Curso: Ciência de Dados
Alunas:
Camila Perazzo
Sara Coutinho

Análise Exploratória

17 de abril de 2023

VISÃO GERAL

A atividade de análise exploratória tem como objetivo a aplicação prática dos conhecimentos adquiridos ao longo do curso de ciência de dados até o momento, bem como o uso do ciclo de vida da ciência de dados. Na atividade em questão, os dados explorados se referem aos sinais vitais de uma população, especificamente, sobre o batimento cardíaco, a pressão arterial e a temperatura corporal.

DIÁRIO DE BORDO

- ***06/04/2023: 10:30 às 11:00*** - Reunião de Alinhamento
- ***08/04/2023: 8:30 às 9:30*** - Compreensao das atividades, entendimento do negócio, coleta dos dados
- ***10/04/2023: 17:00 às 19*** - Amostragem e aplicação de métodos estatísticos
- ***15/04/2023 : 13:30 às 14:30*** - Correção de arquivo e ajustes dos cálculos preenchimento valores nulos
- ***15/04/2023 : 18:00 às 20:30*** - Aplicação dos gráficos de correlação e média
- ***17/04/2023: 17:30 às 23:00*** - Finalização do projeto

OBJETIVOS

O objetivo dessa atividade prática de análise exploratória é investigar e compreender os dados disponibilizados acerca dos sinais vitais humanos de uma determinada população, com o intuito de identificar padrões, tendências, relações e possíveis anomalias presentes nos dados. Ademais, o processo da análise exploratória também pode ser utilizado para identificar problemas nos dados, como valores ausentes ou discrepantes, que possam afetar a qualidade das análises posteriores.

ATIVIDADES

De acordo com o exercício proposto, segue abaixo a lista definindo as atividades solicitadas e em qual momento do ciclo de vida de vida do processo de Ciência de Dados foi resolvido :

1) Preparação dos dados - Está na etapa de preparação dos dados

2) Amostragem - Está na etapa de Análise Exploratória

3) Correlação das variáveis - Está na etapa de Análise Exploratória

4) Estatística descritiva - Está na etapa de Análise Exploratória

ESPECIFICAÇÕES

Entendimento do negócio

Nessa etapa, para o entendimento do negócio, será aplicado a ferramenta 5W2H: What, Who, Where, When, Why, How, How Much, na qual, é um conjunto de questões utilizado para compor planos de ação de maneira rápida e eficiente através da respostas captadas.

- *What* - **Monitoramento de sinais vitais**.
- *Where* - Não definido.
- *When* - Deadline da atividade.
- *Who* - Reportar para o professor da disciplina.
- *Why* - O monitoramento de sinais vitais é importante para a manutenção da saúde da população.
- *How* - A partir de dados obtidos por sensores, a respeito dos sinais vitais, é possível detectar padrões de comportamentos normais dos pacientes a partir de data mining e fazer o monitoramento da saúde dos pacientes.
- *How much* - Não definido, mas a parte de custo seria relativo ao custo para obtenção do arquivo com dados.

Conforme passado na questão, as variáveis dos dados são as descritas abaixo e com os respectivos limites de variação:

0<= BATIMENTO CARDÍACO < 100

0<= PRESSÃO ARTERIAL < 20

0<= TEMPERATURA CORPORAL < 40

Coleta dos dados

Como os dados trabalhados são de batimento cardíaco, pressão arterial e temperatura corporal, espera-se que eles sejam dados quantitativos contínuos para essas três variáveis, pois dessa forma será possível fazer as previsões e classificar o paciente entre saudável ou não saudável.

Preparação dos Dados

Questão 1

1. **Detecção dos valores fora das faixas normais**

   Com base no enunciado, admite-se que o valores são considerados normais quando estão dentro da margem abaixo:

   0<= BATIMENTO CARDÍACO < 100

   0<= PRESSÃO ARTERIAL < 20

   0<= TEMPERATURA CORPORAL < 40

2. **Substituição dos dados fora da faixa de valores normais**

   No caso, esses valores anormais foram substituídos pela média aritmética do valor anterior e posterior do dado anormal.

3. **Exclusão de valores nulos e duplicados**

   Com o intuito de garantir a qualidade dos resultados da análise dos dados, os valores nulos e duplicados serão excluídos, pois eles podem gerar erros, além de distorcer as estatísticas descritivas e correlações entre variáveis. Já dados duplicados podem gerar redundância nas análises e distorcer as estatísticas descritivas e correlações, pois são contabilizados mais de uma vez.

Amostra do conjunto de dados

	Batimento	Pressao	Temperatura
0	66.535898	10.267949	36.832602
1	66.935822	10.467911	36.888632
2	67.428850	10.714425	36.629427
3	68.000000	11.000000	36.877926
4	68.631919	11.315960	36.892643
...	...	...	...
19	69.669352	12.314528	41.011285
20	68.288729	12.505198	37.117865
21	67.653875	12.402058	37.073562
22	67.435384	12.207166	36.627888
23	69.468733	12.029474	36.699478

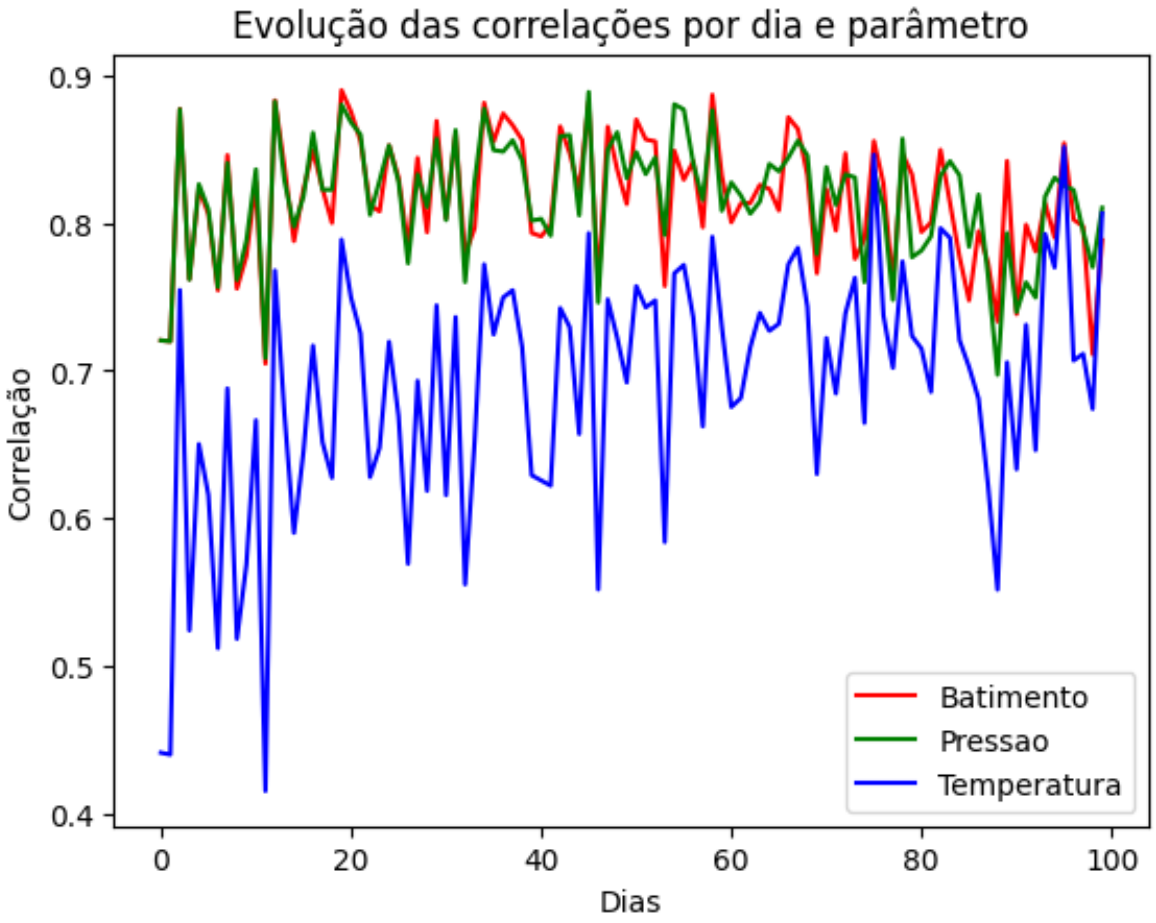
2400 rows × 3 columns

Análise Exploratória dos Dados

Questão 2

A análise exploratória é uma ferramenta importante porque permite que os dados sejam compreendidos em profundidade antes da aplicação de técnicas mais complexas de análise. Ela ajuda a identificar possíveis padrões e tendências nos dados, e também pode ser usada para detectar problemas de qualidade nos dados, como valores ausentes ou inconsistentes. Assim, no caso em questão, foram coletados dados de pressão, batimento e temperatura a cada uma hora ao longo de 100 dias, e abaixo estão alguns dos gráficos construídos durante a análise exploratória desses dados e os insights constatados ao longo da análise.

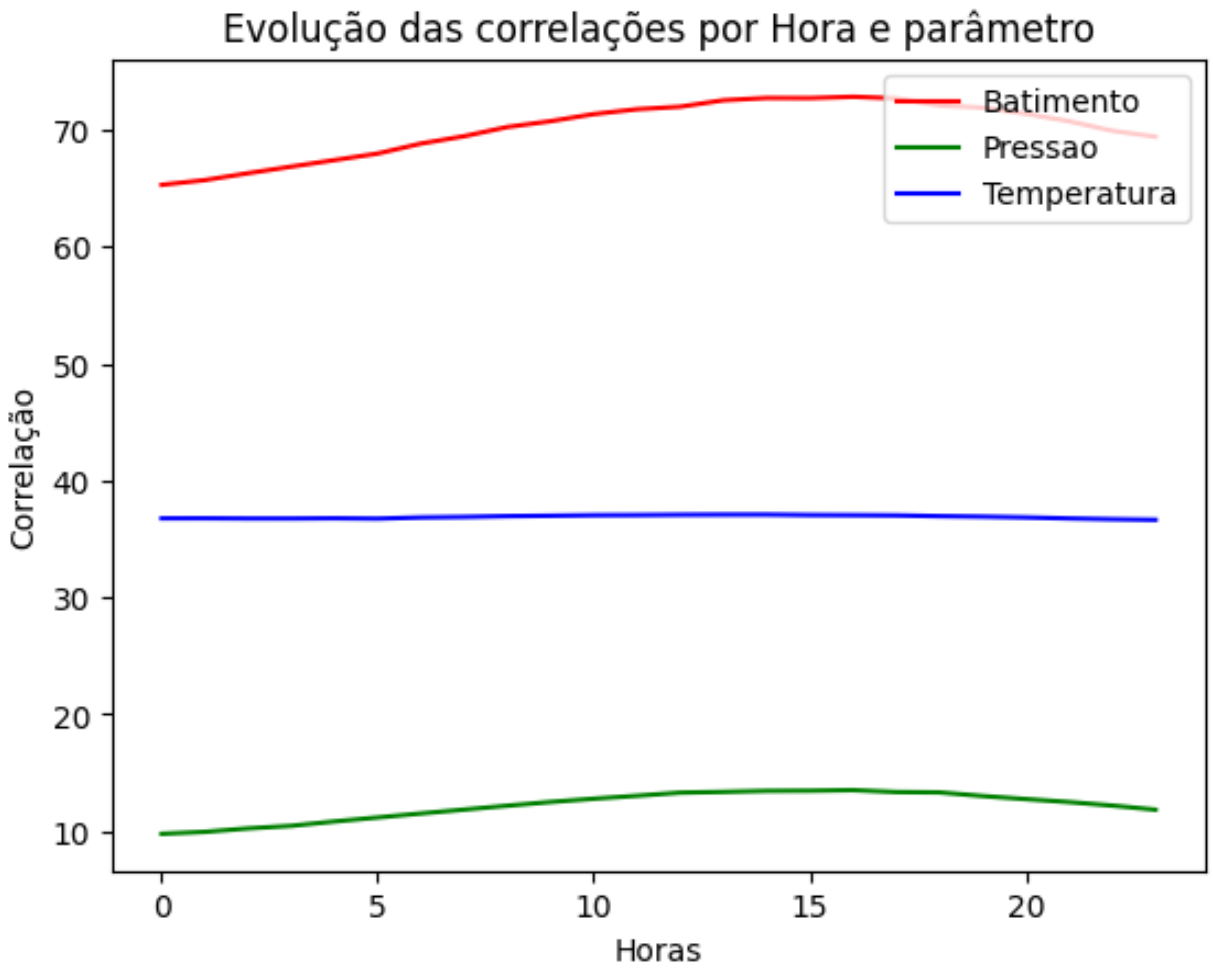
Evolução da correlações por dia e parâmetro



Por esse gráfico, é possível concluir que os valores do Batimento e Pressão estão mais próximos entre si, e crescem e decrescem semelhantemente no decorrer dos dias. Além disso, é observado um decrescimento dos valores de correlações de Batimento e Pressão quando os dias se aproximam de 100. Já a correlação da temperatura possui valores mais distantes das demais correlações obtidas e ao invés de decrescer próximo aos 100 dias, a correlação cresce. Isso indica que as medições de temperatura cresceram no decorrer do tempo. Já as de batimento e pressão não.

Analisando o comportamento das variáveis por Hora, construímos a matriz de correlação de cada variável entre si dentro dos grupos de amostras de tamanho 100 por cada Hora.

Evolução da correlações por hora e parâmetro



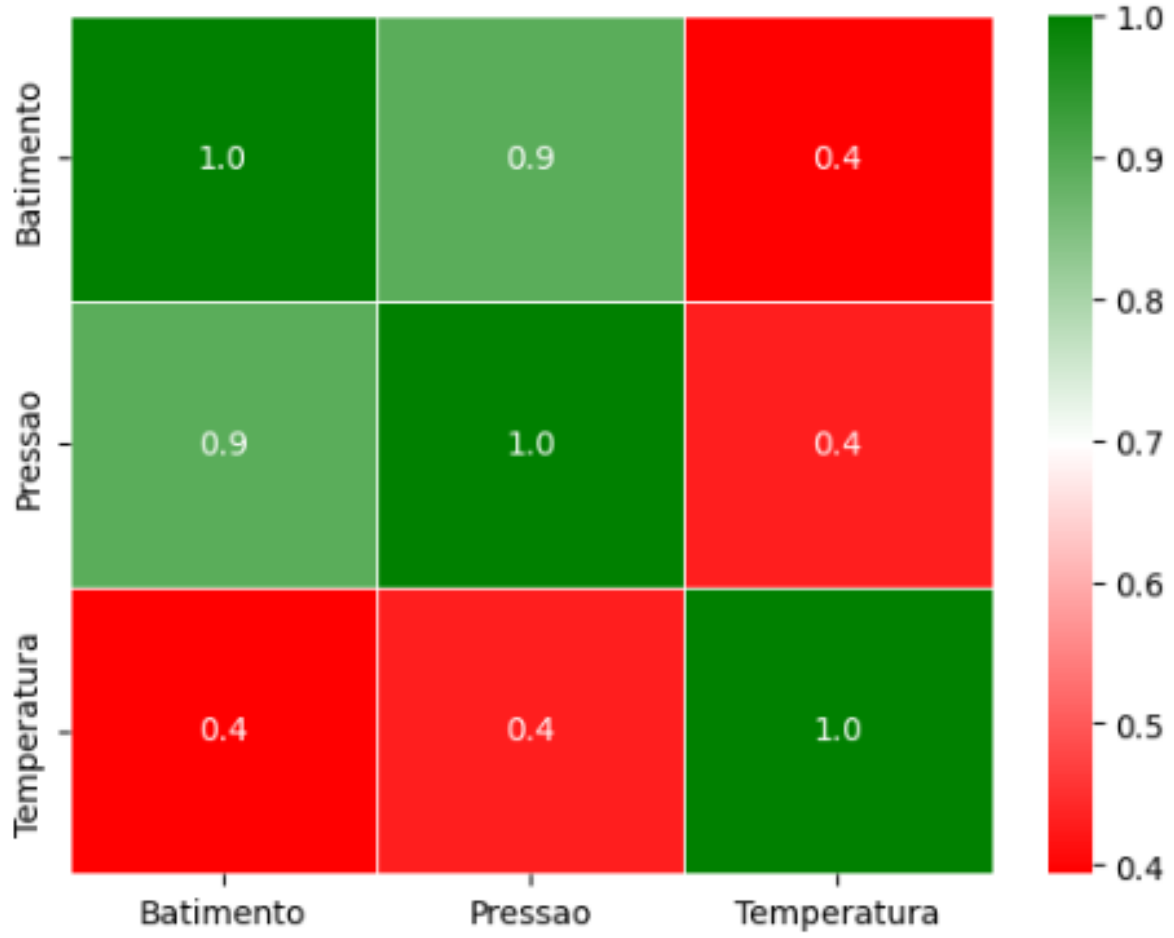
Por esse gráfico vemos que a correlação do valor médio das variáveis Batimento e Pressão apresentam também um comportamento semelhante por Hora. Em contrapartida, o valor médio da Temperatura apresenta um comportamento constante por Hora.

Comparando o comportamento das variáveis considerando as amostras por dia e as amostras por hora, concluiu-se que em ambos os casos tanto o Batimento quanto a Pressão possuem comportamentos semelhantes. Em compensação, a Temperatura decresceu ao longo dos dias. Contudo, a temperatura se mostrou constante ao longo das horas.

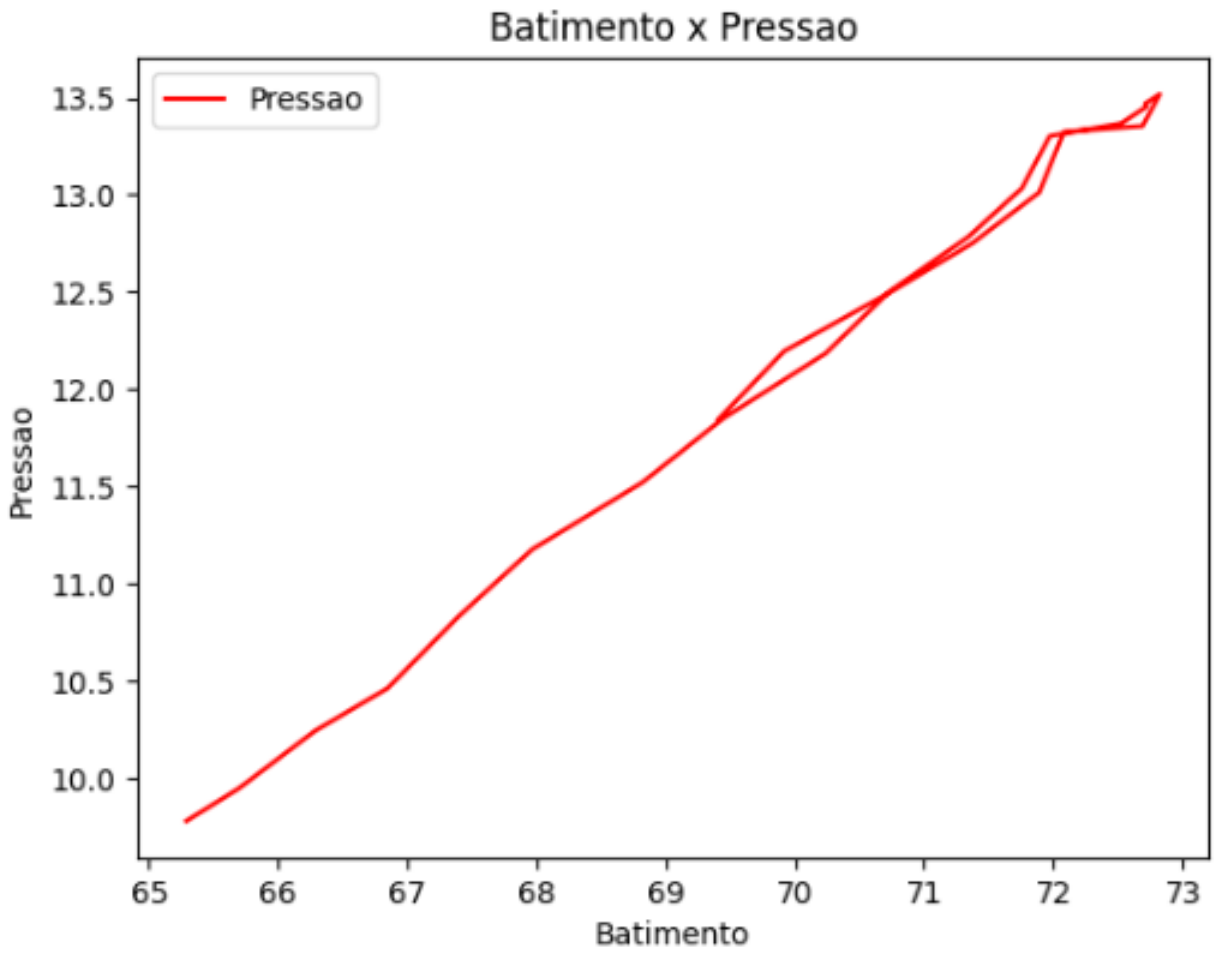
Em termos práticos, é interessante que o paciente averigue sua temperatura dia após dia já que esse é um fator que varia por dia, mas não importa a hora do dia em que ela for medida. Além disso, a hora do dia que o batimento e a pressão apresentam um aumento são em torno das 15 às 20 horas do dia. Por isso, é importante ter em mente essa possível variação existente ao longo do dia também no processo de medição dessas variáveis no paciente.

Para analisar os dados no geral, plotou-se uma matriz sem considerar o processo de amostragem. Foi utilizada a correlação de spearman a fim de que não se assumisse um comportamento linear a priori das variáveis.

Mapa de calor apresentando a correlações das variáveis



Nessa matriz também percebe-se que as variáveis Pressão e Batimento são fortemente correlacionadas, enquanto que com a Temperatura a correlação delas é mais fraca.

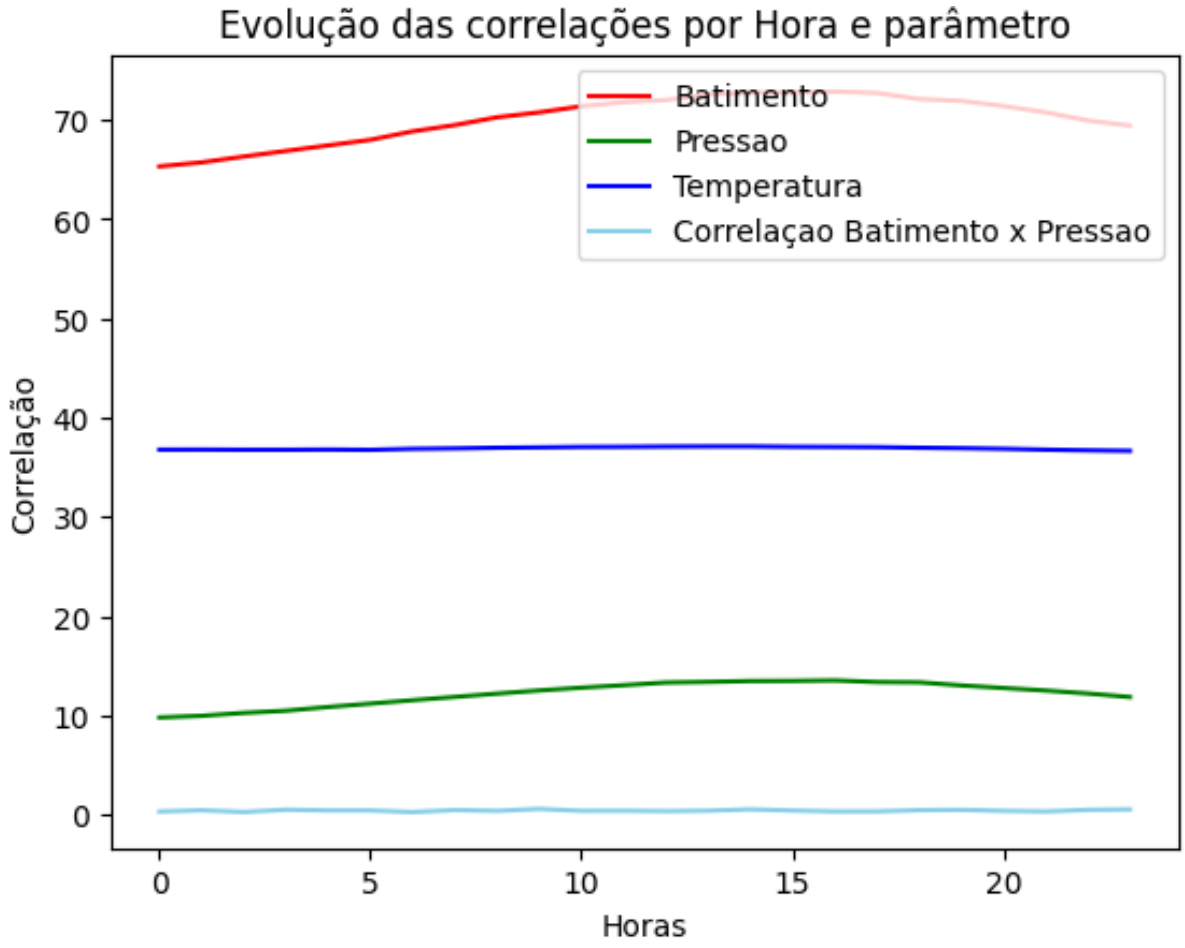


De fato, as variáveis apresentam um comportamento praticamente linear e com tendência crescente confirmando a correlação forte e positiva das variáveis.

Questão 3

Conforme visto na questão 2, as correlações mais fortes são Batimento e Pressão. Considerou-se verificar o novo padrão de aprendizado a partir da correlação dessas duas variáveis. Dessa forma, verificou-se esse comportamento por hora e por dia.

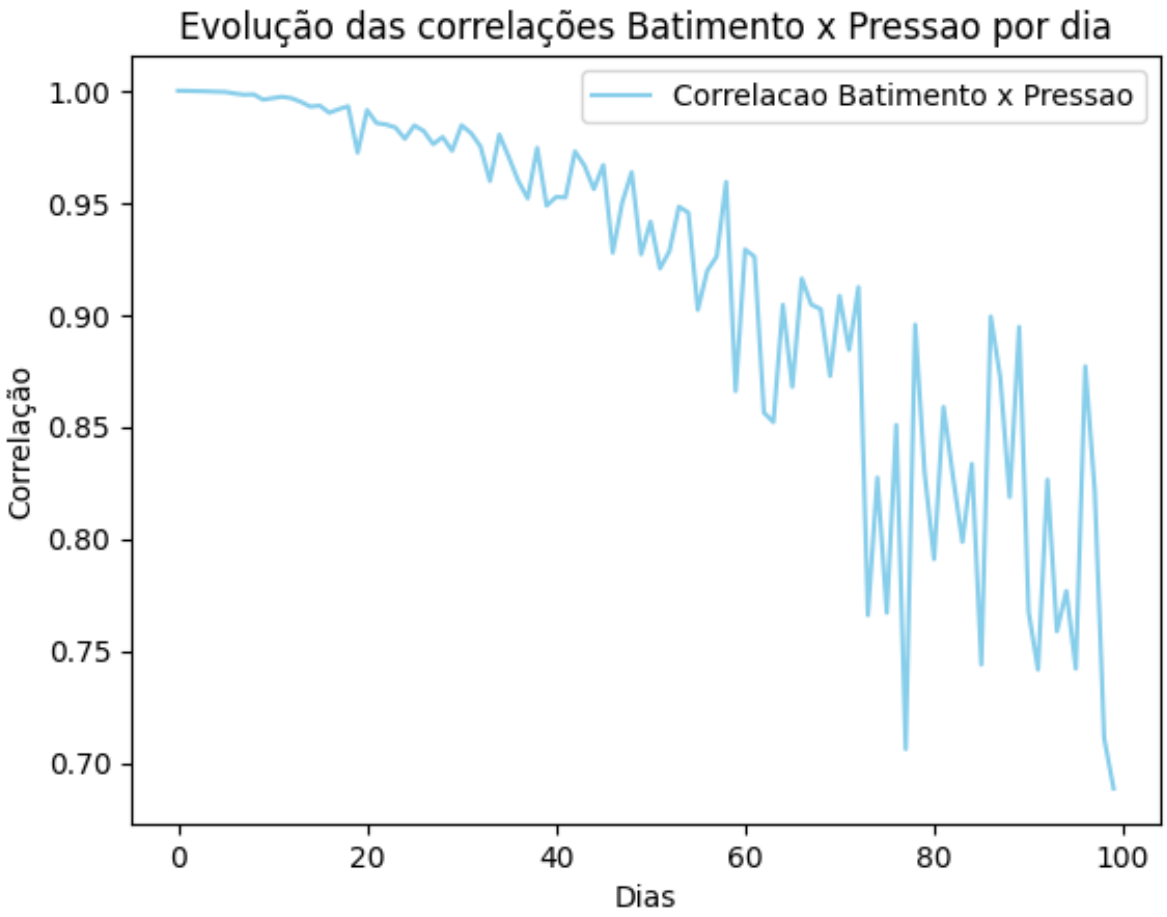
Evolução da correlações por hora e parâmetro (Batimento X Pressão)



Nesse gráfico observa-se que a correlação entre Batimento e Pressão por hora resultou numa curva praticamente reta, não demonstrando o comportamento dessas variáveis. Por isso, pode ser mais interessante utilizar uma das variáveis para representar as duas, já que elas estão fortemente correlacionadas, do que utilizar a correlação delas para o aprendizado. Além disso, é importante notar que essa correlação não muda com o tempo, confirmando que pode valer a pena utilizar apenas uma delas para medição ao longo do tempo.

No caso prático, poderia-se medir apenas a pressão ou o batimento do paciente e a partir disso constatar a saúde do paciente com relação a essas duas variáveis observando apenas uma delas, em termos de hora.

Evolução da correlações por dia e parâmetro (Batimento X Pressão)



No caso da média correlação Batimento x Pressão por dia, observou-se um decrescimento desse valor ao longo dos dias. Na tabela de correlações também é verificado que as correlações dos primeiros dias são maiores e dos últimos dias menores. Apesar das variáveis serem fortemente correlacionadas no geral e em amostras por hora, quando analisadas por dia possuem um decrescimento na correlação.

Em termos práticos, isso pode ser importante na análise do paciente, pois ao passar dos dias, pode ser preciso mensurar as 3 variáveis, uma vez que a correlação entre as duas mais fortes decai.

Questão 4

Nessa etapa foram verificados os resultados para as medidas de posição (média, moda, mediana, histograma) e medidas de dispersão (variância e desvio padrão).

	Batimento	Pressao	Temperatura
count	2400.000000	2400.000000	2400.000000
mean	70.110302	12.184503	36.891543
std	2.596138	1.260424	0.202142
min	62.677785	8.539146	36.456422
25%	68.295517	11.287009	36.749488
50%	70.618385	12.454484	36.894476
75%	72.236410	13.239359	37.010353
max	74.000000	14.000000	37.480333

Esta tabela apresenta os valores obtidos para a média de cada variável na linha denominada *mean*.

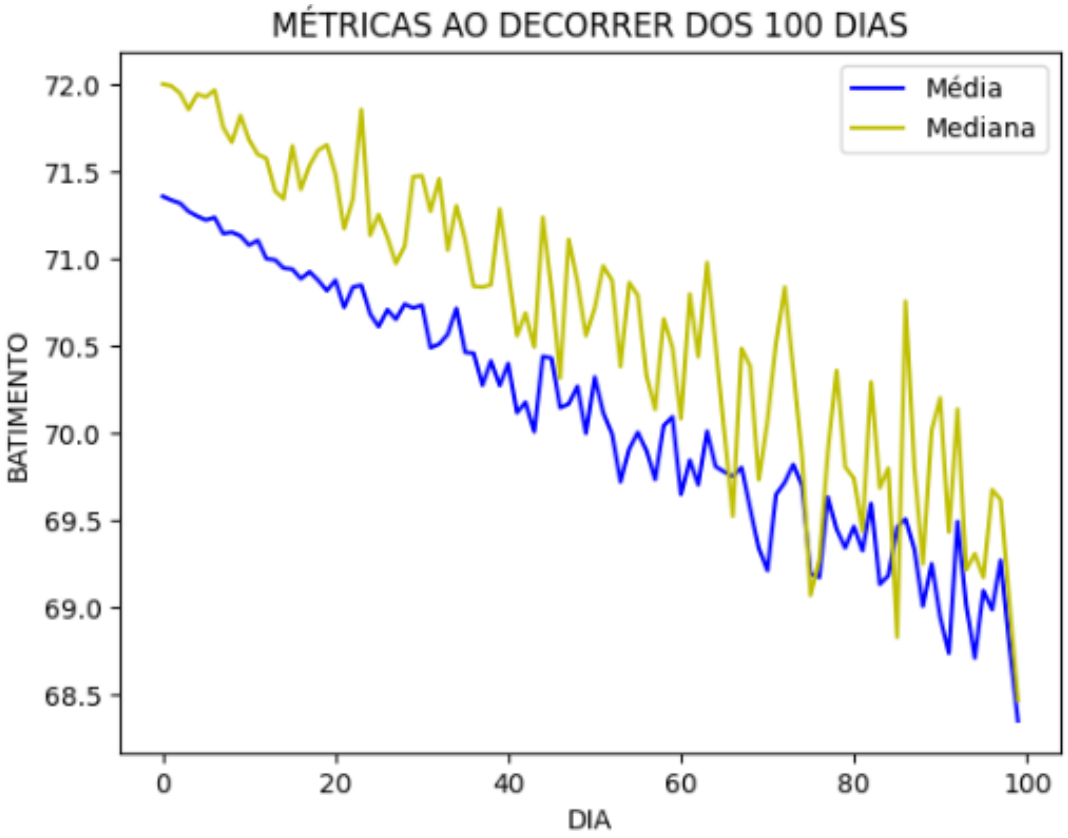
```
#Calculo mediana
print('Mediana batimento: ',round(stats.median(data['Batimento']),2))
print('Mediana Pressao: ',round(stats.median(data['Pressao']),2))
print('Mediana Temperatura: ',round(stats.median(data['Temperatura']),2))
```

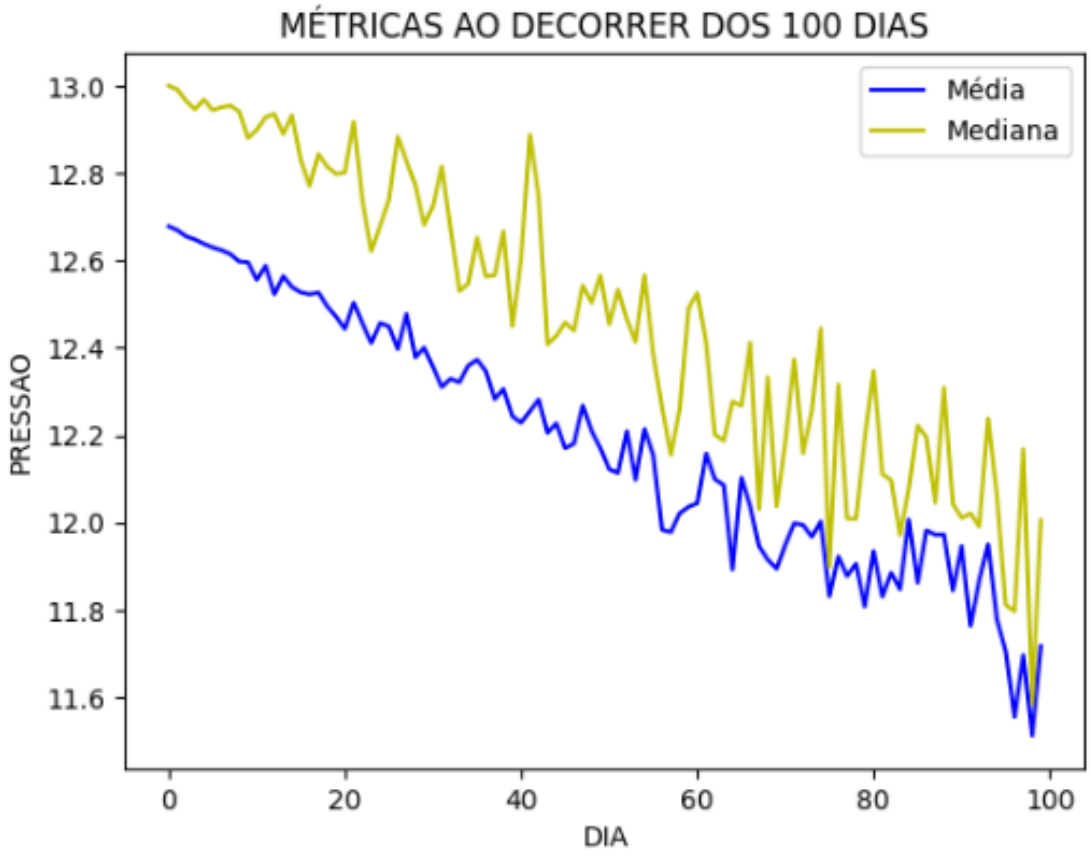
```
Mediana batimento:  70.62
Mediana Pressao:  12.45
Mediana Temperatura:  36.89
```

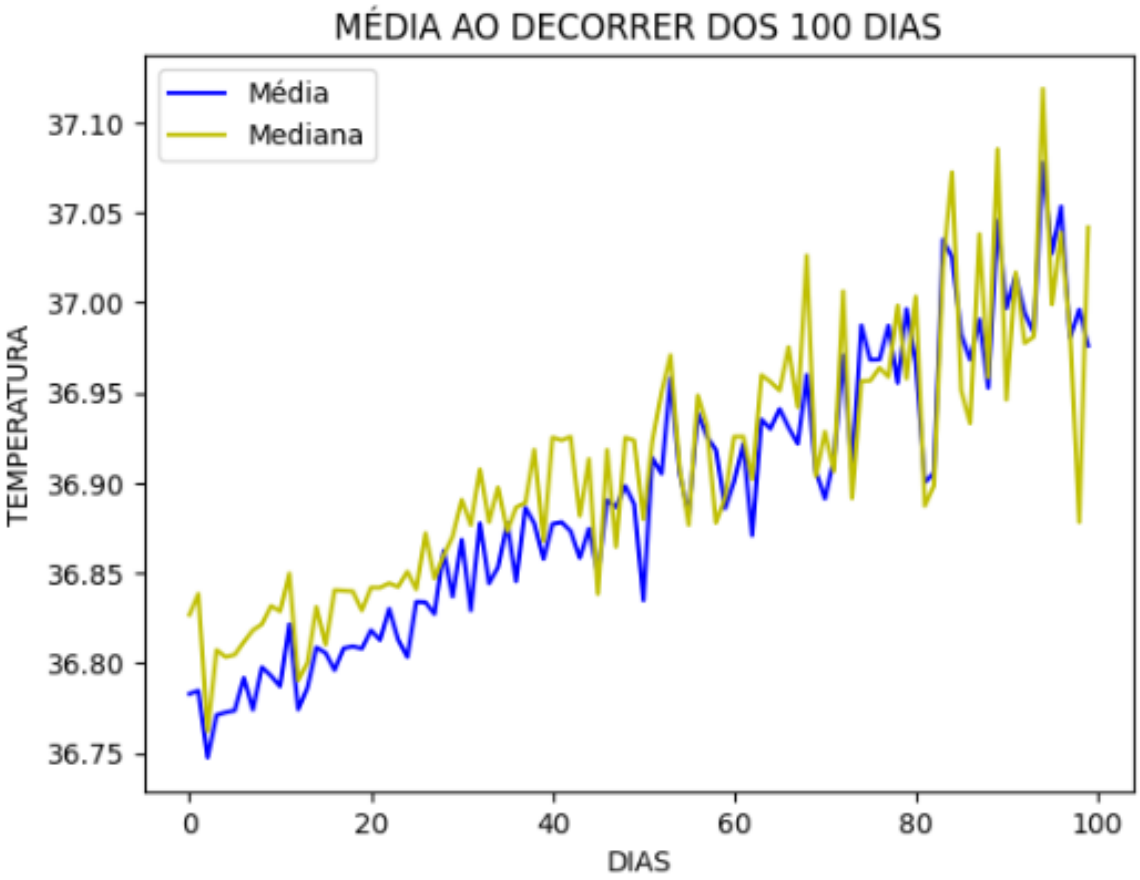
```
#Calculo moda - valida por dia
print('Moda batimento: ',round(stats.mode(data['Batimento']),2))
print('Moda Pressao: ',round(stats.mode(data['Pressao']),2))
print('Moda Temperatura: ',round(stats.mode(data['Temperatura']),2))
```

```
Moda batimento:  70.69
Moda Pressao:  12.35
Moda Temperatura:  36.83
```

São apresentados também os valores de mediana e moda. Nas 3 variáveis tanto médica,quanto mediana e moda são valores próximos. A seguir é verificado o comportamento da média e da mediana ao longo dos dias para as 3 variáveis.

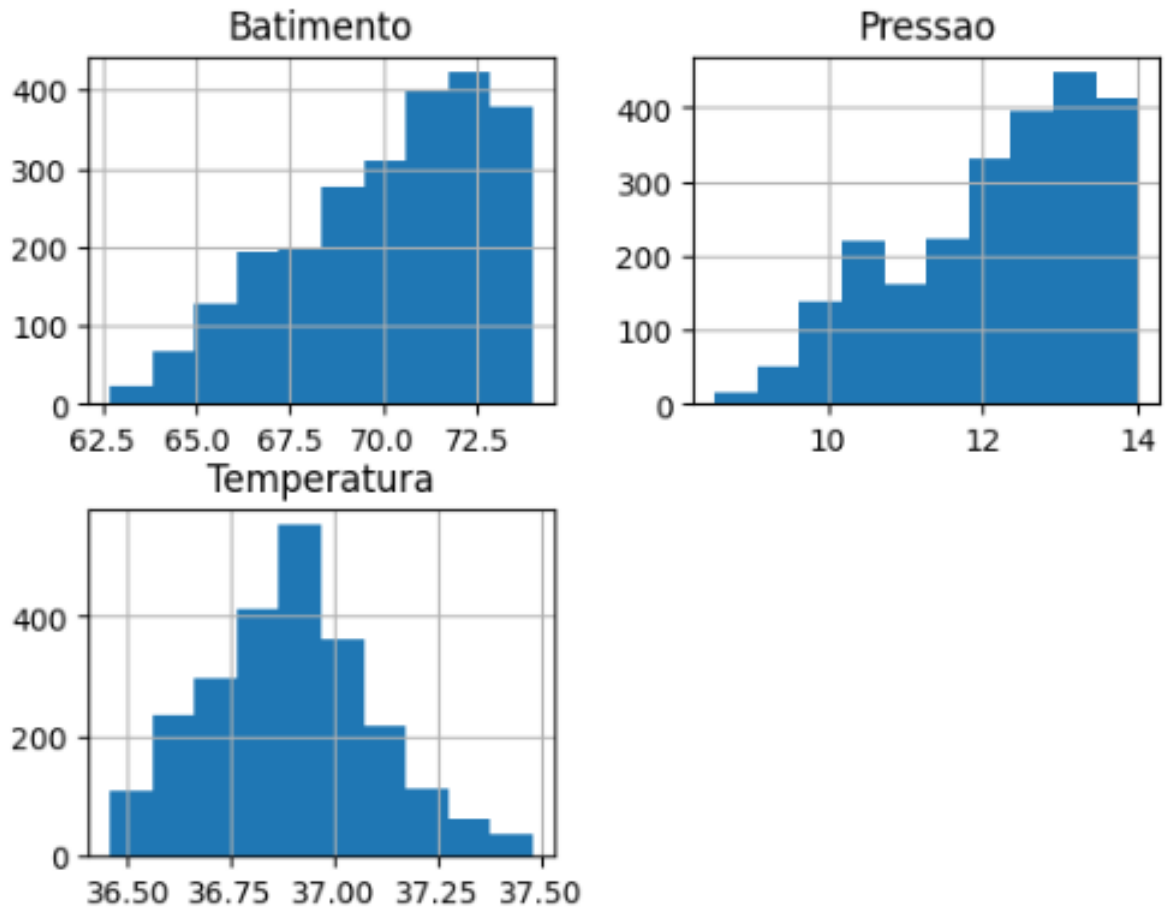



MÉTRICAS AO DECORRER DOS 100 DIAS
Média
Mediana
PRESSAO
DIA
13.0
12.8
12.6
12.4
12.2
12.0
11.8
11.6
0
20
40
60
80
100



Nos 3 casos a média, mediana e moda apresentaram comportamentos similares. Para a Pressão e Batimento eles decaíram e para temperatura cresceram ao longo dos dias. Nos 3 gráficos a mediana apresenta geralmente valores superiores a média. A seguir os histogramas de cada variável.

Histogramas



Observa-se pelos gráficos que a distribuição dos dados para o batimento e pressão são parecidos e possuem uma assimetria à direita. Em contrapartida, a distribuição da temperatura é mais simétrica, com um pico próximo ao centro e é diferente das variações das demais variáveis. Pelos passos anteriores constatou-se que a mediana é maior que a média. Isso pode ser observado sobretudo pela assimetria dos gráficos de batimento e pressão à direita.

Pelos histogramas observou-se os valores para cada uma das variáveis:

- Batimento: Valores geralmente maiores de 60 e menores de 80 bpm.
- Pressão: Valores geralmente maiores de 80 e menores ou iguais a 150 mmHg.
- Temperatura: Valores geralmente maiores que 36 e menores ou iguais a 37,5°C.

Considerando os intervalos de valores comumente adotados pela medicina, tem-se:

- Batimento : de 50 a 100 bpm.
- Pressão: de 96 a 128 mmHg.

- Temperatura: de 36 a 37,2°C.

Para a construção do sistema foi considerado os valores comumente adotados pela medicina. Pelo histograma observa-se que alguns valores já trazem resultados de estados de alerta enquanto outros são referentes a pacientes saudáveis.

Sistema de monitoramento dos sinais vitais

```
#Sistema considerando valores normais adotados na medicina
def main():
  try:
    #entradas
    b, p, t = input("Digite o batimento, a pressao e a temperatura do paciente: ").split()
    #processamento
    b, p, t = float(b), int(p), int(t)
    if (b < 50.0) or (b > 100.0):
      print("Alarme de emergência")
    elif (p < 9.6) or (p > 12.8):
      print("Alarme de emergência")
    elif (t < 36.0 ) or (t > 37.2):
      print("Alarme de emergência")
    else:
      print("O paciente esta saudavel!")
  except:
    print("Numeros invalidos. Favor, repetir a operaçao!")


if __name__ == "__main__":
    main()
```

```
Digite o batimento, a pressao e a temperatura do paciente: 60 12 36
O paciente esta saudavel!
```

O sistema é capaz de classificar o paciente como saudável ou emitir um alerta de caso o valor capturado para uma das variáveis não esteja de acordo com os valores normais esperados para o paciente considerando os intervalos comumente adotados pela medicina. Observa-se que esses valores estão de acordo com os histogramas obtidos também, englobando a maior parte dos valores medidos.

Caso ocorra um erro de digitação ou sejam digitados caracteres ao invés de valores válidos para a lógica do sistema, o sistema apresenta a mensagem de números inválidos e solicita ao usuário repetir a operação.

Conclusão

Por fim, após a realização da análise exploratória, foi possível constatar que a pressão variou entre 80 e 150 mmHg, o batimento entre 60 e 80 bpm e a temperatura entre 36° e 37,5° Celsius. Ademais, houve uma forte correlação positiva entre a pressão e o batimento, uma correlação fraca positiva entre a temperatura e a pressão e uma correlação fraca negativa entre a temperatura e o batimento. A pressão e o batimento mantiveram-se relativamente constantes, enquanto a temperatura aumentou ao longo dos dias. Assim, foi experimentada na prática a importância dos dados para auxiliar na tomada de decisões, e no caso explorado, para auxiliar, especificamente, na tomada de decisões em relação à saúde e ao bem-estar das pessoas monitoradas.