УНИВЕРЗИТЕТ “СВ. КИРИЛ И МЕТОДИЈ” – СКОПЈЕ

ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНИЖЕНЕРСТВО

СЕМИНАРСКА РАБОТА

по предметот

БИЗНИС СТАТИСТИКА

Тема

АНАЛИЗА НА ПОДАТОЦИ КАКО МУЗИКАТА ВЛИЈАЕ ВРЗ МЕНТАЛНОТО ЗДРАВЈЕ

МЕНТОР:

д-р Верица Бакева

ИЗРАБОТИЛА:

Елена Колевска (223002)

Скопје, 2023

ВОВЕД

Бизнис статистиката е област на изучување која вклучува собирање, анализа, толкување, презентација и организација на податоци за да се донесат точни бизнис одлуки. Истата користи статистички методи и техники за да се разберат и решат проблеми поврзани со бизнис одлуки, маркетинг, финансии, економија и други области.

Оваа проектна задача ги истражува врските помеѓу музиката и менталното здравје, фокусирајќи се на резултатите од истражувањето преку анкета. Анкетата имаше за цел да го истражи влијанието на музиката врз различни аспекти на менталното благостворење, вклучувајќи намалување на стресот, подобрување на расположението и механизми за справување. Во оваа работа се анализираат податоците собрани од различни групи испитаници, користејќи статистички методи за откривање на трендови, шаблони и врски помеѓу консумацијата на музика и показателите за ментално здравје. Ова истражување има за цел да идентификува какви, доколку ги има, корелации постојат помеѓу музичкиот вкус на поединецот и неговото ментално здравје.

Во оваа задача ќе бидат применети знаењата стекнати по предметот Бизнис статистика како од предавањата, така и од аудиториските вежби преку користење на програмата R и RStudio.

ПОДАТОЧНО МНОЖЕСТВО

Истражување за музика и ментално здравје

Податочното множество за поврзаноста на музиката и менталното здравје вклучува 737 редови и 33 колони, кои ги опфаќаат различните жанрови музика и влијанието врз менталното здравје на поединецот.

Содржи информации музичката позадина и навики за слушање, колку често слушаат 16 музички жанрови и рангирање анксиозност, депресија, несоница и опсесивно компулсивно растројство на скала од 0 до 10.

Оваа проектна задача ќе биде фокусирана на бројот на часови во денот кога поединецот слуша музика и анксиозноста како и нивната поврзаност.

А. ПРВ ДЕЛ

А.1 ТАБЕЛА СО РАСПРЕДЕЛБА НА ЧЕСТОТИ

При креирање табела на распределба на честори, првиот чекор е податоците да се поделат во интервали. Иако премногу интервали не се корисни за сумирање на податоците, поради големиот број податоци (736) според двете формули се добива број поголем од 15. Затоа, за најдобра прегледност на резултатите јас се одлучив да ги поделам податоците на 23 интервали.

Средната вредност на интервалите ја добив кога го поделив збирот на првата и последната вреедност со два.

Релативната фреквенција ја добив како количник од честотата на секој интервал со вкупниот број податоци (736)

Кумулативната фреквенција ја добив со собирање на последователните честоти.

Па, ја добив оваа табела.

	Interval	Frequency	Midpoint	Relative_Frequency.Var1	Relative_Frequency.Freq	Cumulative_Frequency
(0.857,1.71]	(0.857,1.71]	134	1.2857143	(0.857,1.71]	0.182065217	165
(1.71,2.57]	(1.71,2.57]	179	2.1428571	(1.71,2.57]	0.243206522	344
(2.57,3.43]	(2.57,3.43]	120	3.0000000	(2.57,3.43]	0.163043478	464
(3.43,4.29]	(3.43,4.29]	83	3.8571429	(3.43,4.29]	0.112771739	547
(4.29,5.14]	(4.29,5.14]	55	4.7142857	(4.29,5.14]	0.074728261	602
(5.14,6]	(5.14,6]	47	5.5714286	(5.14,6]	0.063858696	649
(6,6.86]	(6,6.86]	0	6.4285714	(6,6.86]	0.000000000	649
(6.86,7.71]	(6.86,7.71]	15	7.2857143	(6.86,7.71]	0.020380435	664
(7.71,8.57]	(7.71,8.57]	29	8.1428571	(7.71,8.57]	0.039402174	693
(8.57,9.43]	(8.57,9.43]	3	9.0000000	(8.57,9.43]	0.004076087	696
(9.43,10.3]	(9.43,10.3]	20	9.8571429	(9.43,10.3]	0.027173913	716
(10.3,11.1]	(10.3,11.1]	1	10.7142857	(10.3,11.1]	0.001358696	717
(11.1,12]	(11.1,12]	9	11.5714286	(11.1,12]	0.012228261	726
(12,12.9]	(12,12.9]	0	12.4285714	(12,12.9]	0.000000000	726
(12.9,13.7]	(12.9,13.7]	1	13.2857143	(12.9,13.7]	0.001358696	727
(13.7,14.6]	(13.7,14.6]	1	14.1428571	(13.7,14.6]	0.001358696	728
(14.6,15.4]	(14.6,15.4]	2	15.0000000	(14.6,15.4]	0.002717391	730
(15.4,16.3]	(15.4,16.3]	1	15.8571429	(15.4,16.3]	0.001358696	731
(16.3,17.1]	(16.3,17.1]	0	16.7142857	(16.3,17.1]	0.000000000	731
(17.1,18]	(17.1,18]	1	17.5714286	(17.1,18]	0.001358696	732
(18,18.9]	(18,18.9]	0	18.4285714	(18,18.9]	0.000000000	732
(18.9,19.7]	(18.9,19.7]	0	19.2857143	(18.9,19.7]	0.000000000	732
(19.7,20.6]	(19.7,20.6]	1	20.1428571	(19.7,20.6]	0.001358696	733
(20.6,21.4]	(20.6,21.4]	0	21.0000000	(20.6,21.4]	0.000000000	733
(21.4,22.3]	(21.4,22.3]	0	21.8571429	(21.4,22.3]	0.000000000	733
(22.3,23.1]	(22.3,23.1]	0	22.7142857	(22.3,23.1]	0.000000000	733
(23.1,24]	(23.1,24]	3	23.5714286	(23.1,24]	0.004076087	736

Табела 1: Број часови поминати во слушање музика на ден

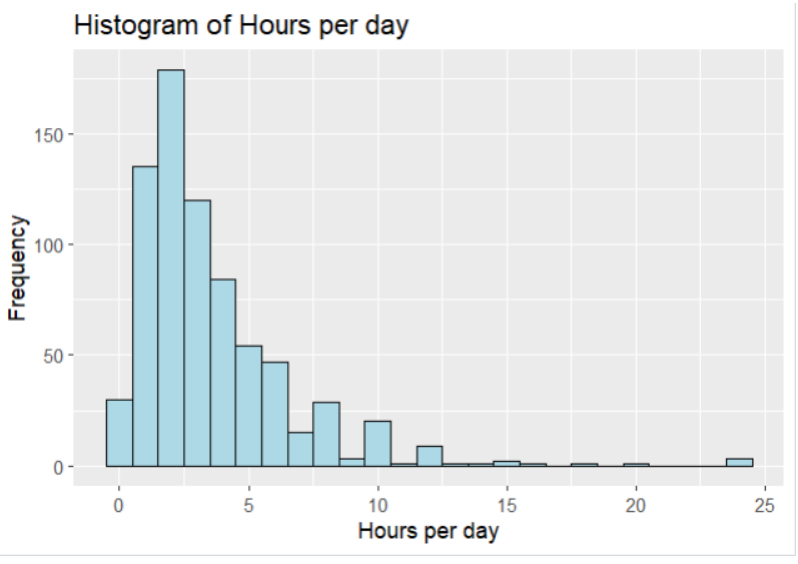
Со спроведување на истите чекори ја добив и табелата за ниво на анксиозност во текот на денот на испитаниците.

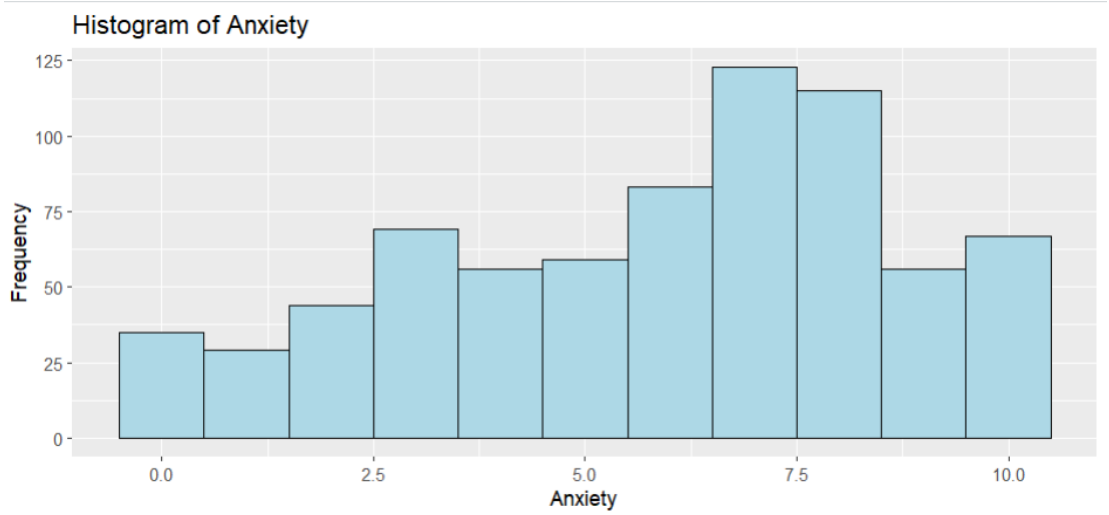
	Interval	Frequency	Midpoint	Relative_Frequency.Var1	Relative_Frequency.Freq	Cumulative_Frequency
(0,0.393]	(0,0.393]	0	0.1964286	(0,0.393]	0.000000000	0
(0.393,0.786]	(0.393,0.786]	0	0.5892857	(0.393,0.786]	0.000000000	0
(0.786,1.18]	(0.786,1.18]	29	0.9821429	(0.786,1.18]	0.041369472	29
(1.18,1.57]	(1.18,1.57]	0	1.3750000	(1.18,1.57]	0.000000000	29
(1.57,1.96]	(1.57,1.96]	0	1.7678571	(1.57,1.96]	0.000000000	29
(1.96,2.36]	(1.96,2.36]	44	2.1607143	(1.96,2.36]	0.062767475	73
(2.36,2.75]	(2.36,2.75]	0	2.5535714	(2.36,2.75]	0.000000000	73
(2.75,3.14]	(2.75,3.14]	69	2.9464286	(2.75,3.14]	0.098430813	142
(3.14,3.54]	(3.14,3.54]	0	3.3392857	(3.14,3.54]	0.000000000	142
(3.54,3.93]	(3.54,3.93]	0	3.7321429	(3.54,3.93]	0.000000000	142
(3.93,4.32]	(3.93,4.32]	56	4.1250000	(3.93,4.32]	0.079885877	198
(4.32,4.71]	(4.32,4.71]	0	4.5178571	(4.32,4.71]	0.000000000	198
(4.71,5.11]	(4.71,5.11]	59	4.9107143	(4.71,5.11]	0.084165478	257
(5.11,5.5]	(5.11,5.5]	0	5.3035714	(5.11,5.5]	0.000000000	257
(5.5,5.89]	(5.5,5.89]	0	5.6964286	(5.5,5.89]	0.000000000	257
(5.89,6.29]	(5.89,6.29]	83	6.0892857	(5.89,6.29]	0.118402282	340
(6.29,6.68]	(6.29,6.68]	0	6.4821429	(6.29,6.68]	0.000000000	340
(6.68,7.07]	(6.68,7.07]	122	6.8750000	(6.68,7.07]	0.174037090	462
(7.07,7.46]	(7.07,7.46]	0	7.2678571	(7.07,7.46]	0.000000000	462
(7.46,7.86]	(7.46,7.86]	1	7.6607143	(7.46,7.86]	0.001426534	463
(7.86,8.25]	(7.86,8.25]	115	8.0535714	(7.86,8.25]	0.164051355	578
(8.25,8.64]	(8.25,8.64]	0	8.4464286	(8.25,8.64]	0.000000000	578
(8.64,9.04]	(8.64,9.04]	56	8.8392857	(8.64,9.04]	0.079885877	634
(9.04,9.43]	(9.04,9.43]	0	9.2321429	(9.04,9.43]	0.000000000	634
(9.43,9.82]	(9.43,9.82]	0	9.6250000	(9.43,9.82]	0.000000000	634
(9.82,10.2]	(9.82,10.2]	67	10.0178571	(9.82,10.2]	0.095577746	701
(10.2,10.6]	(10.2,10.6]	0	10.4107143	(10.2,10.6]	0.000000000	701

Табела 2: Ниво на анксиозност во текот на денот на испитаниците.

А.2 ХИСТОГРАМ

- Крајните точки (или средината) на интервалите се прикажани на хоризонталната оска.
- На вертикалната оска се честотите.
- Над секој интервал се црта столб со висина која одговара на вредноста на честотите во соодветниот интервал.

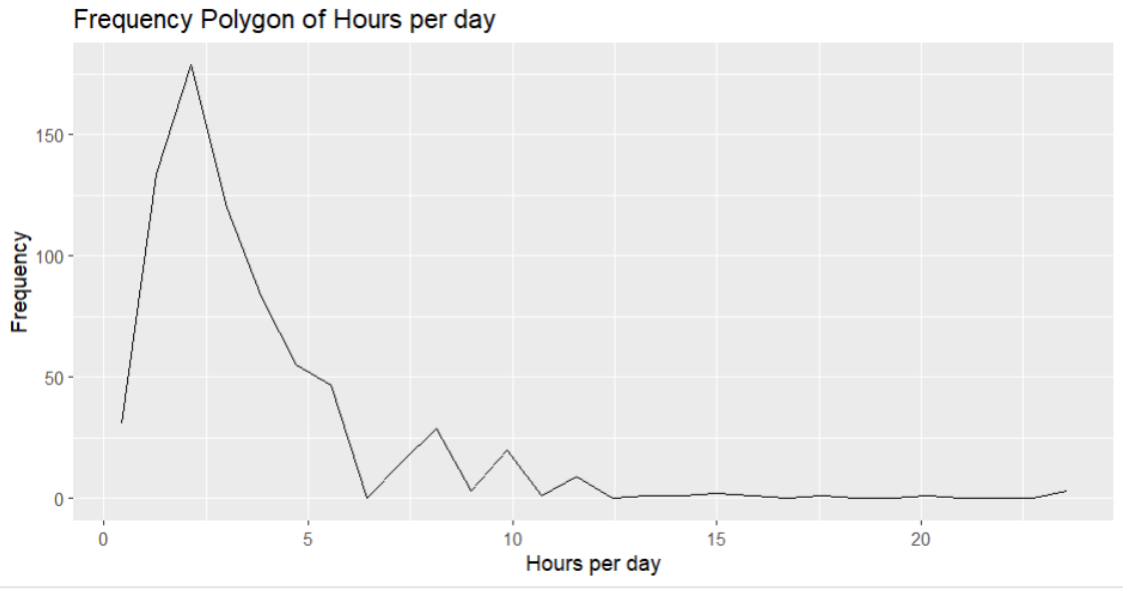


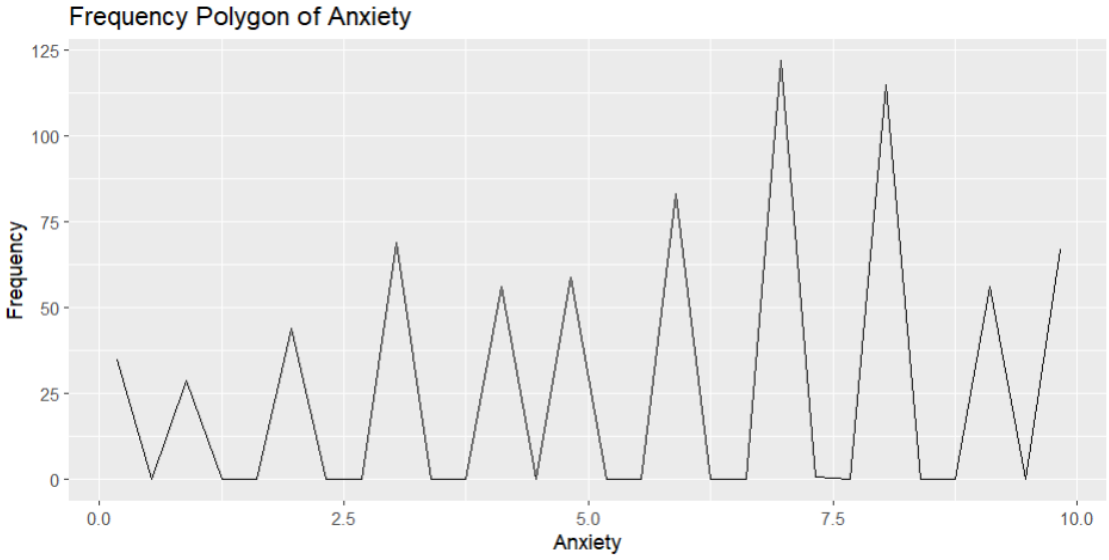


А.3 ПОЛИГОН НА ЧЕСТОТИ

Бидејќи веќе ги имаме хистограмите, можеме да ги добиеме и полигоните на честоти.

Полигонот на честоти се добива со поврзување на точките со координати (средна точка на интервалот, фреквенција на интервалот) во секој од интервалите.





А.4 СТЕБЛО-ЛИСТ ДИЈАГРАМ

За да се направи овој дијаграм секоја вредност се дели на два дела.

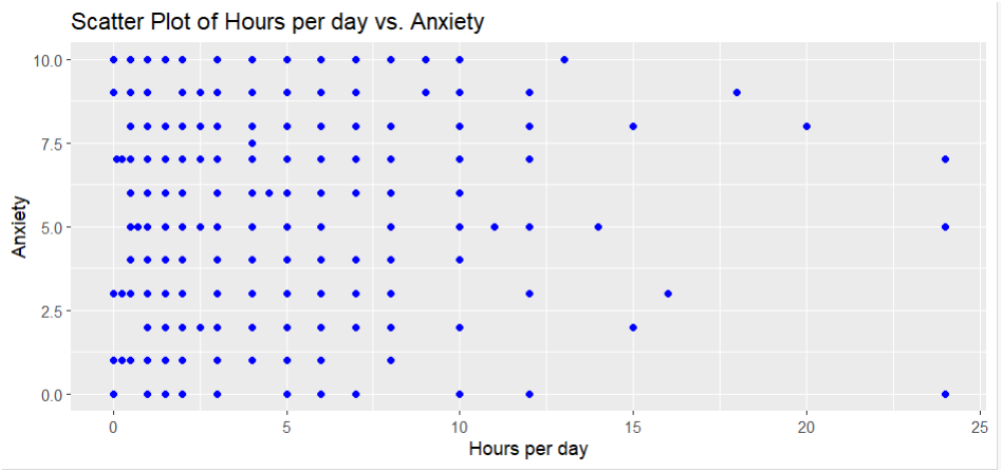
- Првиот дел од бројот се нарекува стебло, а вториот дел лист.

Ги избрав податоците од колоната Age, односно годините на секој испитаник.

```
1 | 8888898899876675579868487799768787878976689596679889969499587699797838986797889996887869679986888746875657989995744878685667775377598398768957667660879659267484394485783527778449869625463687465848894598999857985788898788897868866889348868788869776957689798567778888888889789676986979777996938478992 | 1016034333233254962264205801013233283192498157526115618217770811566206185226301905167359500322110510340043825737300211926221532732311311235435090174703043136620131145203423732828205094158752400820240255141992676734391241101101504678110122063176624183007640117110300136391326119 | 7626617320614522424851030571610530111824113022825323222801984384562402310014726075 | 213119242009034901983206892234 | 33694774888516670396 | 3104491370000340577 | 2301 | 09
```

А.5 ГРАФИК НА РАСЕЈУВАЊЕ ЗА ПОДАТОЦИТЕ ОД ДВЕТЕ ОБЕЛЕЖЈА

Треба да се провери дали има некаква врска помеѓу двете обележја за истиот примерок (736 испитаници)



За да можеме да дискутираме за поврзаноста помеѓу двете обележја потребно е да се најде коефициентот на корелација.

Коефициентот на корелација ја мери јачината на линеарната врска меѓу две квантитативни обележја.

За коефициентот на корелација добив 0.05 што значи дека постои слаба или незначајна врска помеѓу часовите на ден поминати слушајќи музика и анксиозноста. Ова значи дека врската помеѓу двете обележја не е доволно силна за да може да се донесат некои заклучоци и точни предвидувања потпирајќи се на оваа корелација.

А.6 МОДА, МЕДИЈАНА И ПРОСЕК

Мода- вредност од примерокот која има најголема честота.

Медијаната на примерокот е број што стои на средина на подредениот примерок.

Просек (аритметичка средина) е количник од збирот на сите податоци со нивниот број

- ❖ Часови на ден поминати слушајќи музика
  - Мода: 2
  - Медијана: 3
  - Просек: 3.6 часа
- ❖ Ниво на анксиозност
  - Мода: 7
  - Медијана: 6
  - Просек: 5.8

А.7 КВАРТИЛИ, ОПСЕГ И ИНТЕРКВАРТАЛЕН РАСПОН

Квартили (квартал) се вид на перцентили кои го делат примерокот на четири дела, или четвртини, со приближно еднаква големина. За да се пресметаат кварталите, податоците мора да бидат подредени во неопаѓачки редослед.

- ✓ Првиот квартил (Q_1) и тоа е вредност таква што приближно 25% од податоците во подредениот примерок се лево од него, а приближно 75% се десно. Познат е и како долен квартил.
- ✓ Вториот квартил (Q_2) е медијана на примерокот, така што приближно 50% од податоците во подредениот примерок се лево од него и приближно 50% се десно.
- ✓ Третиот квартил (Q_3) и тоа е вредност таква што приближно 75% од податоците во примерокот се лево од него и приближно 25% се десно. Познат е и како горен квартил.

Опсег се пресметува како разлика помеѓу најголемата и најмалата набљудувана вредност.

Интерквартални распон (IQR) е мерка на варирање на податоците. Се пресметува како разлика меѓу третиот и првиот квартал.

- ❖ Часови на ден поминати слушајќи музика
  - Q_1: 3 Q_2: 2 Q_3: 5
  - Опсег: 24-0=24
  - Интерквартален опсег: 5-2=3

- ❖ Ниво на анксиозност
  - Q_1: 4 Q_2: 6 Q_3: 8
  - Опсег: 10
  - Интерквартален опсег: 4

А.8 ДИСПЕРЗИЈА И СТАНДАРДНА ДЕВИЈАЦИЈА

Дисперзијата (варијансата) на примерок е мерка за колку се различни податоците во обележјето од средната вредност.

Стандардна девијација на примерок - го покажува варирањето на податоците во обележјето околу просекот на примерокот. Таа се дефинира како квадратен корен од дисперзијата.

- ❖ Часови на ден поминати слушајќи музика
  - Дисперзија: 9.169988
  - Стандардна девијација: 3.028199
- ❖ Ниво на анксиозност
  - Дисперзија: 7.801153
  - Стандардна девијација: 2.793054

А.8 КОЕФИЦИЕНТ НА КОРЕЛАЦИЈА

Коефициентот на корелација - ја мери јачината на линеарната врска меѓу две квантитативни променливи.

Коефициентот на корелација веќе го пресметавме за да извлечеме заклучок за графикот на расејување и тој е 0.0493189 или приближно 0.05. Ова значи дека дека постои слаба или незначајна врска помеѓу часовите на ден поминати слушајќи музика и нивото на анксиозност.

Б. ВТОР ДЕЛ

Б.1 ИНТЕРВАЛ НА ДОВЕРБА НА МАТЕМАТИЧКО ОЧЕКУВАЊЕ

Треба да се пресмета очекуваното време на слушање музика во текот на денот со 95% интервал на доверба. Веќе знаеме дека просечното време на слушање на 736 испитаници е 3.6 часа, а стандардната девијација е 3.028199. Времето на слушање музика има нормална распределба.

- Бидејќи стандардната девијација е позната. За да се пресмета 95% интервал на доверба на математичко очекување ја користиме формулата:

Интервал на доверба = (просек на примерокот – маргина на грешка, просек на примерокот + маргина на грешка)

```
  > # Specify the column name
  > column_name <- "Hours per day"
>
  > # Extract the column
  > hours_per_day <- data[[column_name]]
>
  > # Calculate the mean and standard error
  > mean_hours_per_day <- mean(hours_per_day)
> std_error <- sd(hours_per_day) / sqrt(length(hours_per_day))
>
  > # Specify the confidence level (e.g., 95%)
  > confidence_level <- 0.95
>
  > # Calculate the margin of error
  > margin_of_error <- qt((1 + confidence_level) / 2, df = length(hours_per_day) - 1) * std_error
>
  > # Calculate the confidence interval
  > confidence_interval <- c(mean_hours_per_day - margin_of_error, mean_hours_per_day + margin_of_error)
>
  > # Print the confidence interval
  > cat("Confidence Interval (", confidence_level * 100, "%): [", confidence_interval[1], ",", confidence_interval[2], "]\n")
Confidence Interval ( 95 %): [ 3.353624 , 3.791892 ]
```

Б.2 ТЕСТИРАЊЕ ХИПОТЕЗИ

Врз основа на случаен примерок со обем n = 736 за нормално распределено обележје X, добиена е дисперзија 9.169988. Дали со ниво на значајност α = 0.05, може да се заклучи дека дисперзијата на обележјето е поголема од 15?

Најпрво ги поставуваме хипотезите:

H_0: Дисперзијата е 15 (DX=15)

H_a: Дисперзијата е поголема од 15 (DX > 15)

Па, ја пресметуваме вредноста на хи-квадрат тест статистиката. Ја одредуваме критичната вредност од хи-квадрат распределбата со n - 1 степени на слобода и ниво на значајност α.

- Ако хи-квадрат статистиката е поголема од критичната вредност, H_0 се отфрла.
- Ако хи-квадрат статистиката не е поголема од критичната вредност, H_0 не се отфрла.

```r
# Specify the column name
column_name <- "Hours per day"

# Extract the column
hours_per_day <- data[[column_name]]

# Specify the null and alternative hypotheses
# Null Hypothesis (H0): Variance is 15 or less (σ^2 ≤ 15)
# Alternative Hypothesis (HA): Variance is greater than 15 (σ^2 > 15)

# Calculate sample size and sample variance
n <- length(hours_per_day)
sample_variance <- var(hours_per_day)

# Define the null variance and calculate the test statistic
null_variance <- 15
test_statistic <- (n - 1) * sample_variance / null_variance

# Define the significance level (α)
alpha <- 0.05

# Calculate the critical value from the chi-squared distribution
critical_value <- qchisq(1 - alpha, df = n - 1, lower.tail = FALSE)

# Compare the test statistic with the critical value
if (test_statistic > critical_value) {
  cat("Reject the null hypothesis. Variance is greater than 15.\n")
} else {
  cat("Fail to reject the null hypothesis. Variance is 15 or less.\n")
}
```

Б.3 ТЕСТ НА РАСПРЕДЕЛБА

Нека X е обележјето – време поминато слушајќи музика на 736 испитаници. Треба да се тестира хипотезата H0 : X ~ B(736, p), каде што параметарот p е непознат и треба да се оцени. За да најдеме оценка на непознатиот параметарот p, тргнуваме од просекот за кој знаеме дека е добар точкаст оценувач на математичкото очекување EX.

За да ја тестираме хипотезата H0: X ~ B(736, p), каде што параметарот p е непознат, и да пронајдеме проценка на непознатиот параметар p користејќи ја средната вредност на примерокот како точкаст оценувач, треба:

Да ги поставиме хипотезите:

H_0: X ~ B(736, p)

H_a: X нема распределба B(736, p)

Да го пресметаме (просекот) оценувачот за непознатиот параметар p.

Со користење на хи-квадрат тест статистиката, добиваме:

```
# Specify the column name
> column_name <- "Hours per day"
>
  > # Extract the column
  > hours_per_day <- data[[column_name]]
>
  > # Calculate the sample mean
  > sample_mean <- mean(hours_per_day)
>
  > # Define the expected probability (p) for the binomial distribution
  > expected_p <- 0.5
>
  > # Calculate the observed sum of values
  > observed <- sum(hours_per_day)
>
  > # Calculate the expected sum based on the length of the data and expected probability
  > expected <- length(hours_per_day) * expected_p
>
  > # Perform the chi-squared goodness-of-fit test
  > chi_sq_test <- chisq.test(x = c(observed, expected))
>
  > # Interpret the test results
  > if (chi_sq_test$p.value > 0.05) {
    +     cat("The null hypothesis is accepted. The feature follows a B(736, 0.5) distribution.\n")
    + } else {
      +     cat("The null hypothesis is rejected. The feature does not follow a B(736, 0.5) distribution.\n")
      + }
The null hypothesis is rejected. The feature does not follow a B(736, 0.5) distribution.
```

Б.4 ТЕСТИРАЊЕ ХИПОТЕЗИ ЗА НЕЗАВИСНОСТ НА ДВЕ ОБЕЛЕЖЈА

Ова тестирање **не може** да се направи. Зошто?

Ако обележјето часови поминати во слушање музика и нивото на анксиозност се непрекинати квантитативни променливи, Пирсоновиот хи-квадрат тест за независност не е применлив. Овој тест е специјално дизајниран за тестирање на независноста помеѓу категоричните променливи.

Во случај на две непрекинати квантитативни променливи, можеме да ја истражиме нивната врска користејќи корелациона или регресиона анализа. Корелационата анализа помага да се одреди силата и насоката на линеарната врска помеѓу променливите, додека регресионата анализа може да ја процени врската и да обезбеди увид во моќта на предвидување на едната променлива врз основа на другата.

Б.5 РЕГРЕСИОНА АНАЛИЗА

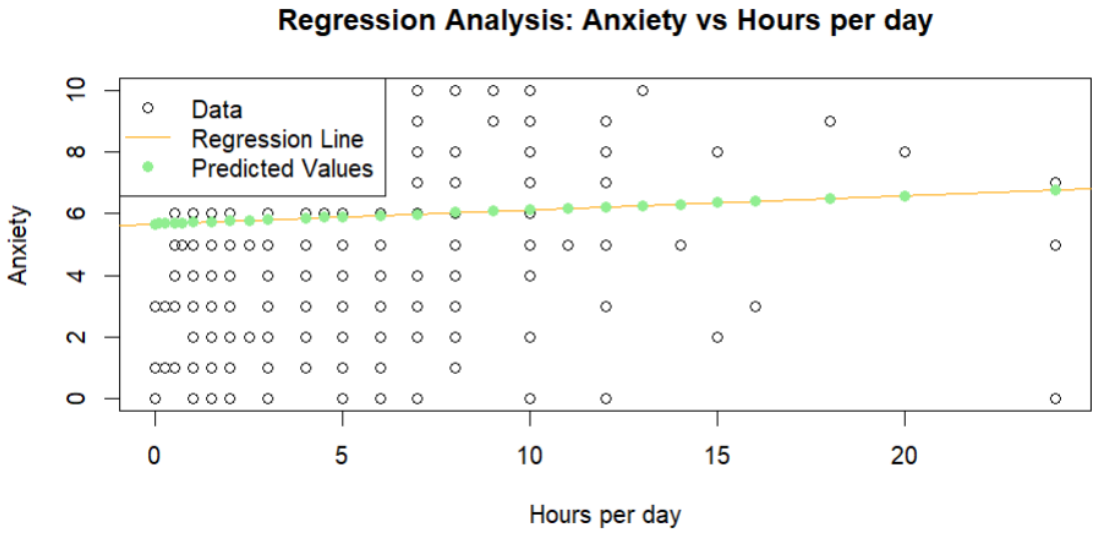
Регресионата анализа се користи за одредување на видот на врската обележјата и главната цел кога се користи овој метод е да се предвиди или процени вредноста на едната променлива за дадена вредност на другата променлива.

Во овој код се извршува линеарна регресија помеѓу "Anxiety" и "Hours per day" користејќи функцијата lm(). Се предвидуваат вредностите на "Anxiety" со користење на моделот кој го креиравме. Се креира график кој ги прикажува податоците за "Anxiety" во зависност од "Hours per day". Се додава линија на регресија на графикот (со abline()

функцијата). Се додаваат предвидените вредности на "Anxiety" на графикот (со points() функцијата).

```
> # Perform linear regression
> model <- lm(anxiety ~ hours_per_day)
>
> # Predicted values
> predicted_values <- predict(model)
>
> # Plot the regression analysis
> plot(hours_per_day, anxiety, xlab = "Hours per day", ylab = "Anxiety",
+      main = "Regression Analysis: Anxiety vs Hours per day")
>
> # Add regression line
> abline(model, col = "orange")
>
> # Add predicted values
> points(hours_per_day, predicted_values, col = "lightgreen", pch = 16)
>
> # Add legend
> legend("topleft", legend = c("Data", "Regression Line", "Predicted Values"),
+        col = c("black", "orange", "lightgreen"), pch = c(1, NA, 16), lty = c(NA, 1, NA))
>
> # Print regression summary
> summary(model)

Call:
lm(formula = anxiety ~ hours_per_day)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7669 -1.9480  0.2794  2.1543  4.3249

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.67511    0.15921  35.646   <2e-16 ***
hours_per_day 0.04549    0.03400   1.338    0.181
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.792 on 734 degrees of freedom
Multiple R-squared:  0.002432,  Adjusted R-squared:  0.001073
F-statistic:  1.79 on 1 and 734 DF,  p-value: 0.1814
```

На крај, овие информации ни даваат информации за како "Hours per day" влијае врз "Anxiety" и дали има статистички значајна врска помеѓу нив. Во овој случај, нашиот модел покажува дека нема значителна врска меѓу овие две променливи.



ЕПИЛОГ

Целта на мојата семинарска работа е да се направи анализа на податоци како музиката влијае врз менталното здравје. Преку разни методи и задачи ја испитав нивната зависност и дојдов до одредени заклучоци.

Во оваа семинарска работа ги користев моите знаења по предметот Бизнис статистика како и вештините за работа со програмата R и RStudio кои ги стекнав пред се' преку гледање едукативни видеа на YouTube и преку истражување на Интернет.

Во овој документ се запишани сите вредности и визуелизации кои ги добив со помош на кодот кој го напишав во Rstudio. Документот со мојот код можете да го пронајдете во истиот репозиториум како и овој документ.