now loading

blanktar.jp

目黒研究室 基礎技術講座

blanktar.jp

講座の流れ

- 11月 2日 **プロトタイピングを支える技術**
  卒研で使うかもしれない技術の紹介。

- 11月16日 **デザインからプロトタイピングまで**
  要件を決めて、プロトタイプを作って動かすまでの流れ。

- 12月14日 **最低限の統計学**
  教授にボコボコにされないためのデータの集め方、使い方。

講座の流れ

- 11月 2日 **プロトタイピングを支える技術**
  卒研で使うかもしれない技術の紹介。

- 11月16日 **デザインからプロトタイピングまで**
  要件を決めて、プロトタイプを作って動かすまでの流れ。

- 12月14日 **最低限の統計学**
  教授にボコボコにされないためのデータの集め方、使い方。

講座の流れ

- 11月 2日 **プロトタイピングを支える技術**
  卒研で使うかもしれない技術の紹介。

- 11月16日 **デザインからプロトタイピングまで**
  要件を決めて、プロトタイプを作って動かすまでの流れ。

12月21日

- ~~12月14日~~ **最低限の統計学**
  教授にボコボコにされないためのデータの集め方、使い方。

この講座について

卒業研究で役に立ちそうな知識をお届けする講座。全3回。

ぶっちゃけ時間が無いのでざっくりやります。
詳しく知りたい方は直接聞いてください。

スライドは補足資料付きで公開してあります。

goo.gl/h8Kp7T

目黒研究室 基礎技術講座 第三回

最低限の統計学

blanktar.jp

TL;DR.

最初の5分で言いたいこと全部言います。

TL;DR.

あとは寝てても良いよ。

TL;DR.

~~あとは寝てても良いよ。~~

Too Long; Don't Read.

統計学

うわぁ数学だぁ…

目黒研でよく聞く会話例

※誇張されています

目黒研でよく聞く会話例

「SNS利用率のアンケートを twitterで拡散してもらいました。」

目黒研でよく聞く会話例

「SNS利用率のアンケートを
twitterで拡散してもらいました。」

↓

SNS利用率: 100%

目黒研でよく聞く会話例

「SNS利用率のアンケートを
twitterで拡散してもらいました。」

↓

めっちゃ普及しててすごい！

目黒研でよく聞く会話例

※誇張されています

目黒研でよく聞く会話例

「この講座をどう思いますか？
1.役に立つ 2.為になる 3.楽しい 4.面白い」

目黒研でよく聞く会話例

「この講座をどう思いますか？
1.役に立つ 2.為になる 3.楽しい 4.面白い」

↓

肯定的回答: 100%

目黒研でよく聞く会話例

「この講座をどう思いますか？
1.役に立つ 2.為になる 3.楽しい 4.面白い」

↓

みんなが喜んでるめっちゃ良い講座だ！

目黒研でよく聞く会話例

※誇張されています

目黒研でよく聞く会話例

「先生、10人にアンケートを取ってきました！」

目黒研でよく聞く会話例

「先生、10人にアンケートを取ってきました！」
「うーん、最低でも50人は欲しいな」

目黒研でよく聞く会話例

「先生、10人にアンケートを取ってきました！」
「うーん、最低でも50人は欲しいな」
「はーい」

目黒研でよく聞く会話例

「先生、10人にアンケートを取ってきました！」
「うーん、最低でも50人は欲しいな」
「はーい」

↓

なんで50？

友達のアニオタ50人に、アニメが性犯罪を増長するかを問う

いやちょっと待て

それおかしくね？

雑なアンケートでも、やれば卒業出来ます。

でもそれじゃ

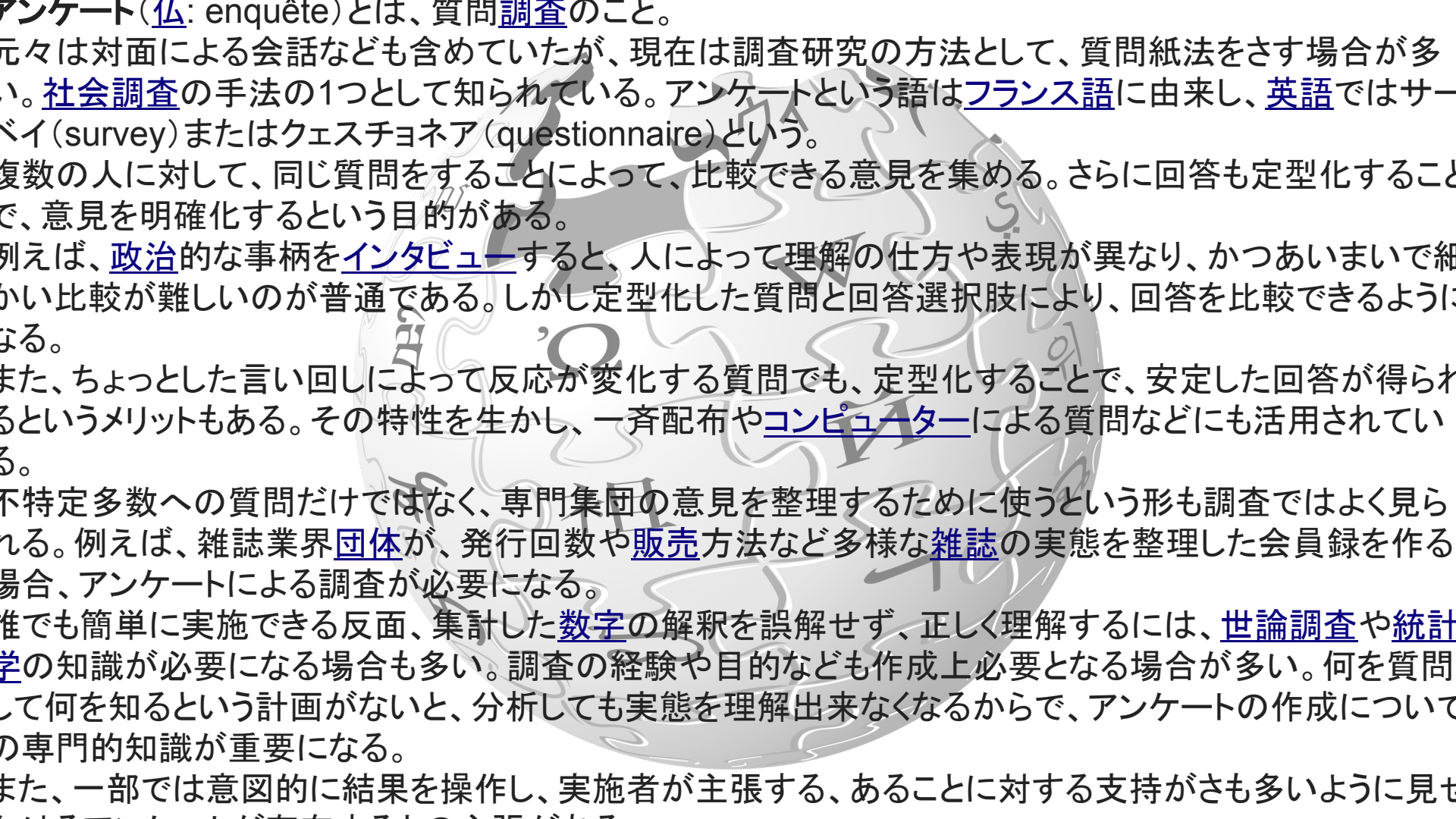
何も分からない

もしもちゃんとやりたいのなら

最低限の統計は抑えておこう

そもそも

アンケートって何だ

アンケート（仏: enquête）とは、質問調査のこと。
元々は対面による会話なども含めていたが、現在は調査研究の方法として、質問紙法をさす場合が多い。社会調査の手法の1つとして知られている。アンケートという語はフランス語に由来し、英語ではサーベイ（survey）またはクェスチョネア（questionnaire）という。
複数の人に対して、同じ質問をすることによって、比較できる意見を集める。さらに回答も定型化することで、意見を明確化するという目的がある。
例えば、政治的な事柄をインタビューすると、人によって理解の仕方や表現が異なり、かつあいまいで細かい比較が難しいのが普通である。しかし定型化した質問と回答選択肢により、回答を比較できるようになる。
また、ちょっとした言い回しによって反応が変化する質問でも、定型化することで、安定した回答が得られるというメリットもある。その特性を生かし、一斉配布やコンピューターによる質問などにも活用されている。
不特定多数への質問だけではなく、専門集団の意見を整理するために使うという形も調査ではよく見られる。例えば、雑誌業界団体が、発行回数や販売方法など多様な雑誌の実態を整理した会員録を作る場合、アンケートによる調査が必要になる。
誰でも簡単に実施できる反面、集計した数字の解釈を誤解せず、正しく理解するには、世論調査や統計学の知識が必要になる場合も多い。調査の経験や目的なども作成上必要となる場合が多い。何を質問して何を知るという計画がないと、分析しても実態を理解出来なくなるからで、アンケートの作成について  の専門的知識が重要になる。
また、一部では意図的に結果を操作し、実施者が主張する、あることに対する支持がさも多いように見せ

何人かの人に
選択肢を提示して
意見を聞いてみる

何人かの人に

選択肢を提示して

意見を聞いてみる

何人か = 全員じゃなくても良い

選択肢 = 全部聞かなくても良い

味噌汁を作る

↓

味見をする

味噌汁を作る

茶碗で味見をするか？

味見をする

味噌汁を作る

スプーンで十分

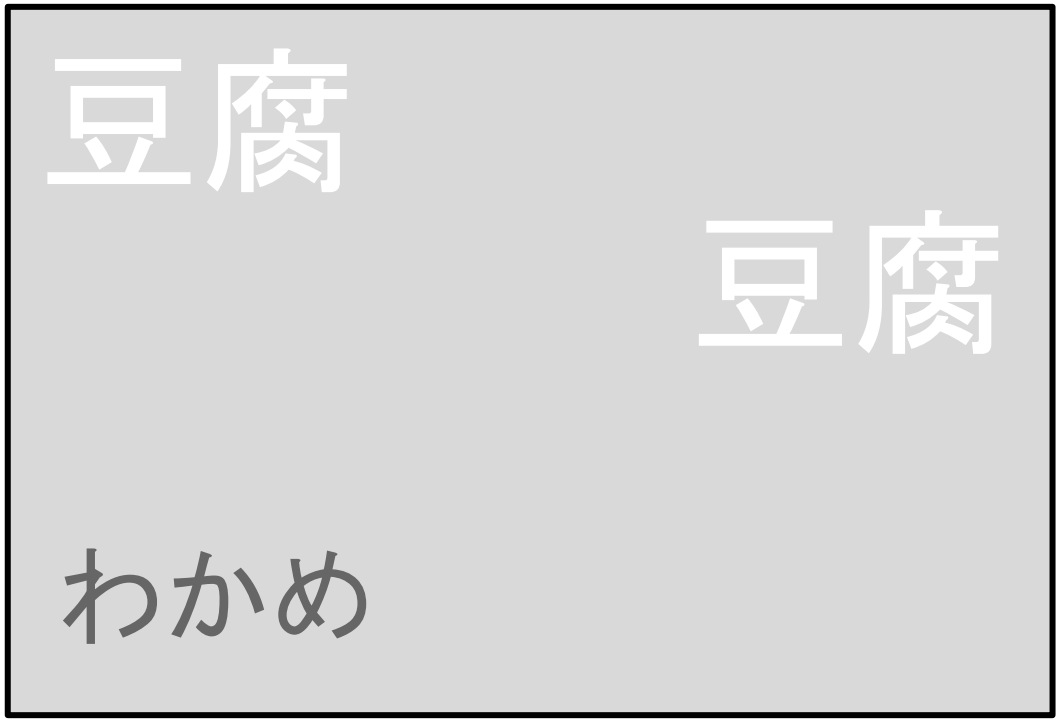
味見をする

スプーンで十分ではない場合

スプーンで十分ではない場合

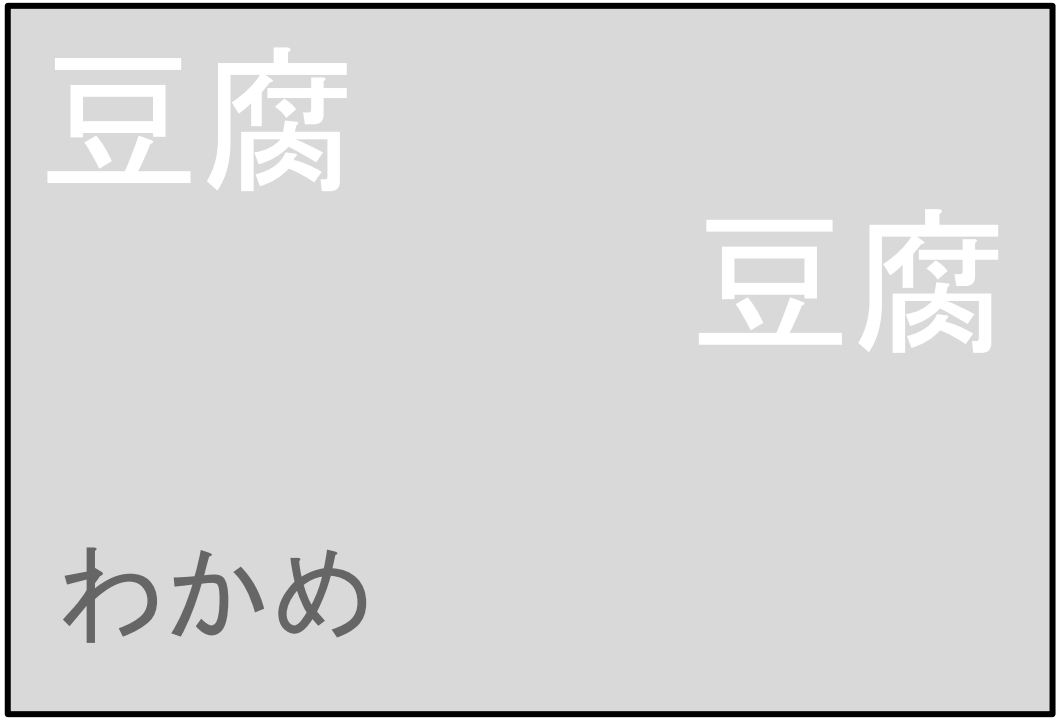
味噌汁を全く混ぜていないとき。

1. 鍋に水を入れる



1. 鍋に水を入れる
2. 具材を煮込む

味噌



1. 鍋に水を入れる
2. 具材を煮込む
3. 味噌を

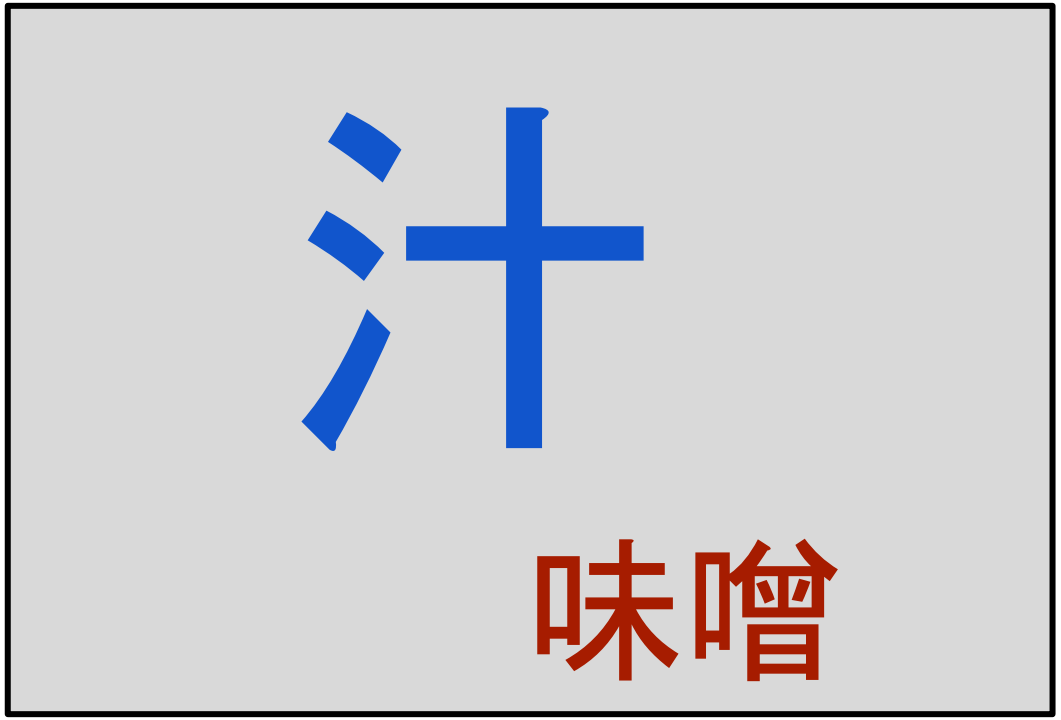
豆腐

豆腐

わかめ

味噌

1. 鍋に水を入れる
2. 具材を煮込む
3. 味噌を
4. ＼ドボーン／



1. 鍋に水を入れる
2. 具材を煮込む
3. 味噌を
4. ＼ドボーン／
5. 味噌と汁。

驚くほど
味がしない

死ぬほど濃い

1. 鍋に水を入れる
2. 具材を煮込む
3. 味噌を
4. ＼ドボーン／
5. 味噌と汁。
6. 味見をします

目黒研でよく聞く会話例

目黒研でよく聞く会話例

「SNS利用率のアンケートを twitterで拡散してもらいました。」

目黒研でよく聞く会話例

「この講座をどう思いますか？
1.役に立つ 2.為になる 3.楽しい 4.面白い」

目黒研でよく聞く会話例

「先生、10人にアンケートを取ってきました！」
「うーん、最低でも50人は欲しいな」
「はーい」

目黒研でよく聞く会話例

「先生、小匙で味見をしました！」
「うーん、最低でも大匙は使ってほしいな」
「はーい」

混ざってない味噌汁を
どんぶりで味見する

よく混ぜてお飲みください。

前半戦終了

ここから詳細に入ります

前半戦終了

言いたいことはほぼ全部言ってしまった

前半戦終了

細かいテクニックの話をします

前半戦終了

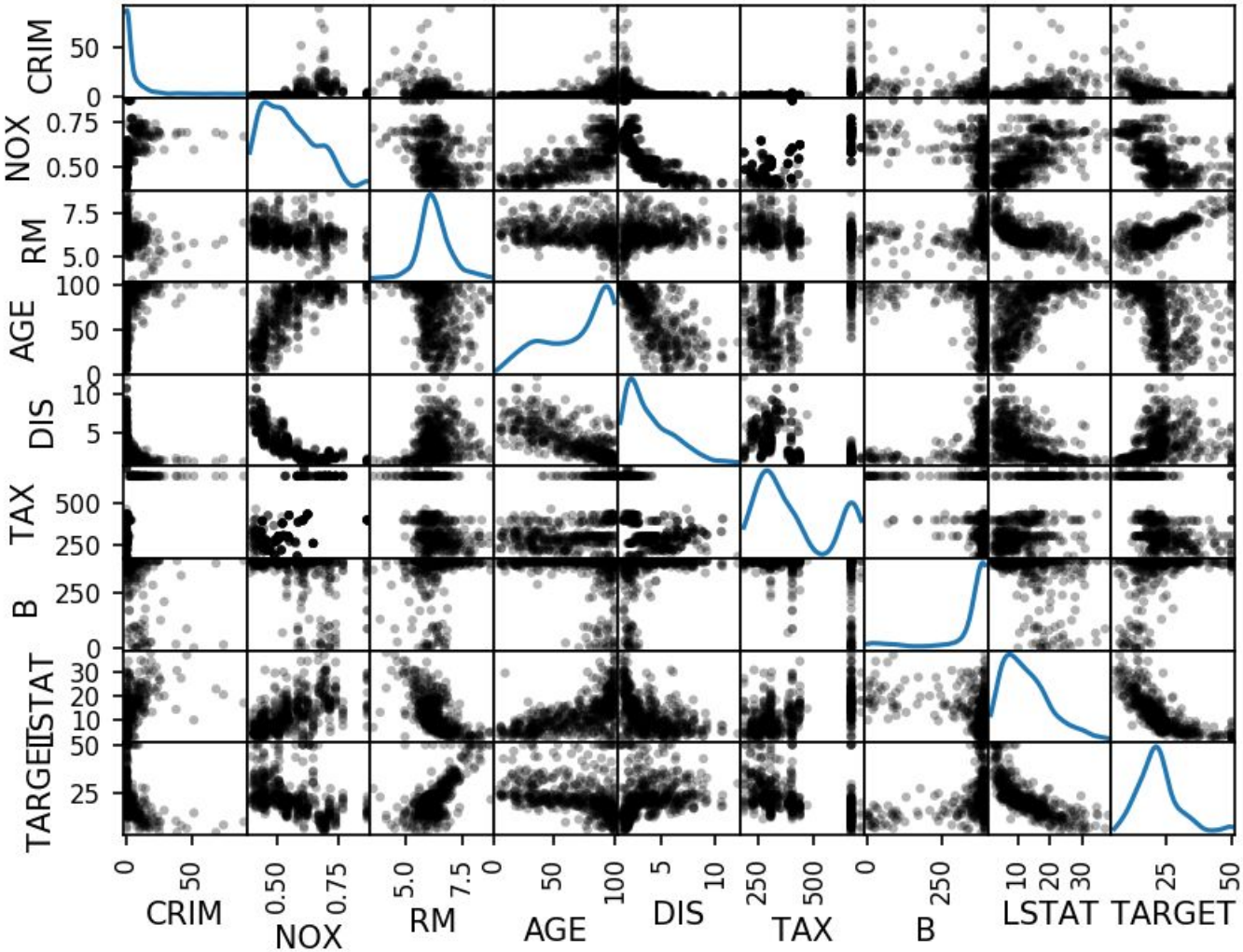
最後に味見スプーンの大きさの話をするよ

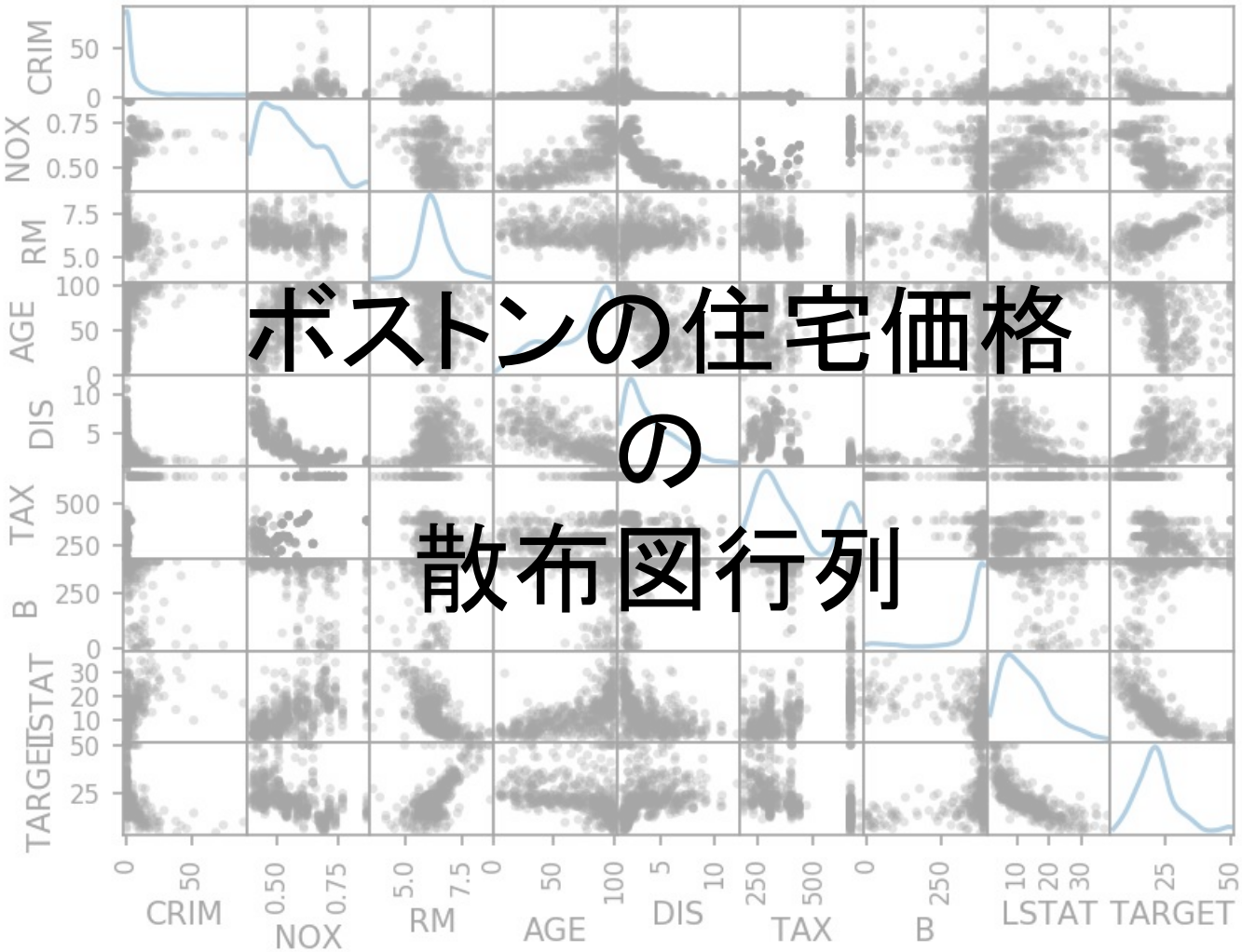
標本抽出、とかググると味噌汁の混ぜ方が色々出てくる

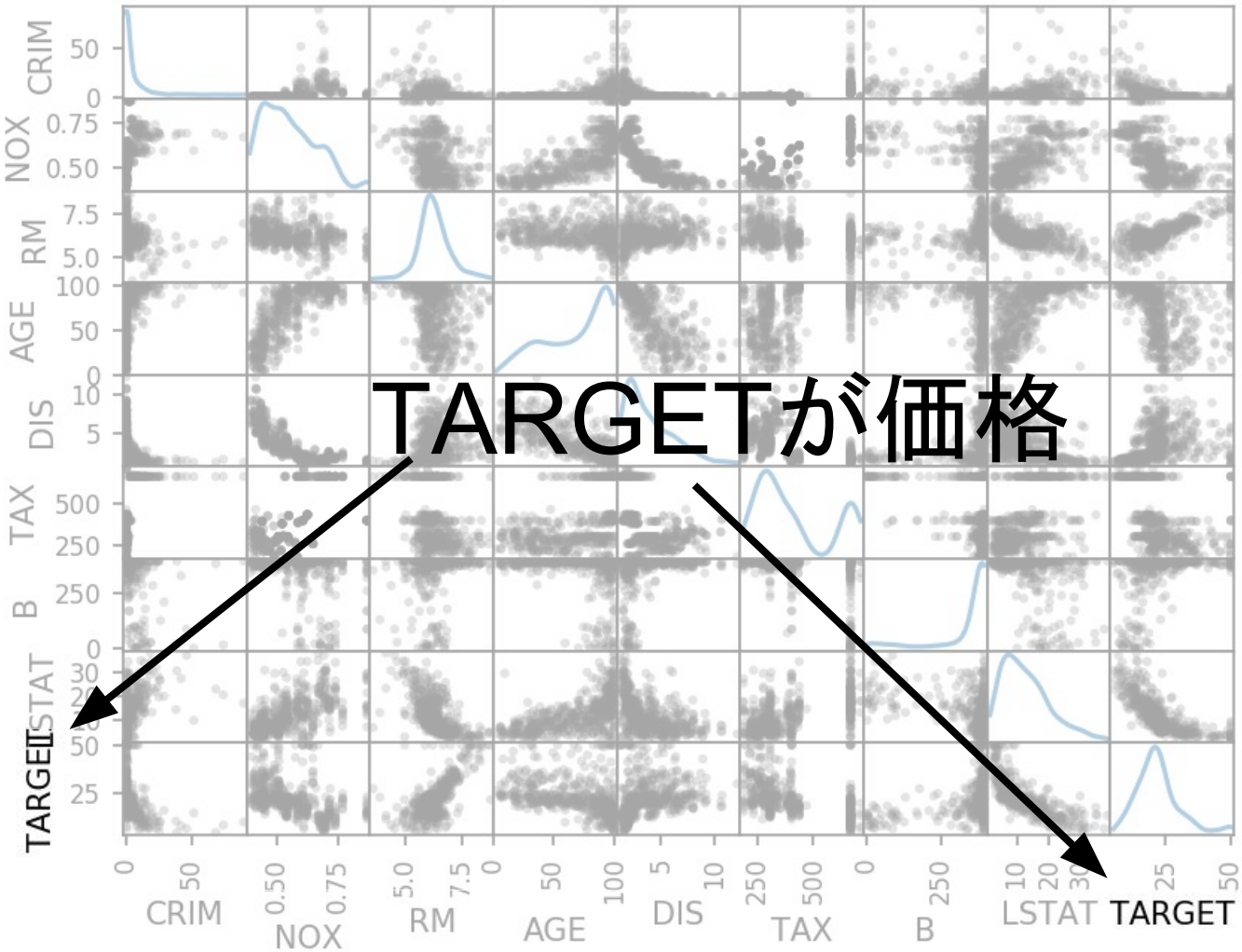
データを集めた後
まず最初にすること

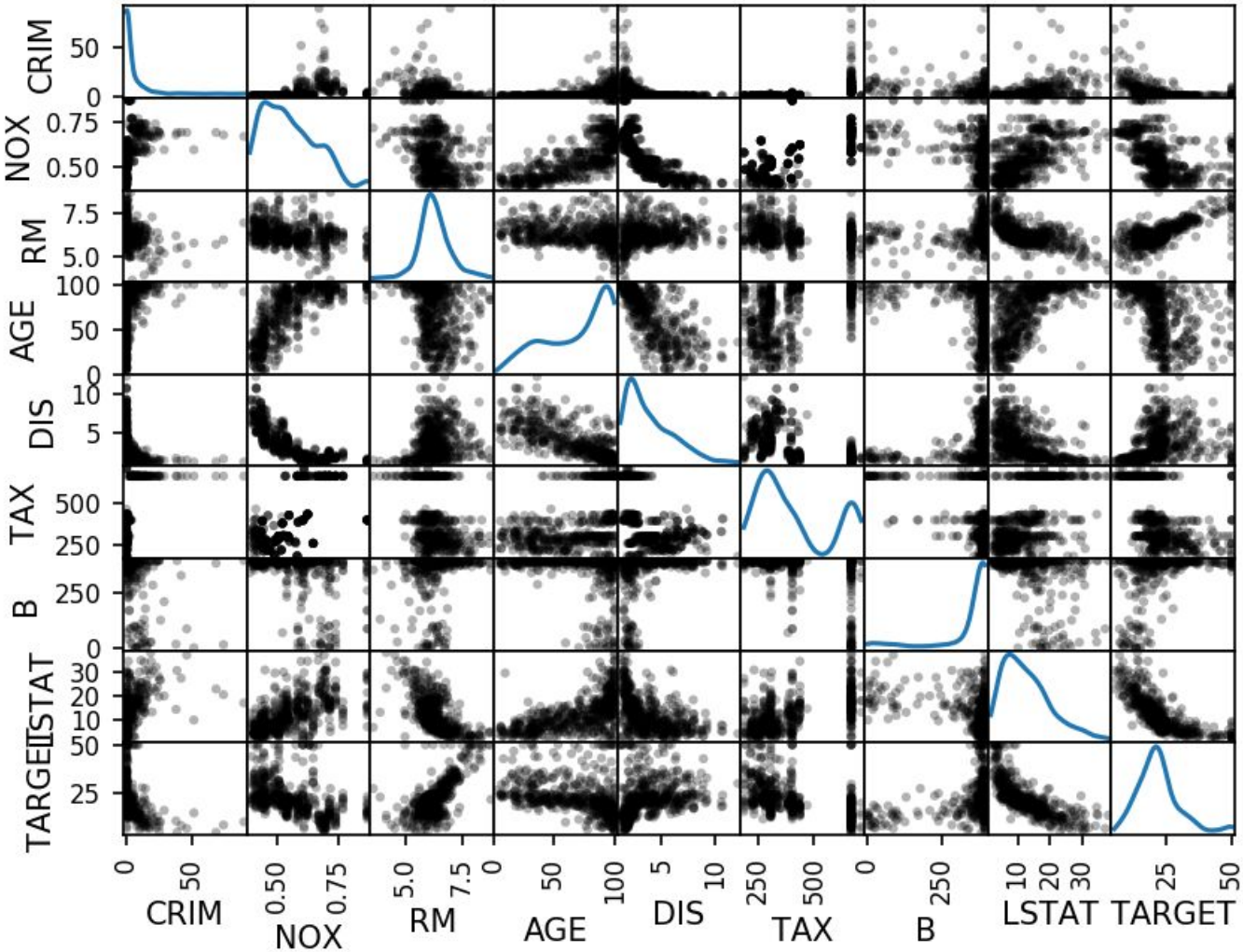
散布図を描く

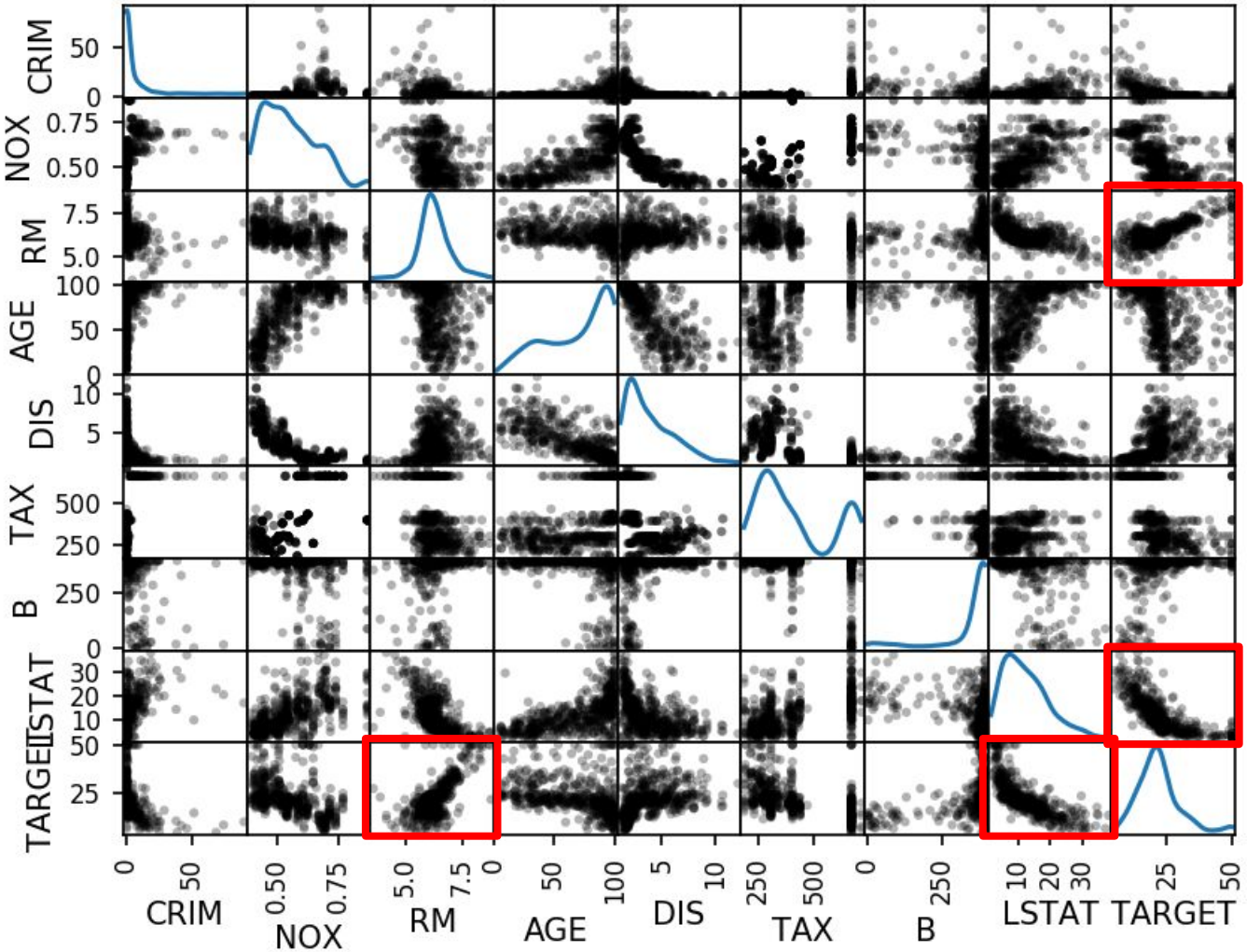
のが多分良い

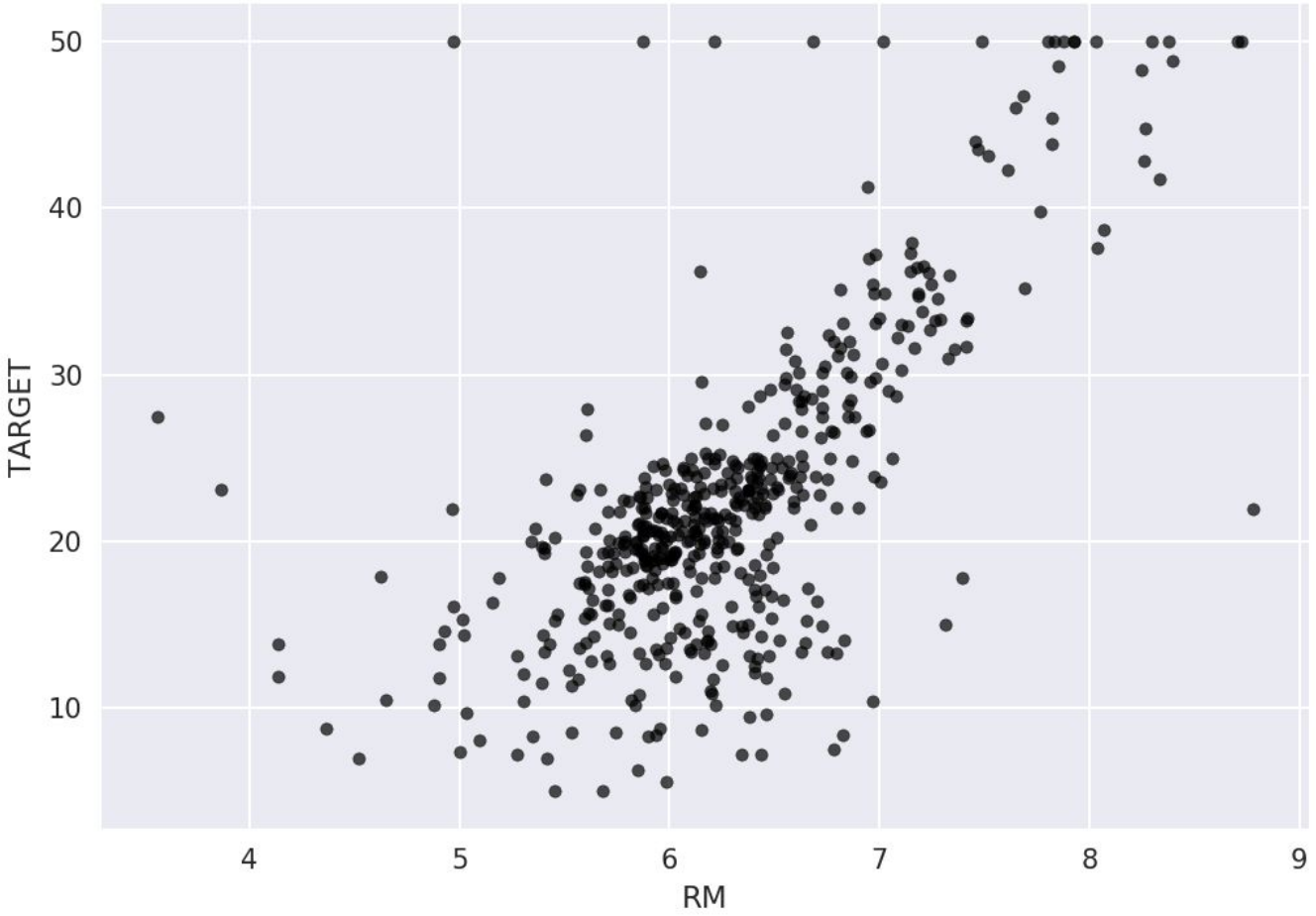

CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET

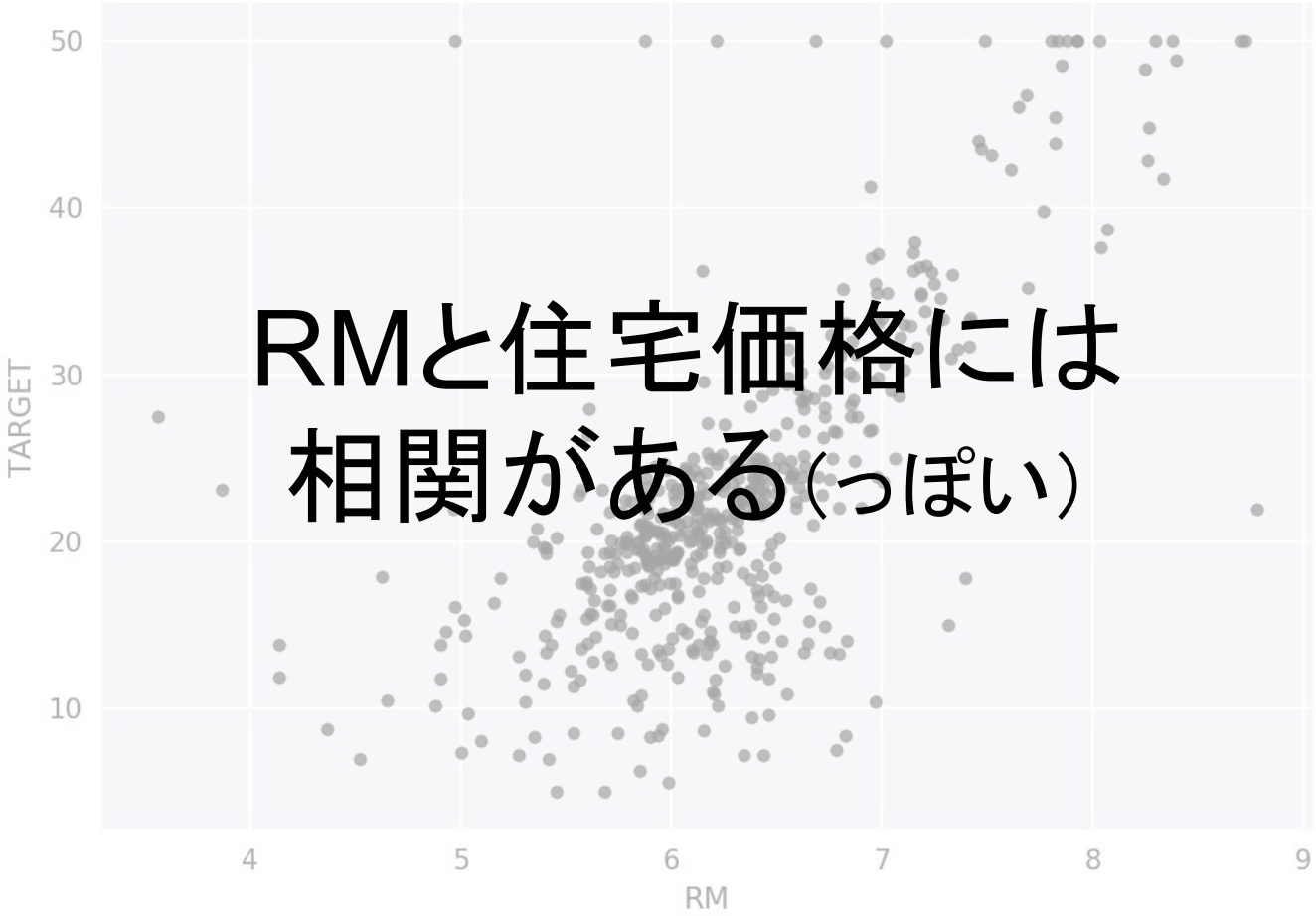

CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET
ボストンの住宅価格
の
散布図行列


TARGETが価格
CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET

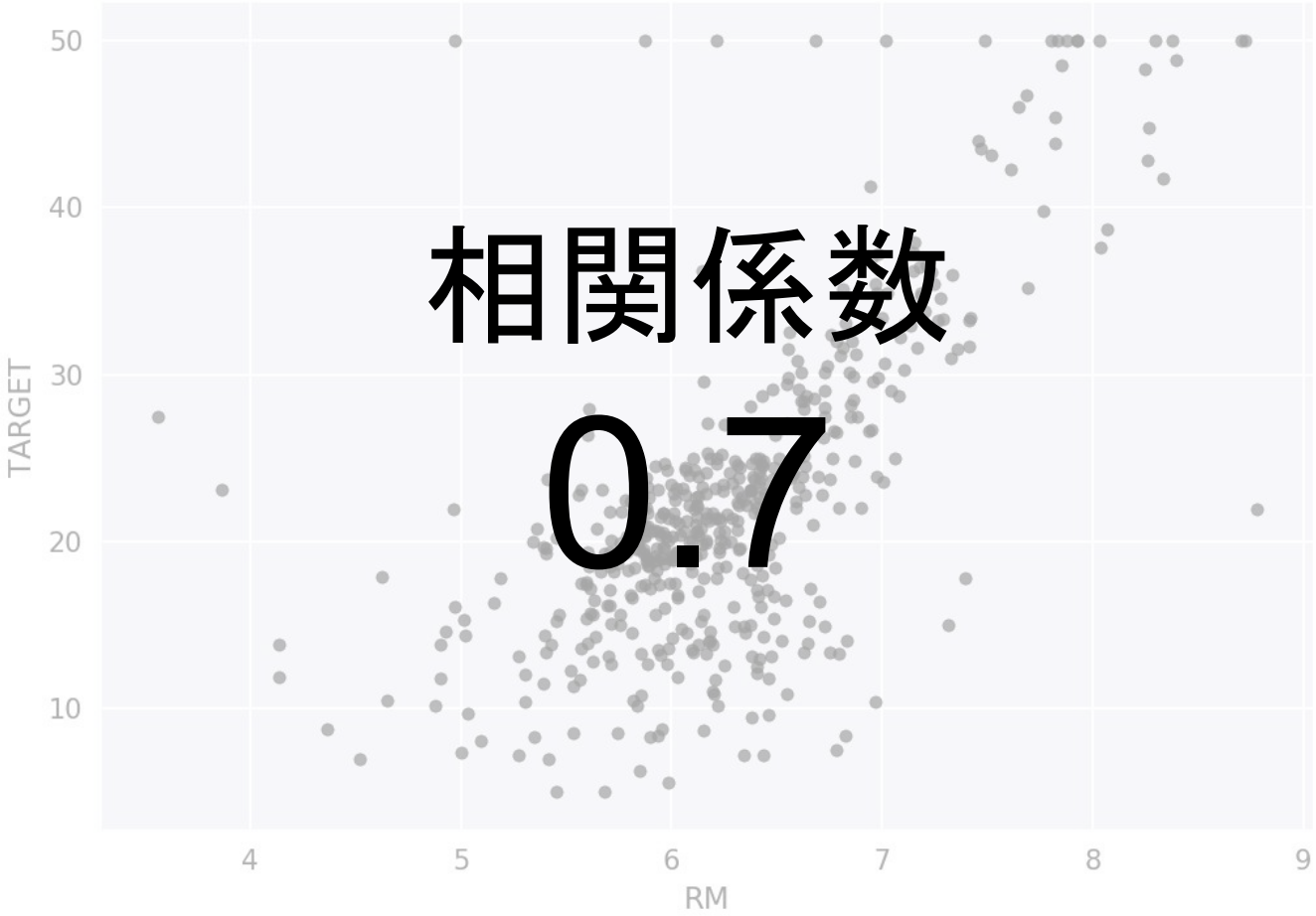

CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET


CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET
0
50
0.50
0.75
5.0
7.5
100
5
10
250
500
30
20
25


50
40
30
20
10
TARGET
4
5
6
7
8
9
RM

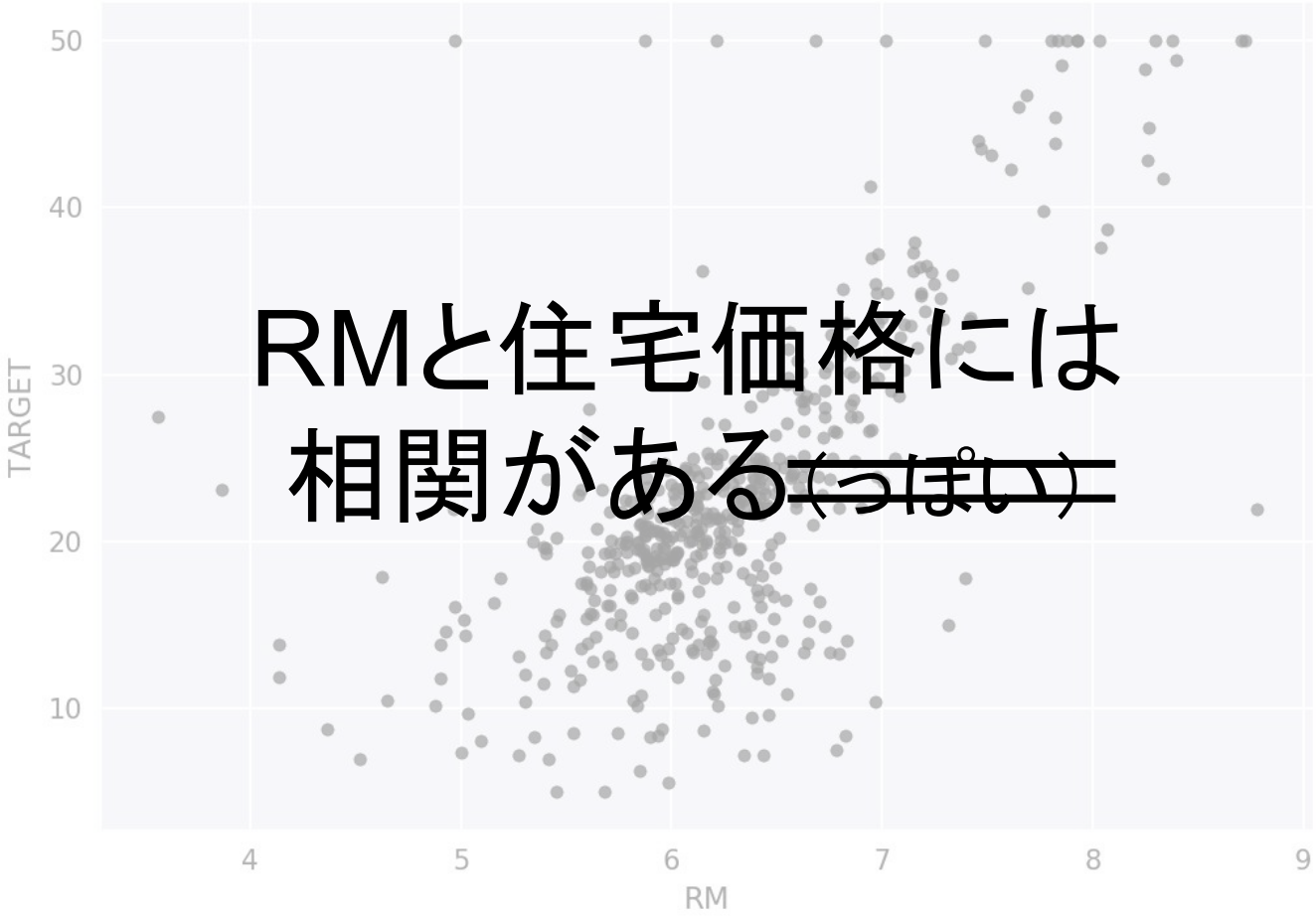


RMと住宅価格には相関がある（っぽい）

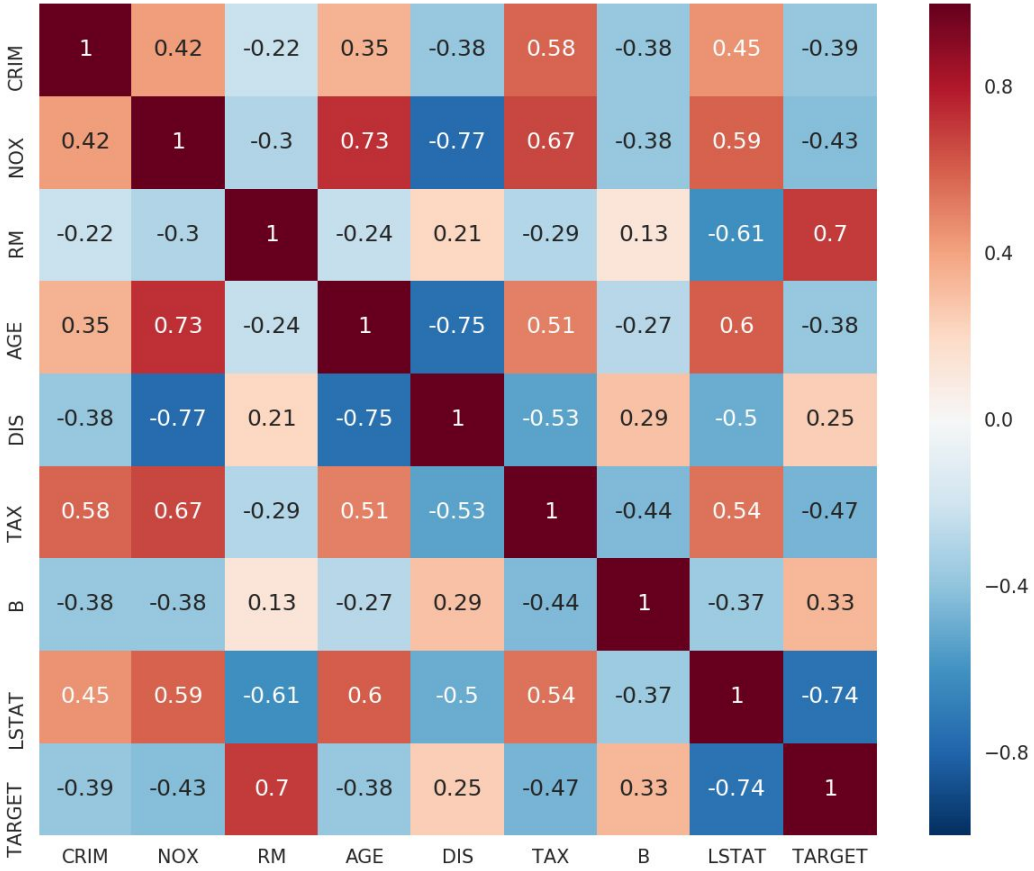


相関係数

0.7


RMと住宅価格には
相関がある(っぽい)
TARGET
RM
50
40
30
20
10
4
5
6
7
8
9

相関係数のヒートマップ



散布図と相関係数を出しとくととりあえずそれっぽい

注意

注意

相関係数が高いからといって関連があるとは限らない

アイスクリームの売上

水難事故の発生件数

アイスクリームを売らなければ
水難事故が無くなる（？）

疑似相関

相関があるけど関連が無い状態

偏相関というのを使えば見抜ける

（詳しいやり方は面倒なのでやりません。ググってね。）

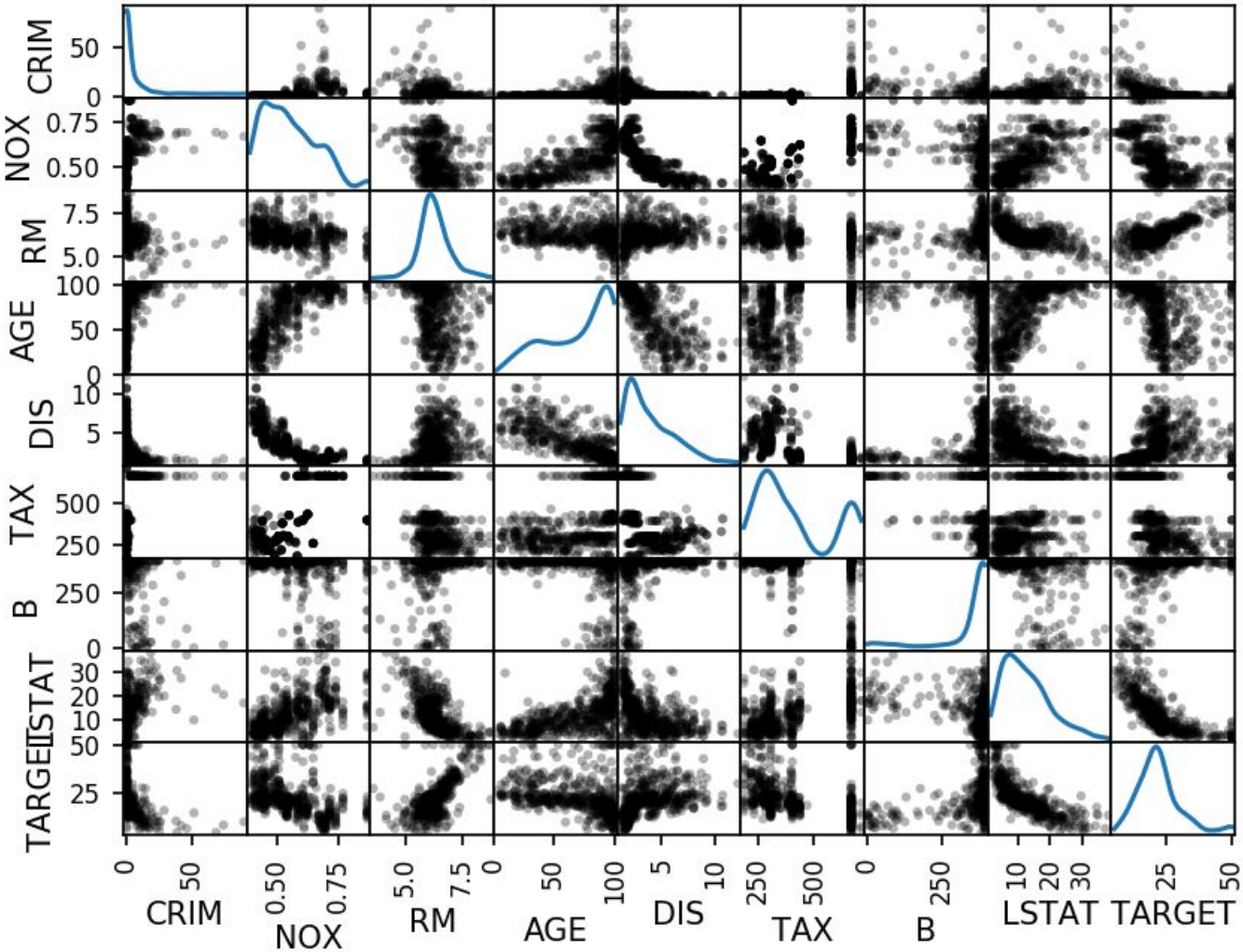
工科大で単位を落とす人のほぼ100%がPCを所有しています。PCを捨てよう。

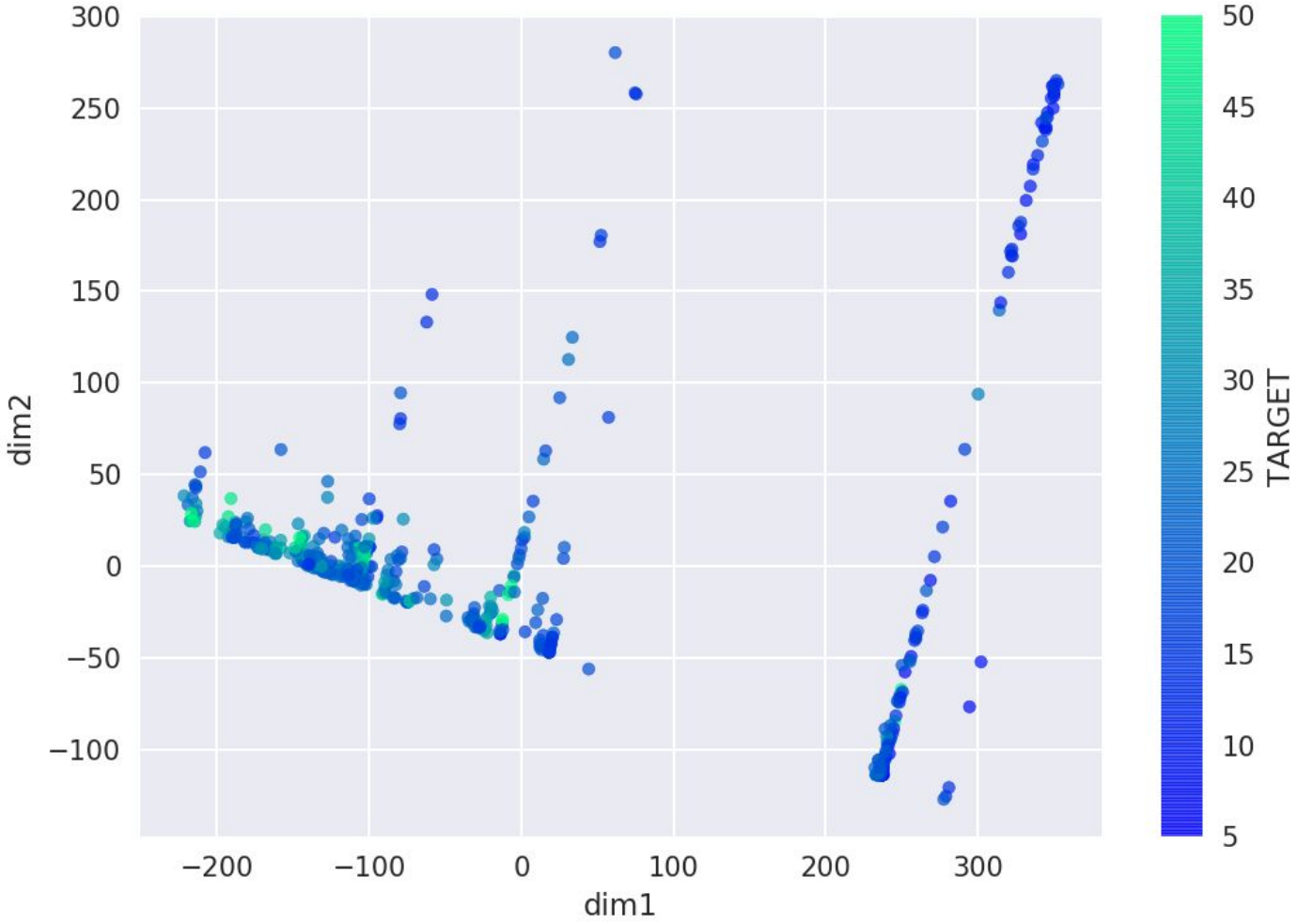
では、何がどのくらい住宅価格に影響するのか

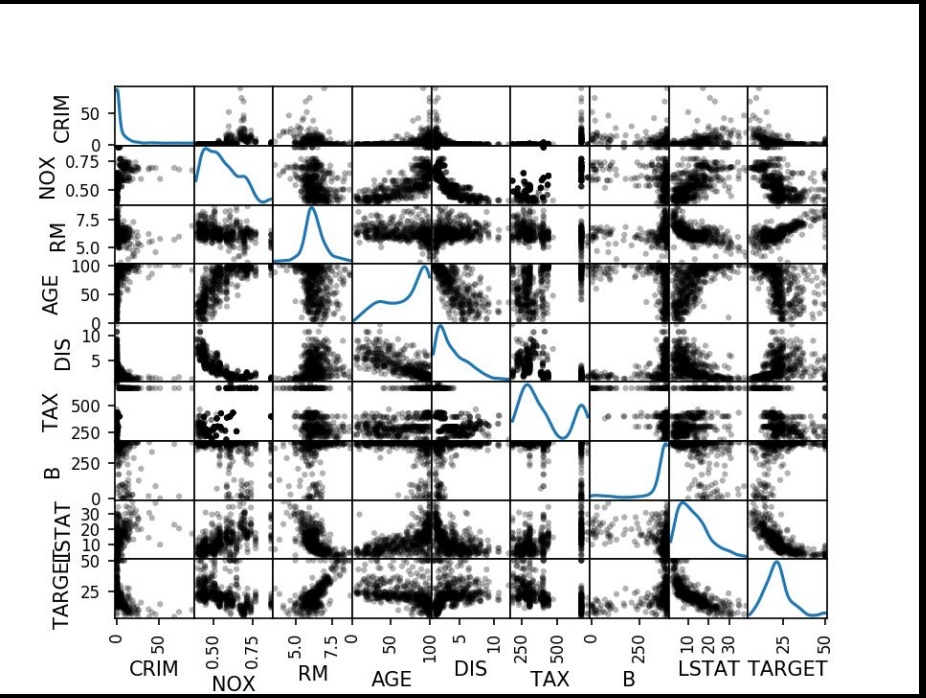
主成分分析

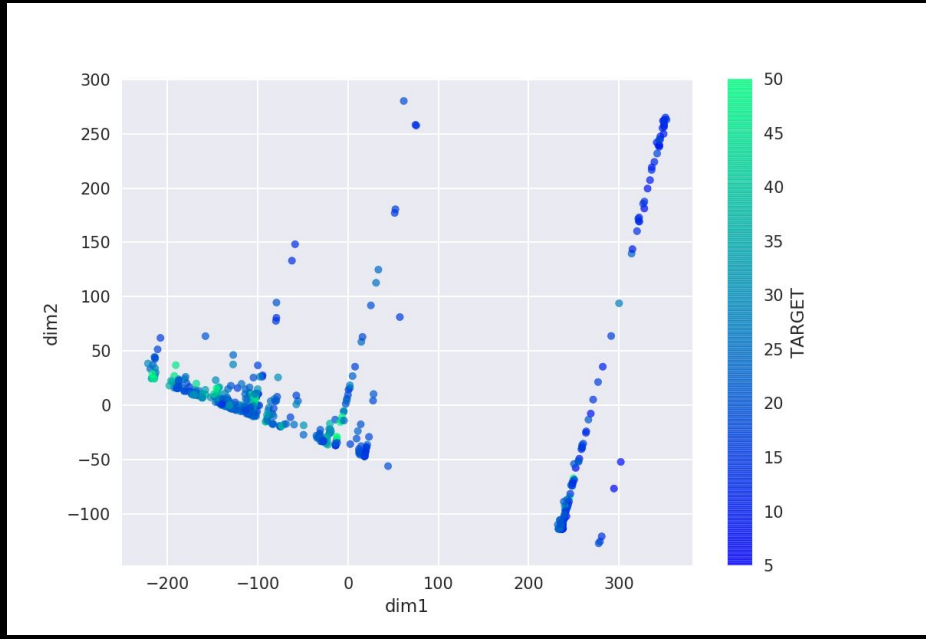
[0.81630913 0.16594063]

[[6.92865615e+02 3.93170379e-01 -1.00726940e+00 7.27642768e+01 -5.70745515e+00
8.22952576e+02 -2.87081217e+02 2.02510727e+01]
 [3.93170379e-01 6.67189406e+02 -1.57197800e-02 1.12521690e+00 -8.82111638e-02
1.28404784e+01 -3.94985955e+00 3.09483179e-01]
 [-1.00726940e+00 -1.57197800e-02 6.67224859e+02 -2.92951508e+00 2.29460770e-01
-3.39010006e+01 8.27038646e+00 -7.90623139e-01]
 [7.27642768e+01 1.12521690e+00 -2.92951508e+00 8.76189633e+02 -1.63817857e+01
2.39274749e+03 -7.00913736e+02 5.72396721e+01]
 [-5.70745515e+00 -8.82111638e-02 2.29460770e-01 -1.63817857e+01 6.68467388e+02
-1.87420194e+02 5.54585585e+01 -4.49032643e+00]
 [8.22952576e+02 1.28404784e+01 -3.39010006e+01 2.39274749e+03 -1.87420194e+02
2.83496085e+04 -6.78488545e+03 6.45983549e+02]
 [-2.87081217e+02 -3.94985955e+00 8.27038646e+00 -7.00913736e+02 5.54585585e+01
-6.78488545e+03 8.31849344e+03 -2.31768522e+02]
 [2.02510727e+01 3.09483179e-01 -7.90623139e-01 5.72396721e+01 -4.49032643e+00
6.45983549e+02 -2.31768522e+02 6.83158346e+02]]

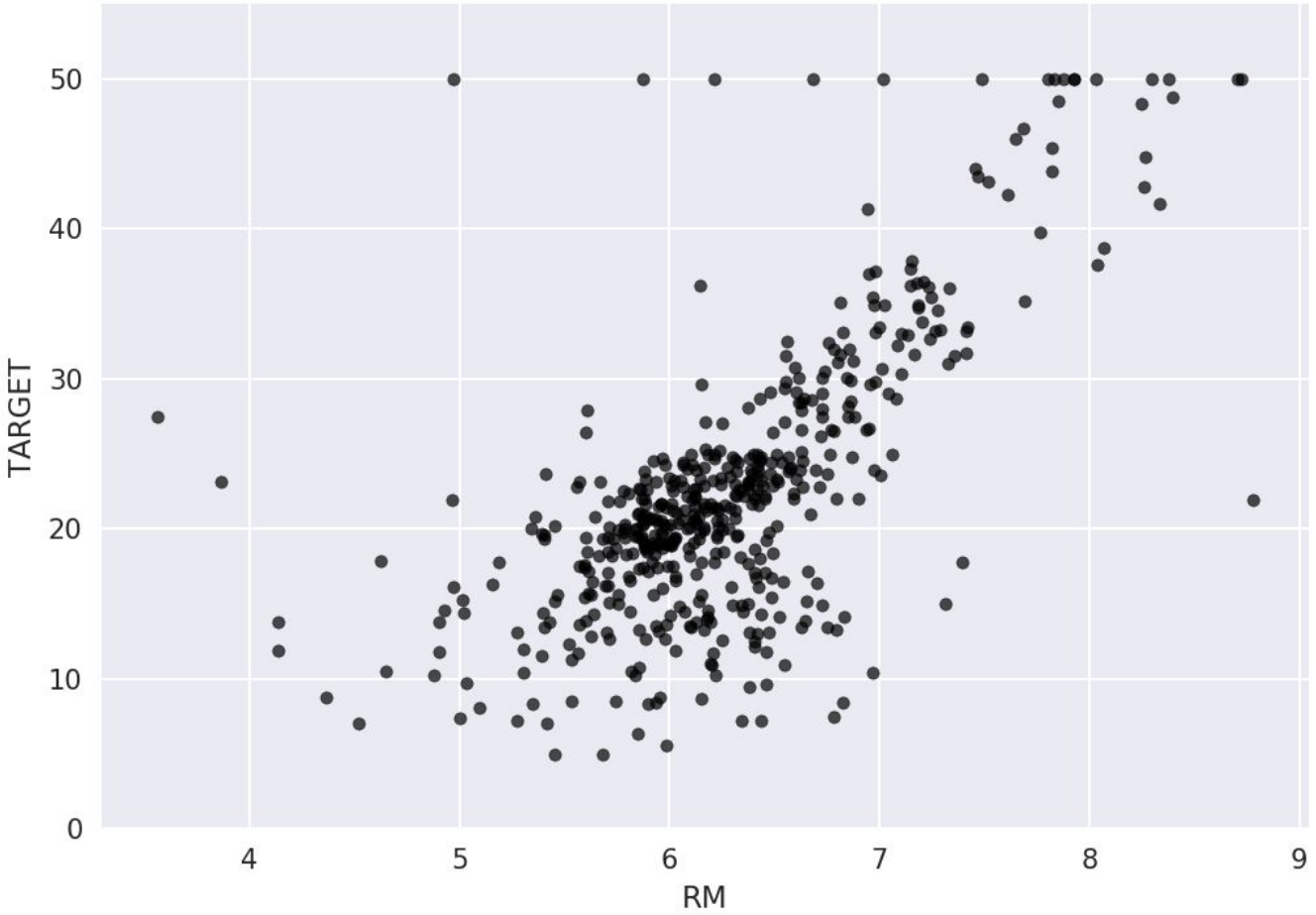

CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET

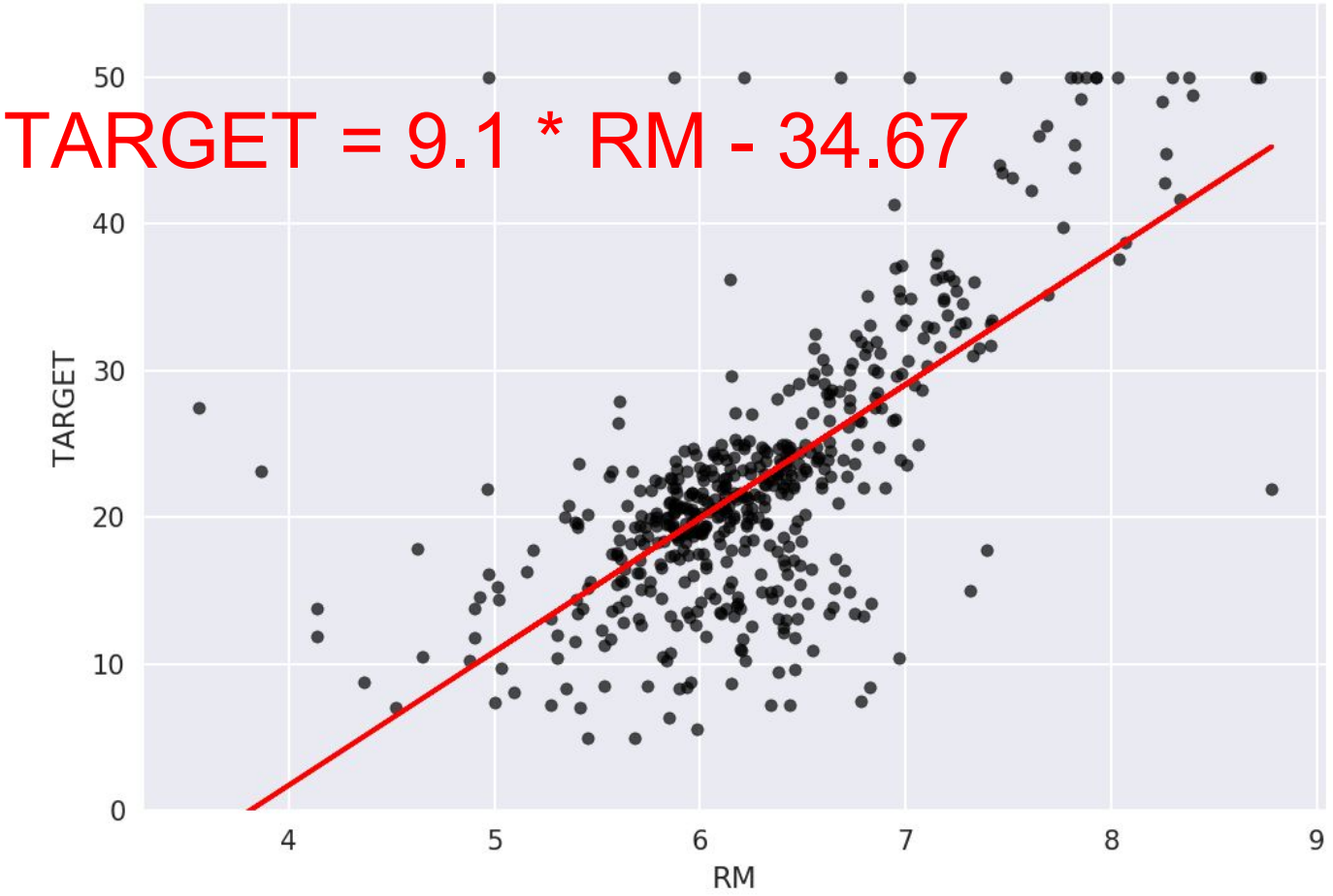

300
250
200
150
100
50
0
−50
−100
dim2
−200
−100
0
100
200
300
dim1
50
45
40
35
30
25
20
15
10
5
TARGET

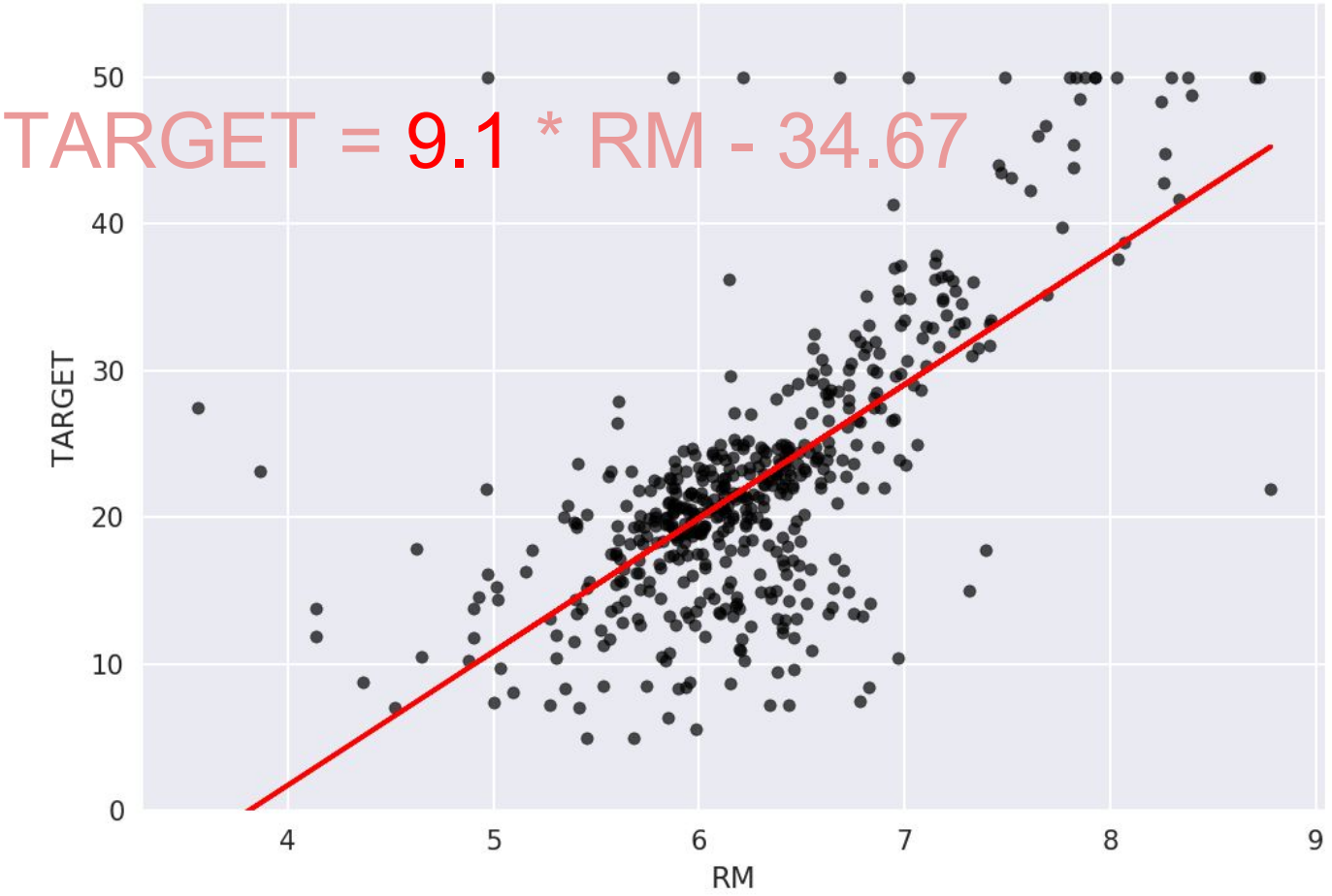

CRIM
NOX
RM
AGE
DIS
TAX
B
LSTAT
TARGET


dim1
dim2
TARGET

重回帰分析


TARGET
RM
0
10
20
30
40
50
4
5
6
7
8
9


TARGET = 9.1 * RM - 34.67
TARGET
RM
0
10
20
30
40
50
4
5
6
7
8
9


TARGET = 9.1 * RM - 34.67
TARGET
RM
0
10
20
30
40
50
4
5
6
7
8
9

$$
\begin{aligned}
\text{TARGET} = &-0.059 * \text{CRIM} + \\
&-6.684 * \text{NOX} + \\
&\ 5.171 * \text{RM} + \\
&-0.019 * \text{AGE} + \\
&-1.091 * \text{DIS} + \\
&-0.005 * \text{TAX} + \\
&\ 0.009 * \text{B} + \\
&-0.544 * \text{LSTAT} + 5.364
\end{aligned}
$$

TARGET = -0.059 * CRIM +

-6.684 * NOX + ← 大気汚染の度合い

5.171 * RM + ← 部屋の数

-0.019 * AGE +

-1.091 * DIS + ← 雇用施設との距離

-0.005 * TAX +

0.009 * B +

-0.544 * LSTAT + 5.364

TARGET = -0.059 * CRIM +

-6.684 * NOX + ← 大気汚染の度合い

5.171 * RM + ← 部屋の数

-0.019 * AGE +

-0.961 * DIS + ← 雇用先までの距離

-0.005 * TAX +

0.009 * B +

-0.544 * LSTAT + 5.364

重回帰分析は便利

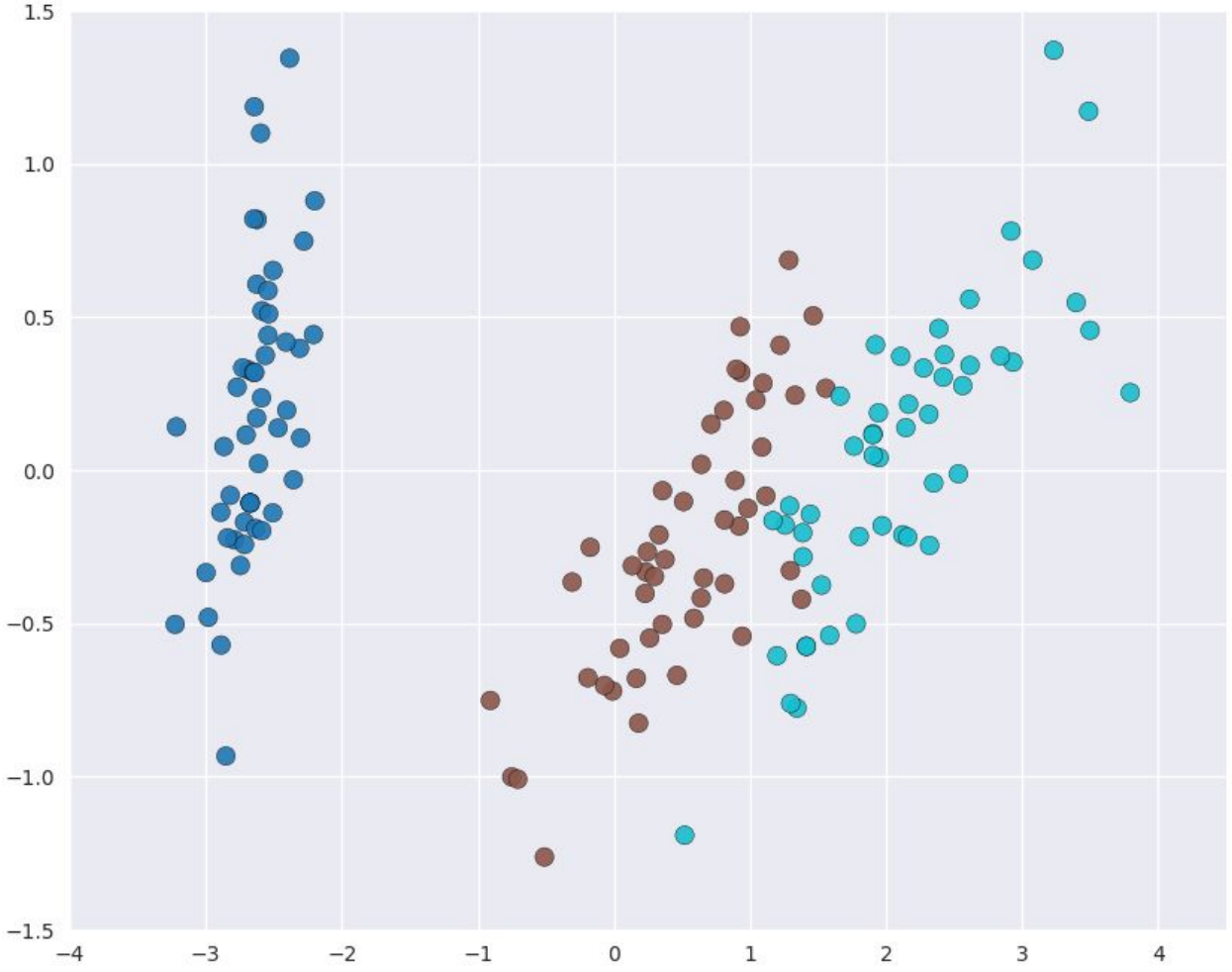
でも疑似相関にはやっぱり注意

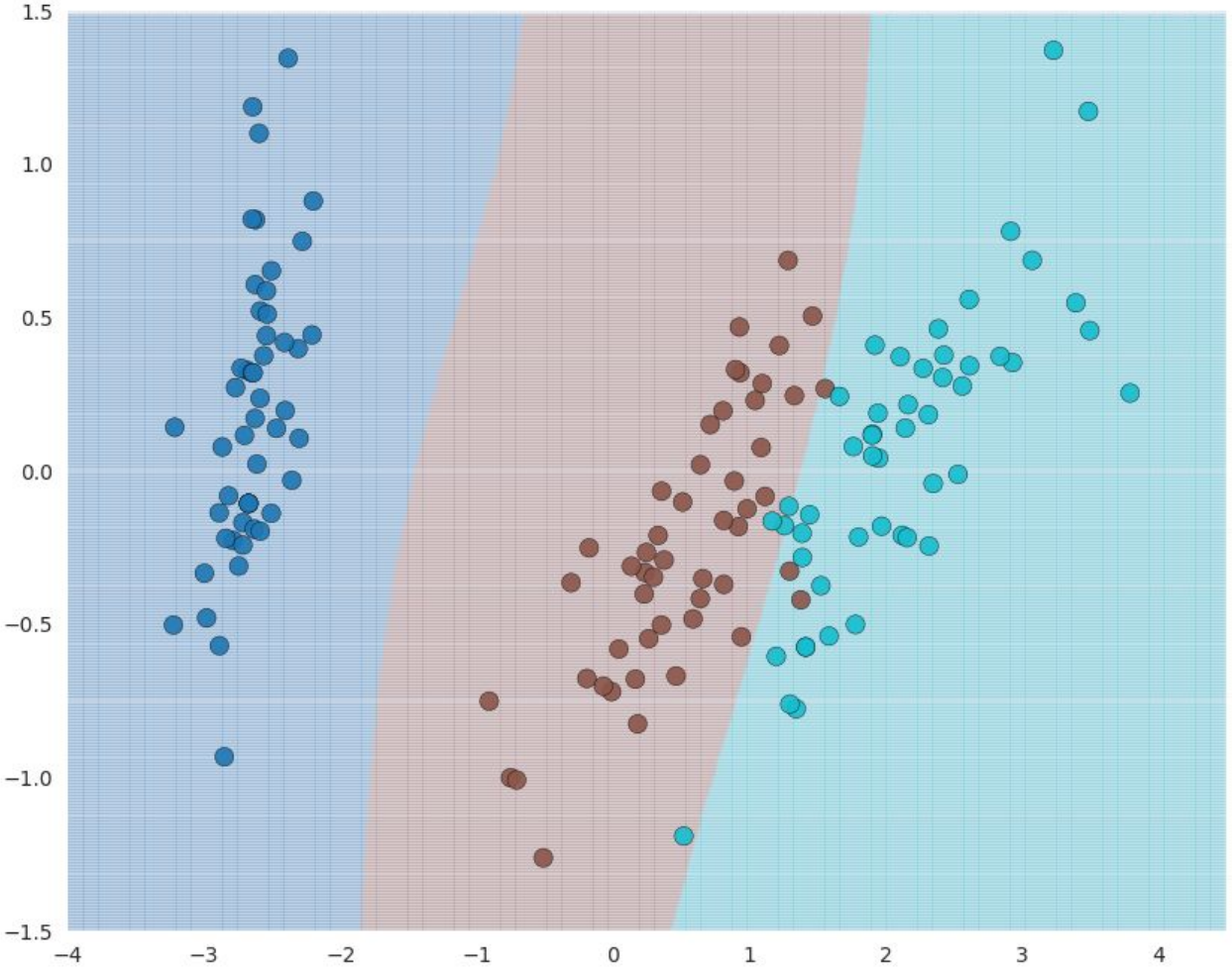
結局のところ、万能のツールは無い

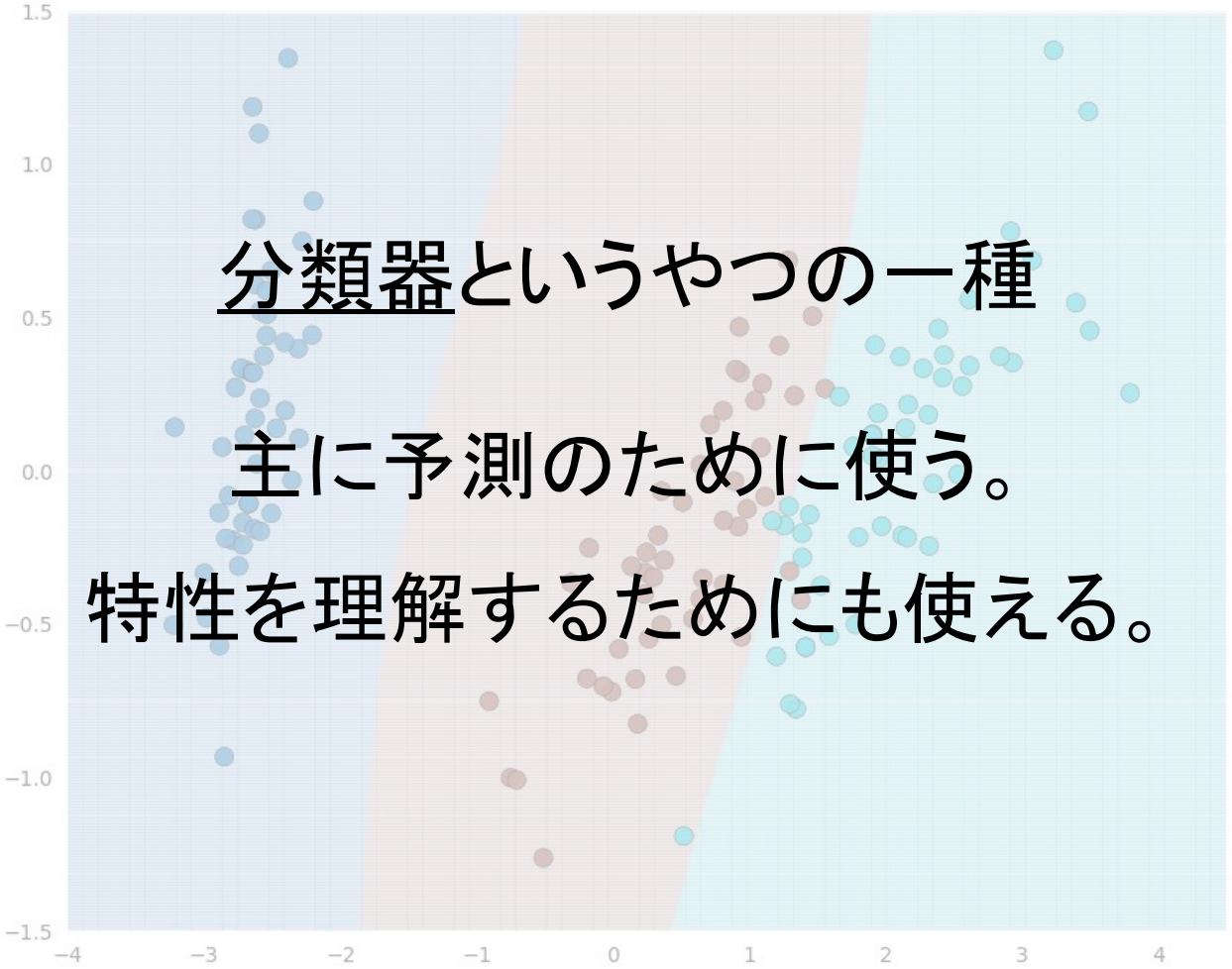
余談

データサイエンティスト曰く

「一番よく使うのはSVM」


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−3
−2
−1
0
1
2
3
4


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−3
−2
−1
0
1
2
3
4



分類器というやつの一種

主に予測のために使う。

特性を理解するためにも使える。

SVMはかなり古くてシンプルなんだけど、これ以上のものはあんまり無い

アンケートの人数と統計的な誤差の話

$$n = \lambda^2 \frac{p(1-p)}{d^2}$$

n：標本数、p：回答比率、d：標本誤差、λ：信頼水準

世論調査とは、一般に、個人を対象として行われる大規模な意識調査のことをいい、国、地方自治体、大学、新聞社・通信社・放送局などの報道機関などに行っています。
世論調査といえば、マスコミなどで行われている内閣支持率や政党支持率などの世論調査が思いつくかと思いますが、国や地方自治体などにおいても世論調査を行っていま
国（内閣府）では、基本的な国民意識の動向や政府の重要施策に関する国民の意識を把握するために世論調査を実施しています。また、地方自治体においても、住民の行政
る要望・意見などを把握したり、条例や計画立案の際の基礎資料とすることなどを目的に行われています。
これらの調査では、調査対象の一部を調べることで調査対象全体を推測する「標本調査」という方法が使われています。標本調査の設計段階においては、調査対象となる集団
）を偏らないよう標本が全国の縮図になるように選ぶ方法や調査の対象者数などは、統計的な理論に基づき決められています。

ここでは、簡単な例として単純無作為抽出（調査対象者を無作為に選出する方法）により支持率などの賛否を問う調査を行う場合について、調査対象者数の決め方を紹介した
ます。
ある高等学校において、「学校生活では、制服と私服のどちらがよいか」について、生徒の意識を調査するために、アンケート調査を行うとします。生徒人数が多いことから、全
て調査することはできないため、標本調査で調査をすることにしました。
学校にある生徒名簿を利用して無作為（ランダム）に調査対象者を決めたとします。このとき、調査に必要な調査対象者数を計算します。

式の導出過程は省きますが、このような賛否を問う調査で必要な調査対象者数は以下の式により算出できます。

要な標本数の計算
回答比率とは、支持率や保有率などの調査対象者の回答比率です。事前に他調査で同様な調査結果がある場合はその比率を用いますが、事前に参考となる結果がない場合
な調査対象者数が最大となる0.5を入れます。
標本誤差には、調査結果で容認できる誤差を入れます。例えば、調査結果の誤差を3%ポイント程度に抑えたいという場合であれば0.03を入れます。
信頼水準とは、正しく判断できる確率をいいます。例えば、信頼水準を95%であれば、母集団（この例では高等学校内の全生徒となります。）の支持率の平均値が95%の確率で「標
均（調査から得られる結果）－標本誤差×1.96～標本平均＋標本誤差×1.96」の範囲に入る可能性を意味しています(※)。

ここでは、回答比率を0.5、標本誤差は5%ポイント、信頼水準を95%（λ＝1.96）として必要とな調査対象者数を計算します。調査に必要な対象者数は、

答比率を0.5、標本誤差は5%ポイント、信頼水準を95%（λ＝1.96）として必要とな調査対象者数を計算
なります。よって、この調査では384人の調査対象者から回答が必要となるわけです。
なお、実際は調査対象者の全員から回答が得られるとは限らないため、想定される回収率を踏まえて、計算で得られた調査対象者数より多めに対象者数を見積もっておく必要
ます。

※)一般的に国などが行っている標本調査は、信頼水準を95%（λ＝1.96）として調査の設計がされています。

工科大生のPC所有率は？

（入学時に買うことになるので99.9%持ってるはず）

工科大生のPC所有率は？

↓

1.54人に聞いてYESならば、
95.9%から100.0%の間であると証明出来る

では、45%から55%の間であると証明するには？

では、45%から55%の間であると証明するには？

↓

384.16人必要（！）

濃過ぎる味噌汁の定義は簡単

ちょうど良い味の定義は難しい

正しい統計を簡単に取りたい場合、

xxな人は50%以上

みたいな証明を目指そう

一応何かツール作ったので計算してみたい人はどうぞ

https://rawgit.com/macrat/meguro-lab-basic-technical-lecture/master/03/calculator.html

アドレスは資料にもあります。

目黒研究室 基礎技術講座 第三回

最低限の統計学

blanktar.jp

Thank you for listening!

Slack登録してない方
もし居たらお願いします

goo.gl/3YRc8d

本日の資料はこっち

goo.gl/h8Kp7T

Slackはこっち

goo.gl/3YRc8d

Thank you for listening!