



**Universidade de Brasília**

DEPARTAMENTO DE ESTATÍSTICA

17 abril 2023

## **Lista 1**

Prof. Dr. George von Borries

Análise Multivariada 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

## 1) Exercício de Leitura:

Ler os seguintes artigos:

- Lozano, T. (2016) Data Science: an Emerging Discipline, CITAS.
- Kiliç, G.G.; Turanli, M.; Özden, Ü.H. (2019) Data Mining and Statistics in Data Science, International Social Science Studies Journal.

Os artigos acima são encontrados em <https://www.semanticscholar.org/paper/DATA-MINING-AND-STATISTICS%3A-WHAT%27S-THE-CONNECTIONFriedman/6b3cb8325ca437b6b005027e2e2c64adb99609c8>

- Zola, A. (2019) Machine Learning Engineer vs. Data Scientist. Disponível em <https://www.springboard.com/blog/data-science/machine-learning-engineer-vs-data-scientist/>

Fazer um resumo de três parágrafos e não mais de uma página indicando semelhanças e diferenças entre Estatística, Mineração de Dados e Ciência de Dados. Você pode incluir referência a textos mais recentes sobre estes temas.

O Artigo Lozano (2016) cita que alguns estatísticos vêem a estatística mais como uma ferramenta multidisciplinar e que deve se ater à cálculos probabilísticos para inferência do que propriamente para ser resignificada como ciência de dados. Citando Friedman (1998), diz que mineração de dados (*data mining*) seria mais especificamente sobre encontrar padrões e correspondências nos dados, com foco em grandes bases de dados. Com citação à Dhar (2013), diz que a grande diferença entre estatística e ciência de dados é o volume da informação. Enquanto estatísticos estão acostumados com um volume de dados que caberia em uma calculadora de bolso, a ciência de dados trabalharia com volume de dado gigantesco, necessitando de grande poder computacional. Cita a limitação de algumas técnicas estatísticas, que tendem a serem sensíveis demais com um volume de dados muito grande (*upscaling problem*). Mas mais importante, cita a comum baixa capacidade de estatísticos de lidarem com dados não numéricos e/ou não estruturados, que compoem grande parte da massa de dados produzido ultimamente. Para este problema, a engenharia de dados é importante para estruturar estes dados e torná-los trabalháveis. O texto do Kiliç (2019) foca mais em detalhar e descrever as técnicas estatísticas utilizadas pela ciência de dados para construção de modelos (PCA, regressão, clusterização e afins), bem como a discussão sobre a aplicabilidade dessas técnicas em grandes bases de dados. Ao fim do artigo, os autores tentam traçar correlações entre estatística e ciência de dados. O último artigo Zola (2019), traz mais uma abordagem mercadológica dos termos e as necessidades do mercado quanto a qualificação dos profissionais que pretendem trabalhar em áreas relacionadas a dados.

Ciencia de dados é grande volume de dados. Estatística tem as técnicas mais sofisticadas, mas por vezes anda devagar em relação as transformações da realidade do exponencial crescimento do poder computacional e das massas de dados. A engenharia de dados é a ciência que transforma um dado desestruturado em algo “trabalhável” com técnicas estatísticas numéricas, enquanto a mineração de dados foca em encontrar dentro de uma grande massa de dados, os fragmentos mais relevantes para uma análise.

A grande semelhança é que no fundo essas ciências são quase a mesma coisa e o profissional que exerce alguma dessas funções têm capacidade de compreender e até realizar as outras ciências correlatas; inclusive, num ambiente pequeno-corporativo, a realidade é que provavelmente será o mesmo profissional à executar todos esses serviços. Em Kiliç (2019), é citado que “O ponto em comum da estatística com a ciência de dados é aprender com dados, ou transformar dados em conhecimento” (Kuonen, 2004 apud Kiliç, 2019)

**2) Escolha uma área de pesquisa de interesse. Pesquise artigos publicados em revista indexadas e descreva (resumidamente) um exemplo incando o tipo de problema (ou problemas) entre os listados abaixo. Inclua referência bibliográfica e indique as características dos dados e estudo que relacionam ao tipo de problema (ou problemas) indicado.**

**a) Análise multivariada clássica ( $n < p$ ).**

Análises fatoriais e escalas psicométricas. “Com o uso dessa técnica, o investigador discerne os aspectos isolados da estrutura latente em uma matriz de dados e, logo depois, indica o quanto que cada variável é explicada por cada dimensão; ou seja, ele é capaz de descrever as correlações entre muitas variáveis observáveis em termos de um número menor de variáveis não observáveis (linearmente relacionadas com as variáveis originais observadas) denominadas dimensões ou variáveis latentes” (VASCONCELOS, Rafaela Balliana Bragança de. (2016) Apud Hair Jr. et al., 2009; Johnson e Wichern, 2007, p.21)[1]

**b) Mineração de dados (*Data Mining*) ( $n$  elevado).**

Uma das problemáticas em que a mineração de dados frequentemente esbarra é na privacidade dos dados. “Given a database containing both”public” information and “sensitive” information, we can use data mining to search for inference paths from the public to the sensitive information. In other words, a rule with antecedents in the public information and consequents in the sensitive information points to a potential problem. Clustering the public information could conceivably bring together the antecedents. This by itself does not pose a problem, but knowing sensitive information about one of the instances could then be conjectured (correctly) to apply to all of them (greatly expanding the “leak” of sensitive information). Proper use of data mining technology can be used to detect such situations. We can then apply techniques from the previous section to ensure that the data cannot be clustered in meaningful ways.” (CLIFTON & MARKS, 1996, p. 18)[2]

**c) Aprendizado estatístico (*Statistical Learning*) ( $p$  elevado).**

Talvez um dos campos que mais difira da análise multivariada clássica, SLT (*Statistical Learning Theory*) trabalha com o cada vez mais demandado desafio de extrair informações de bases de dados com  $p$  elevado e  $n$  “pequeno”. Conforme elocidado pelo Prof. Von Borries em sala de aula, uma das aplicações recorrentes da SLT é em campos ge genética. “SLT effectively describes statistical estimation with small samples. Naturally, as a special case, this theory includes classical statistical methods which are developed for large samples and strict parametric assumptions. SLT is based on principle of empirical risk minimization(ERM) and VC dimension. The empirical risk is the average risk for the training data. This is minimized by choosing the appropriate parameters.” (JUN & OH, 2006, p. 249) [3] [3.1]

[3.1] **Nota:**  $VC = \text{Vapnik-Chervonenkis}$

**d) Reconhecimento de padrões.**

Outro procedimento cada dia mais requisitado na área estatística, o reconhecimento de padrões é essencial quando se trabalha com conjuntos de dados muito grandes ( $n \rightarrow \infty$ ) e/ou desestruturados. O conceito é retirar conclusões de conjuntos em que técnicas clássicas de inferência não funcionam/requisitam poder computacional demasiadamente grande. “A pattern recognition system can be regarded as a process that allows it to cope with real and noisy data. Whether the decision made by the system is right or not mainly depending on the decision make by the human expert.” (LIU et al., 2006, p.58) [4]

**e) *Data Science*.**

*Data science* é como a adequação da estatística clássica ao poder computacional do século XXI. Tomadas de decisões são cada vez mais aperfeiçoadas e possíveis se analisados grandes conjuntos de dados, o que torna o aproveitamento desses para os setores estratégicos cada vez mais indispensável. “Data-driven decisions are better decisions — it’s as simple as that. Using big data enables managers to decide on the

basis of evidence rather than intuition. For that reason it has the potential to revolutionize management.” (MCAFEE et al., 2012, p.64)

#### Referências:

- [1] VASCONCELOS, Rafaela Balliana Bragança de. **Avaliação psicométrica de um instrumento para mensuração de autoeficácia**. Monografia (Bacharelado em Estatística) – Departamento de Estatística, Universidade de Brasília. Brasília, p. 21. 2016.
- [2] CLIFTON, Chris; MARKS, Don. Security and privacy implications of data mining. In: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. 1996. p. 15-19.
- [3] JUN, Sung-Hae; OH, Kyung-Whan. An evolutionary statistical learning theory. International Journal of Computational Intelligence, v. 3, n. 3, p. 249-256, 2006.
- [4] LIU, Jie; SUN, Jigui; WANG, Shengsheng. Pattern recognition: An overview. IJCSNS International Journal of Computer Science and Network Security, v. 6, n. 6, p. 57-61, 2006.
- [5] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. Big data: the management revolution. Harvard business review, v. 90, n. 10, p. 60-68, 2012.

### 3) Descreva cada um dos problemas encontrados na análise multivariada, encontre um exemplo de caso real e indique uma falha nas técnicas estatísticas tradicionais.

#### a) *Mining* (mineração, $n$ muito elevado).

*Data mining* é a mineração de grandes massas de dados, afim de encontrar correlações e padrões de interesse.

O principal problema ocorrido nesses casos é a não identificação do *software* correto para trabalhar com grandes massas de dados. O *R* por exemplo, por suas especificades de programação pouco eficiente e de ser uma linguagem orientada a objetos em memória *RAM*, não costuma lidar bem com volumes de dados  $> 5\text{Gb}$ . Além disso, testes como o  $\chi^2$  tendem a serem sensíveis demais quando  $n \rightarrow \infty$ , rejeitando  $h_0$  sob pequenas flutuações.

#### b) *Scalability* (escalabilidade).

A escalabilidade tende a ter um problema parecido com a mineração de dados, visto que pode esbarrar na sensibilidade de alguns testes para grandes massas de dados. Além disso, como a ciência estatística intrinsecamente trabalha com erro (leia-se; variações, variância), por vezes um resultado encontrado não pode ser replicado exatamente, não garantindo assim a escalabilidade automática.

#### c) *High Dimensional Data* (dados em alta dimensão, $n > p$ ).

Esse talvez seja um dos tópicos mais difíceis para a estatística, ou seja, trabalhar com conjuntos de poucas observações mas muitas variáveis pode ser um desafio no sentido de não haver parâmetros sólidos a serem buscados, além da dificuldade de encontrar padrões e correlações robustas quando não há muito dado para se comparar.

#### d) Pequenas amostras.

O principal diferencial da ciência estatística é se aprimorar em dominar a inferência a partir de amostras. Muitas de nossas técnicas são extremamente sofisticadas e teoricamente carregadas justamente a fim de ser possível analisar evidências populacionais sobre pequenas amostras. Porém, quando  $n \rightarrow 0$ , diversos testes estatísticos como os de normalidade (*Kolmogorov-Smirnov*, *Lilliefors*, etc) tendem a ser muito conservadores com poucos dados; ou seja, não rejeitam  $h_0$  em situações duvidosas.

#### e) *Curse of Dimensionality* (Problema de dimensionalidade).

Similar aos dados em alta dimensão, o problema da dimensionalidade costuma ocorrer quando são estudadas diversas variáveis em uma amostra; o que compromete a replicabilidade do estudo em uma outra amostra de igual tamanho visto que a quantidade de características latentes por vezes torna inviável a homogeneização dessas amostras.

Em análise fatorial, por exemplo, é muito importante a aplicação de técnicas como a análise de componentes principais, entre outras técnicas, com interesse de reduzir a dimensionalidade do conjunto de dados.

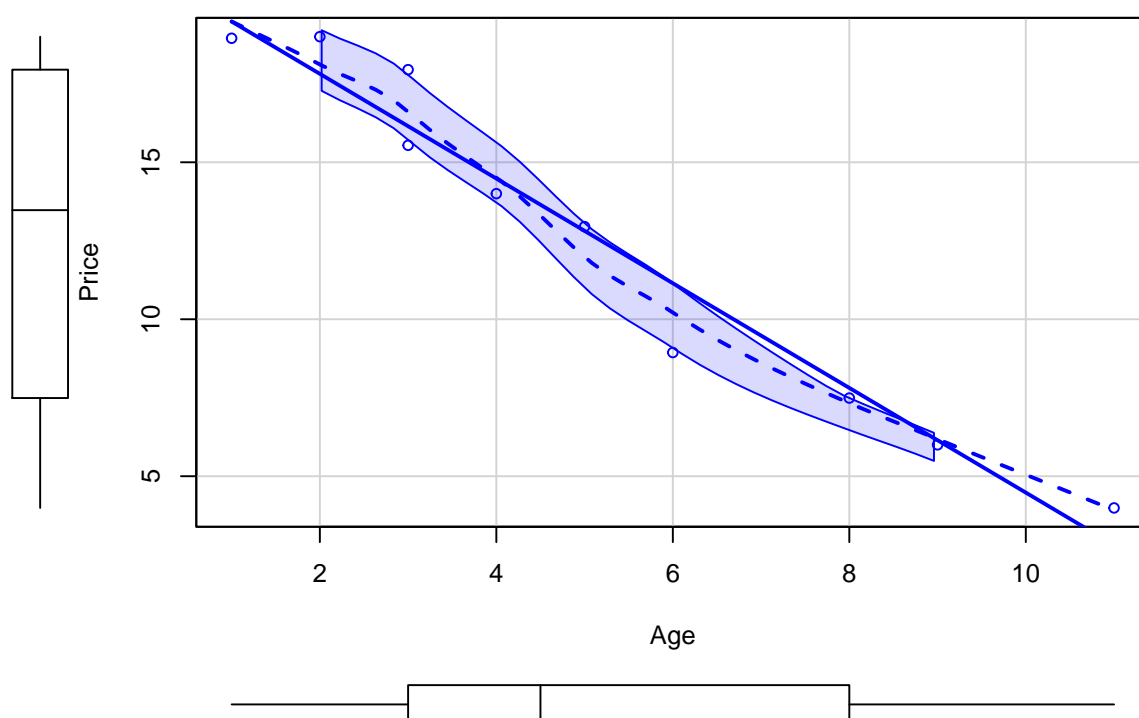
Conforme citado em aula pelo Prof. von Borries, por vezes conjuntos  $k$  dimensionais com  $k \rightarrow \infty$  contém uma quantidade  $\epsilon$  de dimensões que correspondem a 80, 90% da informação, sendo o restante responsável pela menor parte. É trabalho do estatístico analisar e tratar estes conjuntos a fim de tornar a análise trabalhável e replicável dado características latentes de real interesse.

4) Fazer os seguintes exercícios do capítulo 1 de *Johnson e Wichern*: 1.2, 1.6, 1.14 e 1.22.

1.2) A mourning newspaper list the following used-car prices for a foreign compact with age (Age) measured in years and selling price (Price) measured in thousand of dollars:

Age	1.00	2	3.00	3.00	4	5.00	6.00	8.00	9	11.00
Price	18.95	19	17.95	15.54	14	12.95	8.94	7.49	6	3.99

a) Construct a scatter plot of the data and marginal dot diagrams.



b) Infer the sign of the sample covariance  $s_{12}$  from the scatter plot.

Baseado no gráfico de dispersão, a covariância desses dados é alta e negativa.

A covariância para uma amostra é definida como:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

A covariância calculada para este conjunto de dados é -17.71

O que confirma o palpite inferido sobre o gráfico de dispersão.

c) Compute the sample means  $\bar{x}_1$  and  $\bar{x}_2$  and the sample variances  $s_{11}$  and  $s_{22}$ . Compute the sample covariance  $s_{12}$  and the sample correlation coefficient  $r_{12}$ . Interpret these quantities.

	Age	Price
Média	5.20	12.48

	Age	Price
Variância	10.62	30.85

A covariância da amostra  $s_{12}$  observada foi -17.71; enquanto o coeficiente de correlação da amostra  $r_{12}$  observado foi -0.98. Esta covariância é visualmente analisável pelo gráfico de dispersão conforme questão anterior, enquanto o valor observado na correlação indica que ela é muito forte à esquerda, quase absoluta.

- d) Display the sample mean array  $\bar{x}$ , the sample variance-covariance array  $S_n$ , and the sample correlation array  $\mathbf{R}$  using (1-8).

Matriz de médias

	Age	Price
Média	5.2	12.48

Matriz de correlação

	Age	Price
Age	1.0000000	-0.9782684
Price	-0.9782684	1.0000000

Matriz de covariâncias

	Age	Price
Age	10.62222	-17.71022
Price	-17.71022	30.85437

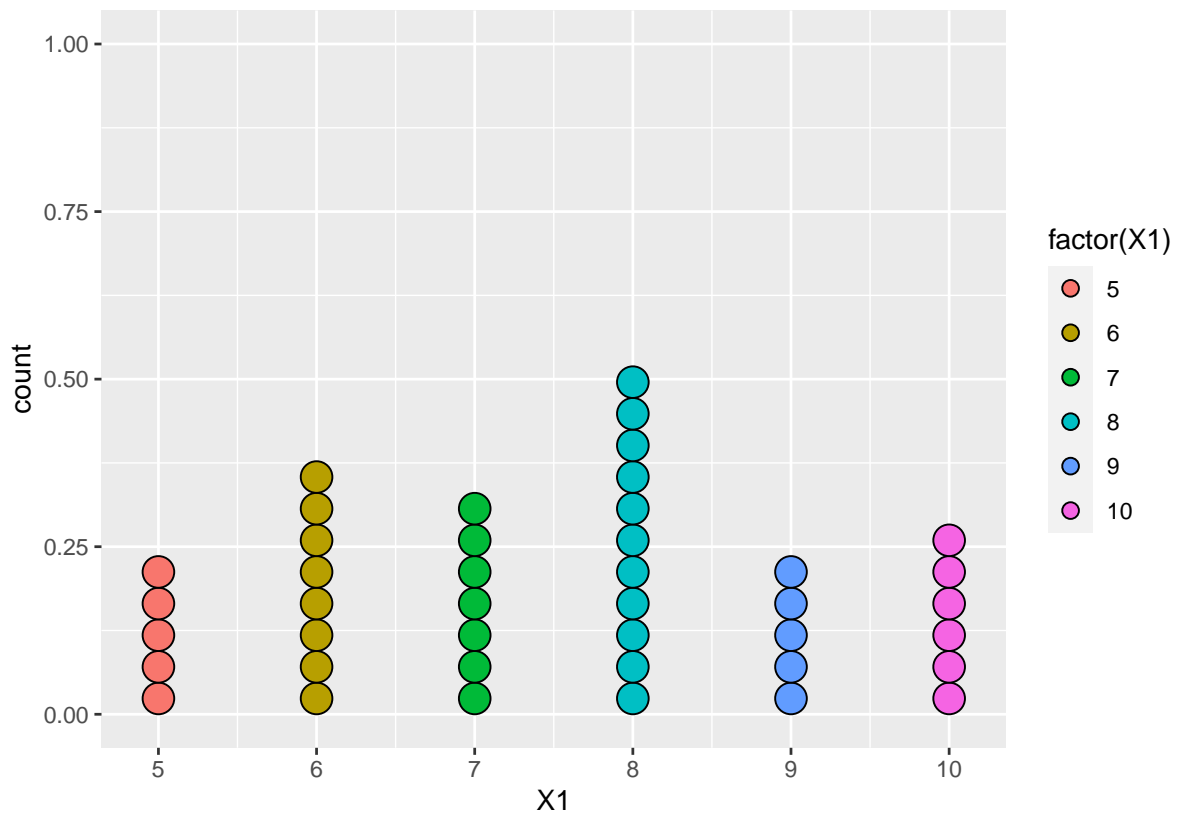
**1.6) The data in Table 1.5 are 42 measurements on air-pollution variables recorded at 12:00 noon in the Los Angeles area on different days.**

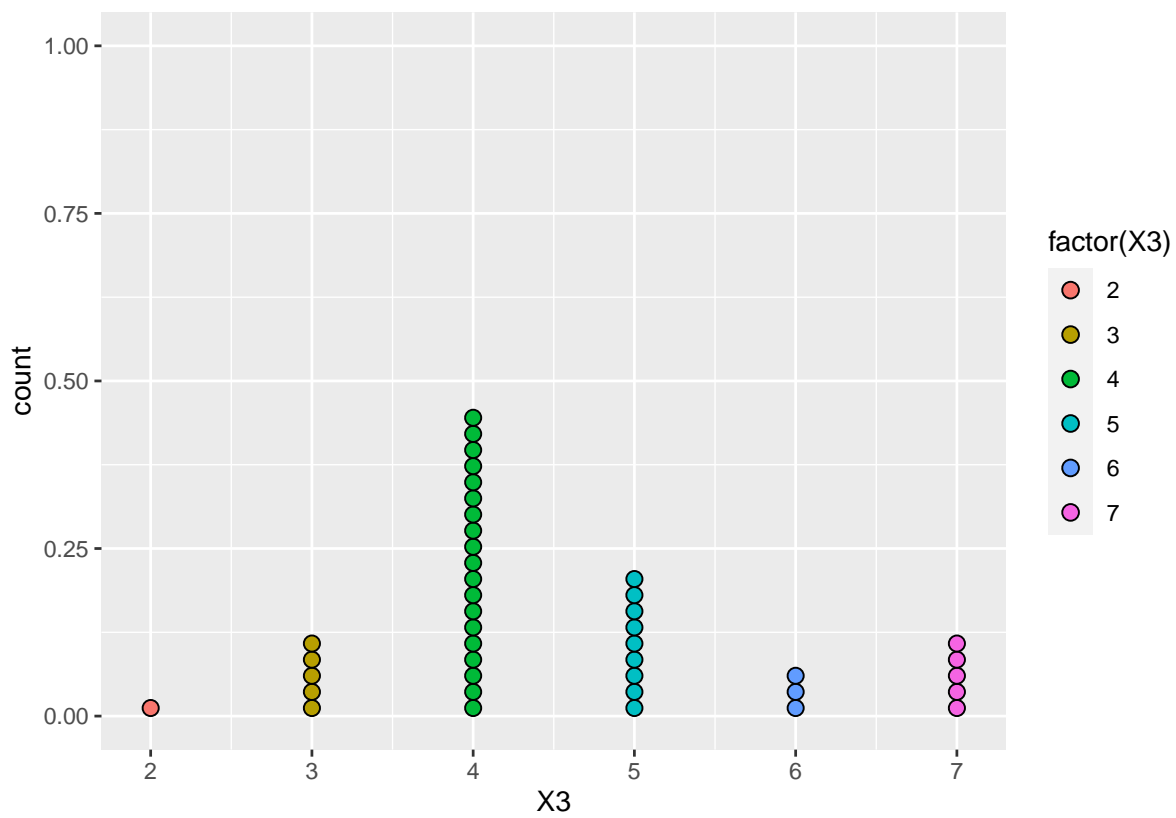
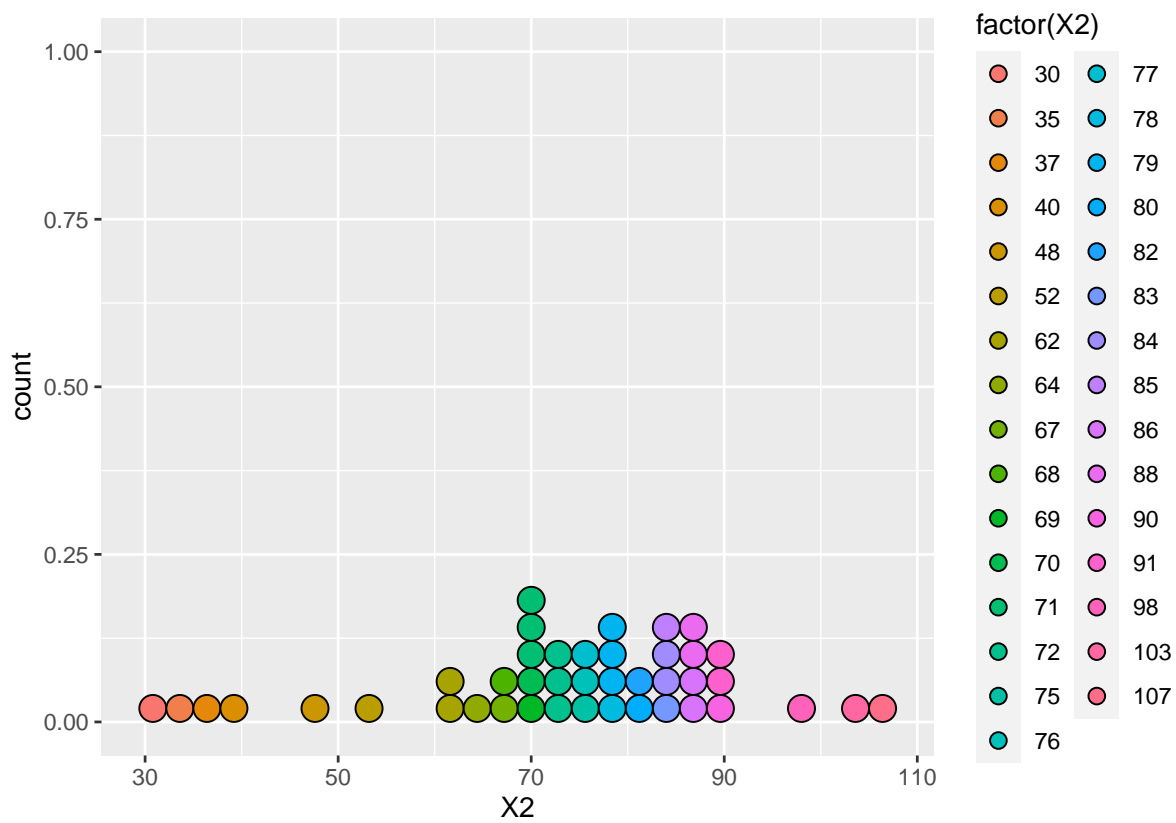
X1	X2	X3	X4	X5	X6	X7
8	98	7	2	12	8	2
7	107	4	3	9	5	3
7	103	4	3	5	6	3
10	88	5	2	8	15	4
6	91	4	2	8	10	3
8	90	5	2	12	12	4
9	84	7	4	12	15	5
5	72	6	4	21	14	4
7	82	5	1	11	11	3
8	64	5	2	13	9	4
6	71	5	4	10	3	3
6	91	4	2	12	7	3
7	72	7	4	18	10	3
10	70	4	2	11	7	3
10	72	4	1	8	10	3
9	77	4	1	9	10	3
8	76	4	1	7	7	3
8	71	5	3	16	4	4
9	67	4	2	13	2	3
9	69	3	3	9	5	3
10	62	5	3	14	4	4

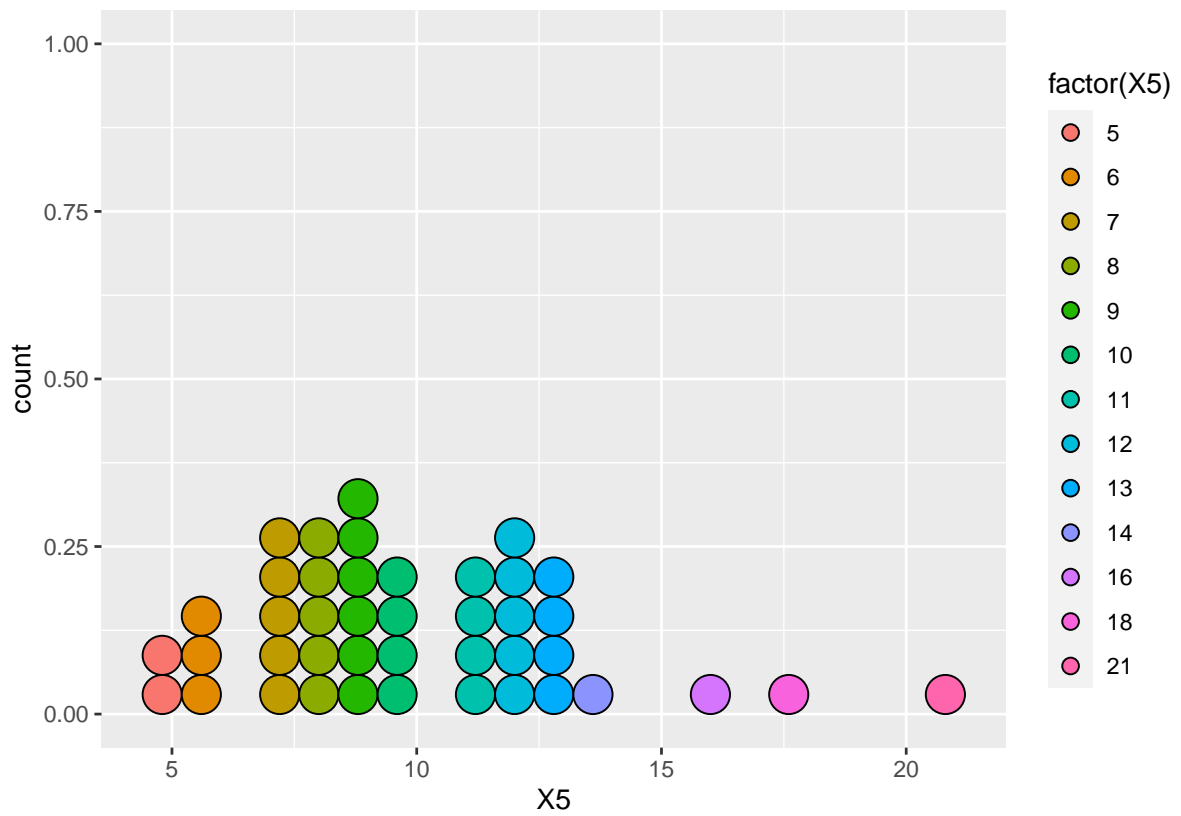
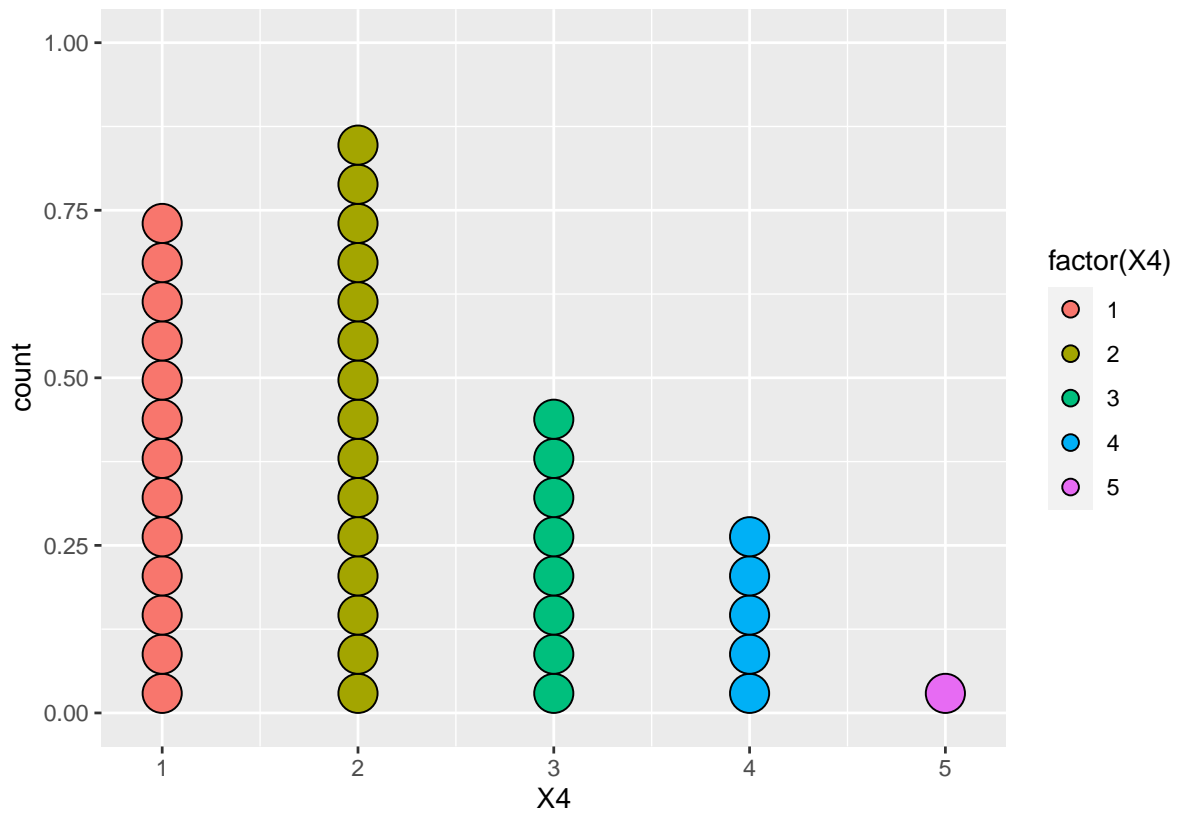


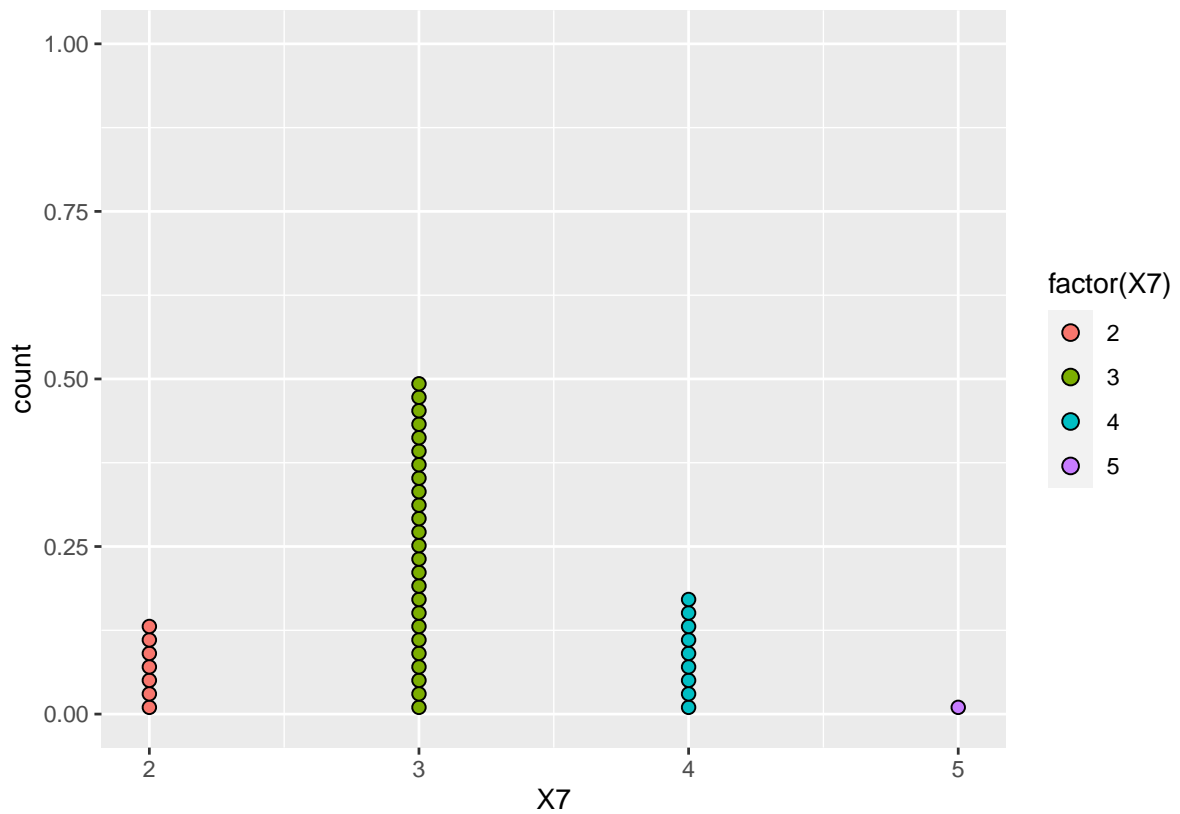
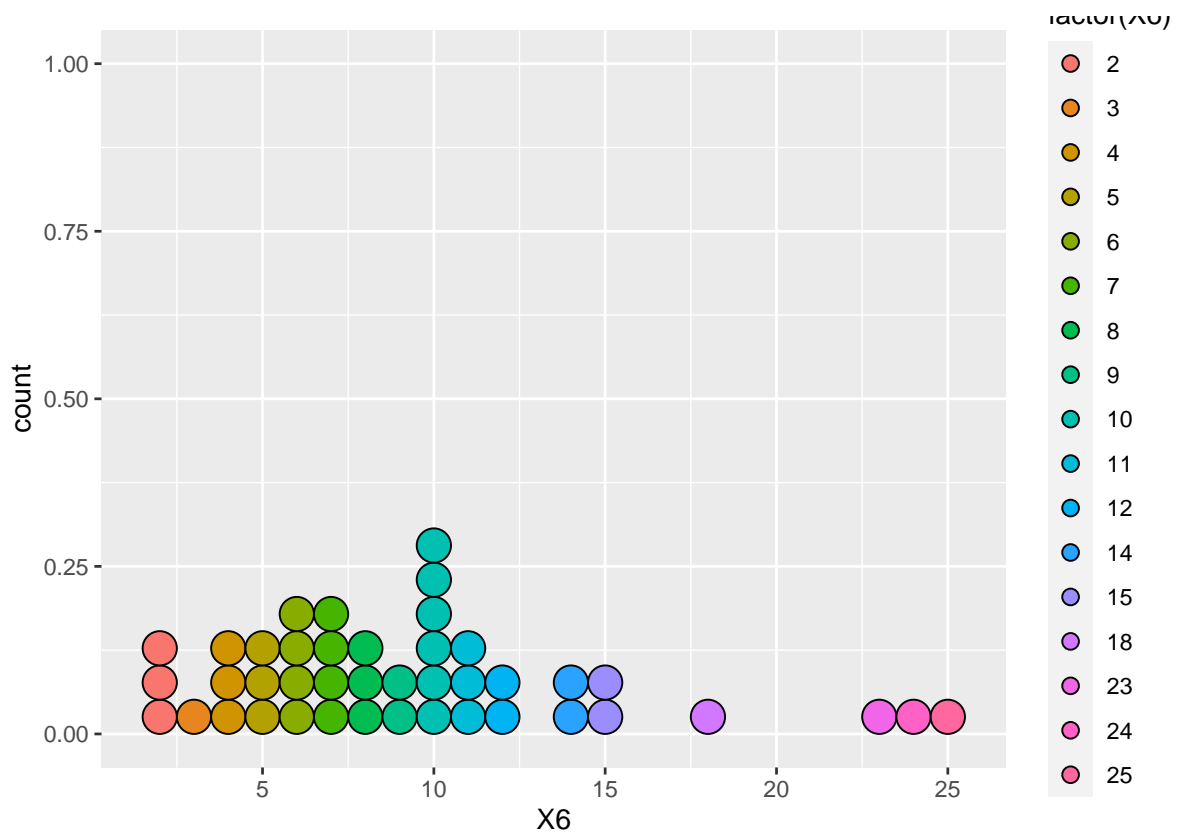
X1	X2	X3	X4	X5	X6	X7
9	88	4	2	7	6	3
8	80	4	2	13	11	4
5	30	3	3	5	2	3
6	83	5	1	10	23	4
8	84	3	2	7	6	3
6	78	4	2	11	11	3
8	79	2	1	7	10	3
6	62	4	3	9	8	3
10	37	3	1	7	2	3
8	71	4	1	10	7	3
7	52	4	1	12	8	4
5	48	6	5	8	4	3
6	75	4	1	10	24	3
10	35	4	1	6	9	2
8	85	4	1	9	10	2
5	86	3	1	6	12	2
5	86	7	2	13	18	2
7	79	7	4	9	25	3
7	79	5	2	8	6	2
6	68	6	2	11	14	3
8	40	4	3	6	5	2

a) Plot the marginal dot diagrams for all the variables.









b) Construct the  $\bar{x}$ ,  $S_n$ , and  $\mathbf{R}$  arrays, and interpret the entries in  $\mathbf{R}$ .  
 Matriz de médias( $\bar{x}$ ):

	Médias
X1	7.500000
X2	73.857143
X3	4.547619
X4	2.190476
X5	10.047619
X6	9.404762
X7	3.095238

	X1	X2	X3	X4	X5	X6	X7
X1	1.0000000	-0.1014419	-0.1938032	-0.2695426	-0.1098249	-0.2535928	0.1560979
X2	-0.1014419	1.0000000	0.1827934	-0.0735691	0.1157320	0.3191237	0.0520104
X3	-0.1938032	0.1827934	1.0000000	0.5021525	0.5565838	0.4109288	0.1660323
X4	-0.2695426	-0.0735691	0.5021525	1.0000000	0.2968981	-0.1339521	0.2347043
X5	-0.1098249	0.1157320	0.5565838	0.2968981	1.0000000	0.1666422	0.4477678
X6	-0.2535928	0.3191237	0.4109288	-0.1339521	0.1666422	1.0000000	0.1544506
X7	0.1560979	0.0520104	0.1660323	0.2347043	0.4477678	0.1544506	1.0000000

Matriz de variâncias( $S_n$ ):

	Variâncias
X1	2.5000000
X2	300.5156794
X3	1.5220674
X4	1.1823461
X5	11.3635308
X6	30.9785134
X7	0.4785134

Matriz de covariâncias(Cov):

	X1	X2	X3	X4	X5	X6	X7
X1	2.5000000	-2.7804878	-0.3780488	-0.4634146	-0.5853659	-2.2317073	0.1707317
X2	-2.7804878	300.5156794	3.9094077	-1.3867596	6.7630662	30.7909408	0.6236934
X3	-0.3780488	3.9094077	1.5220674	0.6736353	2.3147503	2.8217189	0.1416957
X4	-0.4634146	-1.3867596	0.6736353	1.1823461	1.0882695	-0.8106852	0.1765389
X5	-0.5853659	6.7630662	2.3147503	1.0882695	11.3635308	3.1265970	1.0441347
X6	-2.2317073	30.7909408	2.8217189	-0.8106852	3.1265970	30.9785134	0.5946574
X7	0.1707317	0.6236934	0.1416957	0.1765389	1.0441347	0.5946574	0.4785134

Matriz de correlações( $\mathbf{R}$ ):

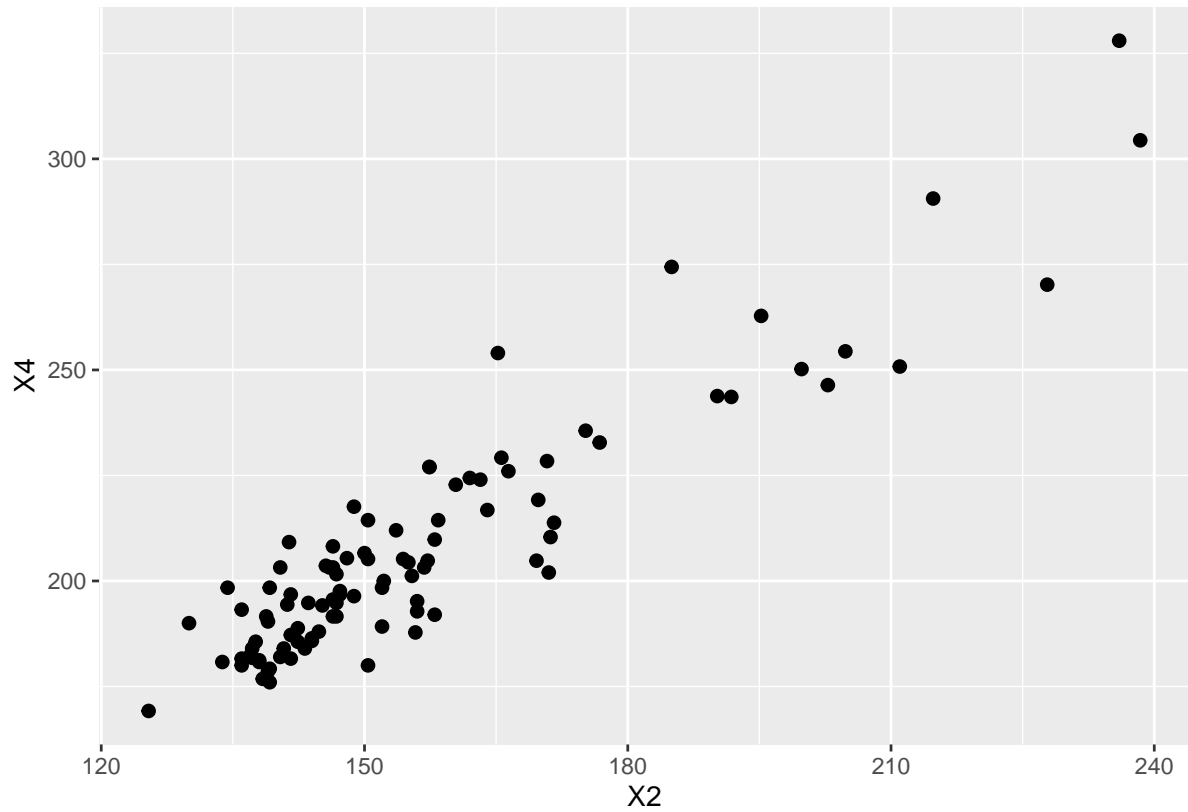
	X1	X2	X3	X4	X5	X6	X7
X1	1.0000000	-0.1014419	-0.1938032	-0.2695426	-0.1098249	-0.2535928	0.1560979
X2	-0.1014419	1.0000000	0.1827934	-0.0735691	0.1157320	0.3191237	0.0520104
X3	-0.1938032	0.1827934	1.0000000	0.5021525	0.5565838	0.4109288	0.1660323
X4	-0.2695426	-0.0735691	0.5021525	1.0000000	0.2968981	-0.1339521	0.2347043
X5	-0.1098249	0.1157320	0.5565838	0.2968981	1.0000000	0.1666422	0.4477678
X6	-0.2535928	0.3191237	0.4109288	-0.1339521	0.1666422	1.0000000	0.1544506
X7	0.1560979	0.0520104	0.1660323	0.2347043	0.4477678	0.1544506	1.0000000

1.14) Table contains some raw data. Two different visual stimuli (S1 and S2) produced responses in both the left eye (L) and the right eye (R) of subjects in the study groups. The values recorded in the table include x1 (subject's age); x2 (total response of both eyes to stimulus S1, that is, S1L + S1R); x3 (difference between responses of eyes to stimulus S1, |S1L-S1R|); and so forth.

X1	X2	X3	X4	X5	X6
18	152.0	1.6	198.4	0.0	0
19	138.0	0.4	180.8	1.6	0
20	144.0	0.0	186.4	0.8	0
20	143.6	3.2	194.8	0.0	0
20	148.8	0.0	217.6	0.0	0
21	141.6	0.8	181.6	0.8	0
21	136.0	1.6	180.0	0.8	0
21	137.6	1.6	185.6	3.2	0
22	140.4	3.2	182.0	3.2	0
22	137.2	0.0	181.8	0.2	0
22	125.4	1.0	169.2	0.0	0
22	142.4	4.8	185.6	0.0	0
22	150.4	0.0	214.4	3.2	0
22	145.6	1.6	203.6	5.2	0
23	147.2	3.2	196.8	1.6	0
23	139.2	1.6	179.2	0.0	0
24	169.6	0.0	204.8	0.0	0
24	139.2	1.6	176.0	3.2	0
24	153.6	0.0	212.0	0.8	0
25	146.8	0.0	194.8	3.2	0
25	139.2	1.6	198.4	3.2	0
25	136.0	1.6	181.6	2.4	0
26	138.8	1.6	191.6	0.0	0
26	150.4	0.0	205.2	0.4	0
26	139.0	1.4	178.6	0.2	0
27	133.8	0.2	180.8	0.0	0
27	139.0	1.8	190.4	1.6	0
28	136.0	1.6	193.2	3.6	0
28	146.4	0.8	195.6	2.8	0
29	145.2	4.8	194.2	3.8	0
29	146.4	0.8	208.2	0.2	0
29	138.0	2.8	181.2	0.4	0
30	148.8	1.6	196.4	1.6	0
31	137.2	0.0	184.0	0.0	0
31	147.2	0.0	197.6	0.8	0
32	144.0	0.0	185.8	0.2	0
32	156.0	0.0	192.8	2.4	0
34	137.0	0.2	182.4	0.0	0
35	143.2	2.4	184.0	1.6	0
36	141.6	0.8	187.2	1.6	0
37	152.0	1.6	189.2	2.8	0
39	157.4	3.4	227.0	2.6	0
40	141.4	0.6	209.2	1.6	0
42	156.0	2.4	195.2	3.2	0
43	150.4	1.6	180.0	0.8	0
43	142.4	1.6	188.8	0.0	0
46	158.0	2.0	192.0	3.2	0
48	130.0	3.6	190.0	0.4	0
49	152.2	1.4	200.0	4.8	0
49	150.0	3.2	206.6	2.2	0

X1	X2	X3	X4	X5	X6
50	146.4	2.4	191.6	2.8	0
54	146.0	1.2	203.2	1.6	0
55	140.8	0.0	184.0	1.6	0
56	140.4	0.4	203.2	1.6	0
56	155.8	3.0	187.8	2.6	0
56	141.6	0.8	196.8	1.6	0
57	144.8	0.8	188.0	0.8	0
57	146.8	3.2	191.6	0.0	0
59	176.8	2.4	232.8	0.8	0
60	171.0	1.8	202.0	3.6	0
60	163.2	0.0	224.0	0.0	0
60	171.6	1.2	213.8	3.4	0
60	146.4	4.0	203.2	4.8	0
62	146.8	3.6	201.6	3.2	0
67	154.4	2.4	205.2	6.0	0
69	171.2	1.6	210.4	0.8	0
73	157.2	0.4	204.8	0.0	0
74	175.2	5.6	235.6	0.4	0
79	155.0	1.4	204.4	0.0	0
23	148.0	0.8	205.4	0.6	1
25	195.2	3.2	262.8	0.4	1
25	158.0	8.0	209.8	12.2	1
28	134.4	0.0	198.4	3.2	1
29	190.2	14.2	243.8	10.6	1
29	160.4	18.4	222.8	31.2	1
31	227.8	90.2	270.2	83.0	1
34	211.0	3.0	250.8	5.2	1
35	204.8	12.8	254.4	11.2	1
36	141.2	6.8	194.4	21.6	1
39	157.4	3.4	227.0	2.6	1
42	166.4	0.0	226.0	0.0	1
43	191.8	35.4	243.6	40.8	1
44	156.8	0.0	203.2	0.0	1
44	202.8	29.2	246.4	24.8	1
44	165.2	18.4	254.0	46.4	1
45	162.0	5.6	224.4	8.8	1
45	138.4	0.8	176.8	4.0	1
45	158.4	1.6	214.4	0.0	1
46	155.4	1.8	201.2	6.0	1
46	214.8	9.2	290.6	0.6	1
47	185.0	19.0	274.4	7.6	1
48	236.0	20.0	328.0	0.0	1
57	170.8	24.0	228.4	33.6	1
57	165.6	16.8	229.2	15.6	1
58	238.4	8.0	304.4	6.0	1
58	164.0	0.8	216.8	0.8	1
58	169.8	0.0	219.2	1.6	1
59	199.8	4.6	250.2	1.0	1

- a) Plot the two-dimensional scatter diagram for the variables  $x_2$  and  $x_4$  for the multiple-sclerosis group. Comment on the appearance of the diagram.



Observando o resultado do gráfico de dispersão, notamos que aparenta existir uma correlação linear positiva entre as duas variáveis.

- b) Compute the  $\bar{x}$ ,  $S_n$ , and  $\mathbf{R}$  arrays for the non-multiple-sclerosis and multiple-sclerosis groups separately.

Matriz de médias das colunas para o conjunto dos que não tem esclerose múltipla

	Médias
X1	37.985507
X2	147.289855
X3	1.562319
X4	195.602899
X5	1.620290

Matriz de médias das colunas para o conjunto dos que tem esclerose múltipla

	Médias
X1	42.06897
X2	178.26897
X3	12.27586
X4	236.93103
X5	13.08276

Matriz de variâncias das colunas para o conjunto dos que não tem esclerose múltipla

	Variâncias
X1	277.632140



Variâncias	
X2	112.294748
X3	1.805030
X4	185.228815
X5	2.355465

Matriz de variâncias das colunas para o conjunto dos que tem esclerose múltipla

Variâncias	
X1	121.1379
X2	844.6808
X3	317.2640
X4	1180.0322
X5	351.0472

Matriz de correlações para o conjunto dos que não tem esclerose múltipla

	X1	X2	X3	X4	X5
X1	1.0000000	0.5402894	0.2394891	0.4573917	0.1267563
X2	0.5402894	1.0000000	0.1240685	0.7404170	0.1255725
X3	0.2394891	0.1240685	1.0000000	0.1222209	0.2431491
X4	0.4573917	0.7404170	0.1222209	1.0000000	0.1125594
X5	0.1267563	0.1255725	0.2431491	0.1125594	1.0000000

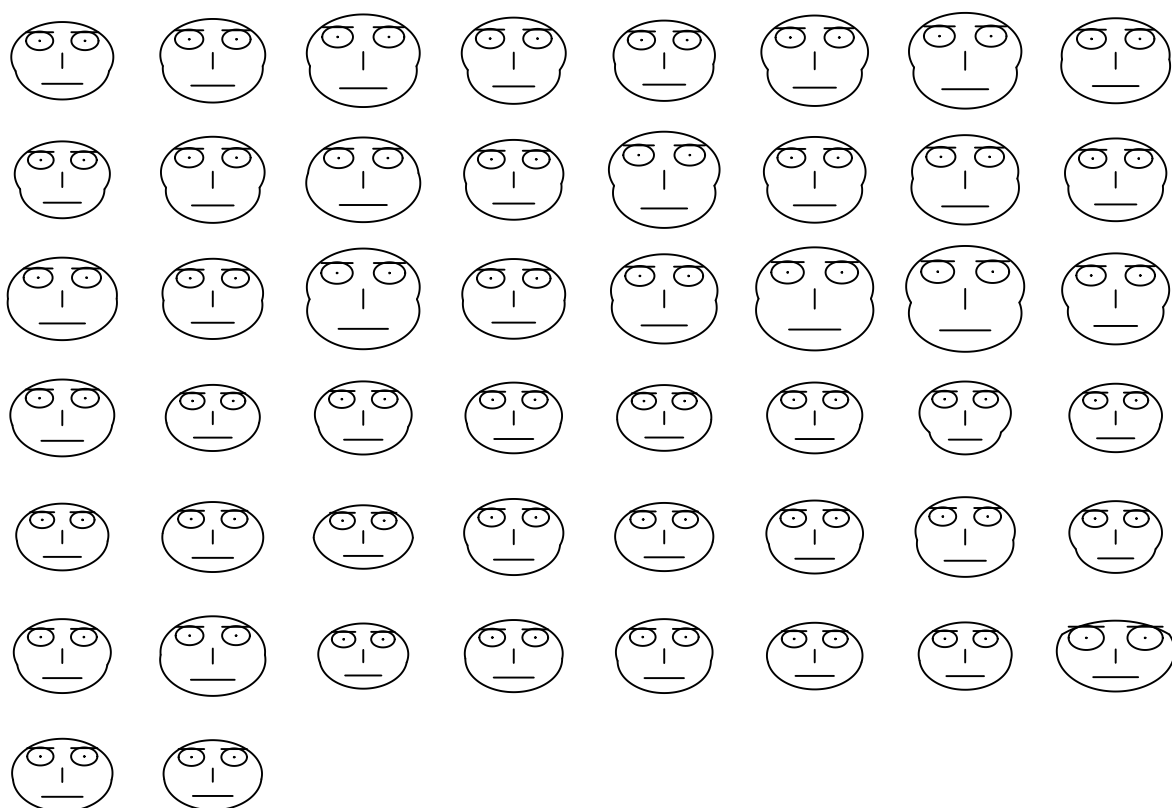
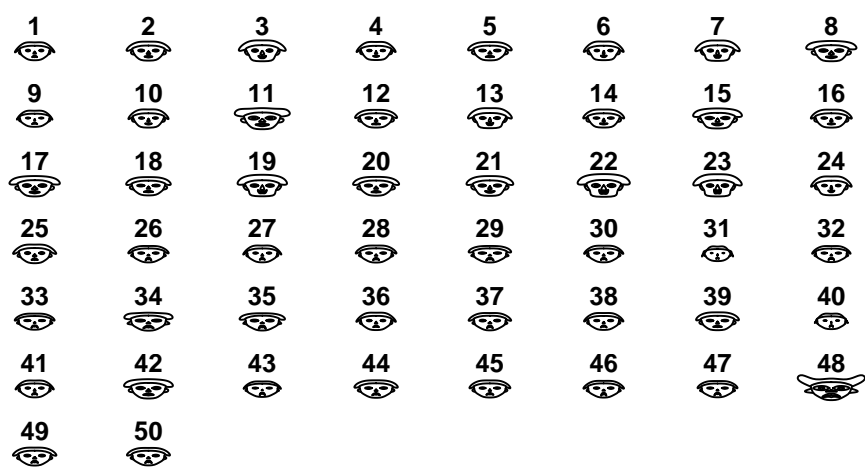
Matriz de correlações para o conjunto dos que tem esclerose múltipla

	X1	X2	X3	X4	X5
X1	1.0000000	0.1650468	-0.1031393	0.1802078	-0.1446065
X2	0.1650468	1.0000000	0.4722334	0.9139010	0.1960632
X3	-0.1031393	0.4722334	1.0000000	0.3797643	0.8909017
X4	0.1802078	0.9139010	0.3797643	1.0000000	0.1260019
X5	-0.1446065	0.1960632	0.8909017	0.1260019	1.0000000

**1.22) Refer to the oxygen-consumption data. Using appropriate computer software,**

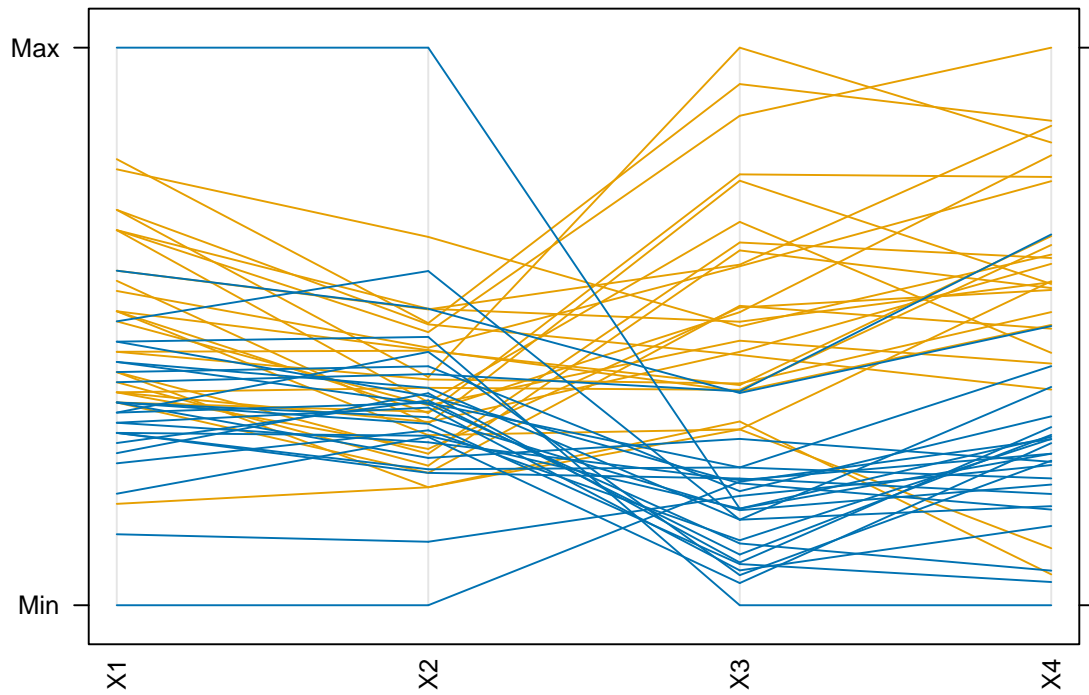
```
##
## -- Column specification -----
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_character()
## )
```

- a) View the entire data set in three dimensions employing various combinations of three variables to represent the coordinate axes. Begin with the  $x_1, x_2, x_3$  space.



##	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 1	5.1	3.5	1.4	0.2	setosa
## 2	4.9	3.0	1.4	0.2	setosa
## 3	4.7	3.2	1.3	0.2	setosa
## 4	4.6	3.1	1.5	0.2	setosa

```
## 5      5.0      3.6      1.4      0.2  setosa
## 6      5.4      3.9      1.7      0.4  setosa
```



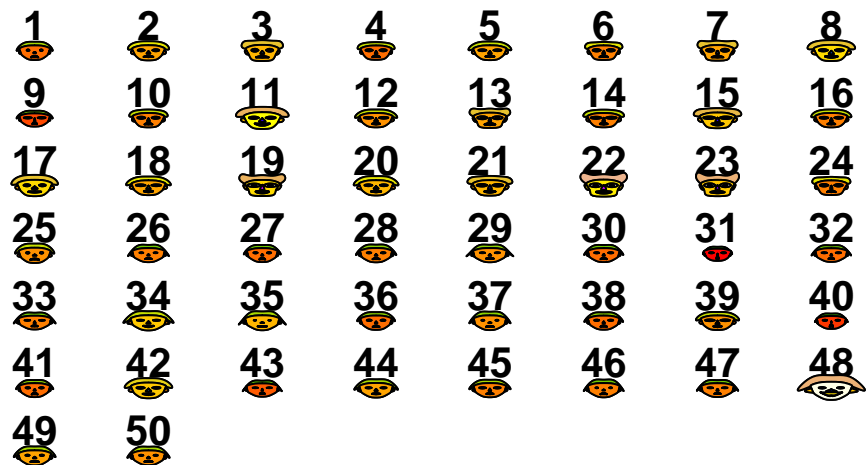
1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50						

```
## effect of variables:
## modified item      Var
## "height of face"  "X1"
```

```

## "width of face" "X2"
## "structure of face" "X3"
## "height of mouth" "X4"
## "width of mouth" "X1"
## "smiling" "X2"
## "height of eyes" "X3"
## "width of eyes" "X4"
## "height of hair" "X1"
## "width of hair" "X2"
## "style of hair" "X3"
## "height of nose" "X4"
## "width of nose" "X1"
## "width of ear" "X2"
## "height of ear" "X3"

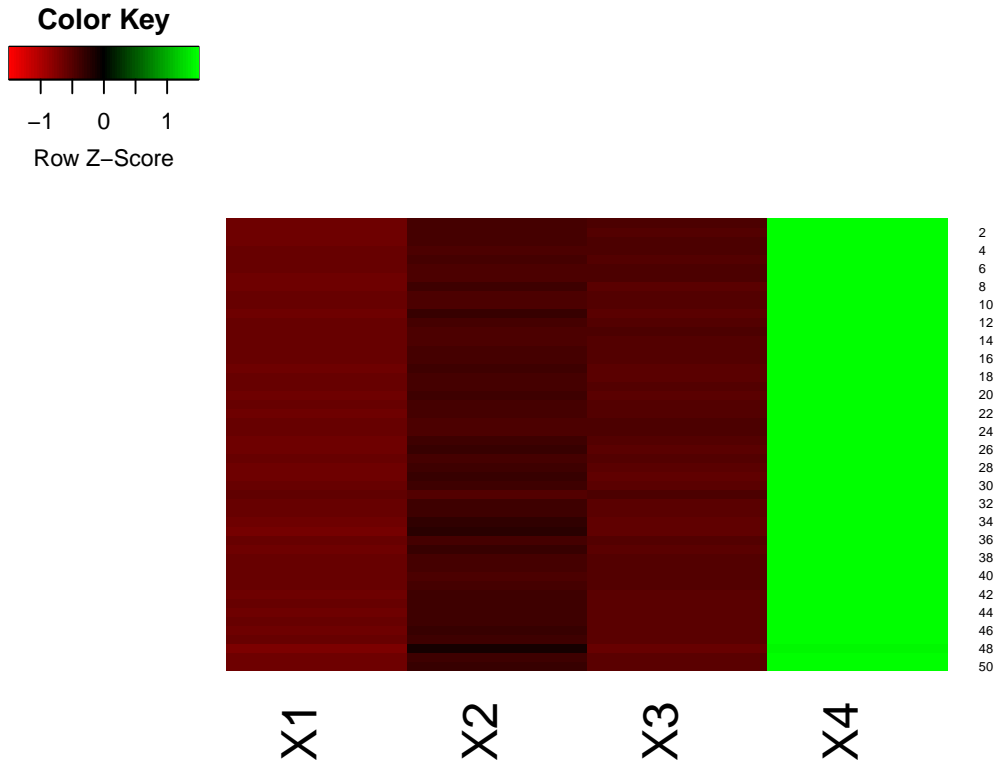
```



```

## effect of variables:
## modified item      Var
## "height of face"  "X1"
## "width of face"   "X2"
## "structure of face" "X3"
## "height of mouth" "X4"
## "width of mouth"  "X1"
## "smiling"         "X2"
## "height of eyes"  "X3"
## "width of eyes"   "X4"
## "height of hair"  "X1"
## "width of hair"   "X2"
## "style of hair"   "X3"
## "height of nose"  "X4"
## "width of nose"   "X1"
## "width of ear"    "X2"
## "height of ear"   "X3"

```



b) Check this data set for outliers.

Primeiramente, irei fixar um nível de significância  $\alpha = 0.05$ . Testarei a normalidade dos dados de cada coluna utilizando o teste de *Shapiro-Wilk*; sob as hipóteses:

$$\begin{cases} H_0 & : \text{Os dados seguem distribuição normal} \\ H_1 & : \text{" " NÃO " " " } \end{cases}$$

Para as colunas em que a hipótese nula não for rejeitada, irei aplicar o teste de *Grubbs* para identificação de *outliers*.

```
##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.96894, p-value = 0.2094

##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.90363, p-value = 0.0006346

##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.94611, p-value = 0.02368

##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.97328, p-value = 0.3131
```

```

##
## Grubbs test for one outlier
##
## data:  .
## G = 3.04029, U = 0.80751, p-value = 0.03592
## alternative hypothesis: highest value 0.66 is an outlier

##
## Grubbs test for one outlier
##
## data:  .
## G = 4.17389, U = 0.63721, p-value = 9.242e-05
## alternative hypothesis: highest value 11.05 is an outlier

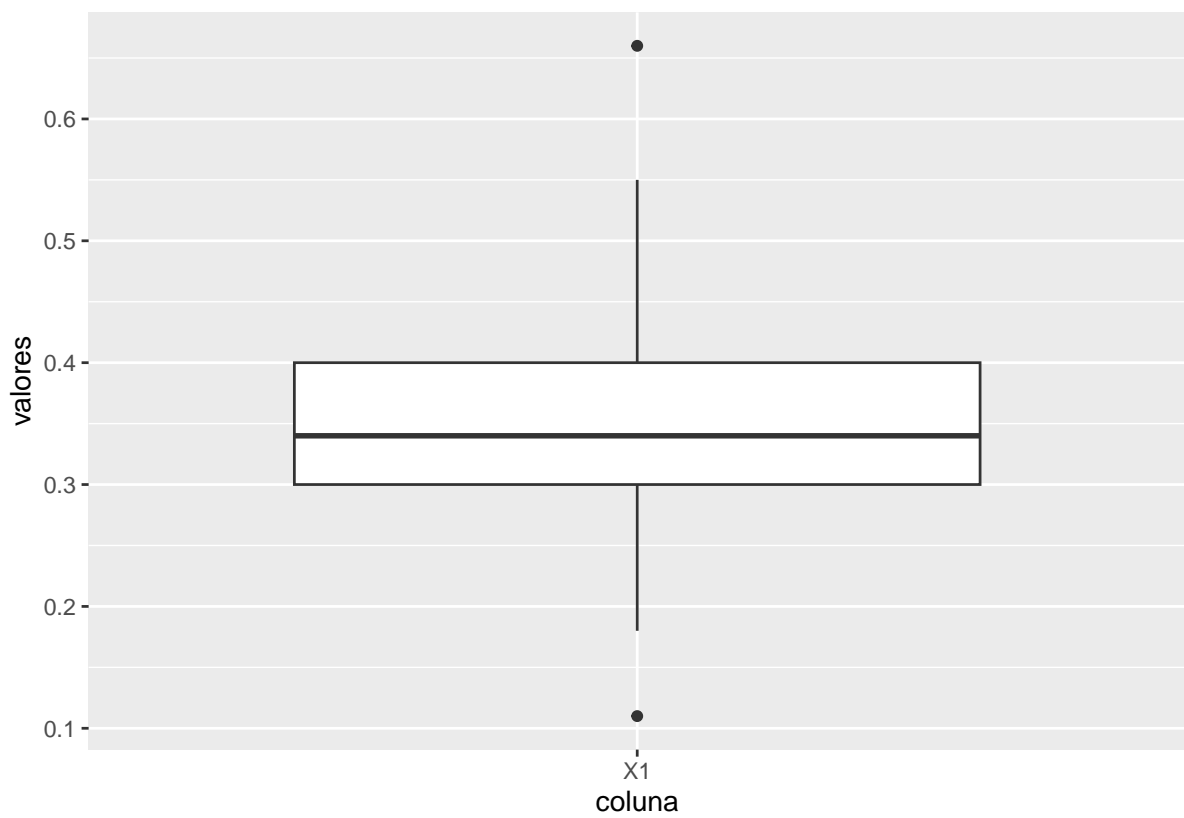
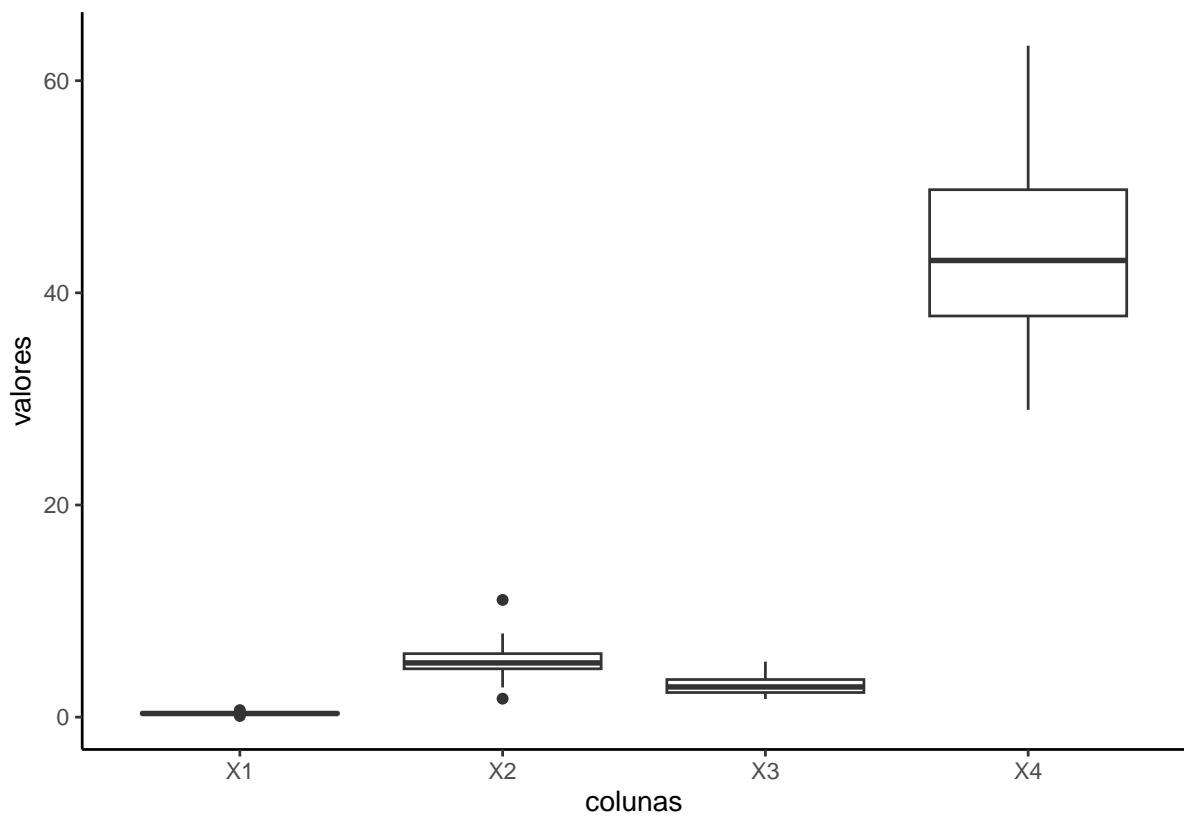
##
## Grubbs test for one outlier
##
## data:  .
## G = 2.55170, U = 0.86441, p-value = 0.2128
## alternative hypothesis: highest value 5.23 is an outlier

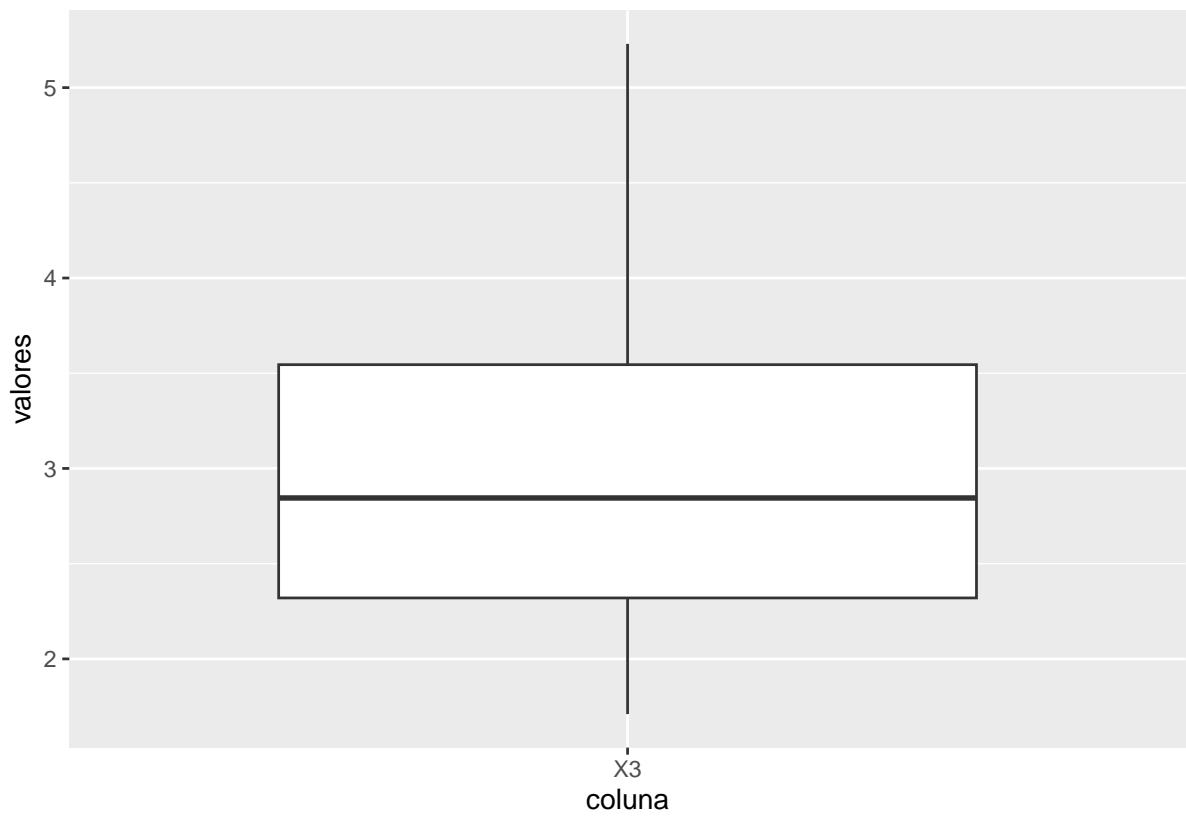
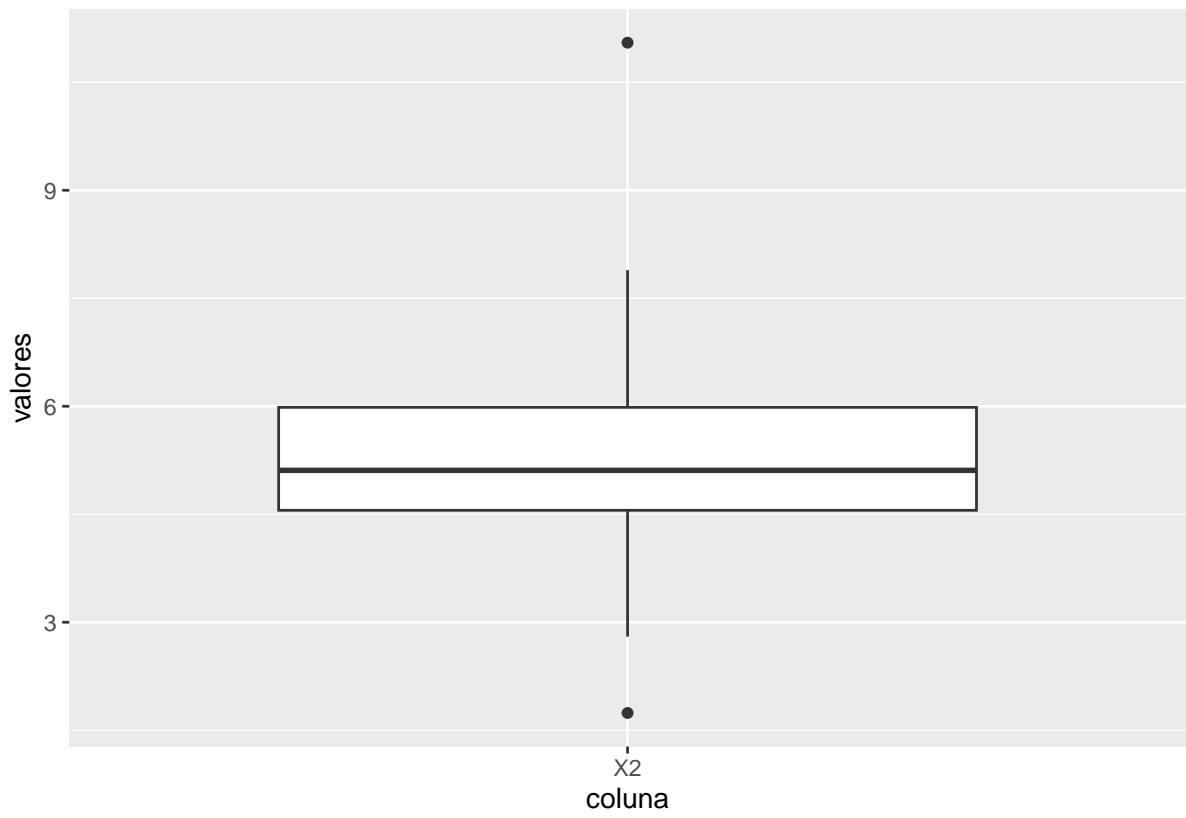
##
## Grubbs test for one outlier
##
## data:  .
## G = 2.31845, U = 0.88806, p-value = 0.4389
## alternative hypothesis: highest value 63.3 is an outlier

```

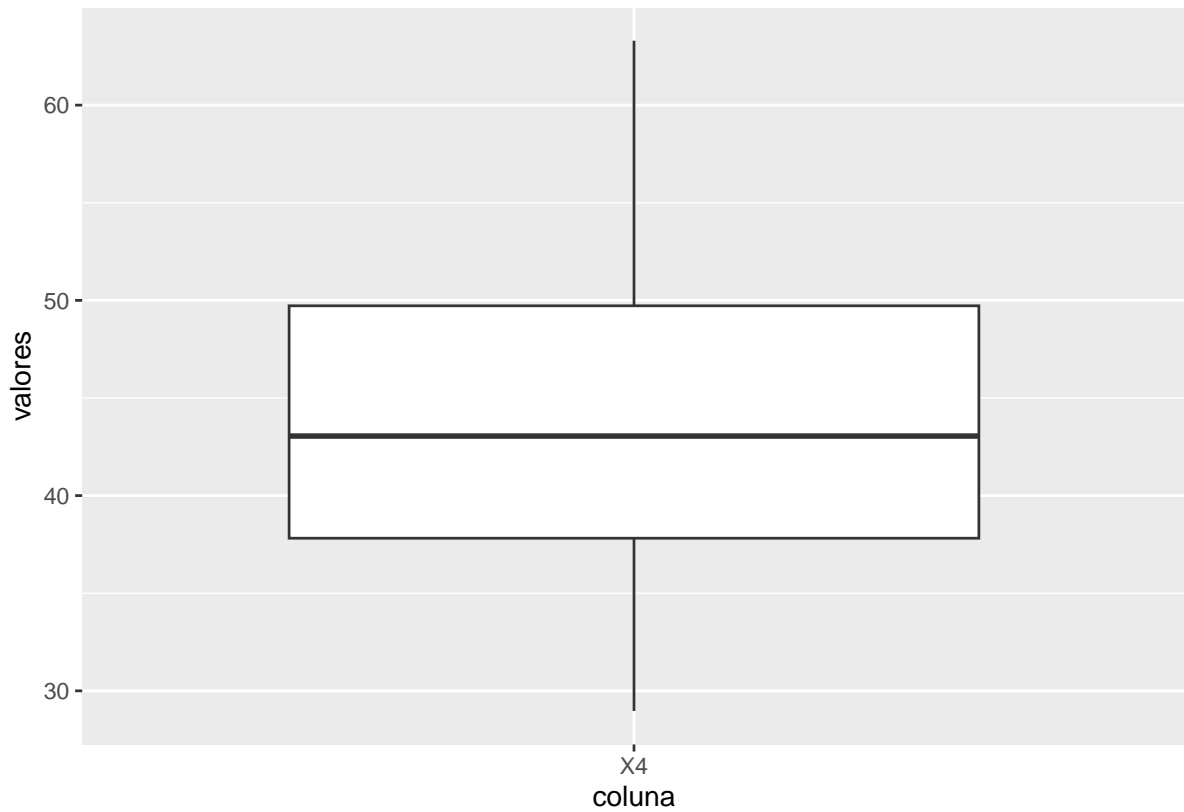
Nas colunas que a hipótese de normalidade não foi rejeitada (colunas  $X_1$  e  $X_4$ ), aplicou-se o teste de *Grubbs* para identificação de *outliers*. Em  $X_1$ , A estatística de teste pontual observada foi  $G = 3.04029$ , e  $p = 0.03592$  é o p-valor. Rejeitamos a hipótese nula pelo p-valor ser menor que 0.05. Portanto, existe ao menos um *outlier* na variável  $X_1$ . Não foram detectados *outliers* na coluna  $X_4$ .

Podemos também fazer *boxplots* para explorar a existência de *outliers*, principalmente para as colunas  $X_2$  e  $X_3$  em que o teste de *Shapiro-Wilk* rejeitou a hipótese nula de normalidade, portanto impossibilitando o uso do teste de *Grubbs*









Pelo *boxplot*, observamos que o teste de *Grubbs* aparenta ter acertado em detectar um *outlier* na coluna  $X_1$ , e na não existência de *outliers* na coluna  $X_4$ . Ainda, é possível perceber que a coluna  $X_2$  aparenta ter um *outlier* pela visualização do *boxplot*, enquanto a coluna  $X_3$ ; não.