DataScientist Society

”情報Ⅱで学んだ内容が実社会でどのように使われているか”の紹介
1つの事例を深く知る
データサイエンス×コピー機製造工場

一般社団法人 データサイエンティスト協会 学生委員会

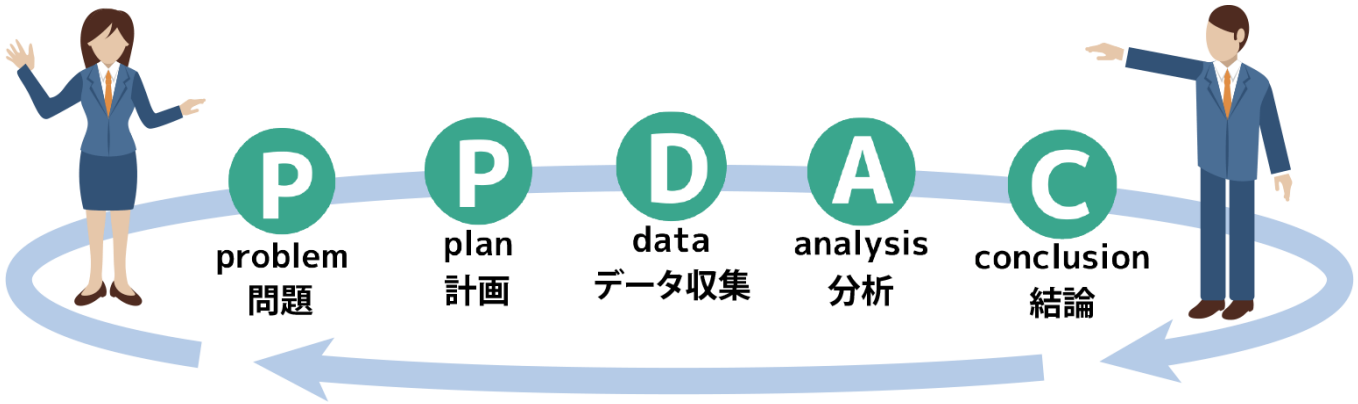
2024年8月29日

園山 将士

実社会での使われ方

データサイエンスはあくまでも”手段”
実社会のどのような課題を解決するかの設計が大事！

データサイエンスを活用した現場改善では、
教育現場でも知られているPPDACサイクルに近いプロセスで進めることが多いため、
今回はPPDACサイクルに則った、工場でのデータサイエンス活用事例を紹介します。



出典：総務省統計局ホームページ
https://www.stat.go.jp/dstart/point/seminar1/01.html

一般社団法人
データサイエンティスト協会

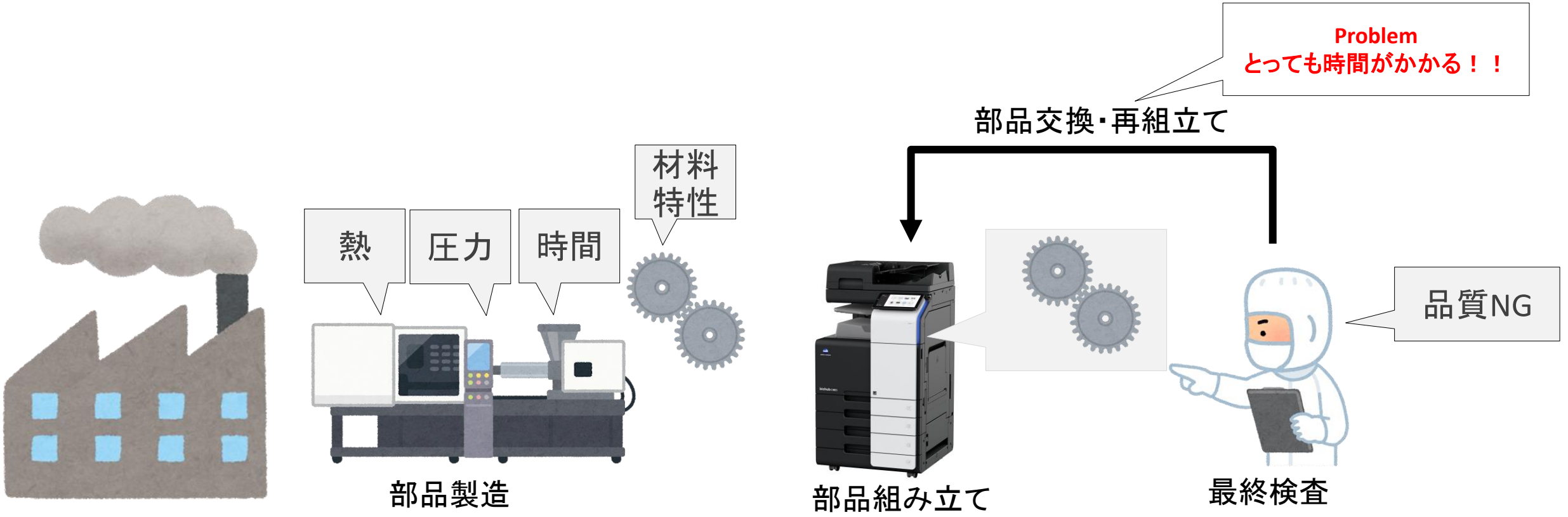
ビジネススキル向上のための課題解決型人材コンテスト2024（応募期間：～6/7まで）

データ活用に必要なスキルはAI/機械学習などのテクノロジーにとどまらず、ビジネスオーナーの課題を適切に理解して実行可能な解決策を提示することにあると言われています。当協会はそれに応える試みとして、実課題と実データを用いてアクションの提案までを行う分析コンテストを実施いたします。他業界のメンバーとチームを組み3か月間、メンターのサポートを受けながら分析プロジェクトを推進することで、ビジネススキルを身につけていただける機会です。分析スキルをさらに業務で活かして行きたいとお考えの方は是非ご参加ください。

出典：データサイエンティスト協会ホームページ
https://www.datascientist.or.jp/news/n-news/post-2890/

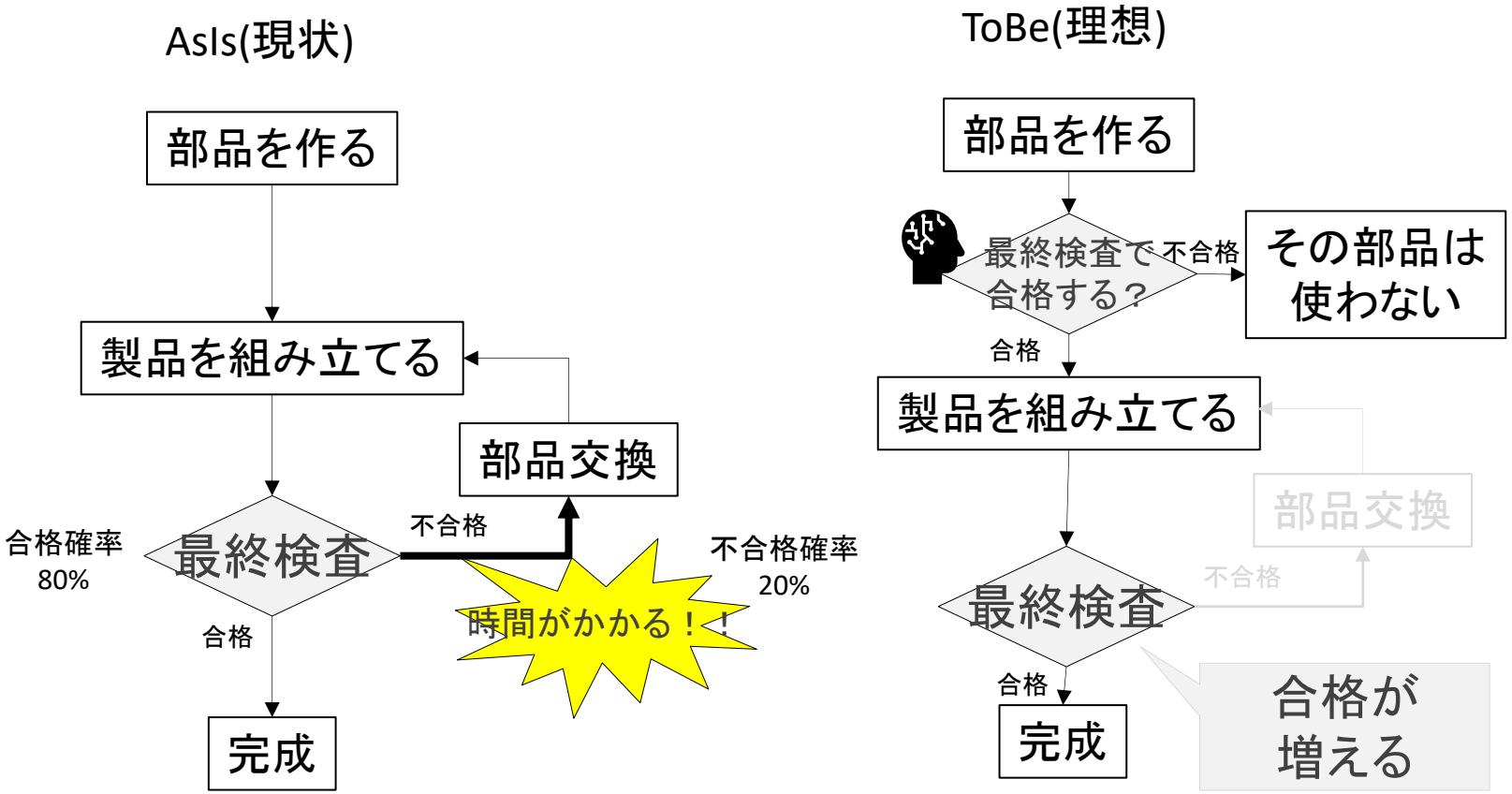
Problem(問題の設定)

Problem：コピー機を組み立てる工場で、ギア部品の不良によって最終検査工程で品質NGが発生する。
その時は工場で部品交換・再組立てを行うが、作業にとても時間がかかるのが問題となっている。



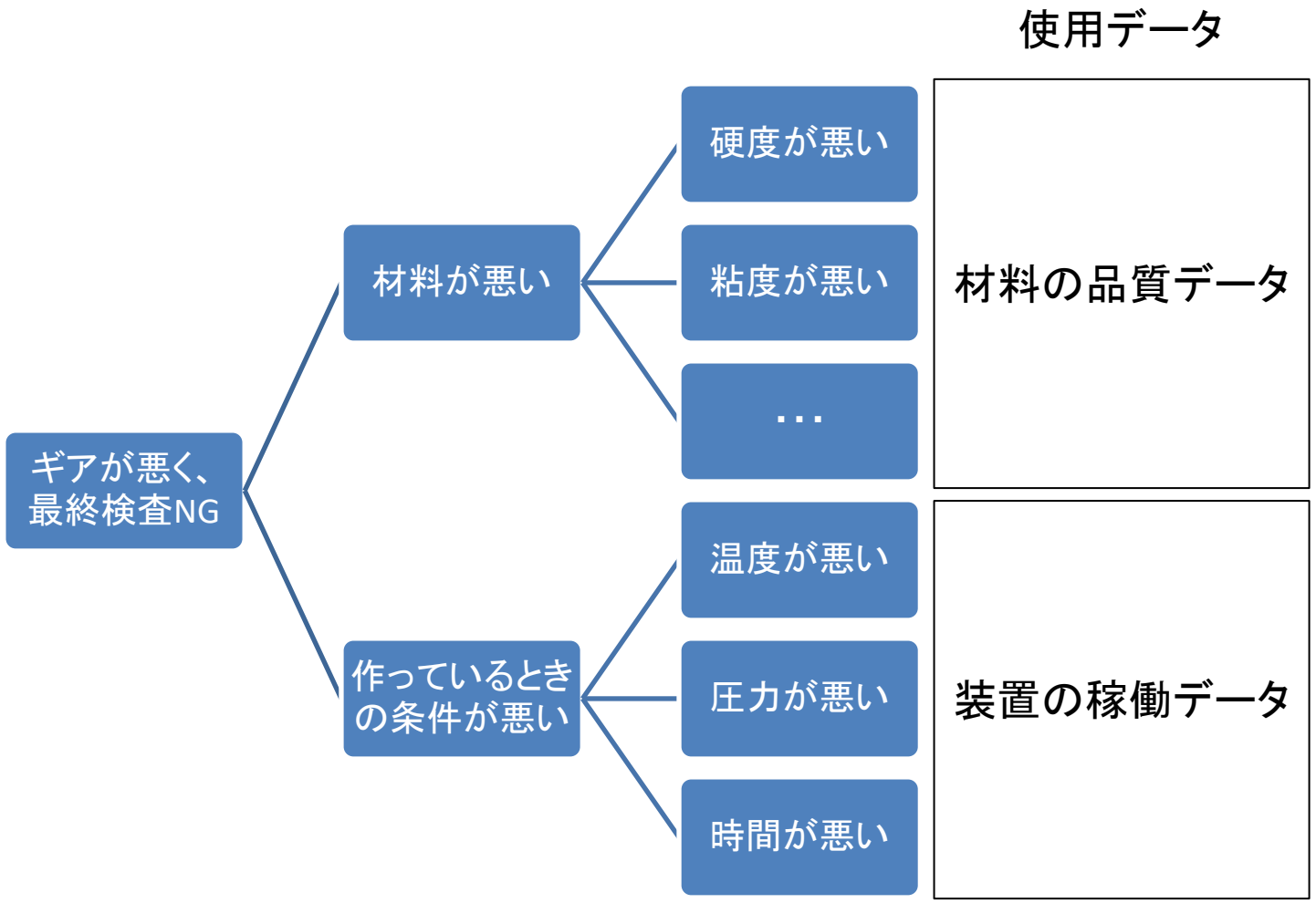
Plan(ワークフロー作成)

Plan：ギア部品を作った時に、最終検査で合格するかどうかを事前に判断することで
最終検査の不合格を減らし、部品交換にかかる時間を削減する。



Plan(要因分析)

Plan：なぜその部品で不合格が発生するかの要因を分解して、使用するデータを特定する。



Data(データの取得)

Data：材料の品質データは、材料メーカからの購入時データを利用。装置の稼働データは装置から取得。
最終検査は1が不合格、0が合格を表している。今回はリアルに近いダミーデータとして、sample_data.csvを準備

材料の品質データ

設定値
硬度：80
粘度：22　g/10min
・・・・

装置の稼働データ

設定値
温度：240℃
圧力：70MPa
時間：18秒

温度	圧力	時間	硬度	粘度	最終検査
239.9739	69.86723	19.73163	77	22	1
239.8371	70.80166	19.6917	77	22	0
240.6744	68.94366	19.28832	77	22	0
241.031	70.1886	18.89654	77	22	0
241.8369	68.5282	18.74179	77	22	0

sample_data.csvはお渡しします！

Analysis(データ可視化)

Analysis：取得したデータの全体像を把握してから詳細データの確認を行い、
目的変数に対して各データがどのような傾向にあるかを分析していく。

全体

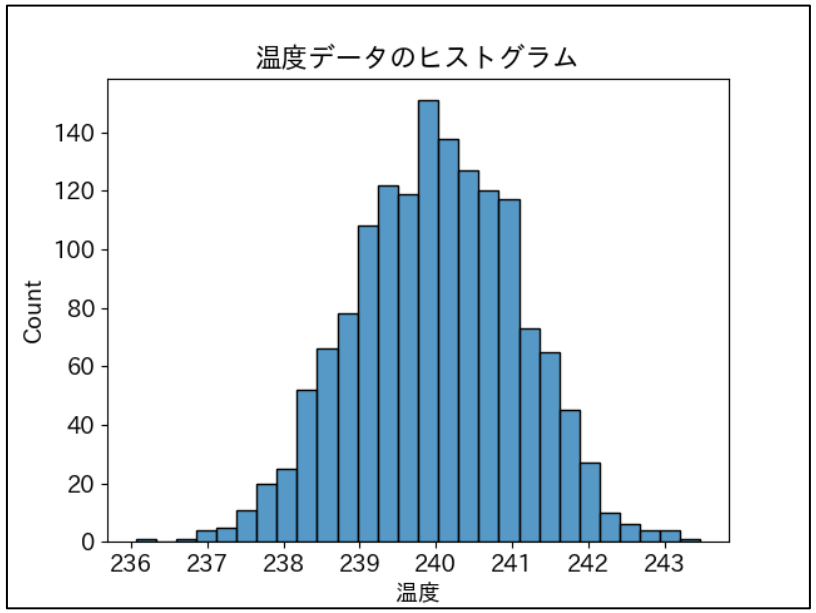
詳細

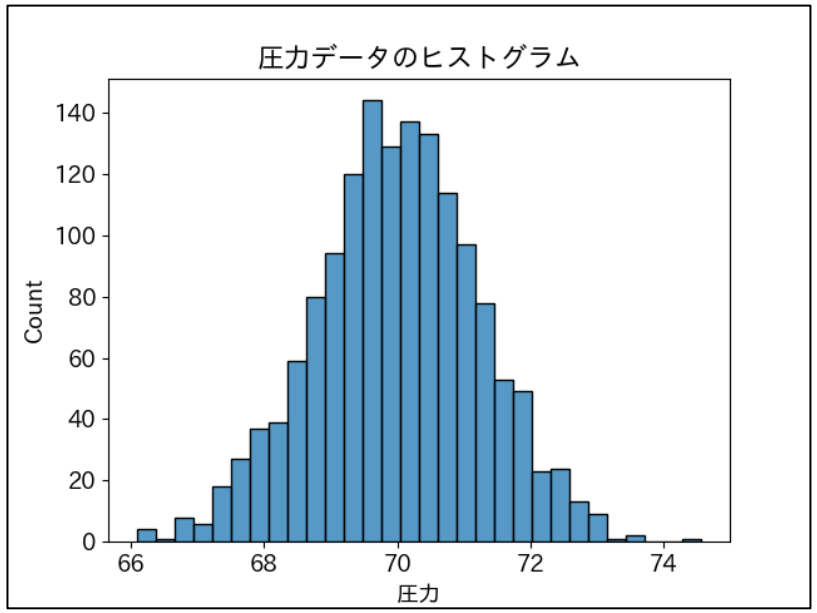
演習：
演習①
・データは何行×何列あるか確認しよう。
・データの最初の5行の値を確認して、どのような列名があるかを確認しよう。
・欠損値が無いかを確認しよう。
・要約統計量を確認しよう。
・最終検査が不合格の件数と割合(%)を確認しよう。
演習②
・(1変量分布)温度・圧力・時間の分布はどのようになっているか？ヒストグラムで確認しよう。
・(1変量分布)硬度・粘度はどの数値がそれぞれいくつあるか？
演習③
・(2変量分布)圧力と温度にはどのような関係があるか？温度をx軸、圧力をy軸の散布図と、相関係数で確認しよう。
演習④
・合格品と不合格品の温度にはどのような違いがあるか？色分けされたヒストグラムで確認してみよう。
・温度で不良品を判別するためにはどうすればよいか考えてみよう。
・合格品と不合格品の時間にはどのような違いがあるか？色分けされたヒストグラムで確認してみよう。
・時間で不良品を判別するためにはどうすればよいか考えてみよう。

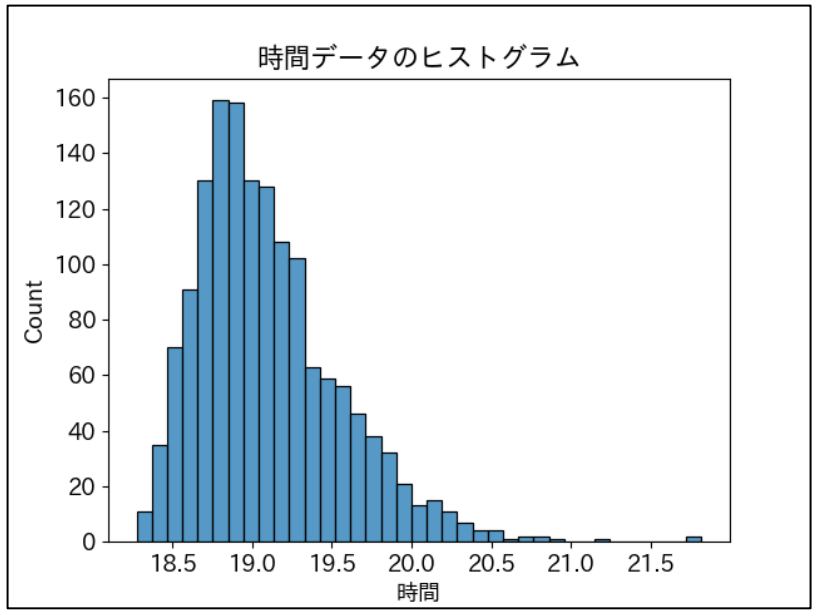
演習の回答pythonコードもお渡しします！

演習答え　例

演習②
・(1変量分布)温度・圧力・時間の分布はどのようになっているか？ヒストグラムで確認しよう。



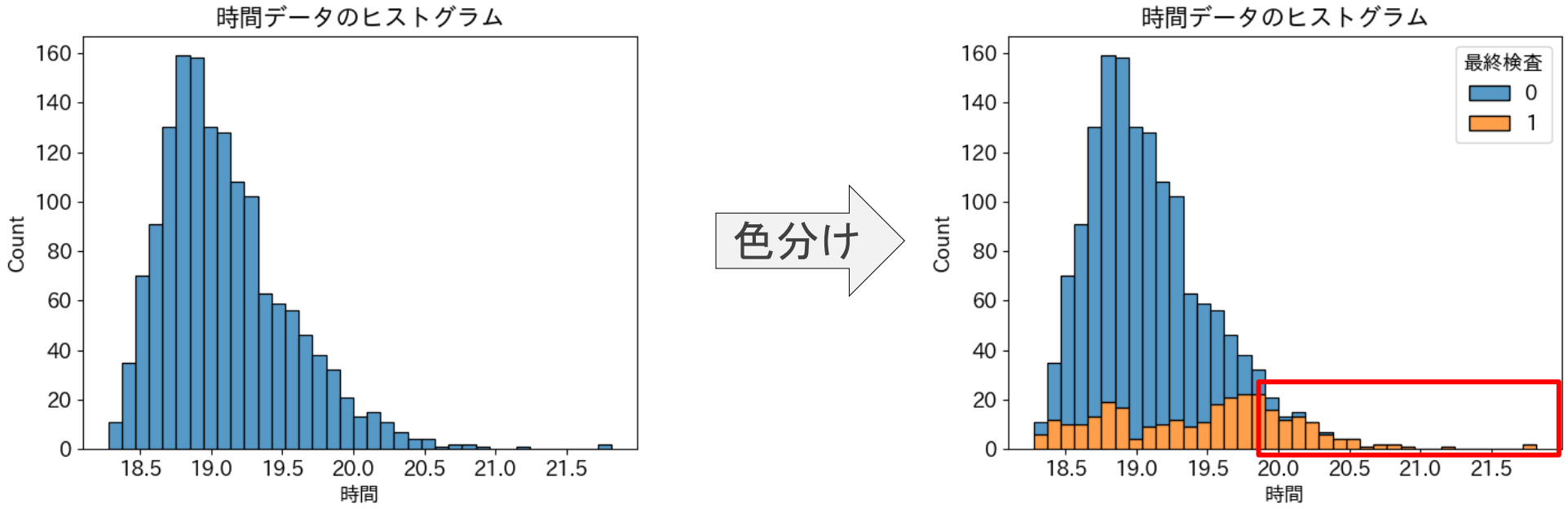




温度：240付近を中心に釣り鐘型の分布
圧力：70付近を中心に釣り鐘型の分布
時間：19付近を中心に、右に裾野が長い分布

演習答え　例

演習④
・合格品と不合格品の時間にはどのような違いがあるか？色分けされたヒストグラムで確認してみよう。
・時間で不良品を判別するためにはどうすればよいか考えてみよう。

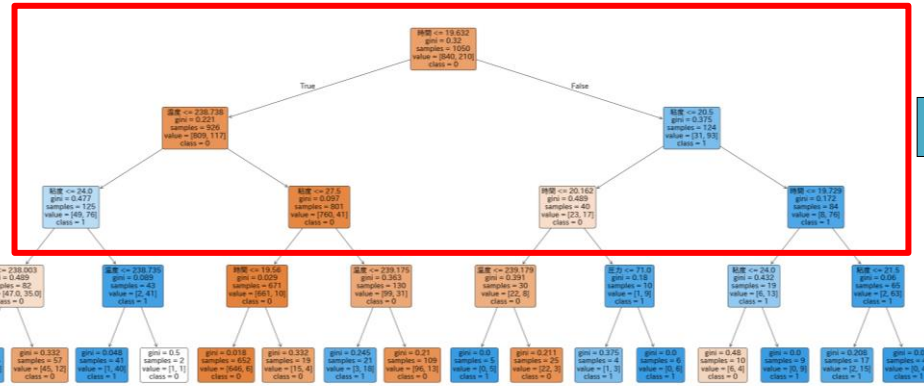


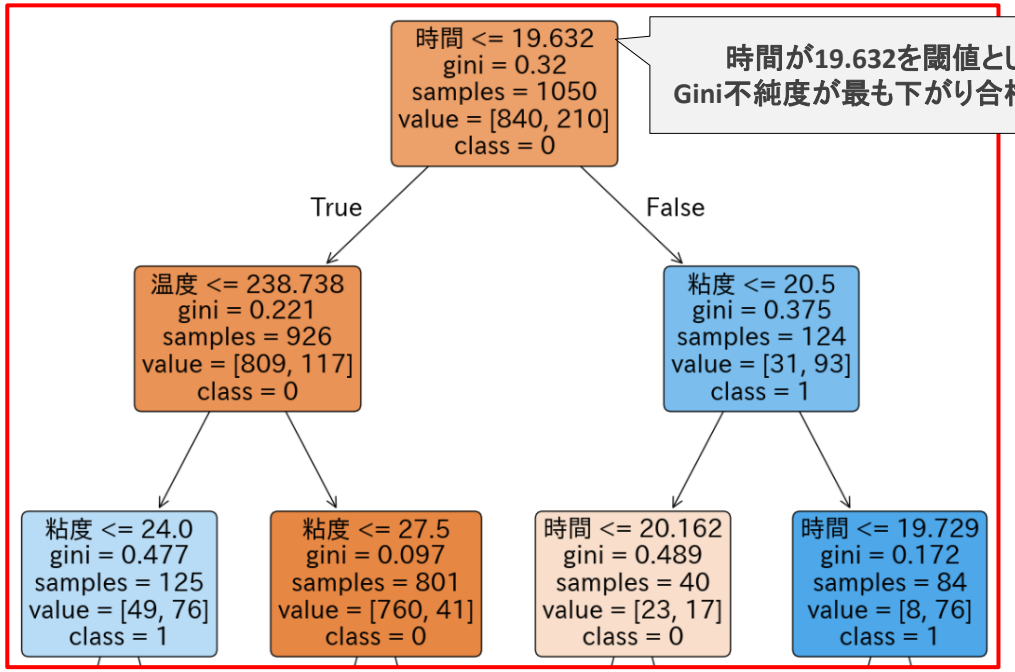
・時間が長い方が、最終検査で不合格品になる割合が高い傾向にある。
・時間が20秒以上であると大半が不合格になるので判別することができるかも。

Analysis(予測モデル構築)

Analysis：最終検査で合格になるかどうかを、決定木を使って予測。

予測するデータ：最終検査の合格/不合格





時間が19.632を閾値としてデータを分けると
Gini不純度が最も下がり合格/不合格を分けられる。

	変数重要度
温度	0.445794
時間	0.381298
粘度	0.171704
圧力	0.001205
硬度	0.000000

リコール0.73 = 最終検査不合格を約73%で検出可能

演習で考えた付近で分けられているかな？

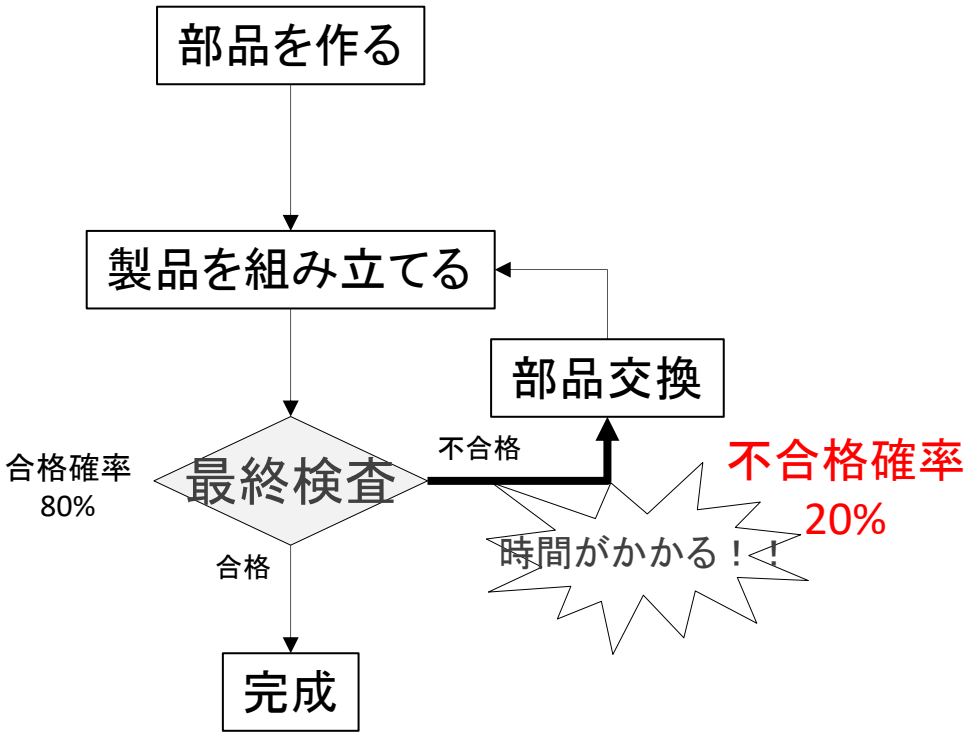
今回作成したモデルにおいては、
温度・時間・粘度の順に重要で、
圧力や硬度がモデルに与える影響は少ない

モデルの作成pythonコードもお渡しします！

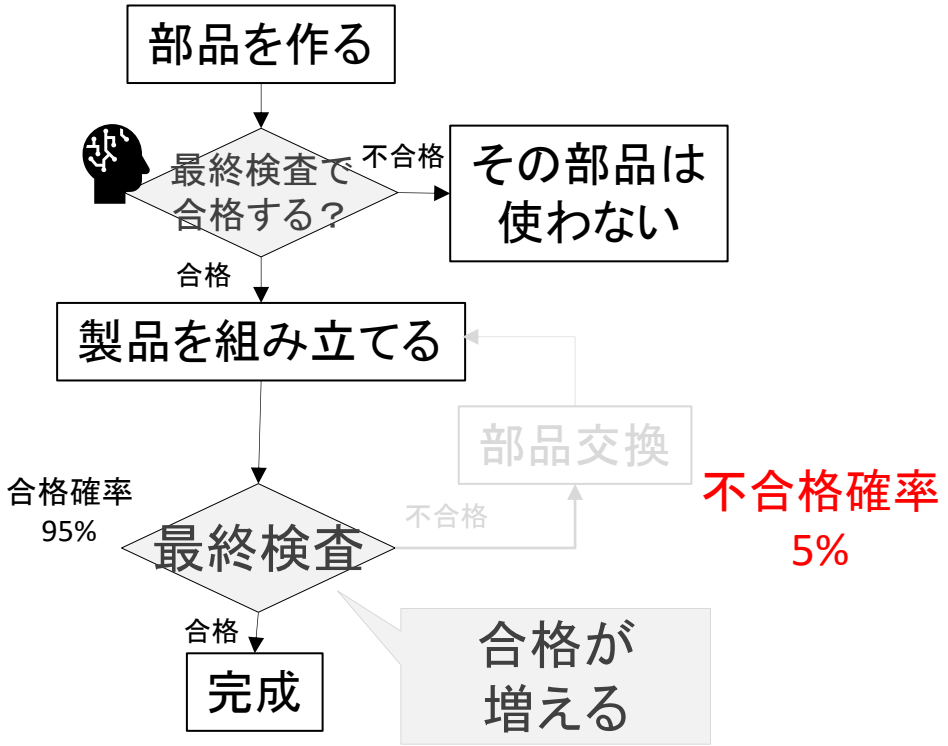
Conclusion(結論)

Conclusion：分析により、最終検査で不合格になる部品を73%で検出可能になり、
それにより最終検査不合格率は20%⇒5%に低減できる。
Problemで立てた“最終検査に時間がかかる”という課題を解決できるので、
このモデルを実際に工場で使うことを提案する。

AsIs(現状)



ToBe(理想)



【情報Ⅱ】第1回全国指導力向上研修会アンケート

スマホで読み取り
ご回答お願いいたします。
匿名回答、2分で終わります！

以上、ご意見とアンケートの回答よろしくお願いします。

演習答え

演習①

・データは何行×何列あるか確認しよう。

```
print(df.shape)
```

✓ 0.0s

```
(1500, 6)
```

1500行×6列

・データの最初の5行の値を確認して、どのような列名があるかを確認しよう。

```
df.head(5)
```

✓ 0.0s

	温度	圧力	時間	硬度	粘度	最終検査
0	239.973925	69.867233	19.731631	77	22	1
1	239.837117	70.801655	19.691697	77	22	0
2	240.674449	68.943656	19.288324	77	22	0
3	241.030988	70.188601	18.896543	77	22	0
4	241.836859	68.528195	18.741790	77	22	0

温度、圧力、時間、硬度、粘度、最終検査の列がある。

演習答え

演習①

・欠損値が無いかを確認しよう。

```
df.isnull().sum()
```

✓ 0.0s

```
温度      0
圧力      0
時間      0
硬度      0
粘度      0
最終検査    0
dtype: int64
```

各列で欠損値は無し。

・要約統計量を確認しよう。

```
df.describe()
```

✓ 0.0s

	温度	圧力	時間	硬度	粘度	最終検査
count	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000
mean	239.999439	69.994083	19.099402	80.333333	22.000000	0.200000
std	1.085978	1.235918	0.452811	2.357809	4.727392	0.400133
min	236.058264	66.085558	18.276073	77.000000	14.000000	0.000000
25%	239.252822	69.198474	18.779735	78.000000	20.000000	0.000000
50%	240.007261	70.008781	19.014476	80.500000	21.500000	0.000000
75%	240.772841	70.823408	19.331784	82.000000	26.000000	0.000000
max	243.475720	74.566506	21.818082	84.000000	29.000000	1.000000

min,maxで明らかにおかしい外れ値が無い事を確認。

・最終検査が不合格の件数と割合(%)を確認しよう。

```
df["最終検査"].value_counts()
```

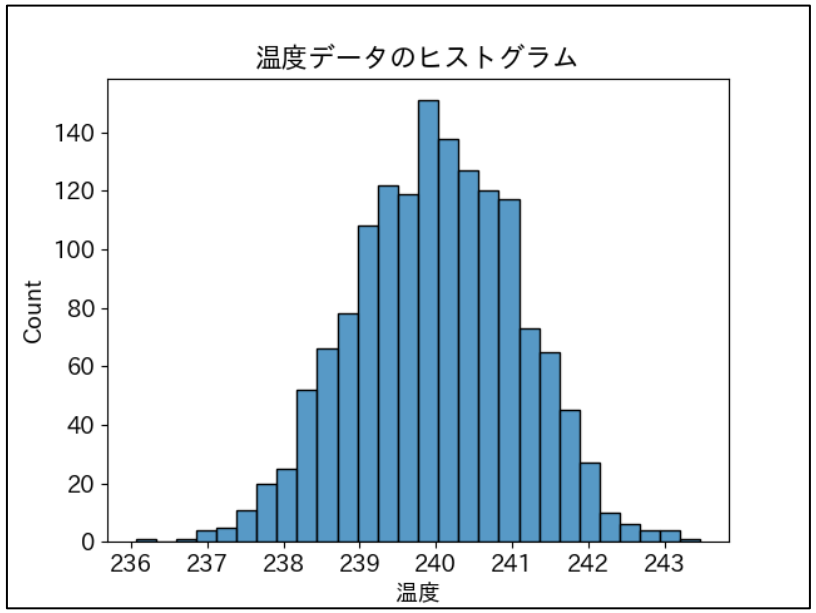
✓ 0.0s

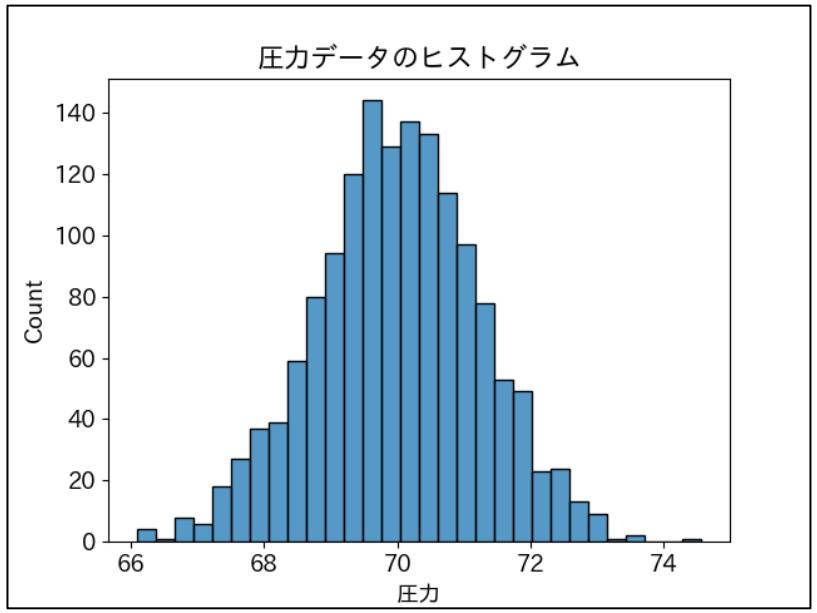
```
最終検査
0    1200
1     300
```

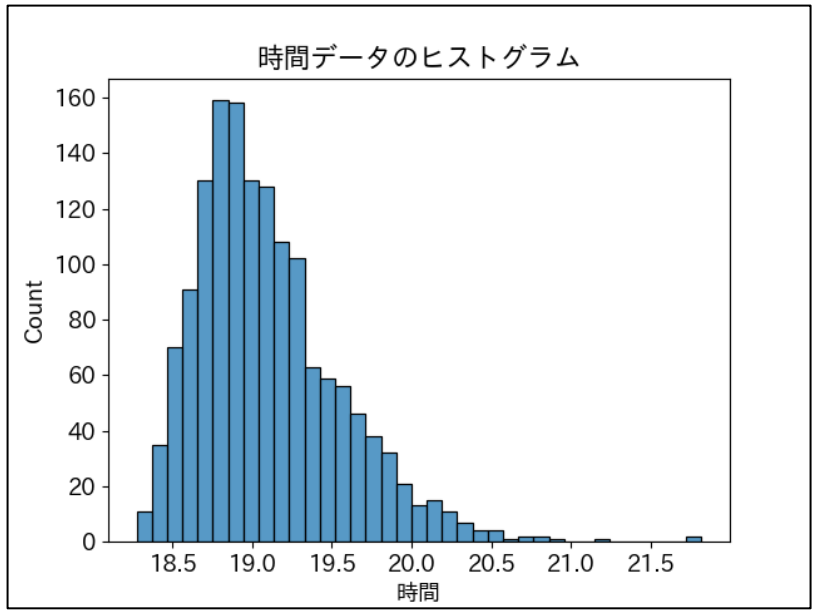
不合格は300件、20%ある。

演習答え

演習②
・(1変量分布)温度・圧力・時間の分布はどのようになっているか？ヒストグラムで確認しよう。







温度：240付近を中心に釣り鐘型の分布
圧力：70付近を中心に釣り鐘型の分布
時間：19付近を中心に、右に裾野が長い分布

演習答え

演習②
・(1変量分布)硬度・粘度はどの数値がそれぞれいくつあるか？

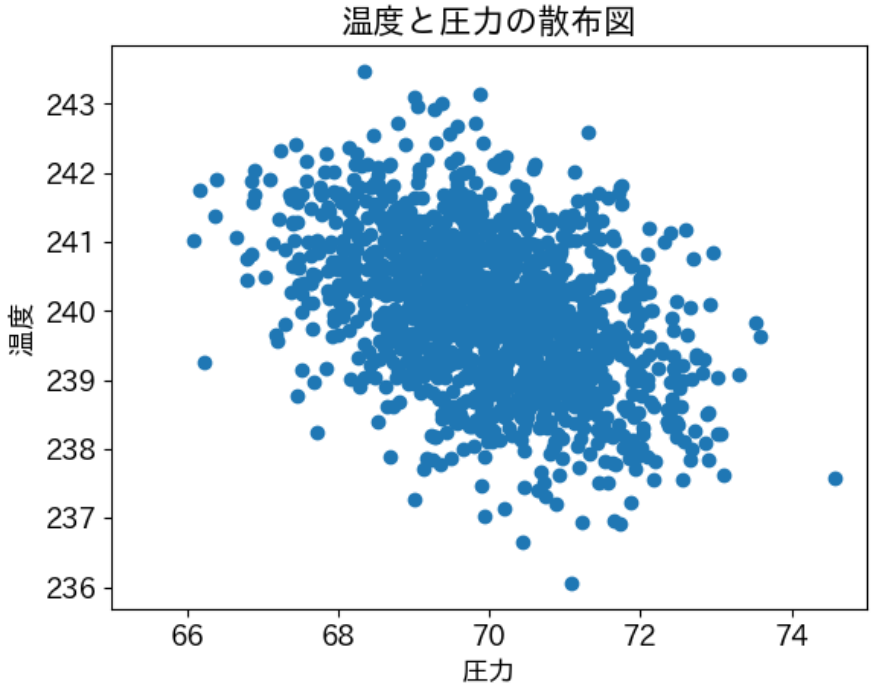
```
df["硬度"].value_counts()
✓ 0.0s
硬度
77    250
78    250
80    250
81    250
82    250
84    250
Name: count, dtype: int64
```

```
df["粘度"].value_counts()
✓ 0.0s
粘度
22    250
20    250
26    250
21    250
29    250
14    250
Name: count, dtype: int64
```

硬度
77,78,80,81,82,84の硬度がそれぞれ250データずつある。
粘度
22,20,26,21,29,14の年度がそれぞれ250データずつある。

演習答え

・(2変量分布)圧力と温度にはどのような関係があるか？温度をx軸、圧力をy軸の散布図と、相関係数で確認しよう。




```
df.corr()
```
✓ 0.0s

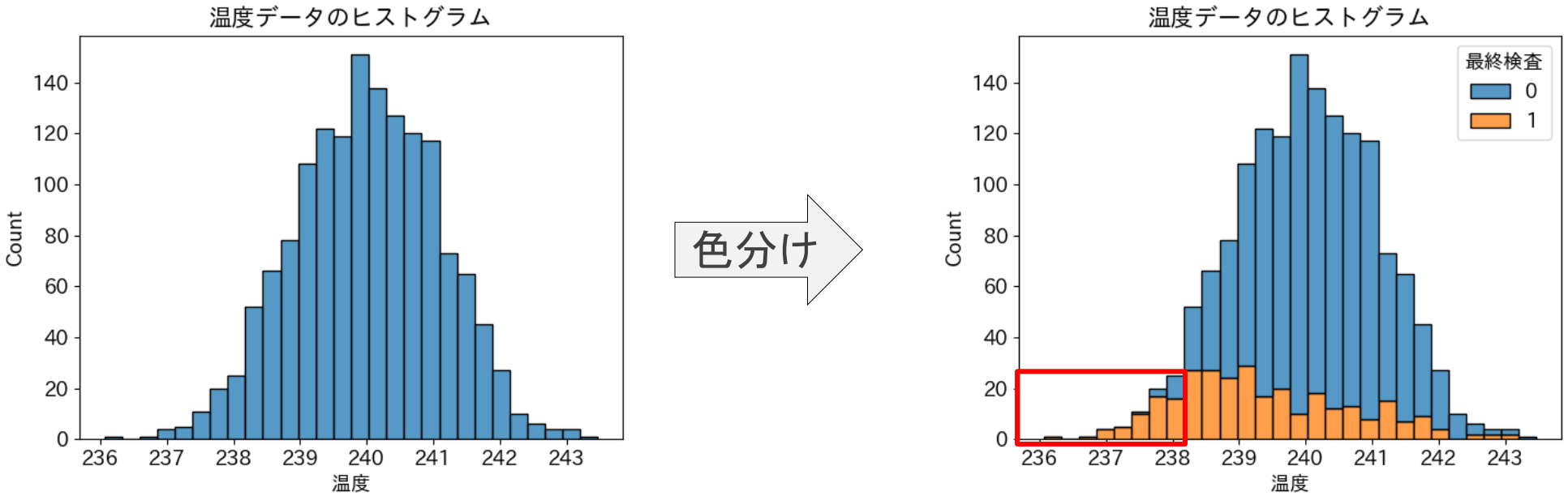
	温度	圧力	時間	硬度	粘度	最終検査
温度	1.000000	-0.413766	0.260059	0.078565	-0.035821	-0.291119
圧力	-0.413766	1.000000	0.341858	-0.078410	0.007121	0.326429
時間	0.260059	0.341858	1.000000	-0.009835	-0.012497	0.403129
硬度	0.078565	-0.078410	-0.009835	1.000000	-0.224440	-0.005657
粘度	-0.035821	0.007121	-0.012497	-0.224440	1.000000	0.284254
最終検査	-0.291119	0.326429	0.403129	-0.005657	0.284254	1.000000

大きく外れているプロットはない。
温度の数値が大きくなると圧力が小さくなる、弱い負の相関がある。

散布図と相関係数は必ずセットで確認する。その理由は以下。
https://bellcurve.jp/statistics/course/9591.html

演習答え

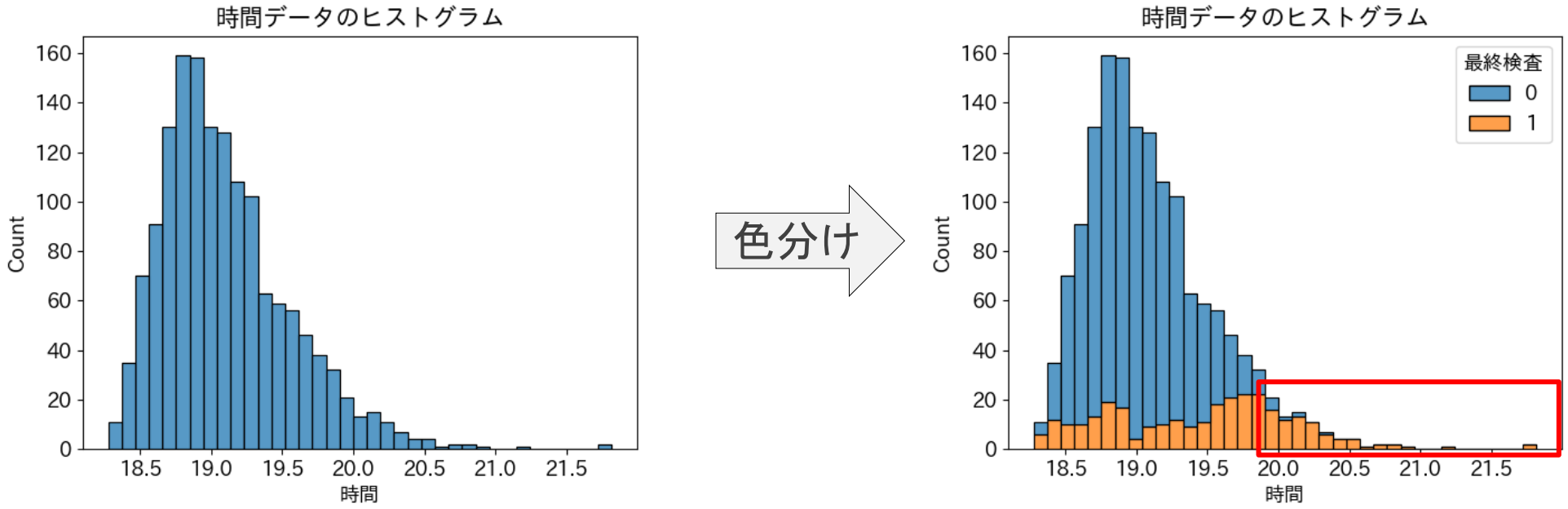
演習④
・合格品と不合格品の温度にはどのような違いがあるか？色分けされたヒストグラムで確認してみよう。
・温度で不良品を判別するためにはどうすればよいか考えてみよう。



・温度が低い方が、最終検査で不合格品になる割合が高い傾向にある。
・温度が238℃以下であれば、大半が不合格になるので判別することができるかも。

演習答え

演習④
・合格品と不合格品の時間にはどのような違いがあるか？色分けされたヒストグラムで確認してみよう。
・時間で不良品を判別するためにはどうすればよいか考えてみよう。



・時間が長い方が、最終検査で不合格品になる割合が高い傾向にある。
・時間が20秒以上であると大半が不合格になるので判別することができるかも。