UNIVERSIDADE DE SÃO PAULO – USP
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DEPARTAMENTO DE MATEMÁTICA APLICADA E ESTATÍSTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

SME0620 - Estatística I

Análise Exploratória e Estatística Descritiva

Professora: Amanda Morales Eudes D´Andrea

Antônio Sebastian Fernandes Rabelo - 10797781
Gabriell Tavares Luna - 10716400
Vitor Oliveira Caires - 10748027

São Carlos
2020

1. Conjunto de dados

Para análise descritiva do conjunto de dados simulados que foi disponibilizado, foram escolhidas 3 informações (variáveis) relacionadas a 50 equipamentos. Os dados faltantes (XX) foram preenchidos de acordo com a tabela 1, portanto todos os dados estão completos.

Tabela 1 -Dados de uma amostra de equipamentos produzidos pela fábrica X.

Identificação do equipamento	Tipo	Número de parafusos utilizados	Custo de fabricação (R$)
1	B	11	153,7
2	B	12	163,7
3	B	12	140,0
4	B	14	155,2
5	C	15	155,8
6	C	15	162,8
7	B	10	154,7
8	C	14	149,1
9	B	11	169,7
10	C	13	161,4
11	A	15	142,7
12	B	12	149,0
13	A	16	152,8
14	B	12	151,3
15	B	12	139,3
16	A	14	165,9
17	B	12	154,6
18	C	13	138,7
19	A	14	166,0
20	B	15	156,1
21	A	15	151,7
22	B	15	153,2
23	B	15	163,1
24	A	11	141,8
25	A	16	140,4
26	A	12	132,2
27	B	12	152,3

28	C	15	131,7
29	A	15	173,2
30	B	16	145,5
31	B	11	158,9
32	B	15	157,4
33	C	14	165,3
34	A	13	143,7
35	B	10	157,1
36	A	13	138,5
37	A	15	163,8
38	C	16	153,2
39	B	13	145,6
40	A	11	150,7
41	A	11	133,2
42	B	11	154,4
43	C	12	144,6
44	B	11	139,3
45	A	12	147,8
46	B	15	165,1
47	A	15	135,7
48	B	11	151,9
49	C	11	146,0
50	B	12	153,5

2. Recursos

A análise dos dados foi feita utilizando o software R, versão 4.0.0 (2020-04-24) -- *"Arbor Day".Copyright (C) 2020 The R Foundation for Statistical Computing. Platform: x86_64-w64-mingw32/x64 (64-bit).* Foi utilizado ainda o Excel para analisar os dados agrupados e para organização das tabelas.

Os comandos utilizados estão disponíveis no GitHub[1] juntamente com o conjunto de dados formatado em tabulações.

Para facilitar, as interpretações são feitas logo após a apresentação dos resultados.

Link para acessar GitHub: github.com/Sebastianfrabelo/Estatistica

3. Análise

De acordo com os dados apresentados das variáveis “Tipo”, “Número de parafusos utilizados” e “Custo de fabricação em reais” são classificadas como variáveis qualitativa nominal, quantitativa discreta e quantitativa contínua respectivamente, sendo possível definir análises diferentes para cada caso.

3.1 “Tipo” - Variável qualitativa nominal

Com o auxílio do software foi criada a tabela de frequência, tabela 2, e também foi implementada uma função para calcular a moda.

Tabela 2 - Tabela de frequência da variável tipo.

Tipo	f_i	F_i	f_{r_i}	F_{r_i}
A	16	16	0,32	0,32
B	24	40	0,48	0,80
C	10	50	0,20	1
Total	50	-	1	-

f_i: frequência absoluta do tipo i

F_i: frequência acumulada para o tipo i

f_{r_i}: frequência relativa do tipo i

F_{r_i}: frequência relativa acumulada do tipo i

Não existe ordem entre os tipos por ser uma variável quantitativa nominal, entretanto foi estipulada a ordem alfabética apenas para cálculo da frequência acumulada.

Com a tabela é possível comprovar o resultado obtido da função para calcular a moda da variável tipo:

$$Moda = B$$

Analisando a tabela 2, percebemos que seria interessante criar o gráfico de Pareto, (Figura 1). Para isso, é necessário a instalação do pacote “qcc”. A importância da tabela é ilustrada ao plotar o gráfico de Pareto em que a tabela é usada como argumento do comando.

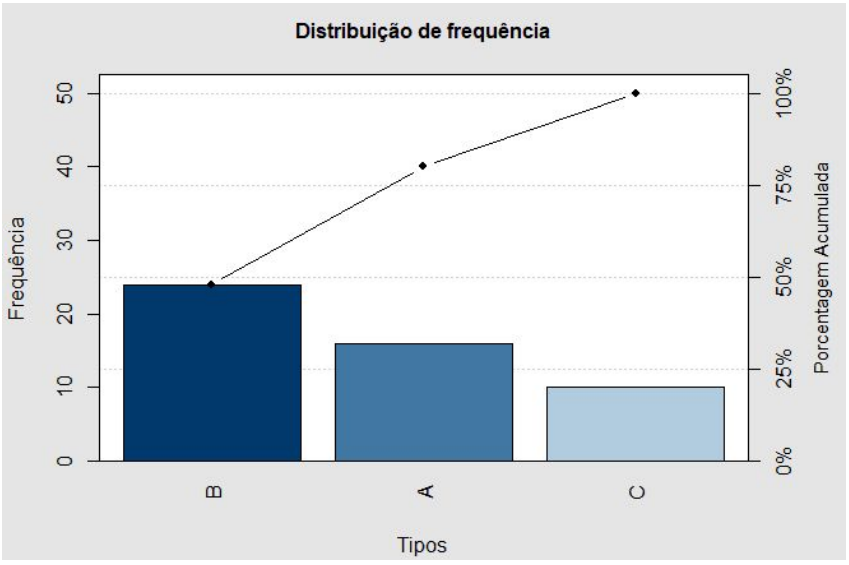

Figura 1 - Gráfico de Pareto para da variável tipo.

Interpretando os resultados, percebemos que aproximadamente metade dos equipamentos são do tipo B, os do tipo C são a menor parte, sendo que acumulando os tipos A e B, correspondem à 80% do equipamentos.

3.2 Variáveis quantitativas

Para análise das variáveis quantitativas foi desenvolvida uma função que fornece as medidas de posição e dispersão organizados nas tabelas 3 e 4.

Tabela 3 - Medidas de posição das variáveis quantitativas.

Variável	Mínimo (min)	Máximo (max)	Média ($\bar{x}$)	Primeiro quartil (Q1)	Mediana (Md)	Terceiro quartil (Q3)
Número de parafusos	10	16	13,12	12	13	15
Custo de fabricação (R$)	131,7	173,2	151,466	143,925	152,55	157,325

É possível afirmar que os dados são **condizentes** pois há pequena variação entre valores de média e mediana para ambas as variáveis, tendo em vista do desvio padrão.

Tabela 4 - Medidas de dispersão das variáveis quantitativas.

Variável	Amplitude (A)	Amplitude interquartil (d_q)	Variância (S^2)	Desvio padrão (S)	Coeficiente de variação (%) (CV)
Número de parafusos	6	3	3,291429	1,814229	13,827969
Custo de fabricação (R$)	41,5	13,4	104,441065	10,219641	6,747152

Sabendo que a média das variáveis é diferente de zero é coerente calcular o coeficiente de variação.

Interpretando as medidas de posição acompanhadas das medidas de dispersão, podemos discutir sobre a coerência dos dados. Destaque para a relevância da variabilidade relativa, pois apesar da variância do custo de fabricação ser maior que a da variável número de parafusos, quando o efeito da unidade de medida de ambas as variáveis é tirado, temos a medida do coeficiente de variação.

Percebe-se então que o número de parafusos varia mais que o custo de produção, pois o coeficiente de correlação do custo de produção é aproximadamente a metade do coeficiente de correlação do número de parafusos. Portanto o coeficiente de variação apresenta melhor a dispersão dos dados.

3.2.1 "Número de parafusos utilizados" - Variável quantitativa discreta

Tabela 5 - Tabela de frequência da variável Número de parafusos utilizados.

Número de parafusos	f_i	F_i	f_{r_i}	F_{r_i}
10	2	2	0,04	0,04
11	10	12	0,2	0,24
12	11	23	0,22	0,46
13	5	28	0,10	0,56
14	5	33	0,10	0,66
15	13	43	0,26	0,92
16	4	50	0,08	1
Total	50	-	1	-

Ainda que o "Número de parafusos utilizados" não seja uma variável qualitativa, é conveniente calcular sua moda pois os elementos se repetem frequentemente, como também pode ser extraído da tabela 5.

$$Moda = 15$$

Foram elaborados ainda o gráfico de barras e o boxplot para a distribuição nas amostras e também em função para variável "tipo".

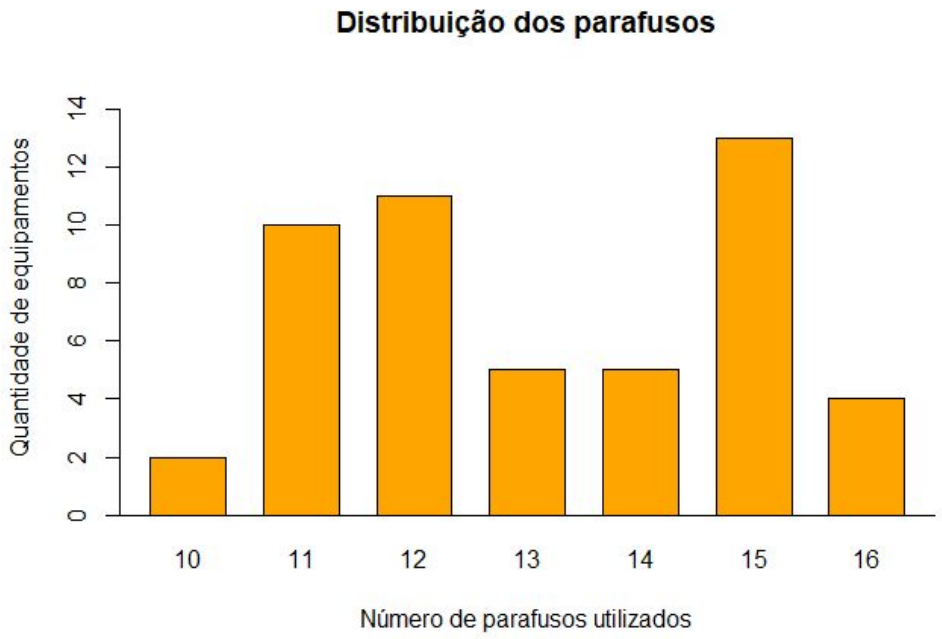

Figura 2 - Gráfico de barras para a variável número de parafusos utilizados.

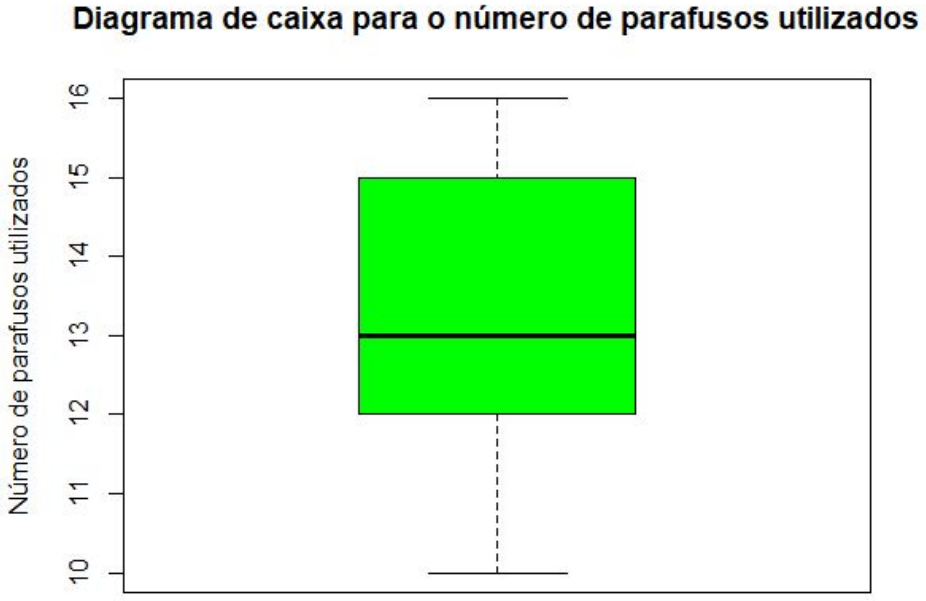

Figura 3 - Diagrama de caixa (Box plot) para distribuição do "Número de parafusos utilizados".

Analisando o diagrama de caixa, os dados nos intervalos entre Q1 e Q2, e entre Q3 e o limite superior, estão mais concentrados que os dados nos intervalos entre Q2 e Q3, e entre o limite inferior e o Q1.

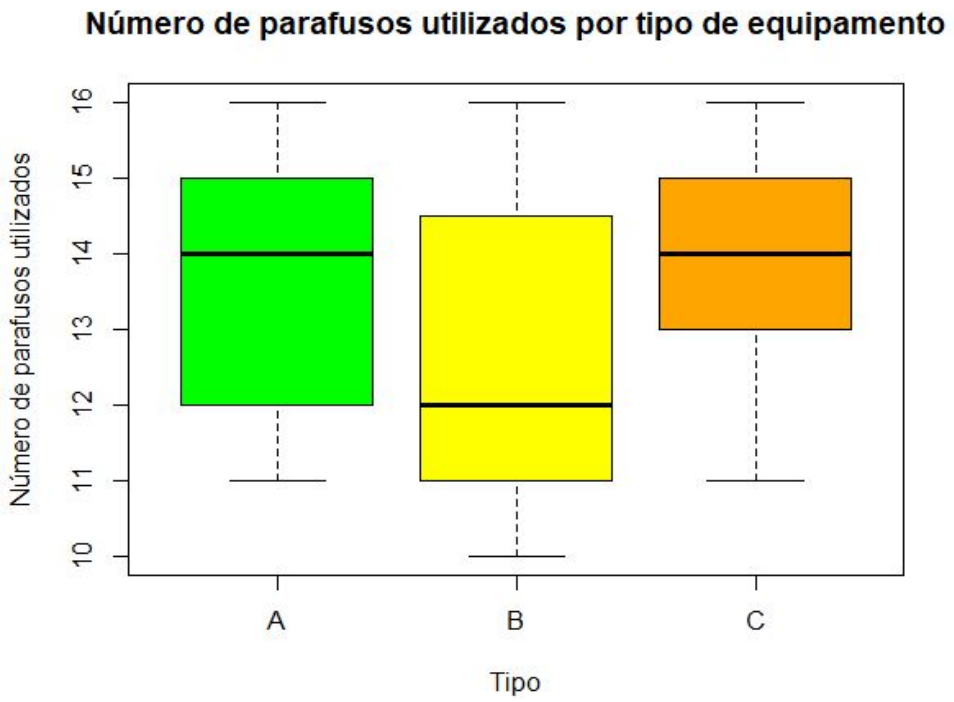

Figura 4 - Distribuição do Número de parafusos utilizados por tipo de equipamento.

Interpretando os resultados, vemos que o número de parafusos utilizados é destacado para as quantidades 11, 12 e 15, para outros valores, é aproximadamente bem distribuído.

É interessante que, ao analisar o diagrama de números de parafusos por tipo, são destacadas as diferenças para o tipo B. Quando comparado com o número de parafusos para os outros tipos, apresenta menor Q1 e menor mediana, os dados entre Q1 e Q2 estão muito mais concentrados que os dados entre Q2 e Q3.

Para os tipos A e C a mediana é maior que comparada a mediana do número de parafusos para todos os equipamentos. Para o tipo C, os valores estão igualmente concentrados entre Q1 e Q2, e entre Q2 e Q3.

Para o tipo A, os dados estão mais dispersos no intervalo entre Q1 e Q2, e estão igualmente concentrados para os os outros intervalos.

3.2.2 "Custo de fabricação (em reais)" - Variável quantitativa contínua

Tratando-se de uma variável quantitativa contínua as amostras foram divididas em sub intervalos regulares, fechados à esquerda e abertos à direita.

Tabela 6 - Tabela de frequência da variável Custo de produção em reais.

Ordem	Custo de produção (R$)	Ponto médio (R$)	f_i	F_i	f_{r_i}	F_{r_i}
1	130 |-- 135	132,5	3	3	0,06	0,06
2	135 |-- 140	137,5	5	8	0,10	0,16
3	140 |-- 145	142,5	6	14	0,12	0,28
4	145 |-- 150	147,5	6	20	0,12	0,40
5	150 |-- 155	152,5	13	33	0,26	0,66
6	155 |-- 160	157,5	6	39	0,12	0,78
7	160 |-- 165	162,5	5	44	0,10	0,88
8	165 |-- 170	167,5	5	49	0,10	0,98
9	170 |--| 175	172,5	1	50	0,02	1
		Total	50	-	1	-

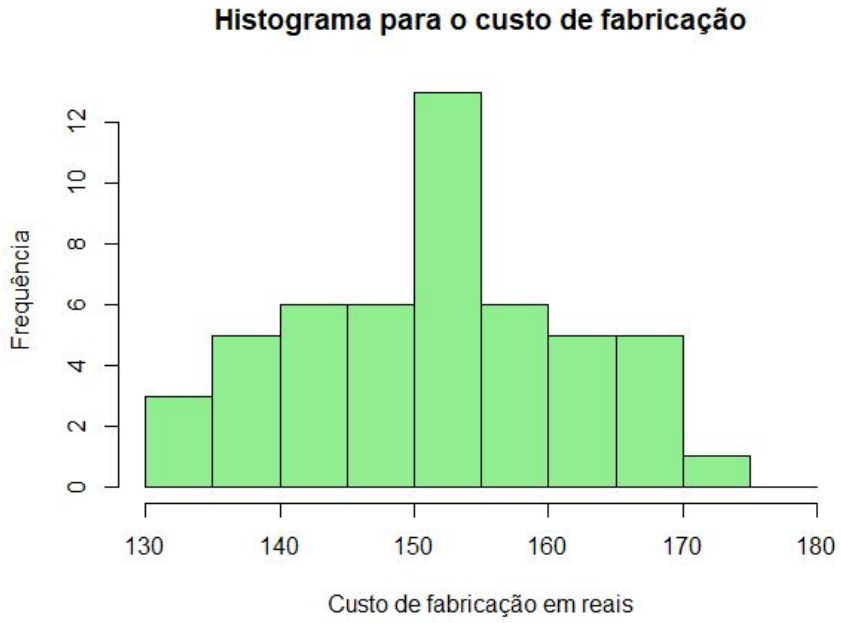

Figura 5 - Histograma do custo de produção em reais..

Para verificação da perda de informações ao agrupar os dados, foram calculadas as medidas de posição e dispersão. Nesse caso não é relevante calcular a densidade de frequência já que os intervalos são regulares.

Tabela 7 - Medidas de dispersão para valores reais e agrupados.

Custo de fabricação (R$)	Média ($\bar{x}$)	Variância (S^2)	Desvio padrão (S)	Coeficiente de variação (%) (CV)
Cálculo para Dados reais	151,466	104,441065	10,219641	6,747152
Cálculo para Dados agrupados	151,5	106,1224	10,30158	6,79972

As unidades de medida foram subtraída por simplicidade, facilitando a interpretação da tabela.

Com base na tabela 7, é possível afirmar que a perda de informações não foi de grande relevância, dado a pequena diferença nas medidas de posição e dispersão. Tendo como base de comparação o desvio padrão calculado para dados reais.

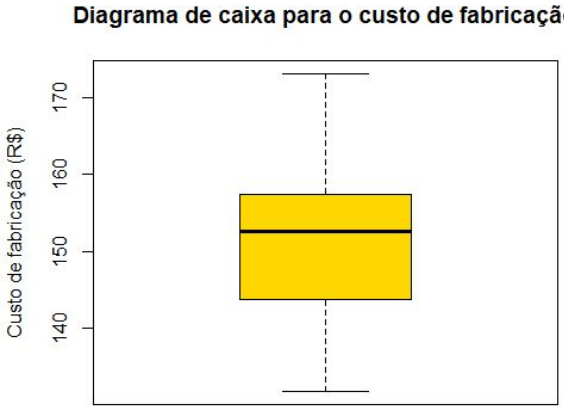

Figura 6 - Box plot para distribuição do custo de fabricação

O diagrama de caixa revela que o dados estão menos dispersos entre e primeiro e o terceiro quartil.

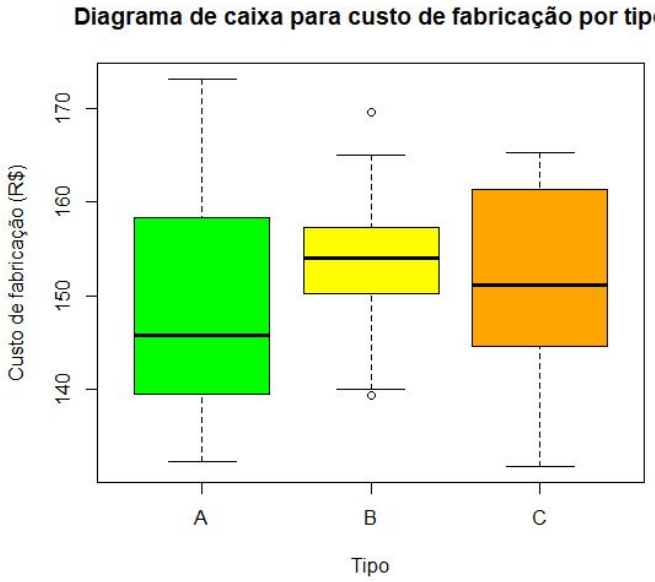

Figura 7 - Distribuição do custo de fabricação por tipo de equipamento.

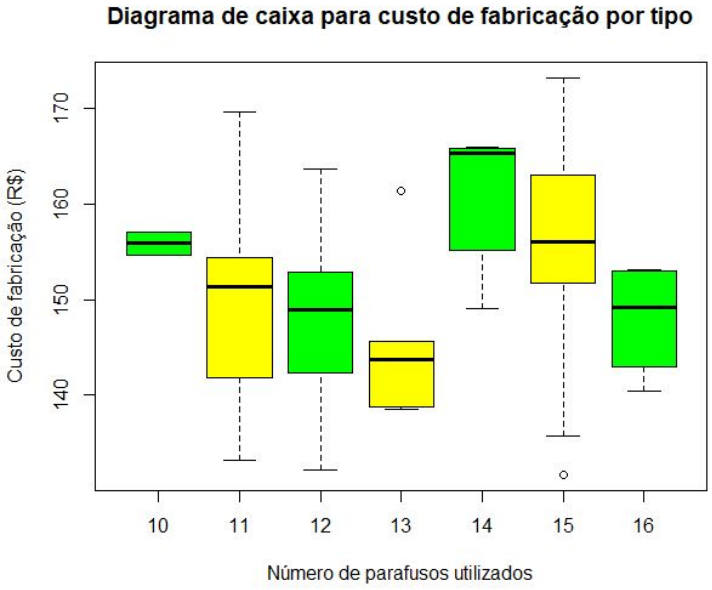

Figura 8 - Distribuição do custo de fabricação por número de parafusos utilizados.

Com base na figura 7, percebemos a diferença na distribuição dos dados para os diferentes tipos de equipamentos, apesar de ser uma análise delicada pois a quantidade de dados analisados em cada grupo é pequena como apresentado na tabela 2, podendo influenciar no aparecimento de dados extremos.

Também é possível perceber as diferenças dos custo para cada quantidade de parafusos utilizados, os diagramas para esses casos apresentam grandes diferenças principalmente da variância e da dispersão dos dados. Nesse gráfico também existem problemas relacionados à pequena quantidade de dados para cada classe, como para 10 e 13 parafusos.

3. 3 Análise bivariada

Para análise bivariada entre as variáveis quantitativas, foram plotados dois gráficos.

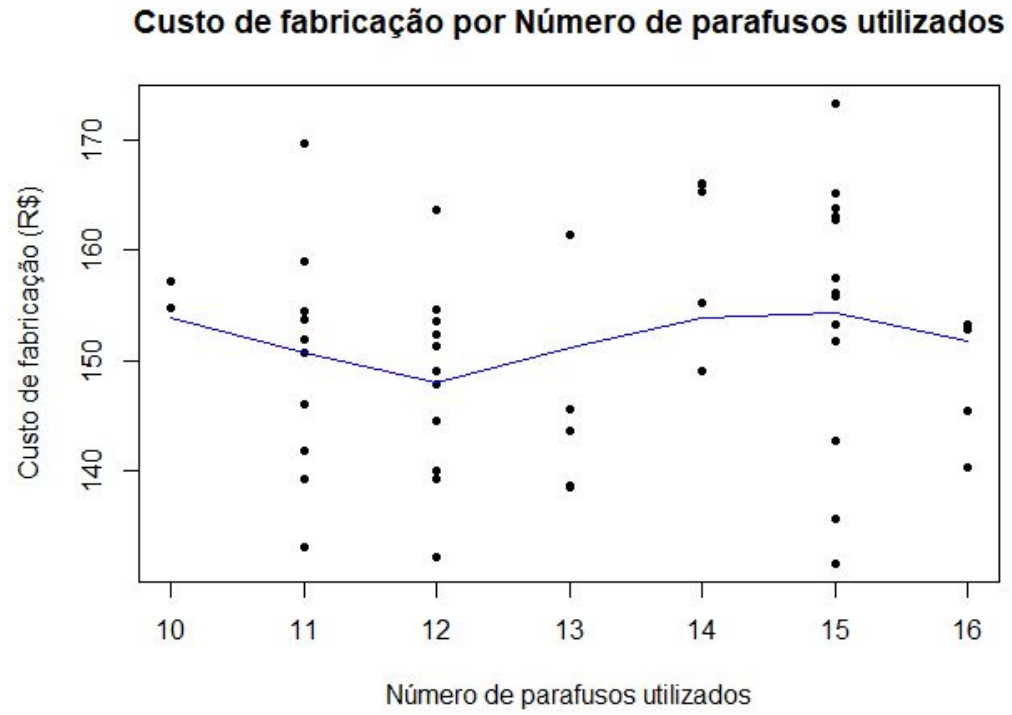

Figura 9 -Número de parafusos utilizados por Custo de fabricação.

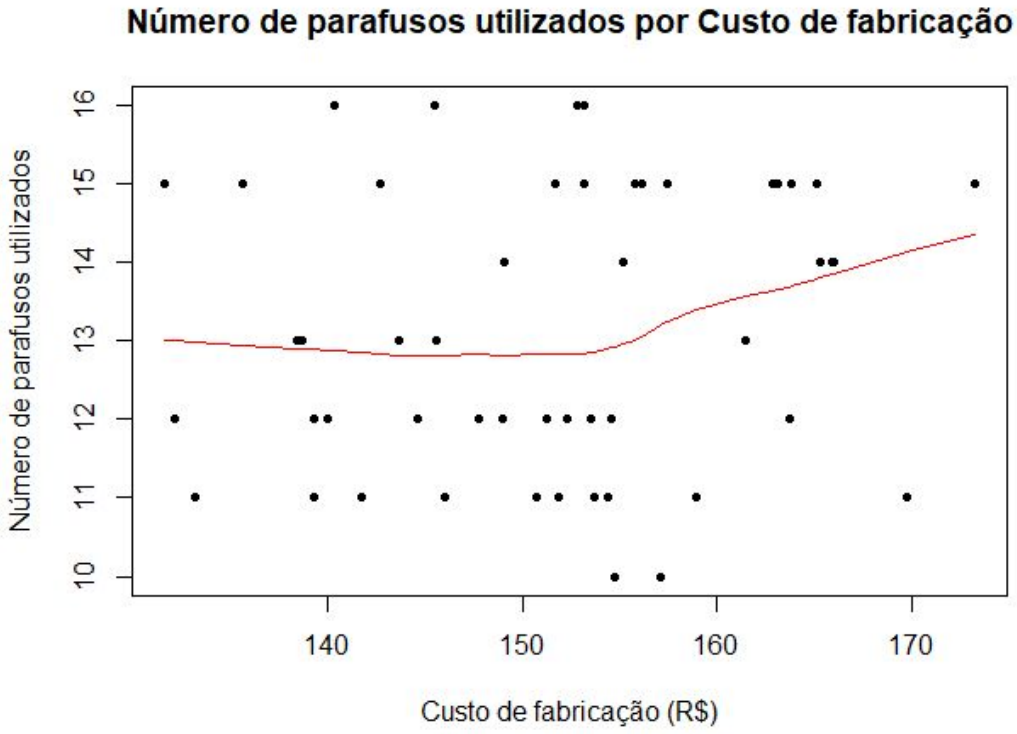

Figura 10 - Custo de fabricação por Número de parafusos utilizados.

Podemos calcular então a Covariância e o Coeficiente de Correlação Linear das variáveis:

- Covariância:

$$\sigma_{xy} = 2.655184$$

- Coeficiente de Correlação Linear:

$$\rho_{xy} = 0.1432078$$

Pode-se observar uma covariância positiva entre as variáveis, ou seja, aumentos causados por uma tendem a implicar em aumentos na outra.

Por outro lado, o Coeficiente de Correlação Linear, que avalia o quanto as variáveis estão correlacionadas linearmente e varia entre $-1 \leq \rho_{xy} \leq 1$, é positivo e próximo de zero. Isso significa que existe uma relação linear relativamente fraca, porém positiva.

Portanto, as variáveis quantitativas crescem proporcionalmente, o que é perceptível também por meio do gráfico da figura 10.

Apesar de que no gráfico da figura 9, esse padrão não ser evidente, não sendo possível apresentar alguma resolução, o que é justificado por serem poucos dados simulados, mostrando a necessidade da escolha de uma amostra de dados representativa, seguindo as necessidade de ser aleatória e grande o suficiente.