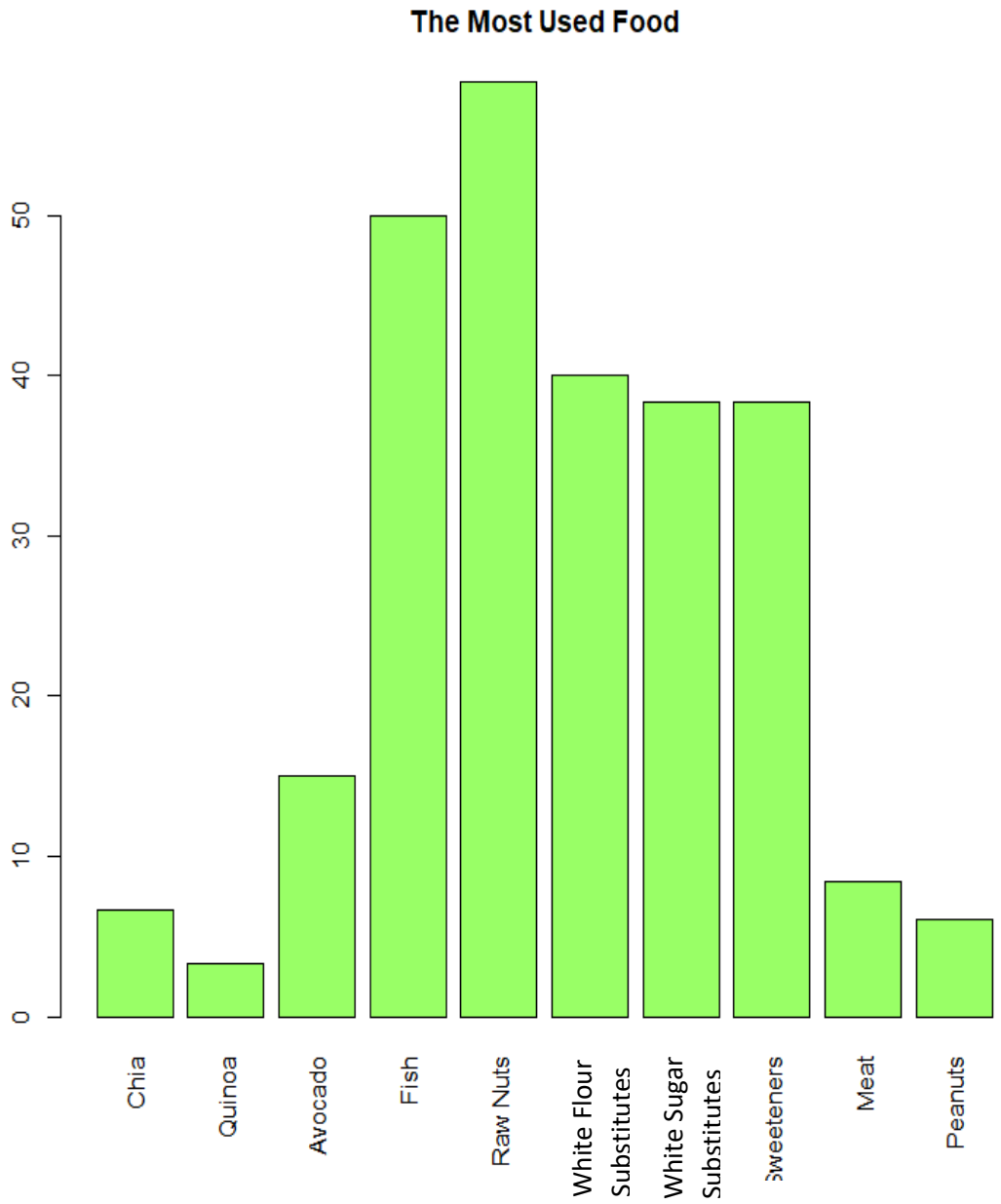
От графиката става ясно, че най-обичаният плод от анкетираните хора е бананът – 30%. Вторият най-обичан плод са черешите – 21%. Третият най-обичан плод са ягодите – 18%. И едва на 2% от анкетираните хора любимият плод са боровинките.



Фиг. 21: Кръгова диаграма, показваща процентите от анкетираните хора, които обичат най-много съответния зеленчук

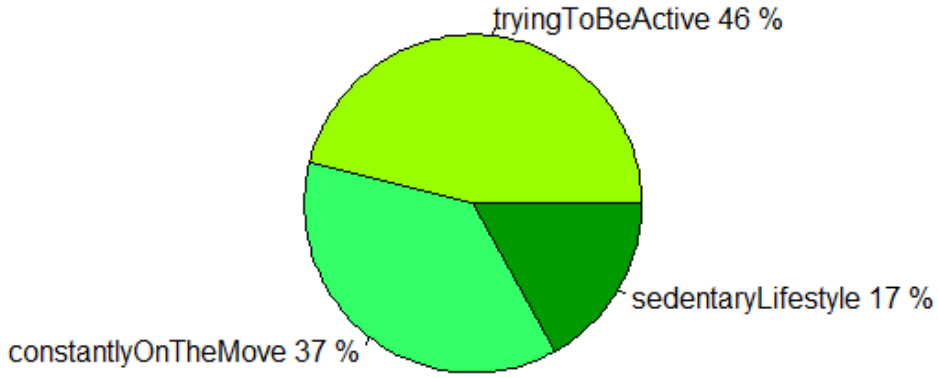
От диаграмата става ясно, че най-обичаният зеленчук е картофът – 39%. Вторият най-обичан зеленчук е краставицата с 23%. Третият и четвъртият

най-обичан зеленчук са доматите и броколите с 13%. Най-малко обичаните зеленчуци са чушките и морковите – само 2% от анкетираните хора са ги отбелязали като любими.



Фиг. 22: Хистограма, показваща най-често използваните храни от анкетираните хора

От диаграмата става ясно, че най-използваната храна са суровите ядки – над 50% от анкетираните хора са ги отбелязали. Доста често използвани храни са рибата и заместителите на бяла захар, заместителите на бяло брашно и подсладителите. Най-рядко използваните продукти са чията и киноата.



Фиг. 23: Кръгова диаграма, показваща процентите от анкетираните хора, които водят съответния тип начин на живот

От графиката става ясно, че най-много от анкетираните хора се опитват да водят активен начин на живот – 46%. Едва 17% водят заседнал начин на живот, а останалите 37% са постоянно в движение.

2.4. Определяне вида на разпределението

За определяне вида на разпределението е приложен тест на Shapiro-Wilcoxon с нулева хипотеза H0 „Разпределението е нормално“ и алтернативна хипотеза „Разпределението не е нормално“, с равнище на значимост p = 0.05. Използвана е вградена функция **shapiro.test(values)**. При **p >= 0.05** приемаме нулевата хипотеза, а в противен случай я отхвърляме и приемаме алтернативната. В следващите две таблици са представени резултати от прилагането на теста върху изследваните данни и заключенията за това дали разпределението е нормално или не.

Извадка	p-value	Нормално ли е разпределението?
allMales$Kcal	0.0002485	Не
allMales$QuantityWater	3.771e-05	Не
allMales$QuantityCarbonatedDrinks	0.0006079	Не
allMales$QuantityFruits	2.628e-06	Не
allMales$QuantityVegetables	5.814e-07	Не
allMales$QuantitySweets	0.000921	Не

Фиг. 24: Таблица с определяне на разпределенията на данните за мъжете

Извадка	p-value	Нормално ли е разпределението?
allFemales$Kcal	2.555e-07	Не
allFemales$QuantityWaterPerDay	1.523e-05	Не
allFemales$QuantityCarbonatedDrinks	8.02e-08	Не
allFemales$QuantityFruits	4.378e-08	Не
allFemales$QuantityVegetables	6.953e-08	Не
allFemales$QuantitySweetsPerDay	2.088e-05	Не

Фиг. 25: Таблица с определяне на разпределенията на данните за жените

3. Многомерен анализ

3.1. Категорийна VS числова

В тази точка ще дадем отговор на поставените в началото на документа въпроси.

1. Съществува ли съществена статистическа разлика между количеството вода, което изпиват мъжете и жените?

И двете разпределения не са нормални, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция **wilcox.test()**. Резултатът е:

```
> wilcox.test(allMales$QuantityWaterPerDay [which(Gender == 'm')], allFemales$QuantityWaterPerDay [which(Gender == 'f')])
```

Wilcoxon rank sum test with continuity correction

data: allMales$QuantityWaterPerDay[which(Gender == "m")] and allFemales$QuantityWaterPerDay[which(Gender == "f")]
W = 837.5**, p-value** = 3.252e-05
alternative hypothesis: true location shift is not equal to 0

Оттук следва, че **има** статистически значима разлика между двете извадки.

2. Съществува ли съществена статистическа разлика между калориите, които изяждат мъжете и жените?

И двете разпределения не са нормални, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция **wilcox.test()**. Резултатът е:

```
> wilcox.test(allMales$Kcal [which(Gender == 'm')], allFemales$Kcal [which(Gender == 'f')])
```

Wilcoxon rank sum test with continuity correction

data: allMales$Kcal[which(Gender == "m")] and allFemales$Kcal[which(Gender == "f")]
W = 756, **p-value** = 0.001698
alternative hypothesis: true location shift is not equal to 0

Оттук следва, че **има** статистически значима разлика между двете извадки.

3. Съществува ли съществена статистическа разлика между количеството зеленчуци, които изяждат мъжете и жените?

И двете разпределения не са нормални, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция **wilcox.test()**. Резултатът е:

```
> wilcox.test(allMales$QuantityVegetablesPerDay [which(Gender == 'm')], allFemales$QuantityVegetablesPerDay [which(Gender == 'f')])
```

Wilcoxon rank sum test with continuity correction

data: allMales$QuantityVegetablesPerDay[which(Gender == "m")] and allFemales$QuantityVegetablesPerDay[which(Gender == "f")]

W = 374, **p-value** = 0.02552

alternative hypothesis: true location shift is not equal to 0

Оттук следва, че **има** статистически значима разлика между двете извадки.

4. Съществува ли съществена статистическа разлика между количеството плодове, които изяждат мъжете и жените?

И двете разпределения не са нормални, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция **wilcox.test()**. Резултатът е:

```
> wilcox.test(allMales$QuantityFruitsPerDay [which(Gender == 'm')], allFemales$QuantityFruitsPerDay [which(Gender == 'f')])
```

Wilcoxon rank sum test with continuity correction

data: allMales$QuantityFruitsPerDay[which(Gender == "m")] and allFemales$QuantityFruitsPerDay[which(Gender == "f")]
W = 343, **p-value** = 0.007508
alternative hypothesis: true location shift is not equal to 0

Оттук следва, че **има** статистически значима разлика между двете извадки.

5. Съществува ли съществена статистическа разлика между количеството газирани напитки, които изпиват мъжете и жените?

И двете разпределения не са нормални, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция **wilcox.test()**. Резултатът е:

```
> wilcox.test(allMales$QuantityCarbonatedDrinksPerDay [which(Gender == 'm')], allFemales$QuantityCarbonatedDrinksPerDay [which(Gender == 'f')])
```

Wilcoxon rank sum test with continuity correction

data: allMales$QuantityCarbonatedDrinksPerDay[which(Gender == "m")] and allFemales$QuantityCarbonatedDrinksPerDay[which(Gender == "f")]
W = 801, **p-value** = 0.0002484
alternative hypothesis: true location shift is not equal to 0

Оттук следва, че **има** статистически значима разлика между двете извадки.

6. Съществува ли съществена статистическа разлика между количеството сладкиши, които изяждат мъжете и жените?

И двете разпределения не са нормални, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция **wilcox.test()**. Резултатът е:

```
> wilcox.test(allMales$QuantitySweetsPerDay [which(Gender == 'm')], allFemales$QuantitySweetsPerDay [which(Gender == 'f')])
```

Wilcoxon rank sum test with continuity correction

data: allMales$QuantitySweetsPerDay[which(Gender == "m")] and allFemales$QuantitySweetsPerDay[which(Gender == "f")]
W = 665, **p-value** = 0.06938
alternative hypothesis: true location shift is not equal to 0

Оттук следва, че **няма** статистически значима разлика между двете извадки.

3.2. Числова VS числова

3.2.1. Корелационен анализ

Тук ще дадем отговори на въпросите 7-22.

7. Каква е зависимостта между количеството плодове и количеството зеленчуци, които изяждат жените?

```
> cor(allFemales$QuantityFruitsPerDay [which(Gender == 'f')],
allFemales$QuantityVegetablesPerDay [which(Gender == 'f')])
```

[1] 0.04972402

От това можем да заключим, че корелацията между двете величини е **слаба**.

8. Каква е зависимостта между количеството плодове и количеството зеленчуци, които изяждат мъжете?

```
> cor(allMales$QuantityFruitsPerDay [which(Gender == 'm')],
allMales$QuantityVegetablesPerDay [which(Gender == 'm')])
```
[1] 0.5838164

От това можем да заключим, че корелацията между двете величини е **значителна**.

9. Каква е зависимостта между количеството вода, което изпиват жените и калориите, които изяждат жените?

```
> cor(allFemales$QuantityWaterPerDay [which(Gender == 'f')], allFemales$Kcal
[which(Gender == 'f')])
```

[1] 0.3773978

От това можем да заключим, че корелацията между двете величини е **умерена**.

10. Как е зависимостта между количеството вода, което изпиват мъжете и калориите, които изяждат мъжете?

```
> cor(allMales$QuantityWaterPerDay [which(Gender == 'm')], allMales$Kcal [which(Gender == 'm')])
```

[1] 0.4273274

От това можем да заключим, че корелацията между двете величини е **умерена**.

11. Каква е зависимостта между калориите и количеството плодове, които изяждат жените?

```
> cor(allFemales$Kcal [which(Gender == 'f')], allFemales$QuantityFruitsPerDay [which(Gender == 'f')])
```

[1] 0.1312205

От това можем да заключим, че корелацията между двете величини е **слаба**.

12. Каква е зависимостта между калориите и количеството плодове, които изяждат мъжете?

```
> cor(allMales$Kcal [which(Gender == 'm')], allMales$QuantityFruitsPerDay [which(Gender == 'm')])
```

[1] 0.2434396

От това можем да заключим, че корелацията между двете величини е **слаба**.

13. Каква е зависимостта между калориите и количеството зеленчуци, които изяждат жените?

```
> cor(allFemales$Kcal [which(Gender == 'f')], allFemales$QuantityVegetablesPerDay [which(Gender == 'f')])
```

[1] 0.2255106

От това можем да заключим, че корелацията между двете величини е **слаба**.

14. Каква е зависимостта между калориите и количеството зеленчуци, които изяждат мъжете?

```
> cor(allMales$Kcal [which(Gender == 'm')], allMales$QuantityVegetablesPerDay [which(Gender == 'm')])
```

[1] 0.05604485

От това можем да заключим, че корелацията между двете величини е **слаба**.

15. Каква е зависимостта между калориите и количеството сладкиши, които изяждат жените?

```
> cor(allFemales$Kcal [which(Gender == 'f')], allFemales$QuantitySweetsPerDay [which(Gender == 'f')])
```

[1] 0.1112472

От това можем да заключим, че корелацията между двете величини е **слаба**.

16. Каква е зависимостта между калориите и количеството сладкиши, които изяждат мъжете?

```
> cor(allMales$Kcal [which(Gender == 'm')], allMales$QuantitySweetsPerDay [which(Gender == 'm')])
```

[1] 0.5951482

От това можем да заключим, че корелацията между двете величини е **значителна**.

17. Каква е зависимостта между калориите и количеството газирани напитки, което изпиват жените?

```
> cor(allFemales$Kcal [which(Gender == 'f')], allFemales$QuantityCarbonatedDrinksPerDay [which(Gender == 'f')])
```

[1] -0.1135772

От това можем да заключим, че корелацията между двете величини е **отрицателна**.

18. Каква е зависимостта между калориите и количеството газирани напитки, което изпиват мъжете?

```
> cor(allMales$Kcal [which(Gender == 'm')],
allMales$QuantityCarbonatedDrinksPerDay [which(Gender == 'm')])
```

[1] 0.438529

От това можем да заключим, че корелацията между двете величини е **умерена**.

19. Каква е зависимостта между количеството газирани напитки, което изпиват жените и количеството сладкиши, което изяждат жените?

```
> cor(allFemales$QuantityCarbonatedDrinksPerDay [which(Gender == 'f')],
allFemales$QuantitySweetsPerDay [which(Gender == 'f')])
```

[1] 0.5094117

От това можем да заключим, че корелацията между двете величини е **значителна**.

20. Каква е зависимостта между количеството газирани напитки, което изпиват мъжете и количеството сладкиши, което изяждат мъжете?

```
> cor(allMales$QuantityCarbonatedDrinksPerDay [which(Gender == 'm')],
allMales$QuantitySweetsPerDay [which(Gender == 'm')])
```

[1] 0.7131997

От това можем да заключим, че корелацията между двете величини е **висока**.

21. Каква е зависимостта между количеството плодове и количеството сладкиши, което изяждат жените?

```
> cor(allFemales$QuantityFruitsPerDay [which(Gender == 'f')],
allFemales$QuantitySweetsPerDay [which(Gender == 'f')])
```

[1] 0.1230566

От това можем да заключим, че корелацията между двете величини е **слаба**.

22. Каква е зависимостта между количеството плодове и количеството сладкиши, което изяждат мъжете?

```
> cor(allMales$QuantityFruitsPerDay [which(Gender == 'm')],
allMales$QuantitySweetsPerDay [which(Gender == 'm')])
```

[1] 0.2393026

От това можем да заключим, че корелацията между двете величини е **слаба**.

3.2.2. Линейна регресия

Прилагам линейна регресия относно въпроси 7 и 20, за да проуча и обобщя връзките между посочените множества от непрекъснати променливи. За останалите въпроси изводите са аналогични.
Прилагам функцията за линейна регресия lm().

7. Каква е зависимостта между количеството плодове и количеството зеленчуци, които изяждат жените?

```
> modelFruitsAndVegetablesFemales <- lm(allFemales$QuantityFruitsPerDay [which(Gender == 'f')] ~ allFemales$QuantityV
egetablesPerDay [which(Gender == 'f')], data = allFemales)
> model

Call:
lm(formula = allMales$Kcal[which(Gender == "m")] ~ allMales$QuantityWaterPerDay[which(Gender ==
    "m")], data = allMales)

Coefficients:
                                    (Intercept)  allMales$QuantityWaterPerDay[which(Gender == "m")]
                                      1360.2484                                              0.3478

> modelFruitsAndVegetablesFemales

Call:
lm(formula = allFemales$QuantityFruitsPerDay[which(Gender ==
    "f")] ~ allFemales$QuantityVegetablesPerDay[which(Gender ==
    "f")], data = allFemales)

Coefficients:
                                          (Intercept)  allFemales$QuantityVegetablesPerDay[which(Gender == "f")]
                                            271.23556                                                    0.05627
```

След като съм построила линеен модел, проверявам до колко този модел описва добре данните и какви са оценките на коефициенти му.

```
> summary(modelFruitsAndVegetablesFemales)

Call:
lm(formula = allFemales$QuantityFruitsPerDay[which(Gender ==
    "f")] ~ allFemales$QuantityVegetablesPerDay[which(Gender ==
    "f")], data = allFemales)

Residuals:
    Min      1Q  Median      3Q     Max
-185.30 -175.46  -35.30  -28.97  478.76

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        271.23556   59.39732   4.566  5.6e-05 ***
allFemales$QuantityVegetablesPerDay[which(Gender == "f")]   0.05627    0.18839   0.299    0.767
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 217 on 36 degrees of freedom
Multiple R-squared:  0.002472,  Adjusted R-squared:  -0.02524
F-statistic: 0.08923 on 1 and 36 DF,  p-value: 0.7669
```

В резултат се показват 3 таблици. Първата от тях е **Residuals** - статистика за остатъците, **Coefficients** - коефициентите и **Residual standard error** - до колко линейната регресия работи добре.

Таблицата **Residuals** - тя дава информация за минималната стойност, 1-вия квартил, медиана (2-рия квартил), 3-тия квартил и максималната стойност.

Таблицата **Coefficients** по редове показва участващите коефициенти. Първата колона показва оценката, втората - стандартната грешка с която се построява доверителен интервал, третата колона представлява частното на оценката и стандартната грешка, а в последната колона се намерат стойностите на p-value.

Първо ще проверя дали коефициентите са статистически значими - дали е необходимо да участват в анализа. За всеки един коефициент проверявам хипотезата дали той е равен на 0. За да бъде значим трябва да бъде отхвърлена тази хипотеза. За да се отхвърли H0, то стойността на pvaue трябва да бъде по-малка от 0.05.

Разглеждам оценките пред коефициента **Intercept**. Оценката е **271.23556**. Той е статистически значим, защото стойността на p-value е **5.6e-05 < 0,05**. Чрез третата таблица проверявам до колко модела описва добре данните. Разглеждам статистиките **Multiple R-squared** или **Adjusted Rsquared**.Те приемат стойности в интервала [0-1].

Стойността на **Adjusted R-squared** е **-0.02524**, тоест **не** може да бъде направен изводът че моделът описва много добре данните.

20. Каква е зависимостта между количеството газирани напитки, което изпиват мъжете и количеството сладкиши, което изяждат мъжете?

```
> modelCarbonatedDrinksAndSweetsMales <- lm(allMales$QuantityCarbonatedDrinksPerDay [which(Gender == 'm')] ~ allM
ales$QuantitySweetsPerDay [which(Gender == 'm')], data = allMales)
> modelCarbonatedDrinksAndSweetsMales

Call:
lm(formula = allMales$QuantityCarbonatedDrinksPerDay[which(Gender ==
    "m")] ~ allMales$QuantitySweetsPerDay[which(Gender == "m")],
    data = allMales)

Coefficients:
                                        (Intercept)  allMales$QuantitySweetsPerDay[which(Gender == "m")]
                                            158.333                                              1.917
```

След като съм построила линеен модел, проверявам до колко този модел описва добре данните и какви са оценките на коефициенти му.

```
> summary(modelCarbonatedDrinksAndSweetsMales)

Call:
lm(formula = allMales$QuantityCarbonatedDrinksPerDay[which(Gender ==
    "m")] ~ allMales$QuantitySweetsPerDay[which(Gender == "m")],
    data = allMales)

Residuals:
    Min      1Q  Median      3Q     Max
-733.33 -158.33   16.67  191.67 1150.00

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      158.3333   125.2828   1.264    0.218
allMales$QuantitySweetsPerDay[which(Gender == "m")]   1.9167     0.3694   5.188 2.05e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 397.4 on 26 degrees of freedom
Multiple R-squared:  0.5087,    Adjusted R-squared:  0.4898
F-statistic: 26.92 on 1 and 26 DF,  p-value: 2.047e-05
```

Разглеждам оценките пред коефициента **Intercept** . Оценката е **158.3333**. Той е статистически незначим, защото стойността на **p-value** е **0.218 > 0,05**. Чрез третата таблица проверявам до колко модела описва добре данните. Разглеждам статистиките **Multiple R-squared** или **Adjusted Rsquared**.
Те приемат стойности в интервала [0-1].

Стойността на **Adjusted R-squared** е **0.4898**, тоест **не** може да бъде направен изводът че моделът описва много добре данните.

За останалите въпроси изводите са аналогични.

4. Изводи

Целта на анализа беше да покажа основните разлики между хранителните навици на мъжете и жените. Като например разликата в калориите, които приемат дневно, разликата в количеството вода, което изпиват и т.н.

След анализа на всички въпроси от анкетата мога да твърдя следните неща:

1. Много хора се опитват да спазват здравословен начин на живот.
2. Жените приемат по-малко калории в сравнение с мъжете.
3. Жените приемат повече плодове и зеленчуци в сравнение с мъжете.
4. Мъжете приемат повече вода на ден в сравнение с жените.
5. Жените приемат по-малко сладкиши в сравнение с мъжете.
6. Малко хора използват навлезлите наскоро здравословни храни като чия, киноа и авокадо, а най-много хора използват типичните са нашата кухня храни като ядки и риба.

Като заключение мога да кажа, че жените са по-заинтересовани на тема здравословен начин на живот и повече ограничават високо калоричните храни и напитки като сладкиши и газирани напитки.