事例から学ぶ統計データ利活用セミナー

講師：請園　正敏

データサイエンス研究所

© Data Science Institute

本セミナーの目的とゴール

目的

Excelを用いて、データ利活用のための統計分析の考え方について実践的に学ぶ

ゴール

データ分析について理解する。

e-Statのデータを分析する方法について理解する。

jSTAT MAPを用いたデータ分析方法について理解する。

© Data Science Institute

目次

１．統計データ利活用方法の基本
- なぜ統計データ分析方法の理解は重要か
- データ利活用のメリット

２．ビジネスにおけるデータ利活用の最前線
- ビジネスにおけるデータ利活用の領域の拡大
- 公的統計データ等を活用した、様々な領域におけるデータ分析事例紹介（民間企業における活用事例の紹介

３．Excelでデータ利活用の基本を学ぶ
- Excelによる統計データ分析の基本
- Excelでのグラフ作成による可視化とデータ観察方法

４．政府統計データの活用～e-Stat編～
- e-Statのデータを用いた具体的な事例分析演習
- e-Statのデータから抽出し、相関、回帰分析

５．政府統計データの活用～jSTAT MAP編～

jSTAT MAPを用いたデータ分析結果の可視化演習

６．質疑応答

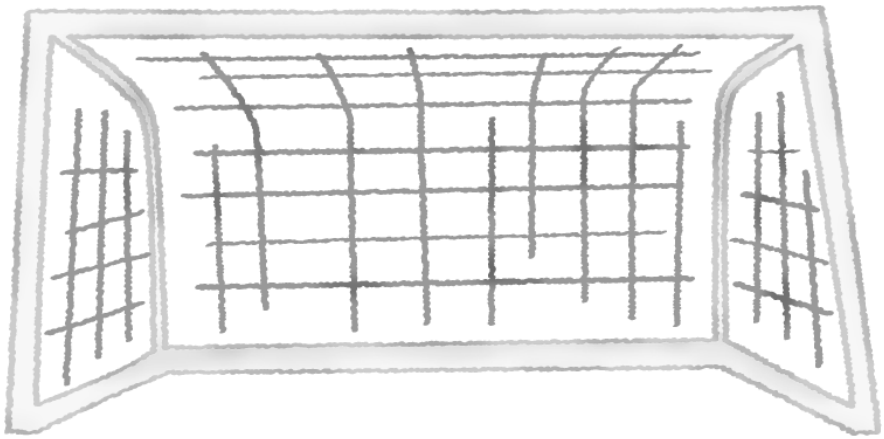
© Data Science Institute

統計データ利活用方法の基本

© Data Science Institute

▪なぜ統計データ分析方法の理解は重要か

さて、PKであなたはどこに蹴りますか・・・・？

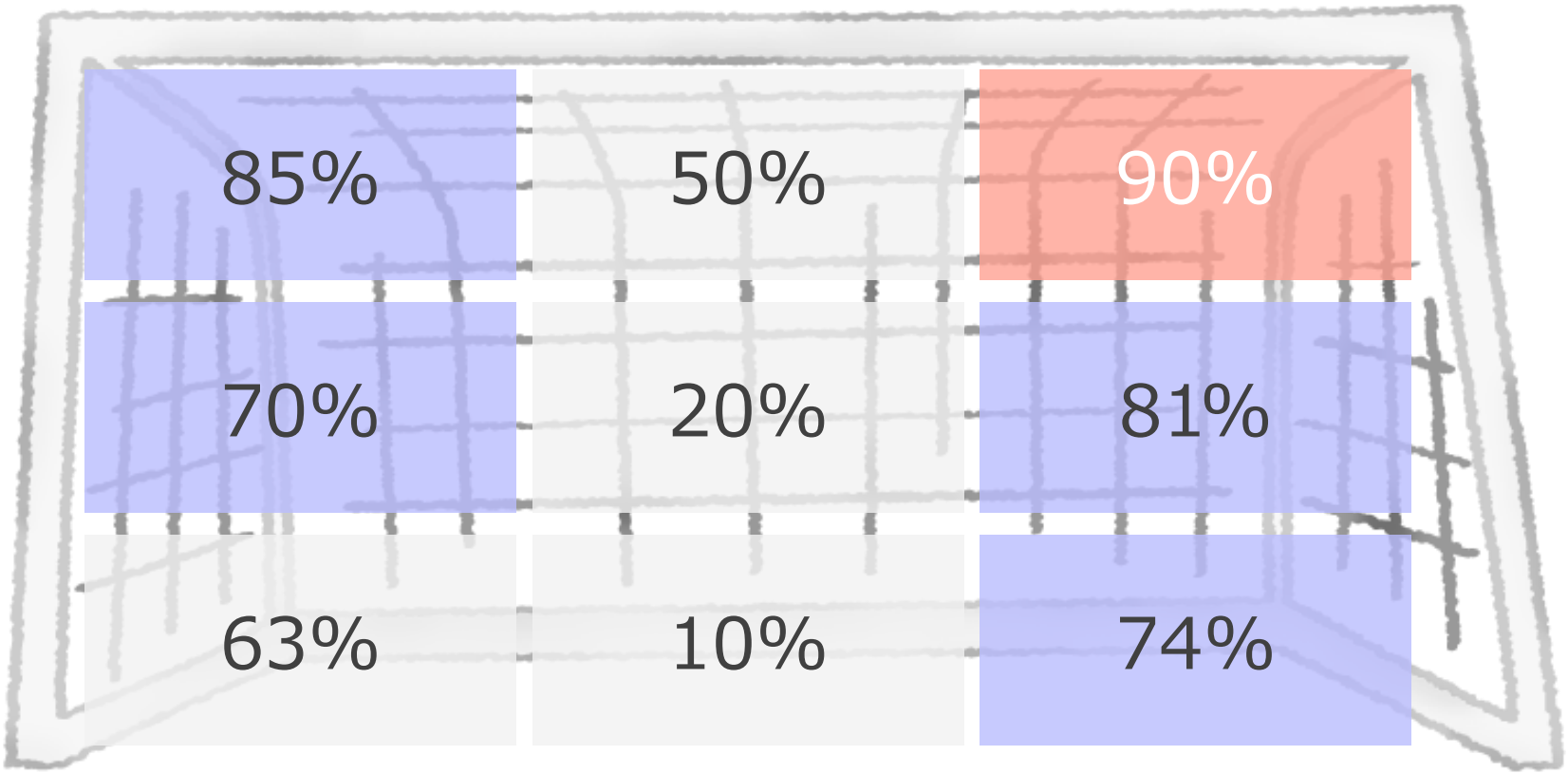




© Data Science Institute

▪なぜ統計データ分析方法の理解は重要か

過去のゴール成功率を見ると・・・？



© Data Science Institute

▪なぜ統計データ分析方法の理解は重要か

■ KKD（勘、経験、度胸）での判断は、主観的な判断や部分最適になりやすい
■ データがあれば、全体感を捉えた上で、客観的な判断が誰にでもできるようになる

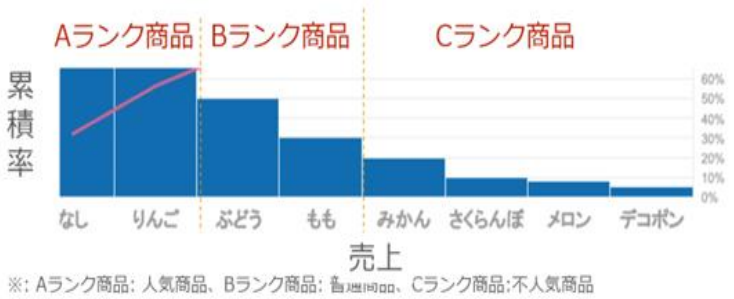
KKD（勘、経験、度胸）の世界



この前も成功したし、
今回も気合いで決め
るぞ！

- 主観的な判断
- 個人のスキルや経験に依存（属人化）
- 部分最適に陥りがち

データがあれば・・・



果物の売り上げとその累積率のデータより、
A、B、Cに区切ってどの商品を仕入れるか決める

- 客観的な判断
- 誰でも同じような判断ができる
- 全体最適を目指しやすい（全体感がわかる）

© Data Science Institute

▪なぜ統計データ分析方法の理解は重要か

- 単にデータがあれば良いわけではない
- データを「分析」することで、人が解釈可能な「情報」となる

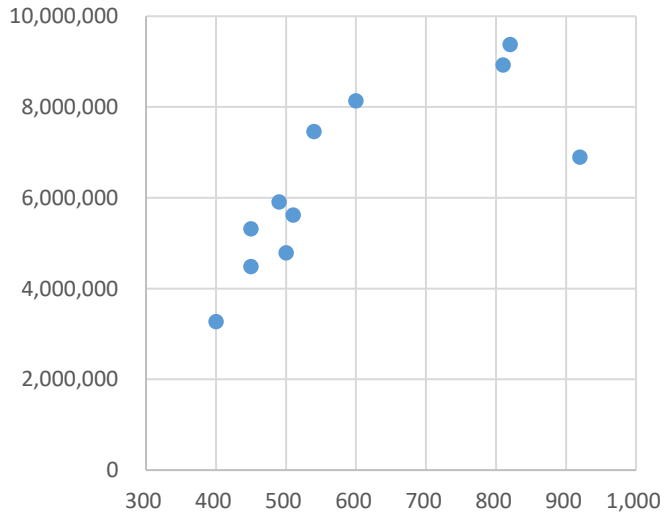
データ

店舗	商圏人口	売場面積	売上金額
熱海店	150	400	3,273,000
安城店	400	450	5,318,000
小田原店	600	500	4,789,000
静岡店	1,000	600	8,139,000
東京店	8,000	920	6,894,000
豊橋店	500	490	5,911,000
名古屋店	2,500	820	9,378,000
浜松店	900	540	7,463,000
富士店	500	510	5,617,000
三島店	400	450	4,481,000
横浜店	3,000	810	8,926,000

分析

＋価値

情報



売場面積広いと売上が上がる
だが最大で600㎡で十分

© Data Science Institute

▪なぜ統計データ分析方法の理解は重要か

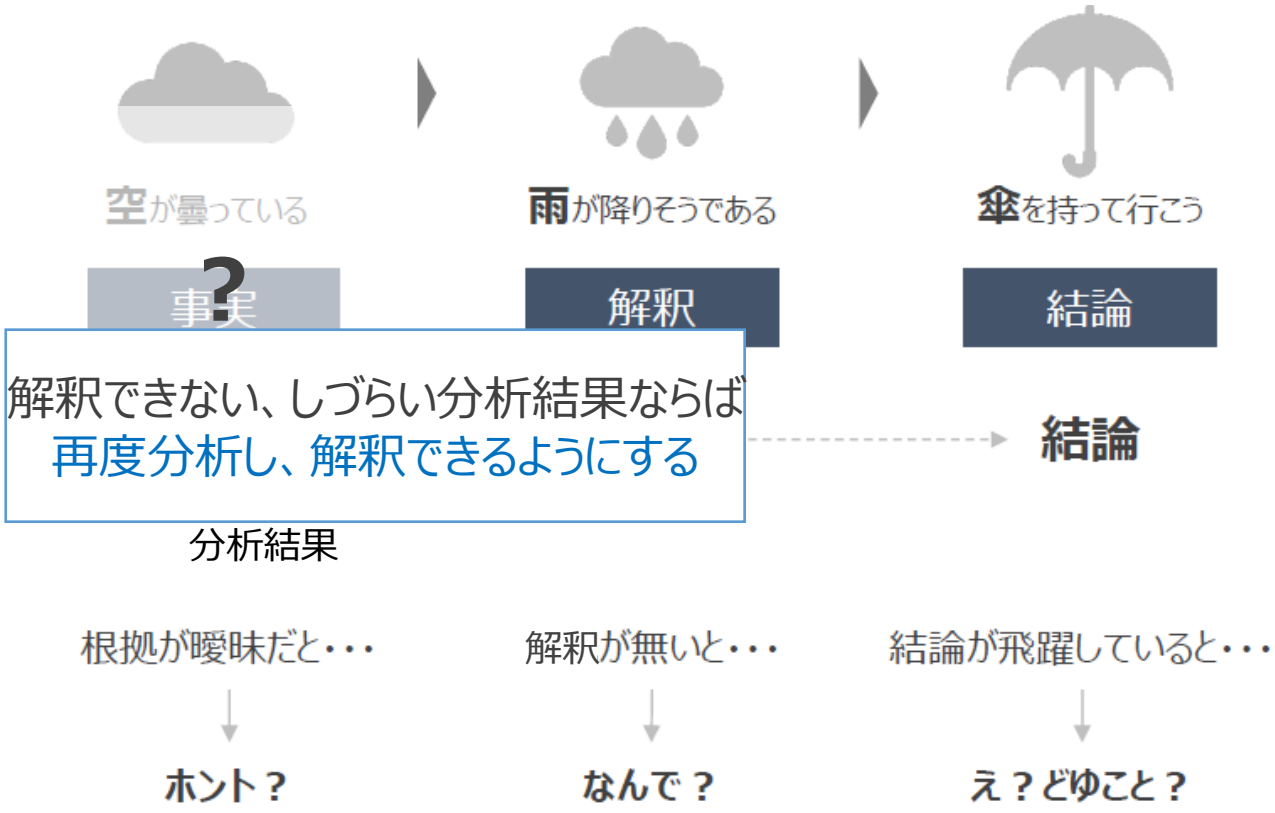
■ データ分析をしたら、その結果を解釈し、最後には、「**意思決定**」が必ずある
■ 結果を解釈することで、分析への理解が深まる



© Data Science Institute

▪なぜ統計データ分析方法の理解は重要か

■ データ分析をしたら、その結果を解釈し、最後には、「**意思決定**」が必ずある
■ 結果を解釈することで、分析への理解が深まる



© Data Science Institute

▪データ利活用のメリット

- データ分析はデータから知りたい情報を獲得する
- データ利活用はデータから収集した情報を活用しビジネスに役立てる
- データ分析は、データ活用の手順のひとつ

売上の向上



業務の効率化・コスト削減



デジタル化による競争力強化・データ利活用に取り組む企業では、労働生産性や売上高にプラスの影響
（中小企業庁の2022年度版中小企業白書）

出典：住友電工情報システム株式会社（https://www.sei-info.co.jp/column/data-utilization/）

© Data Science Institute

▪データ利活用のメリット

- 無駄な業務・工数が発生している箇所をデータ分析で洗い出し、分析結果を基に業務改善
- 蓄積データから売上や必要在庫を予測できれば、不必要な仕入れ数や配置人員を事前に削減

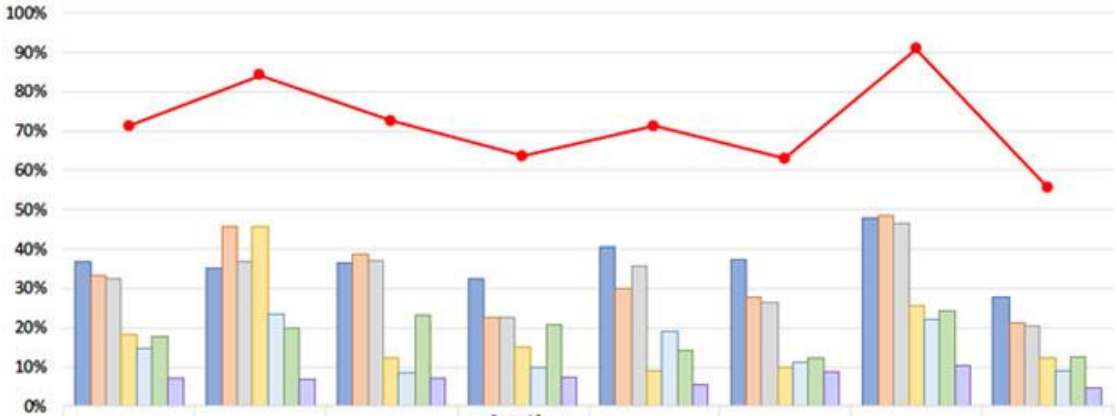
カテゴリ	データの種類
業務データ	・顧客データ ・経理データ ・業務日誌データ
販売記録	・POSデータ ・eコマースにおける販売記録データ
顧客とのコミュニケーション	・電子メール ・CTI音声データ ・固定電話 ・携帯電話
自動取得	・アクセスログ ・動画・映像視聴ログ ・Blog・SNS等記事データ ・GPSデータ ・RFIDデータ ・センサーデータ ・交通量・渋滞情報データ ・気象データ ・防犯・遠隔監視カメラデータ

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

▪データ利活用のメリット

■ 幅広い業界で開発や効率化のためにデータが活用されている



	全体 (n=2003)	製造業 (n=404)	情報通信業 (n=385)	エネルギー・インフラ (n=330)	商業・流通業 (n=481)	サービス業 (n=403)	大企業 (n=886)	中小企業 (n=1117)
経営企画・組織改革	36.7%	35.1%	36.4%	32.4%	40.7%	37.5%	47.9%	27.9%
製品・サービスの企画、開発	33.2%	45.8%	38.7%	22.7%	29.9%	27.8%	48.4%	21.1%
マーケティング	32.2%	36.6%	37.1%	22.7%	35.8%	26.6%	46.7%	20.7%
生産・製造	18.2%	45.5%	12.2%	15.2%	9.1%	9.9%	25.7%	12.3%
物流・在庫管理	14.8%	23.3%	8.6%	10.0%	19.1%	11.2%	22.1%	9.0%
保守・メンテナンス・サポート	17.8%	19.8%	23.1%	20.9%	14.1%	12.4%	24.3%	12.6%
その他（基礎研究、リスク管理等）	7.1%	6.7%	7.3%	7.6%	5.6%	8.7%	10.4%	4.5%
いずれかを利用している	71.2%	84.2%	72.5%	63.6%	71.3%	63.0%	90.9%	55.6%

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

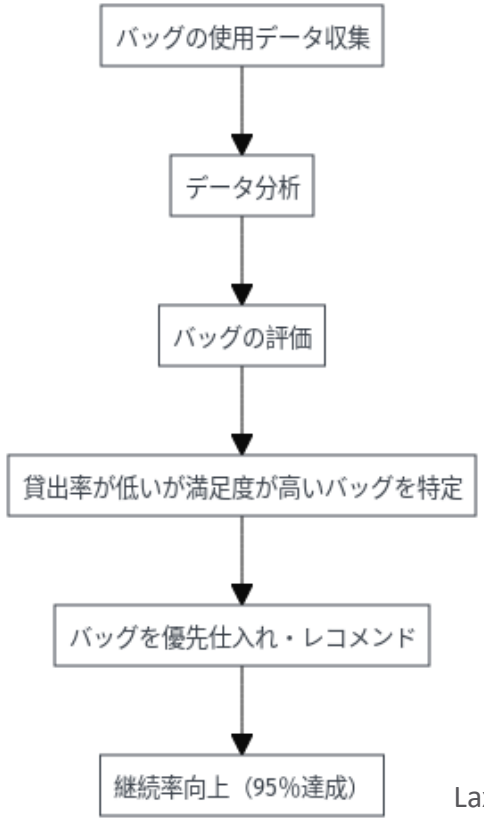
ビジネスにおけるデータ利活用の最前線

© Data Science Institute

▪ビジネスにおけるデータ利活用の領域の拡大

- ラクサス・テクノロジーズ株式会社
- サービス業：ブランドバッグのシェアサービス
- 課題：継続利用者の拡大、バッグが多すぎて顧客が好みに合ったバッグを探せない





Laxus：https://corp.laxus.co/

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

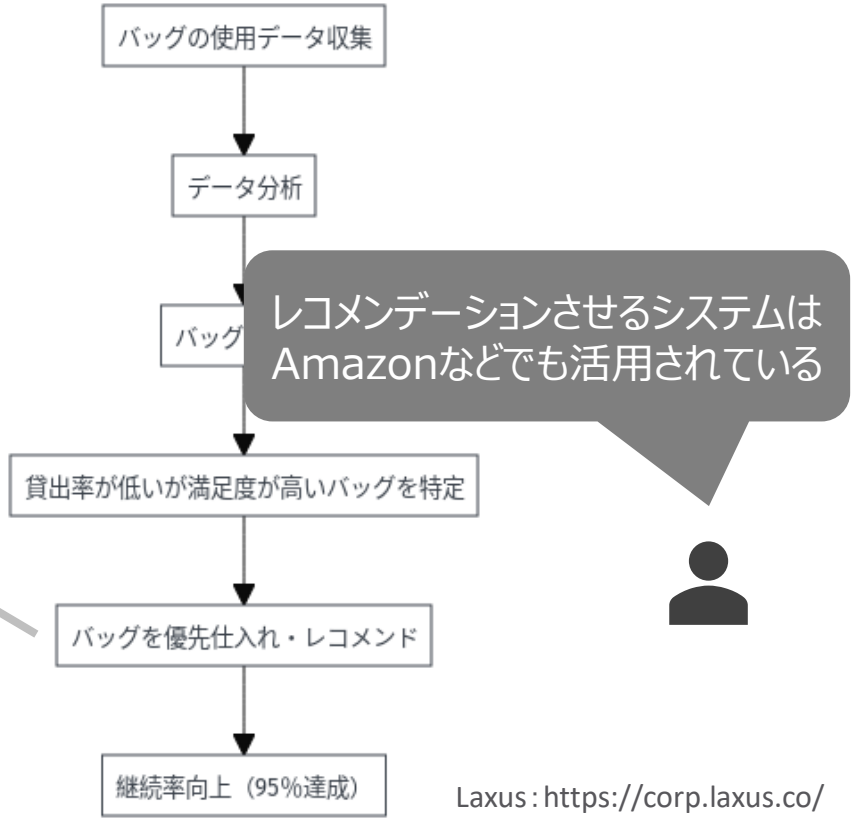
▪ビジネスにおけるデータ利活用の領域の拡大

- ラクサス・テクノロジーズ株式会社
- サービス業：ブランドバッグのシェアサービス
- 課題：継続利用者の拡大、バッグが多すぎて顧客が好みに合ったバッグを探せない

Laxus



長く使われるバッグは満足度が高いという分析結果を仕入れに活用



Laxus：https://corp.laxus.co/

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

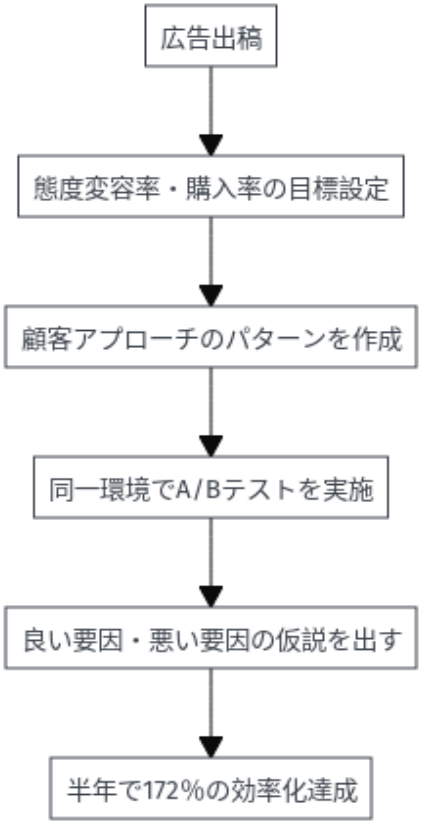
▪ビジネスにおけるデータ利活用の領域の拡大

- 資生堂
- 製造業：化学
- 課題：商品やブランドの価値を伝える場の減少（マス広告の効果下落）

SHISEIDO

活用したデータ

- ウェブ上での行動履歴
- 店舗での購買履歴
- 外部オーディエンスデータ
- リサーチデータ



- 広告出稿から態度変容の起こる割合実際に購入する割合の目標値を設定
- 顧客アプローチのパターンをいくつか作って同一環境でA/Bテスト
- 良い要因・悪い要因の仮説を出すというプロセスを実施

必ず保存されているだろうデータをしっかり分析することによる効果

資生堂：https://corp.shiseido.com/jp/

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

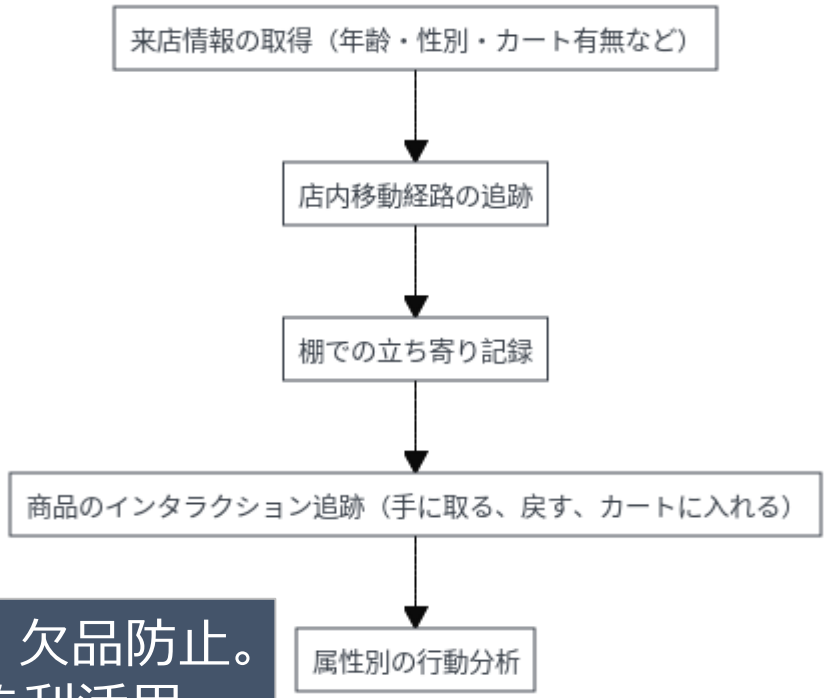
▪ビジネスにおけるデータ利活用の領域の拡大

- スーパー トライアル社
- 小売業：スーパー
- 課題：消費者の属性により、適した広告ができていない。
  棚の欠品が出たら即座に必要な数だけを補充できていない。

TRIAL HLDGS



棚の状況や来店客の画像情報を活用



スタッフの補充作業の軽減、欠品防止。
配置されているカメラデータを利活用。

Trial HLDGS：https://trial-holdings.inc/

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

▪ビジネスにおけるデータ利活用の領域の拡大

- 国分グループ本社
- 食品卸

食のマーケティングカンパニー
KOKUBU



店舗の購買層を分析しただけで売上げ上昇につながる

POSデータ分析（年齢・性別・購買傾向）

↓

商品配置変更の提案（飲用酢を飲料コーナーへ移動）

↓

売り場見直し実施

食酢のコーナーに並ぶことが多かったストレートタイプの飲用酢を、野菜ジュースなどの飲料コーナーに並べた

↓

効果測定（買い上げ率8.8%上昇、売上金額16%増加）

国分グループ：https://www.kokubu.co.jp/

出典：総務省「デジタルデータの経済的価値の計測と活用の現状に関する調査研究の請負報告書」

© Data Science Institute

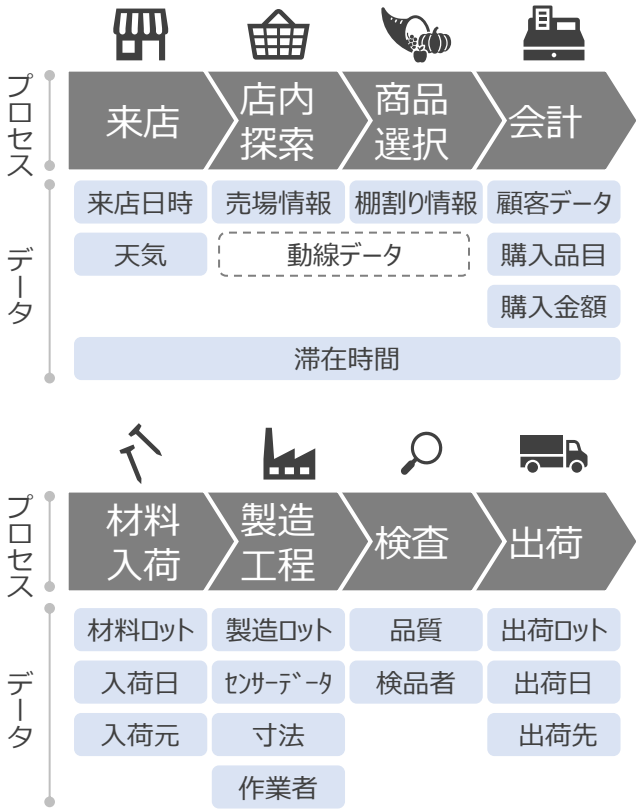
Excelでデータ利活用の基本を学ぶ

© Data Science Institute

▪Excelによる統計データ分析の基本

- 分析する前に明確な目的をたてる
- そのためには対象ビジネスの「プロセス」と「構造」を理解しておくことが重要

ビジネスプロセスを理解する



ビジネス構造を理解する



© Data Science Institute

▪Excelによる統計データ分析の基本

- 分析を開始するためには、データの特徴について把握する必要がある
- データの特徴把握のために代表値がある





店舗	商圏人口	売場面積	売上金額
熱海店	150	400	3,273,000
安城店	400	450	5,318,000
小田原店	600	500	4,789,000
静岡店	1,000	600	8,139,000
東京店	8,000	920	6,894,000
豊橋店	500	490	5,911,000
名古屋店	2,500	820	9,378,000
浜松店	900	540	7,463,000
富士店	500	510	5,617,000
三島店	400	450	4,481,000
横浜店	3,000	810	8,926,000

・
・

代表値 → 平均値

例

売場面積：平均値　590 ㎡
売上金額：平均値　6,380,818円

© Data Science Institute

▪Excelによる統計データ分析の基本

- 分析を開始するためには、データの特徴について把握する必要がある
- データの特徴把握のために代表値がある



データが多すぎて
何が何だか・・・



平均値から各店舗大体
6,380,818円の売上か！

店舗	商圏人口	売場面積	売上金額
熱海店	150	400	3,273,000
安城店	400	450	5,318,000
小田原店	600	500	4,789,000
静岡店	1,000	600	8,139,000
東京店	8,000	920	6,894,000
豊橋店	500	490	5,911,000
名古屋店	2,500	820	9,378,000
浜松店	900	540	7,463,000
富士店	500	510	5,617,000
三島店	400	450	4,481,000
横浜店	3,000	810	8,926,000

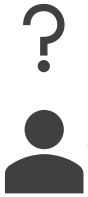
・
・
・

代表値

→ 平均値

例

売場面積：平均値　590 ㎡
売上金額：平均値　6,380,818円



平均で、大体各店舗の売上は分かったけど、
大きく違うところがある

© Data Science Institute

▪Excelによる統計データ分析の基本

■平均値の特徴の理解が必須
■平均値だけではなく、中央値をしっかり確認

金融広報中央委員会の家計の金融行動に関する世論調査により

30代の1人暮らしの
平均貯蓄額：912万円
貯蓄額の中央値：300万円
※貯蓄のない方は除く



平均値はデータをすべて足した後にデータの個数で割った数値
中央値はデータを大きい順に並べたときに、中央に来る数値



平均値は、最小や最大の数値がかけ離れた値によって、データの真ん中から大きくずれる
そのため、順位として真ん中になる中央値を、平均値とあわせて確認する

MUFG：https://www.cr.mufg.jp/

© Data Science Institute

▪Excelによる統計データ分析の基本



データが多すぎて
何が何だか・・・



中央値より平均値の方が大きい！
高い売上を出している店舗がある！

店舗	商圏人口	売場面積	売上金額
熱海店	150	400	3,273,000
安城店	400	450	5,318,000
小田原店	600	500	4,789,000
静岡店	1,000	600	8,139,000
東京店	8,000	920	6,894,000
豊橋店	500	490	5,911,000
名古屋店	2,500	820	9,378,000
浜松店	900	540	7,463,000
富士店	500	510	5,617,000
三島店	400	450	4,481,000
横浜店	3,000	810	8,926,000

・
・
・

代表値

平均値
中央値

例 売場面積：平均値　590 ㎡
中央値　510 ㎡
売上金額：平均値　6,380,818円
中央値　5,911,000円

© Data Science Institute

▪Excelによる統計データ分析の基本

データ

店舗	商圏人口	売場面積	売上金額
熱海店	150	400	3,273,000
安城店	400	450	5,318,000
小田原店	600	500	4,789,000
静岡店	1,000	600	8,139,000
東京店	8,000	920	6,894,000
豊橋店	500	490	5,911,000
名古屋店	2,500	820	9,378,000
浜松店	900	540	7,463,000
富士店	500	510	5,617,000
三島店	400	450	4,481,000
横浜店	3,000	810	8,926,000

AVERAGE（平均値）

MEDIAN（中央値）



	A	B	C	D
1	店舗	商圏人口	売場面積	売上金額
2	熱海店	150	400	3,273,000
3	安城店	400	450	5,318,000
4	小田原店	600	500	4,789,000
5	静岡店	1,000	600	8,139,000
6	東京店	8,000	920	6,894,000
7	豊橋店	500	490	5,911,000
8	名古屋店	2,500	820	9,378,000
9	浜松店	900	540	7,463,000
10	富士店	500	510	5,617,000
11	三島店	400	450	4,481,000
12	横浜店	3,000	810	8,926,000
13				
14	平均値	1,632	590	6,380,818
15	中央値	600	510	5,911,000

© Data Science Institute

▪Excelでの可視化とデータ観察方法

- グラフによるデータの可視化
- 表やグラフにまとめた方がただ代表値を見るより見やすくなる
- どのような観点でまとめるかによって、把握しやすい特徴が異なる点に注意

例

商品Aと商品Bを男性と女性にそれぞれ０~10点で評価してもらった

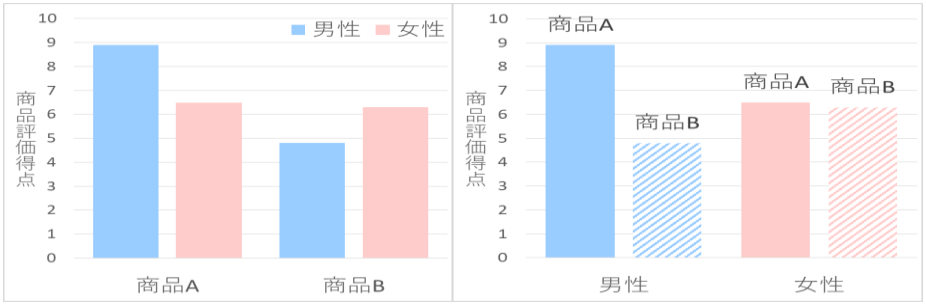
・文章のみ

男性の商品A得点は8.9点、商品Bは4.8点、女性の商品A得点は6.5点、商品Bは6.3点でした。

・表

	商品A	商品B
男性	8.9	4.8
女性	6.5	6.3

・グラフ

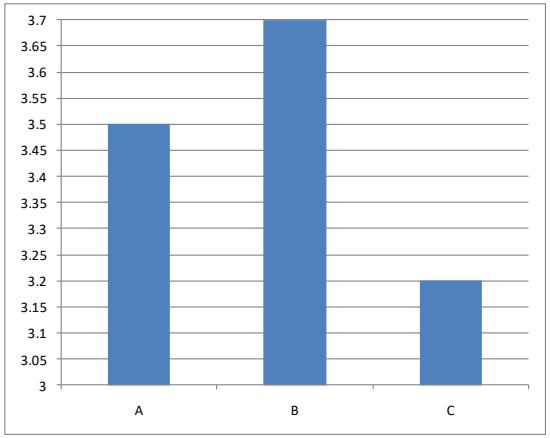


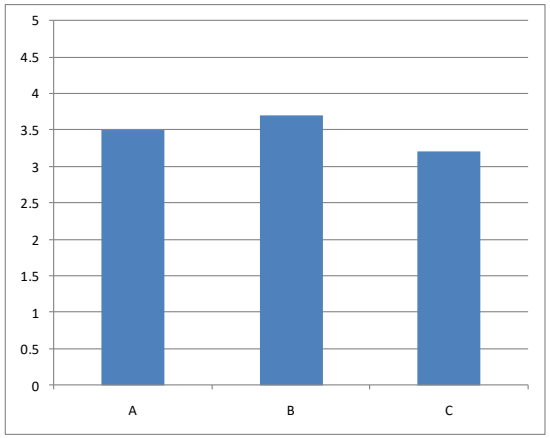
© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ グラフはデータを可視化して分かりやすくするが、誤った解釈に繋がる危険性あり

次の2つのグラフ、左側のグラフは大きな差があるように見える
一方で、右側のグラフはあまり差が無いように見える
しかし、この2つは全く同じデータを示したグラフである





A	3.5
B	3.7
C	3.2

※データの範囲は0～5

異なる部分は軸の値の範囲
左：3～3.7　(0.05刻み)
右：0～5　(0.5刻み)

特に比較をする場合はデータの範囲、単位などに気を付ける

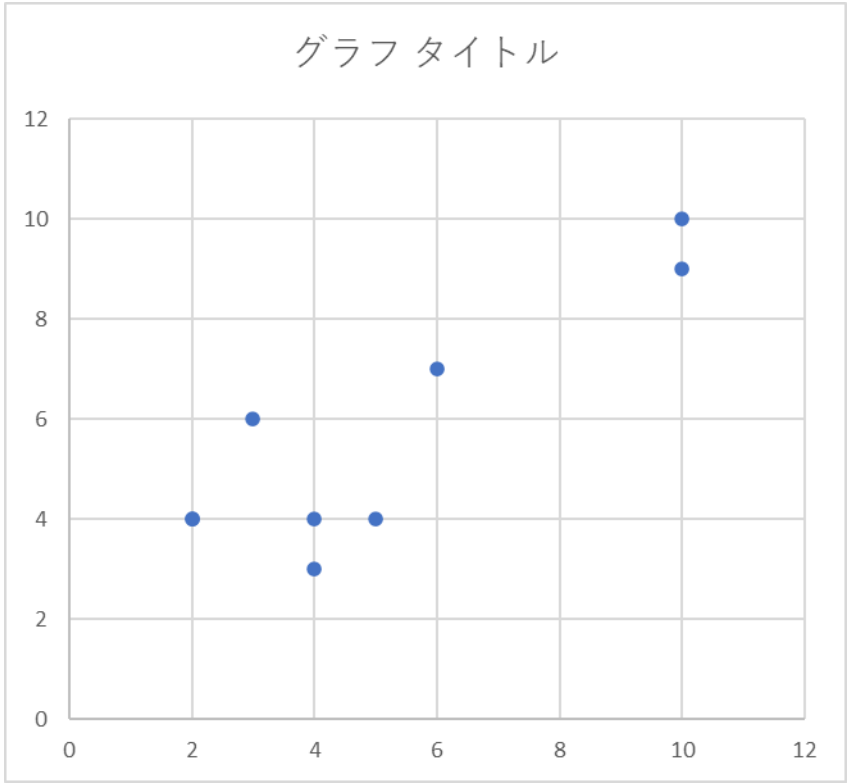
© Data Science Institute

▪Excelでの可視化とデータ観察方法

- 散布図は2変数間の関係性を表す際に用いる
- データのバラツキや、変数間にどのような関係があるのかを端的に示す

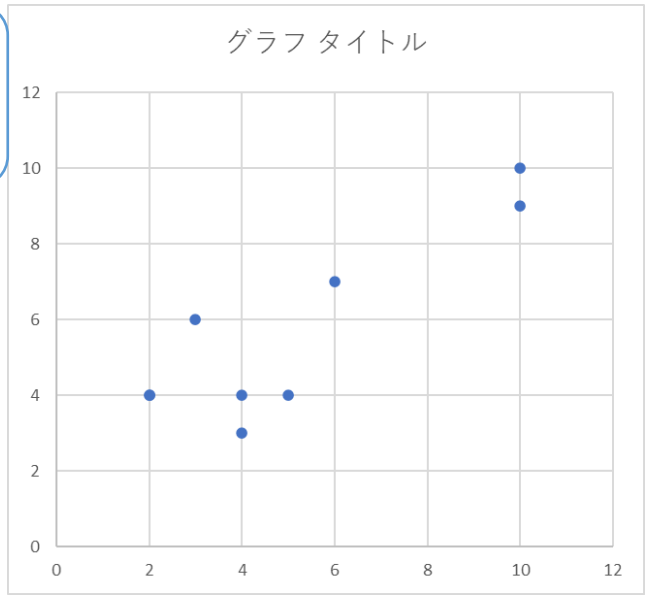
スマートフォン9モデルの
通信性能とアプリ性能の評価

	通信性能	アプリ性能
モデル01	4	3
モデル02	10	10
モデル03	6	7
モデル04	2	4
モデル05	3	6
モデル06	5	4
モデル07	10	9
モデル08	2	4
モデル09	4	4



© Data Science Institute

グラフ作成のアイコンは
「挿入タブ」を選択



行方向

	B	C
	通信性能	アプリ性能
2	モデル01	4	3
3	モデル02	10	
4	モデル03	6	
5	モデル04	2	
6	モデル05	3	6
7	モデル06	5	4
8	モデル07	10	9
9	モデル08	2	4
10	モデル09	4	4

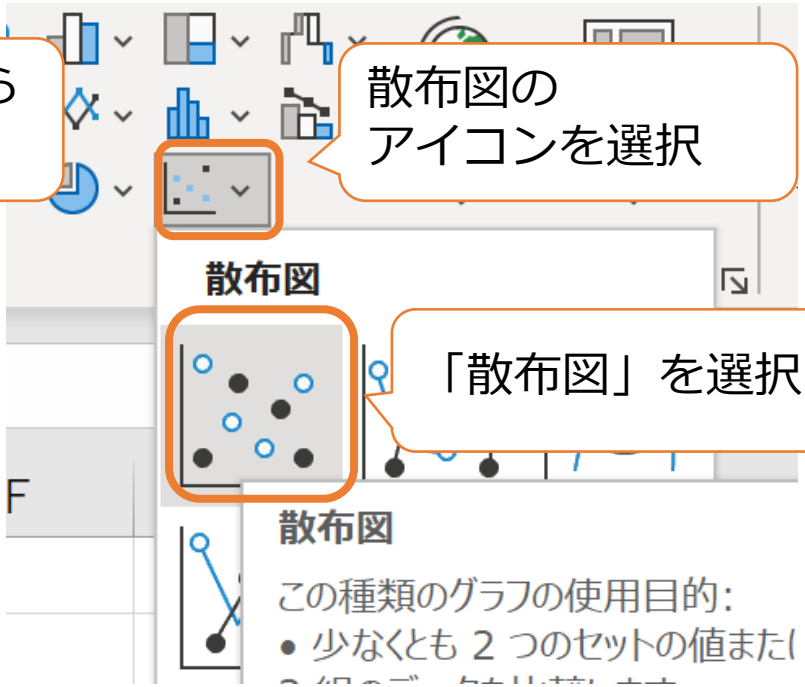
まず、
1列目だけを選択

「Ctrl」キーを押しながら
2つ目の列を選択

列方向

	A	B
		通信性能	アプリ性能
	1	4	3
3	モデル02	10	10
4	モデル03	6	7
5	モデル04	2	4
6	モデル05	3	6
7	モデル06	5	4
8	モデル07	10	9
9	モデル08	2	4
10	モデル09	4	4

Ctrl



散布図の
アイコンを選択

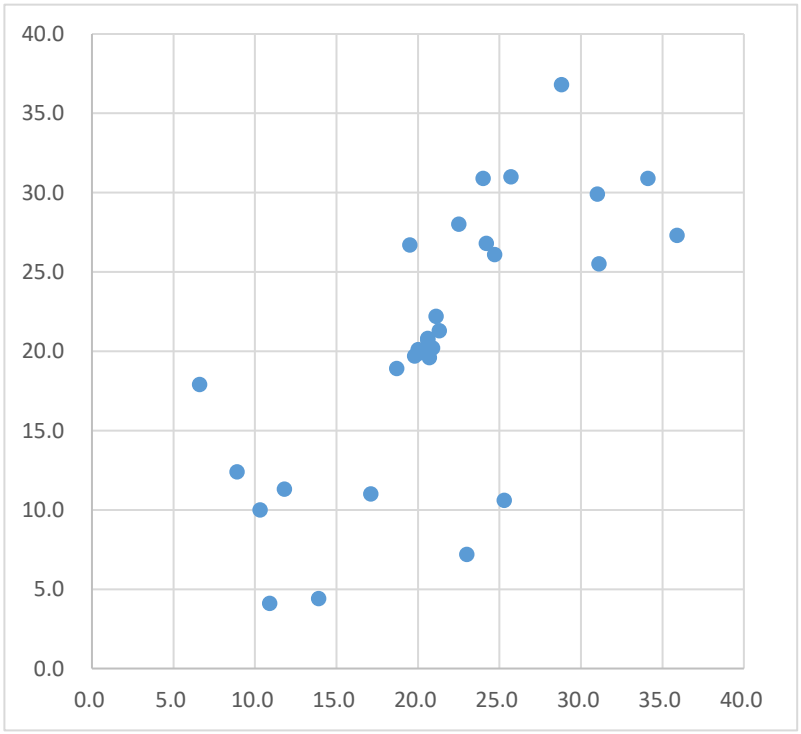
「散布図」を選択

© Data Science Institute

▪Excelでの可視化とデータ観察方法

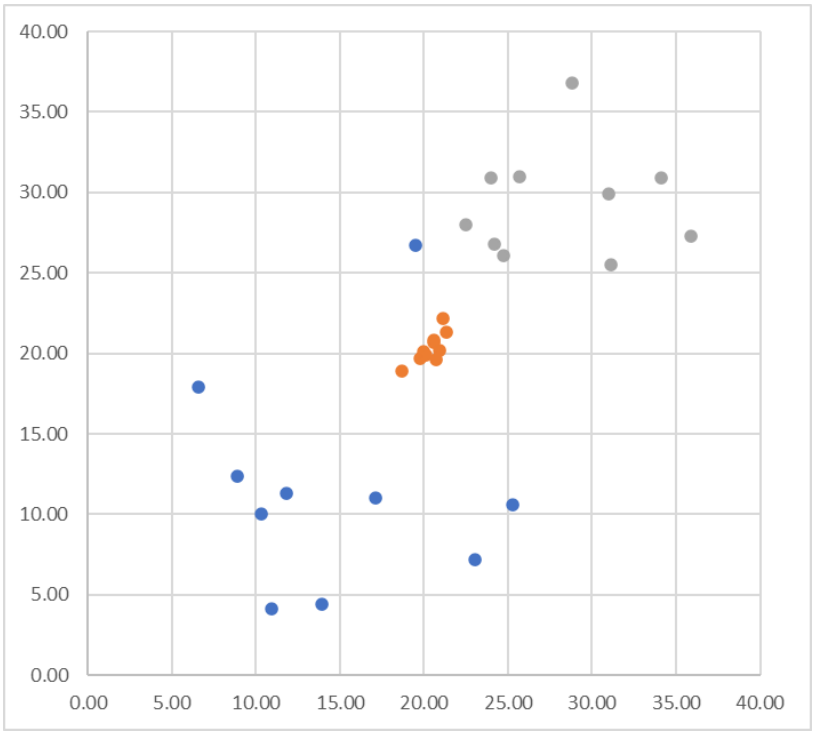
- 層別散布図はカテゴリごとに散布図を作成しそれを重ね合わせたもの
- カテゴリによる分布の違いを捉えることができる

散布図



散布図で見ると関係性がありそう

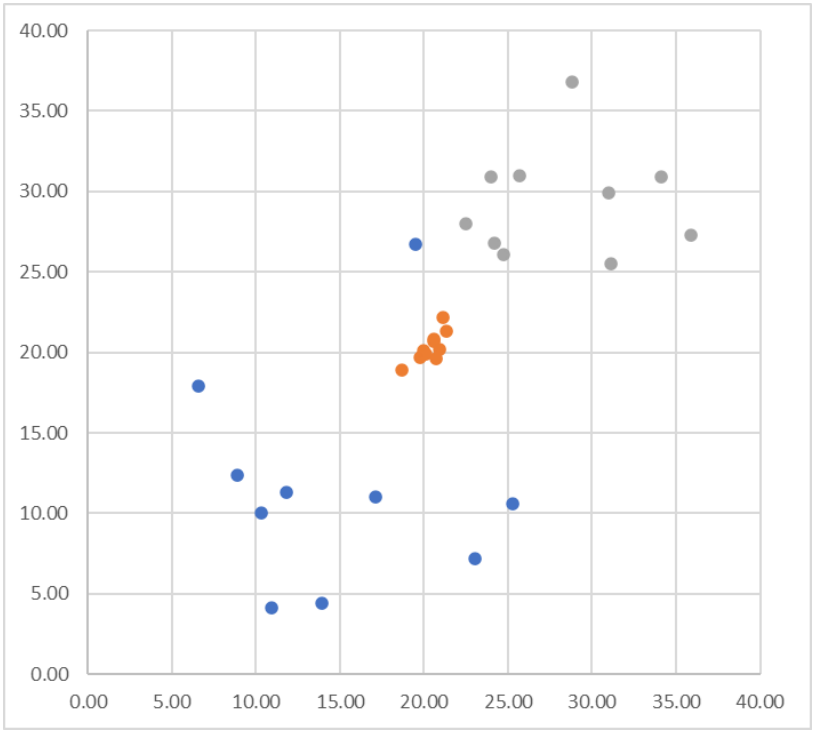
層別散布図



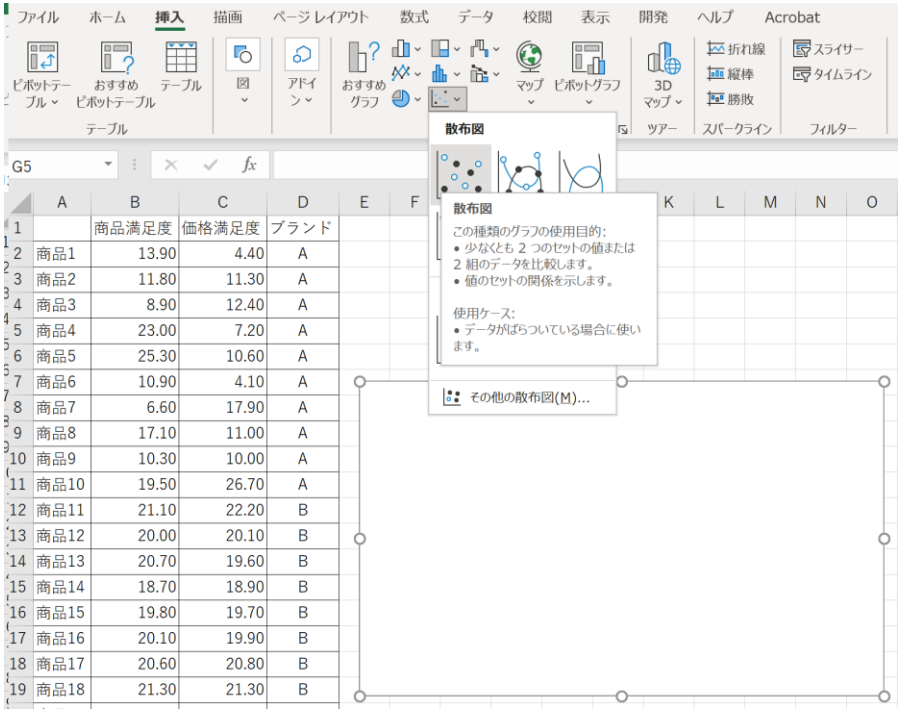
カテゴリで色分けすると関係性がなさそう

© Data Science Institute

○層別散布図：作成方法



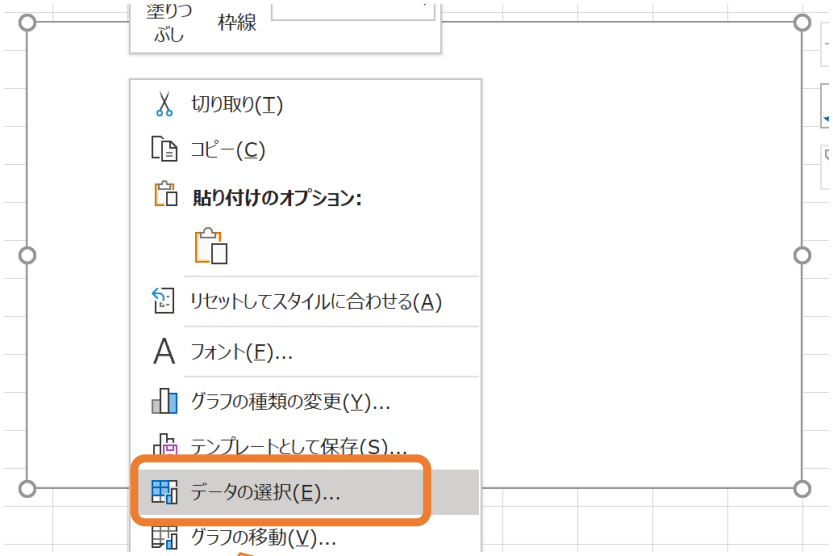
何も入力されていないセルを選択し、
まずは空の散布図を作成する



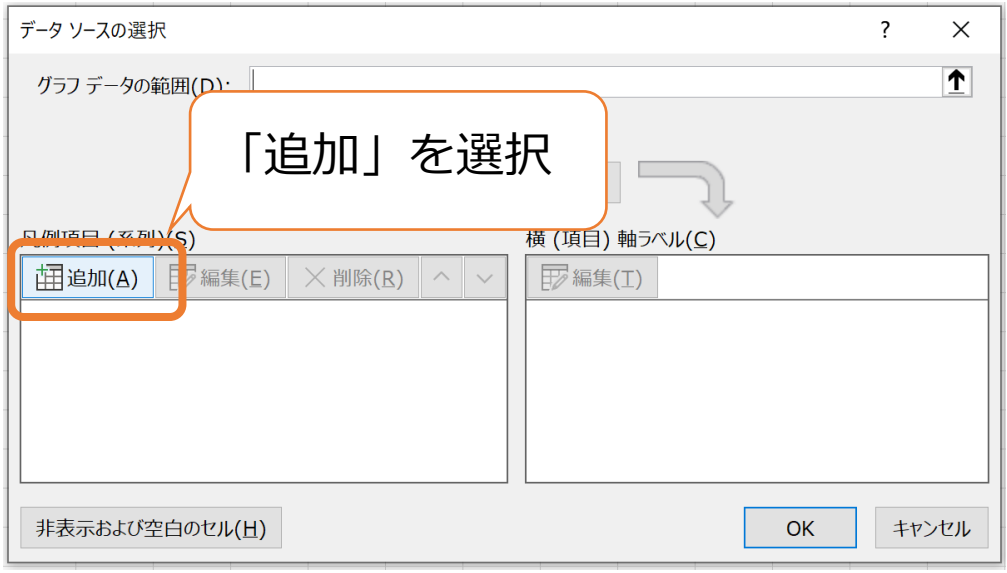
© Data Science Institute

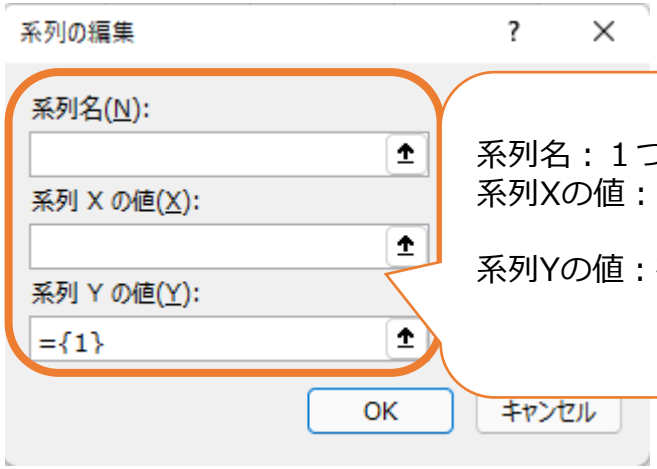
○層別散布図：作成方法

層に指定するカテゴリごとに
データを追加する必要がある



空のグラフの上で右クリック
「データの選択」を選択



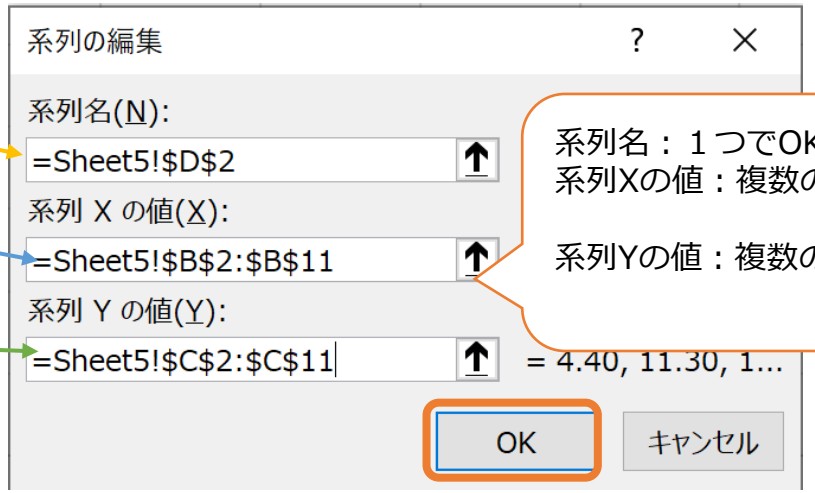


系列名：１つでOK
系列Xの値：複数の値を
まとめて選択
系列Yの値：複数の値を
まとめて選択

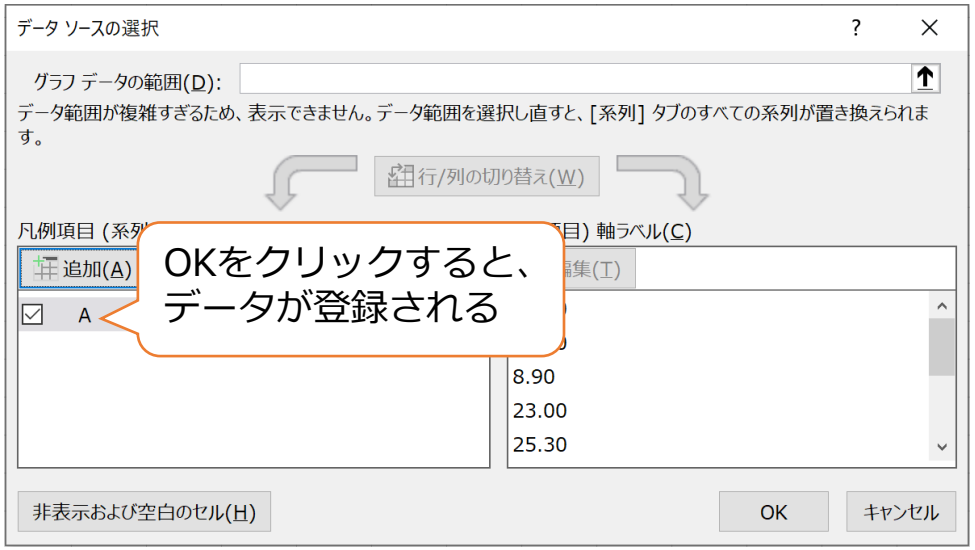
© Data Science Institute

○層別散布図：作成方法

	A	B	C	D
1		商品満足度	価格満足度	ブランド
2	商品1	13.9	4.4	A
3	商品2	11.8	11.3	A
4	商品3	8.9	12.4	A
5	商品4	23	7.2	A
6	商品5	25.3	10.6	A
7	商品6	10.9	4.1	A
8	商品7	6.6	17.9	A
9	商品8	17.1	11	A
10	商品9	10.3	10	A
11	商品10	19.5	26.7	A
12	商品11	21.1	22.2	B
13	商品12	20	20.1	B
14	商品13	20.7	19.6	B
15	商品14	18.7	18.9	B
16	商品15	19.8	19.7	B
17	商品16	20.1	19.9	B
18	商品17	20.6	20.8	B
19	商品18	21.3	21.3	B
20	商品19	20.9	20.2	B
21	商品20	20.6	20.7	B
22	商品21	24.2	26.8	C
23	商品22	22.5	28	C
24	商品23	35.9	27.3	C
25	商品24	24	30.9	C
26	商品25	31	29.9	C
27	商品26	25.7	31	C
28	商品27	24.7	26.1	C
29	商品28	28.8	36.8	C
30	商品29	34.1	30.9	C
31	商品30	31.1	25.5	C



系列名：１つでOK
系列Xの値：複数の値を
まとめて選択
系列Yの値：複数の値を
まとめて選択

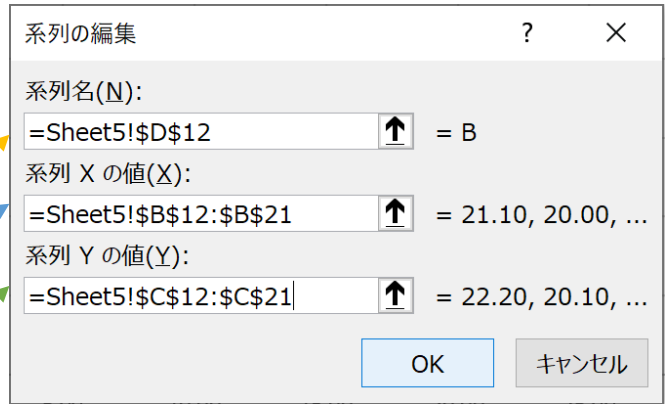


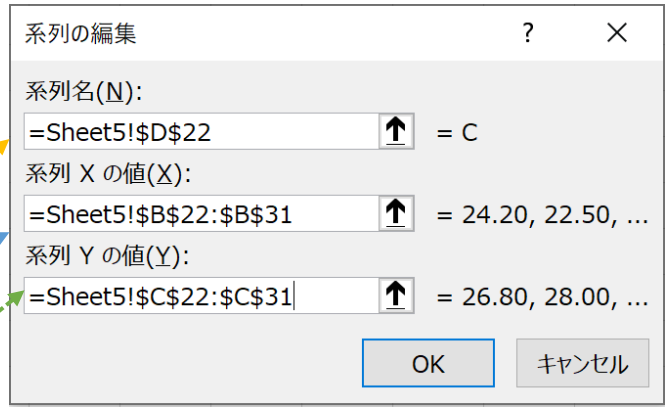
OKをクリックすると、
データが登録される

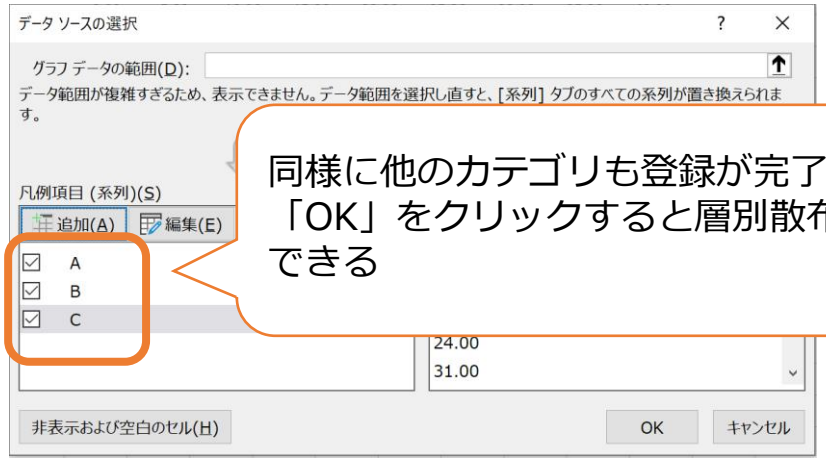
© Data Science Institute

○層別散布図：作成方法

	A	B	C	D
1		商品満足度	価格満足度	ブランド
2	商品1	13.9	4.4	A
3	商品2	11.8	11.3	A
4	商品3	8.9	12.4	A
5	商品4	23	7.2	A
6	商品5	25.3	10.6	A
7	商品6	10.9	4.1	A
8	商品7	6.6	17.9	A
9	商品8	17.1	11	A
10	商品9	10.3	10	A
11	商品10	19.5	26.7	A
12	商品11	21.1	22.2	B
13	商品12	20	20.1	B
14	商品13	20.7	19.6	B
15	商品14	18.7	18.9	B
16	商品15	19.8	19.7	B
17	商品16	20.1	19.9	B
18	商品17	20.6	20.8	B
19	商品18	21.3	21.3	B
20	商品19	20.9	20.2	B
21	商品20	20.6	20.7	B
22	商品21	24.2	26.8	C
23	商品22	22.5	28	C
24	商品23	35.9	27.3	C
25	商品24	24	30.9	C
26	商品25	31	29.9	C
27	商品26	25.7	31	C
28	商品27	24.7	26.1	C
29	商品28	28.8	36.8	C
30	商品29	34.1	30.9	C
31	商品30	31.1	25.5	C







同様に他のカテゴリも登録が完了したら、「OK」をクリックすると層別散布図ができる

© Data Science Institute

○層別散布図：作成方法

商品満足度	A	B	C
13.9	4.4		
11.8	11.3		
8.9	12.4		
23	7.2		
25.3	10.6		
10.9	4.1		
6.6	17.9		
17.1	11		
10.3	10		
19.5	26.7		
21.1		22.2	
20		20.1	
20.7		19.6	
18.7		18.9	
19.8		19.7	
20.1		19.9	
20.6		20.8	
21.3		21.3	
20.9		20.2	
20.6		20.7	
24.2			26.8
22.5			28
35.9			27.3

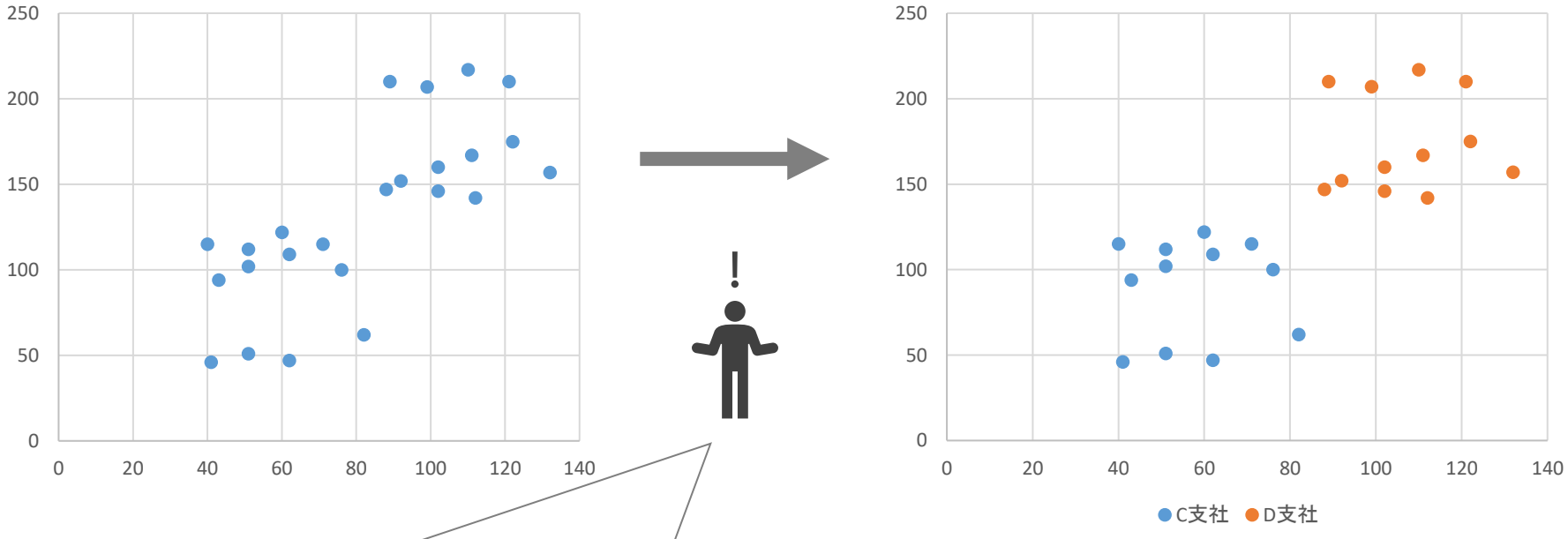


列ごとにカテゴリを分けて
作成することも可能

© Data Science Institute

▪Excelでの可視化とデータ観察方法

- 関係性を見たい目的で、まずは全体のデータの把握として散布図を作成する
- カテゴリごとに層別に分けて確認してから解釈しましょう

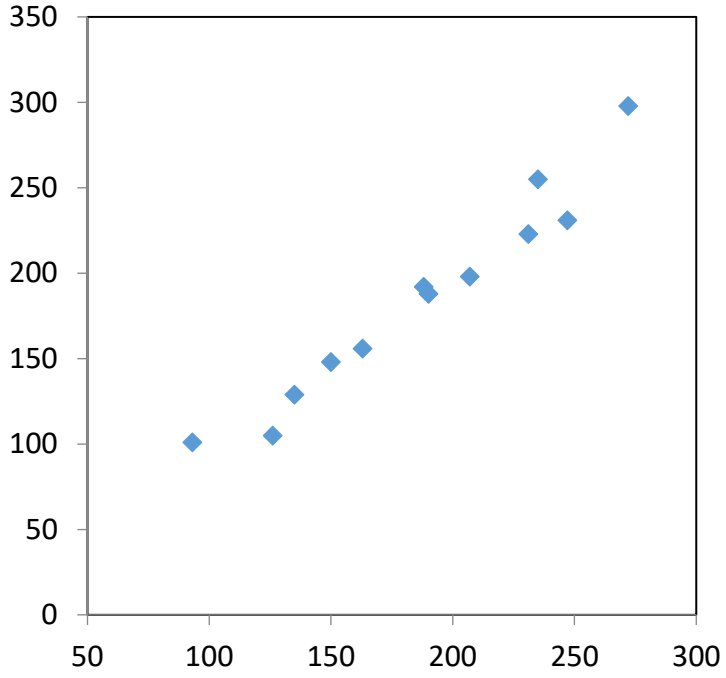


左の散布図だけだと、広告費（横軸）が増えるほど、売上（縦軸）が高いと解釈できそう！
支社ごとに層別散布図にしたら、あまり関係がなさそうなことが分かった

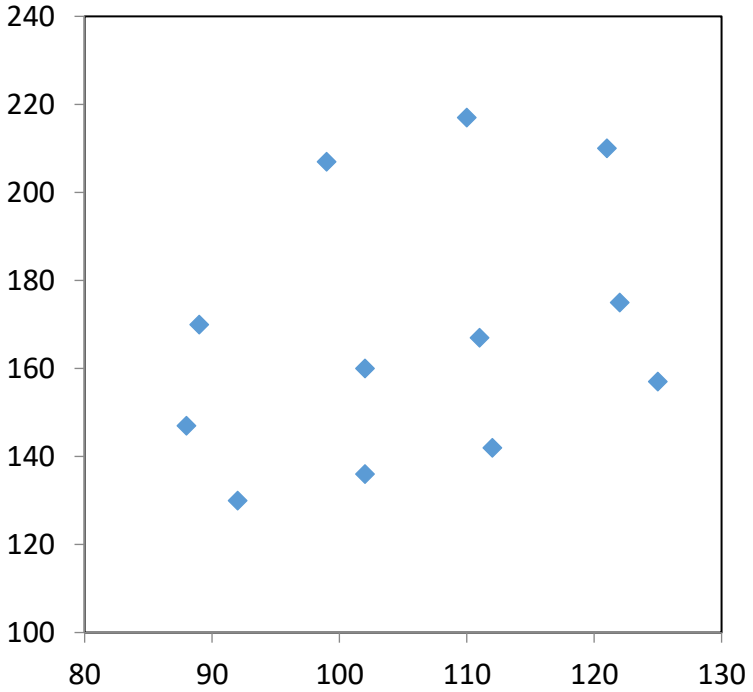
© Data Science Institute

▪Excelでの可視化とデータ観察方法

- 散布図で2変数間の関係性はみれるが、どちらがより関係性が高いかの判断が図からは困難
- 関係性の強度を示す代表値として**相関係数（r）**がある



r ＝0.97



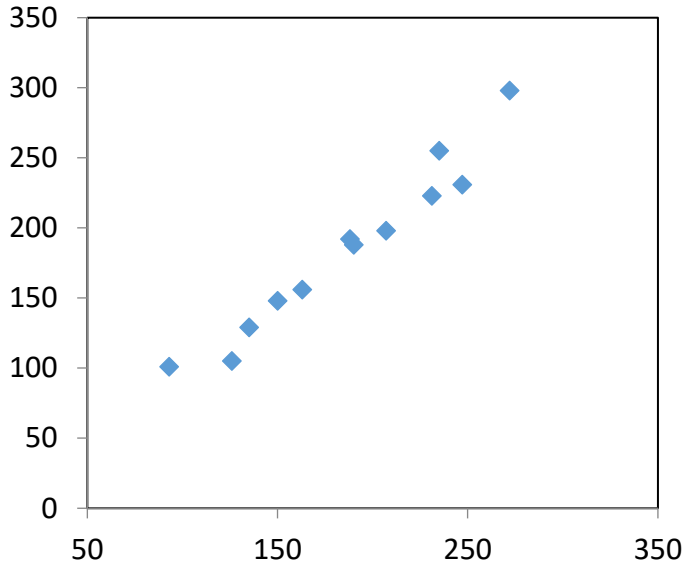
r ＝0.32

$$-1 \leqq r \leqq 1$$

© Data Science Institute

▪Excelでの可視化とデータ観察方法

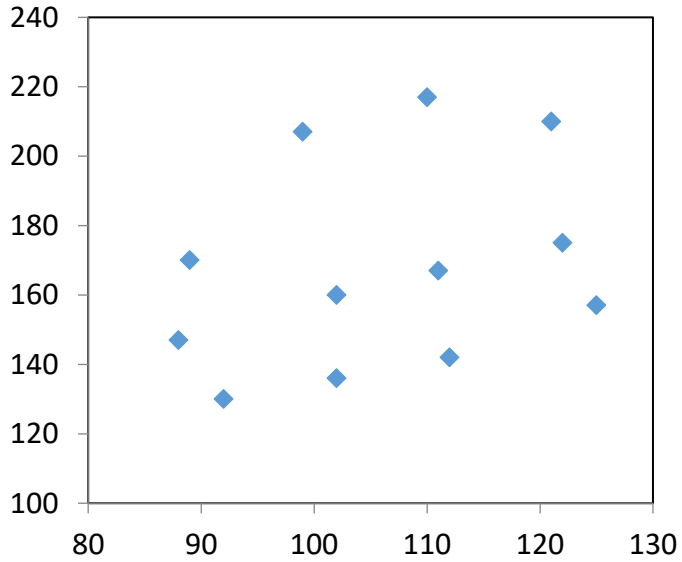
■ 相関係数は二乗することで解釈が容易になる



r ＝0.97

r^2＝0.9409

説明力は94.09%



r ＝0.32

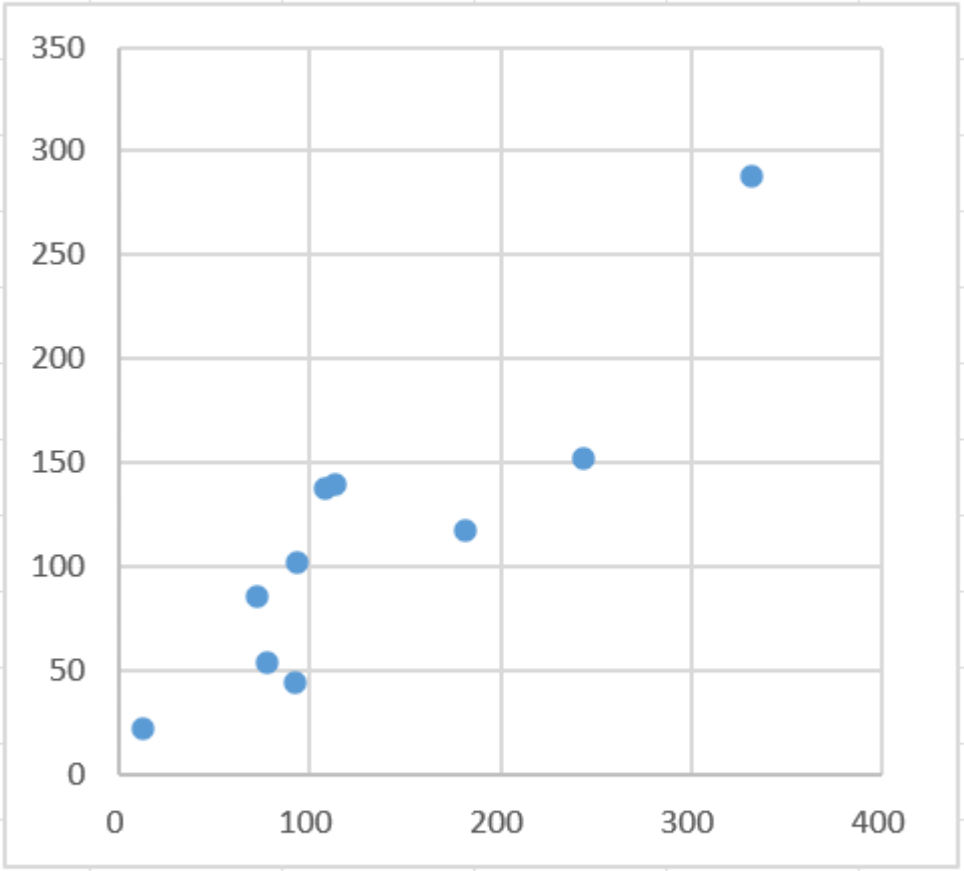
r^2＝0.1024

説明力は10.24%

© Data Science Institute

=CORREL(B2:B11,C2:C11)

支店	広告費	売上高
北海道	92	44
東北	93	102
関東	332	288
北陸	78	54
中部	181	118
近畿	108	138
中国	113	140
四国	72	86
九州	243	152
沖縄	13	22
合計	1,325	1,144

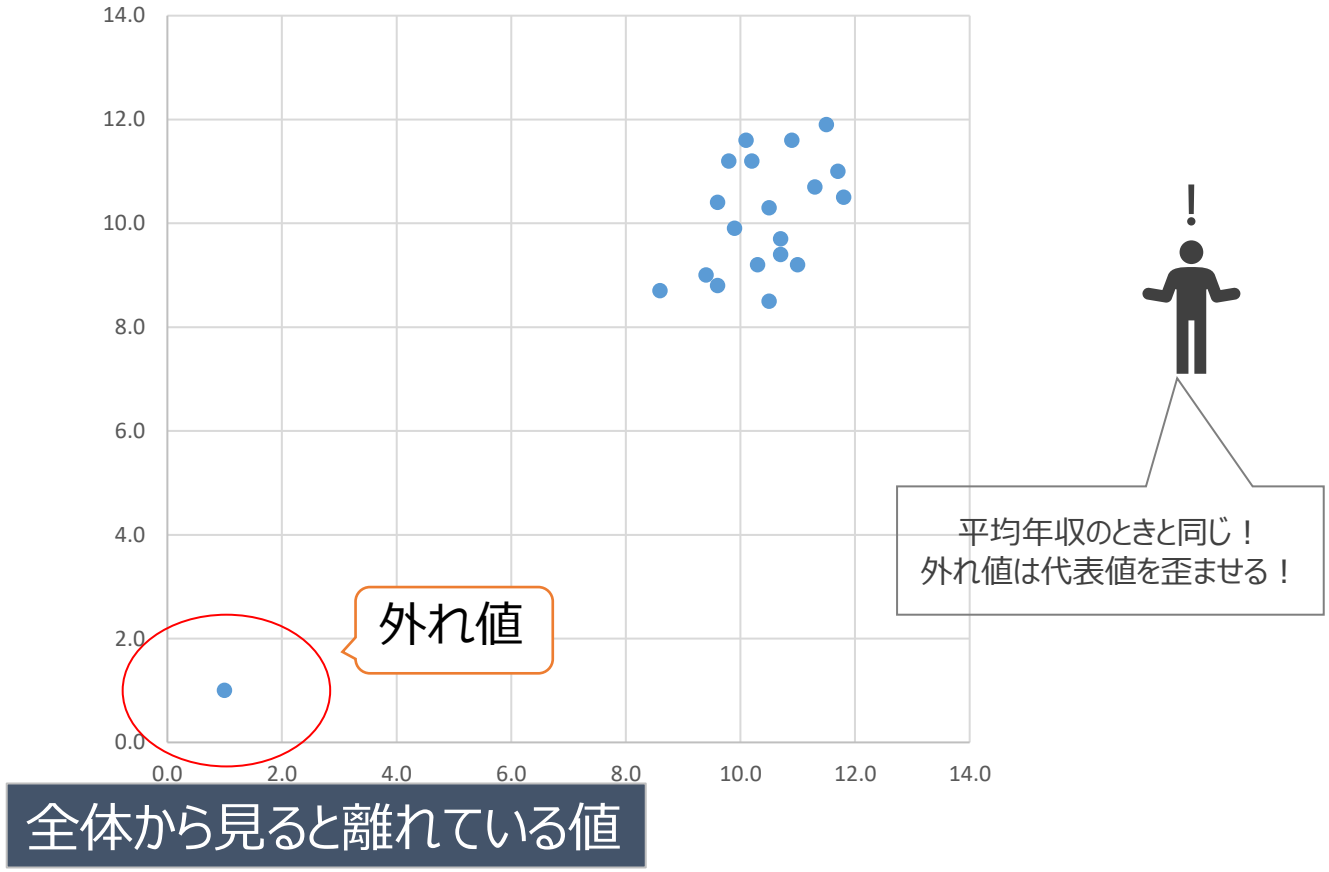


相関係数	0.90087	CORREL

CORREL（相関係数）

© Data Science Institute

商品A	商品B
1.0	1.0
10.5	8.5
8.6	8.7
9.6	8.8
9.4	9.0
11.0	9.2
10.3	9.2
10.7	9.4
10.7	9.7
9.9	9.9
10.5	10.3
9.6	10.4
11.8	10.5
11.3	10.7
11.7	11.0
10.2	11.2
9.8	11.2
10.9	11.6
10.1	11.6
11.5	11.9



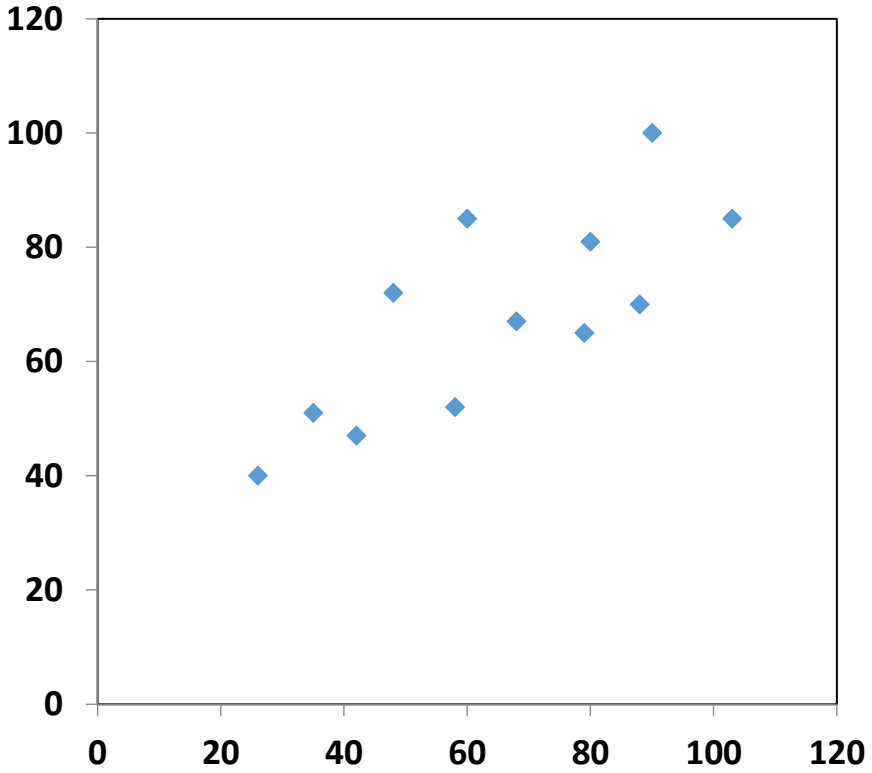
外れ値を含むと相関係数0.9
外れ値を削除すると相関係数0.4

© Data Science Institute

A支社の広告費と売上高

月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	42	26	35	68	103	90	88	58	79	80	48	60
売上高	47	40	51	67	85	100	70	52	65	81	72	85

$r = 0.775$

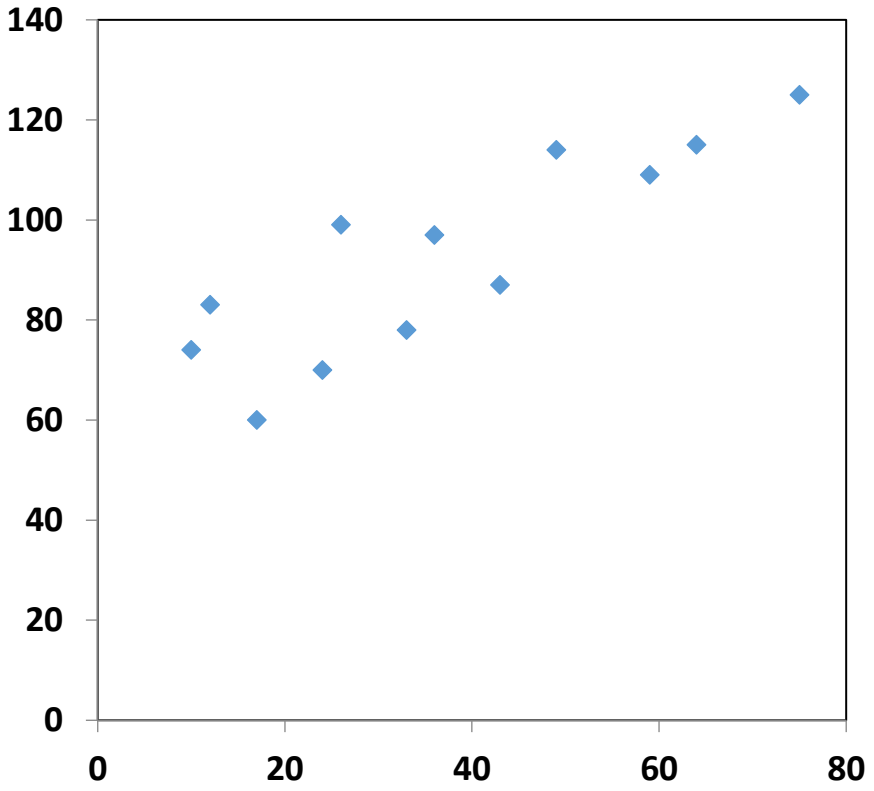


© Data Science Institute

B支社の広告費と売上高

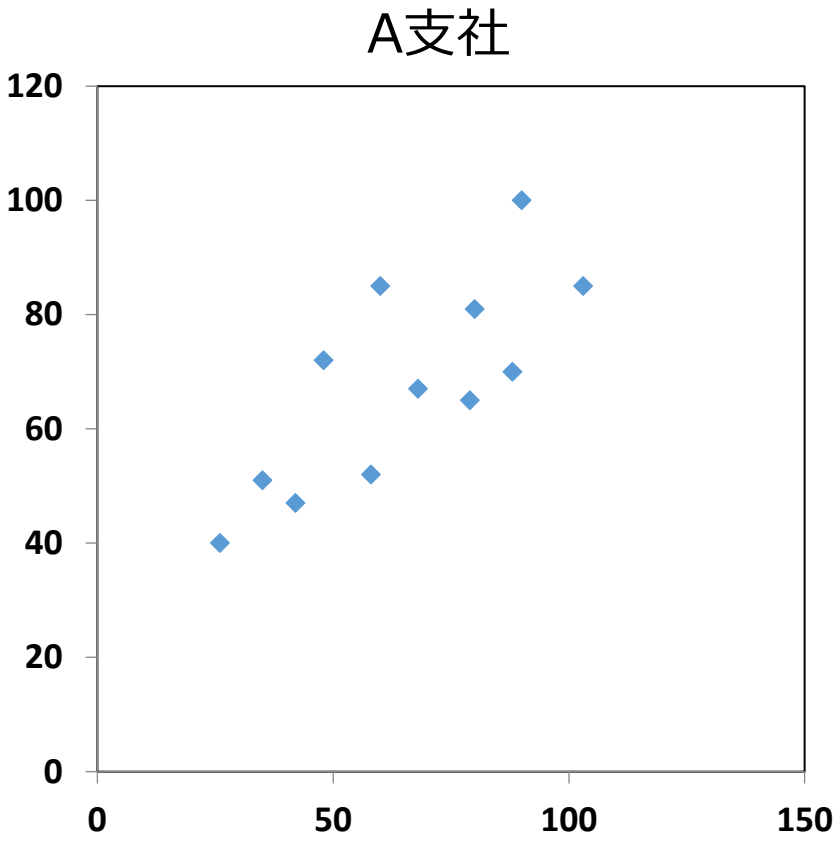
月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	12	10	17	59	75	64	33	24	49	43	36	26
売上高	83	74	60	109	125	115	78	70	114	87	97	99

r ＝0.863

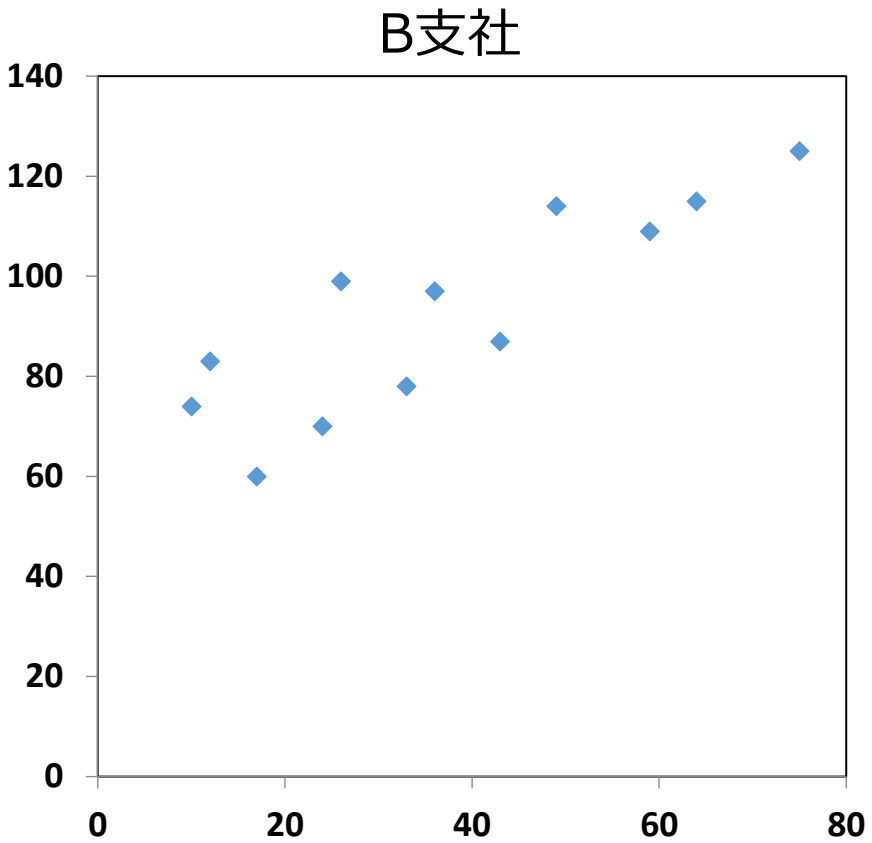


© Data Science Institute

A支社、B支社の広告費と売上高



r ＝0.775



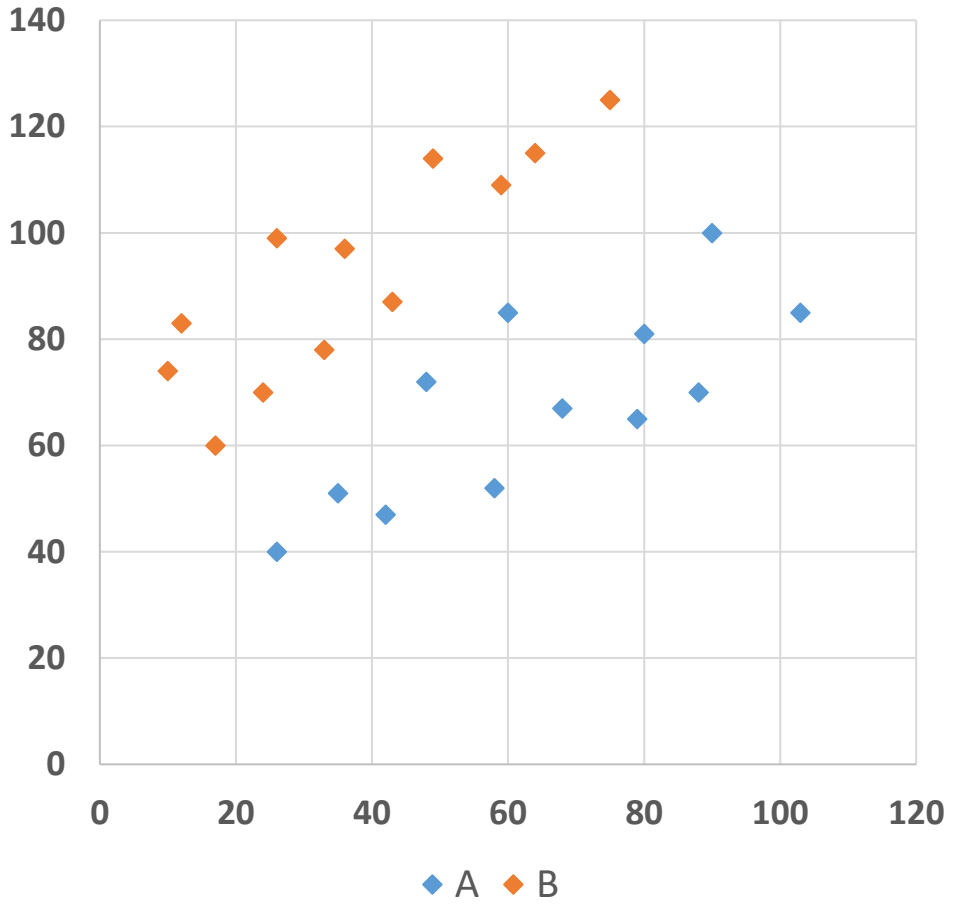
r ＝0.863

（A支社＋B支社）の相関係数は？

© Data Science Institute

（A支社＋B支社）の広告費と売上高

r ＝0.272

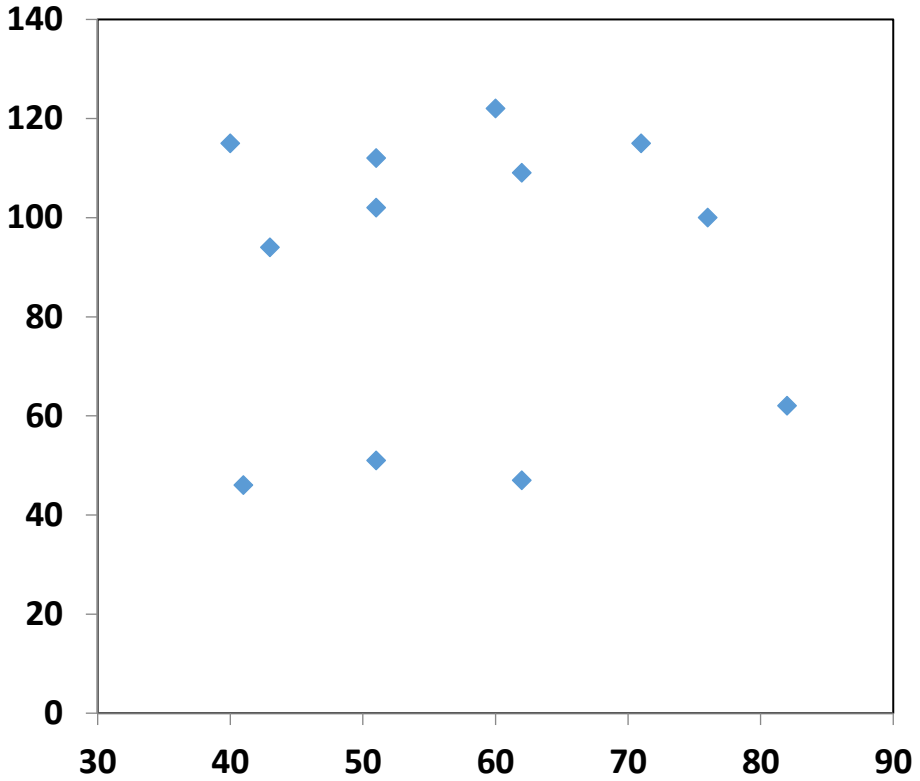


© Data Science Institute

C支社の広告費と売上高

月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	43	51	62	76	41	51	62	82	40	51	60	71
売上高	94	102	109	100	46	51	47	62	115	112	122	115

r＝0.016

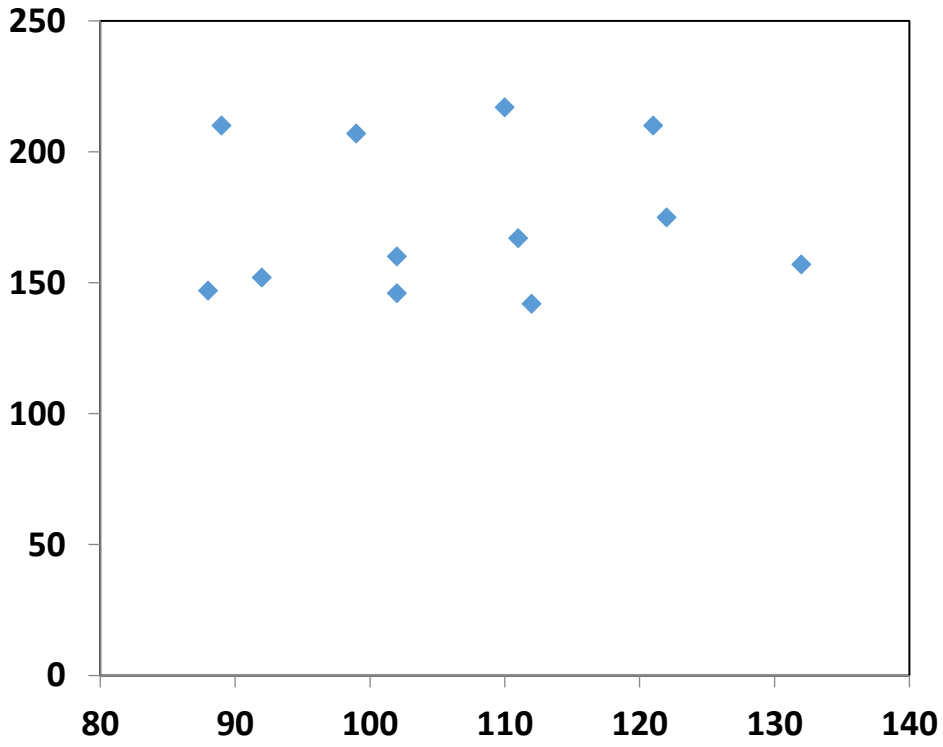


© Data Science Institute

D支社の広告費と売上高

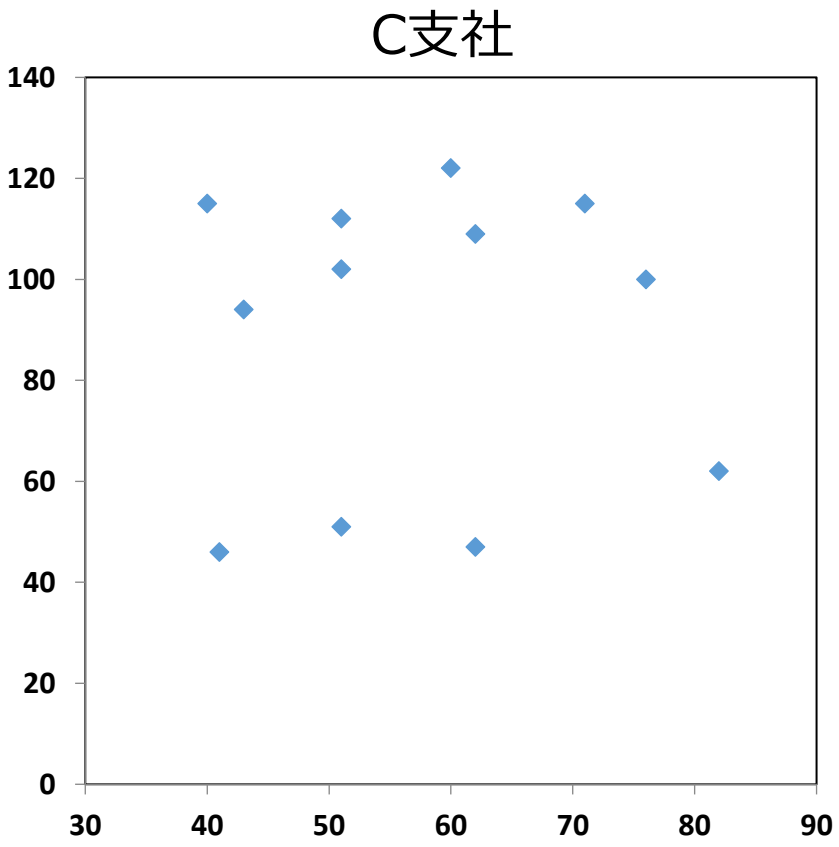
月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	92	102	111	122	88	102	112	132	89	99	110	121
売上高	152	160	167	175	147	146	142	157	210	207	217	210

r ＝0.040

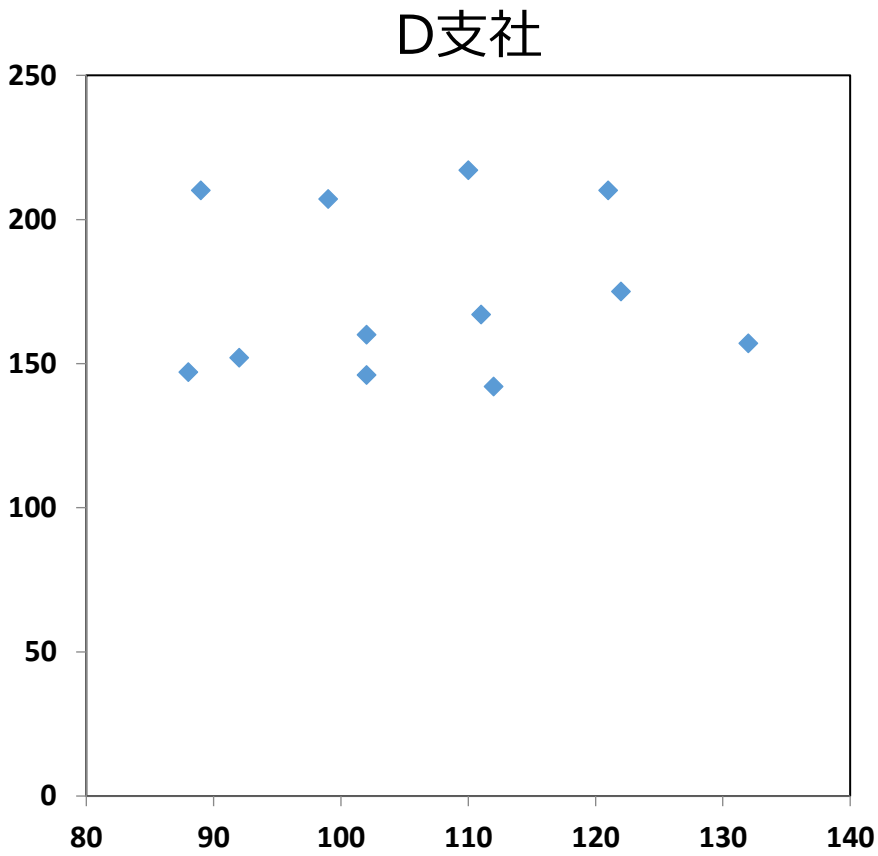


© Data Science Institute

C支社,D支社の広告費と売上高



r＝0.016



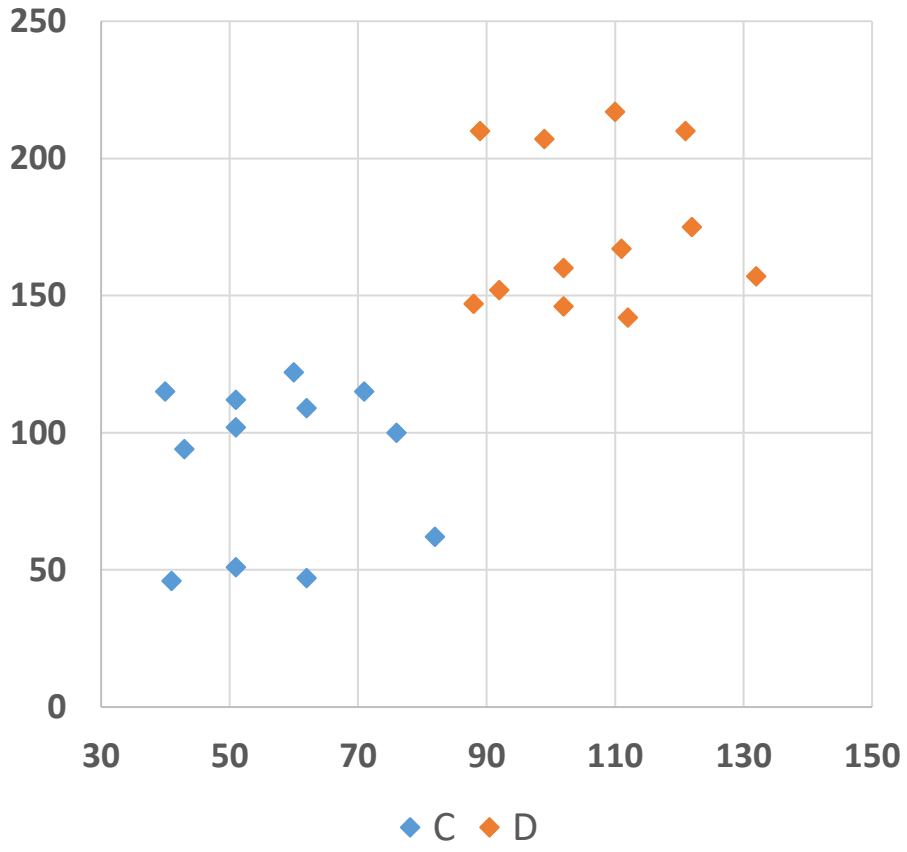
r ＝0.040

（C支社＋D支社）の相関係数は？

© Data Science Institute

（C支社＋D支社）の広告費と売上高

r ＝0.743



© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 複数列にわたる要因の相関係数の算出はCORRELよりも「データ分析」が便利

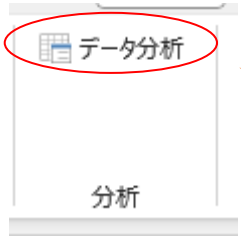


	手になじむ	デザイン性	通信性能	アプリ性能
モデル01	2	8	4	3
モデル02	9	5	10	10
モデル03	2	6	6	7
モデル04	4	4	2	4
モデル05	2	9	3	6
モデル06	4	5	5	4
モデル07	7	3	10	9
モデル08	8	2	2	4
モデル09	3	8	4	4





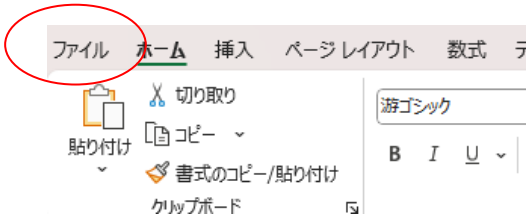
（もしデータ分析表示されてない人は次のスライド51と52を参照）



データ分析を選択

© Data Science Institute

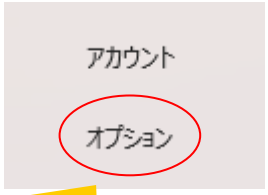
Excelデータ分析表示方法



左上のファイルをクリック



オプションをクリック

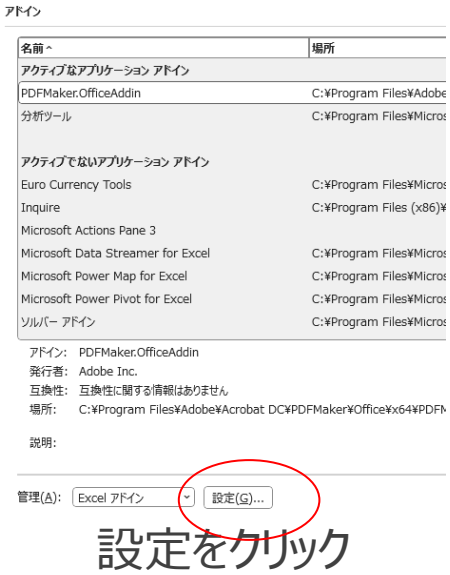




アドインをクリック

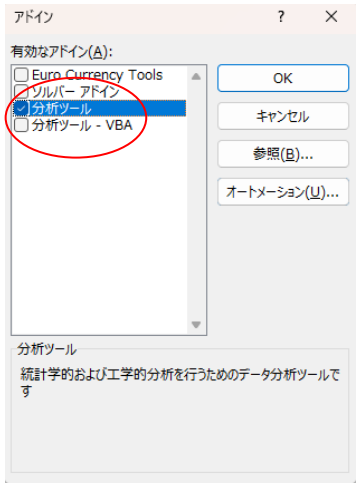
© Data Science Institute

Excelデータ分析表示方法

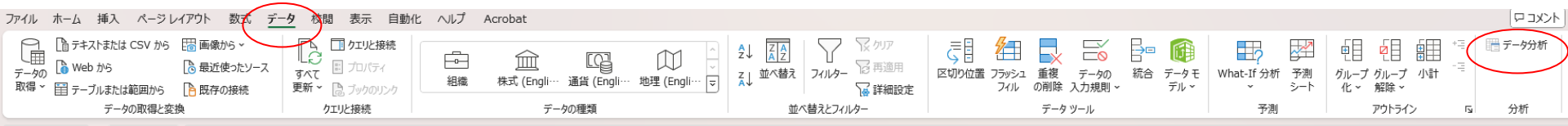


設定をクリック

分析ツールのチェックボックスをクリック



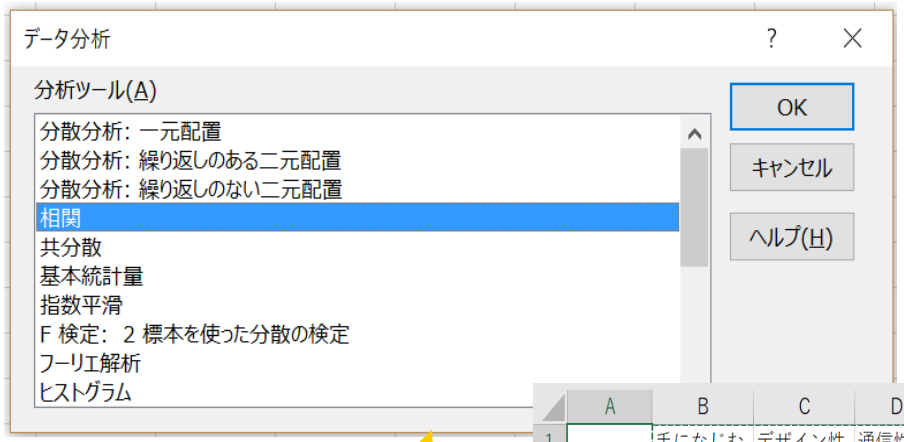
データタブに「データ分析」が表示されるようになる



© Data Science Institute

▪Excelでの可視化とデータ観察方法

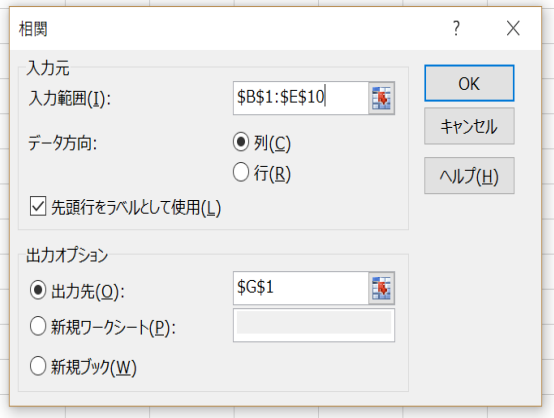
■ 複数列にわたる要因の相関係数の算出はCORRELよりも「データ分析」が便利



分析ツールを用いた相関行列の作成

- 分析ツール「相関」ではデータの総組合せで相関係数を計算する「相関行列」を作成する

	A	B	C	D	E
1		手になじむ	デザイン性	通信性能	アプリ性能
2	モデル01	2	8	4	3
3	モデル02	9	5	10	10
4	モデル03	2	6	6	7
5	モデル04	4	4	2	4
6	モデル05	2	9	3	6
7	モデル06	4	5	5	4
8	モデル07	7	3	10	9
9	モデル08	8	2	2	4
10	モデル09	3	8	4	4
11					
12					
13					



© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 相関は関係性を見るのにとても楽に算出できるが交絡要因に注意が必要



広告費を増やしても売り上げが増えるとは限らない

背後の第3の要因である人口（交絡要因）を検討することにより
正しい因果関係が得られる。

人口の影響を除いた広告費と売上高の関係を検討する必要

© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 交絡要因のあたりのつけ方

◇支店別広告費と売上高

	A	B	C	D	E
1	支店	広告費	売上高	人口	
2	北海道	92	44	5506	
3	東北	93	102	9335	
4	関東	332	288	42604	
5	北陸	78	54	5443	
6	中部	181	118	18127	
7	近畿	108	138	12912	
8	中国	113	140	15554	
9	四国	72	86	3976	
10	九州	243	152	13204	
11	沖縄	13	22	1393	
12	合計	1325	1144	128054	
13					
14					

広告費と売上高	r = 0.901
人口と広告費	r = 0.895
人口と売上高	r = 0.951

人口は交絡要因

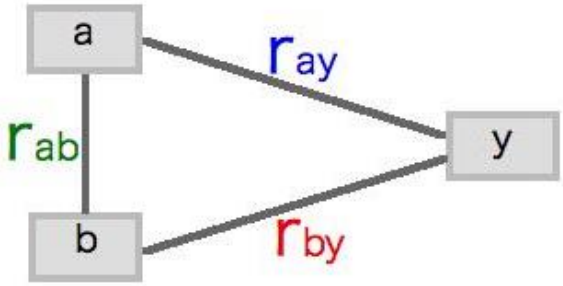
人口の影響を除いたときの広告費と売上高の相関係数

➡ 偏相関係数

© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 偏相関係数の算出は手計算



aの影響を除いた、bとyの偏相関係数

$$\frac{r_{by}-(r_{ay}\times r_{ab})}{\sqrt{1-r_{ay}^{2}}\times\sqrt{1-r_{ab}^{2}}}$$

$$\frac{0.901-(0.895\times0.951)}{\sqrt{1-0.895^{2}}\times\sqrt{1-0.951^{2}}}$$

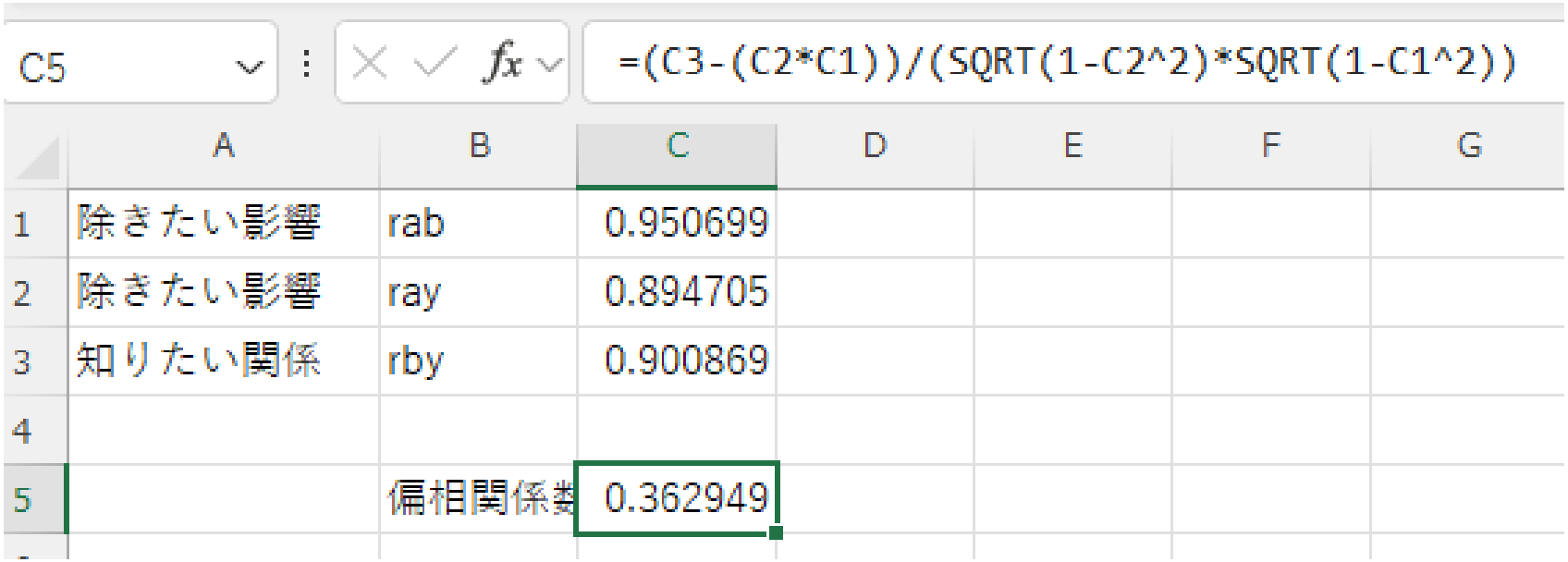
$$= 0.373$$

人口の影響を除いたときの広告費と売上高の
相関係数　：　0.373

© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 偏相関係数の算出は大変なので、先にExcelで作っておくと便利



	A	B	C	D	E	F	G
1	除きたい影響	rab	0.950699				
2	除きたい影響	ray	0.894705				
3	知りたい関係	rby	0.900869				
4							
5		偏相関係数	0.362949				

© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 満足度に影響を与える要因を調べる

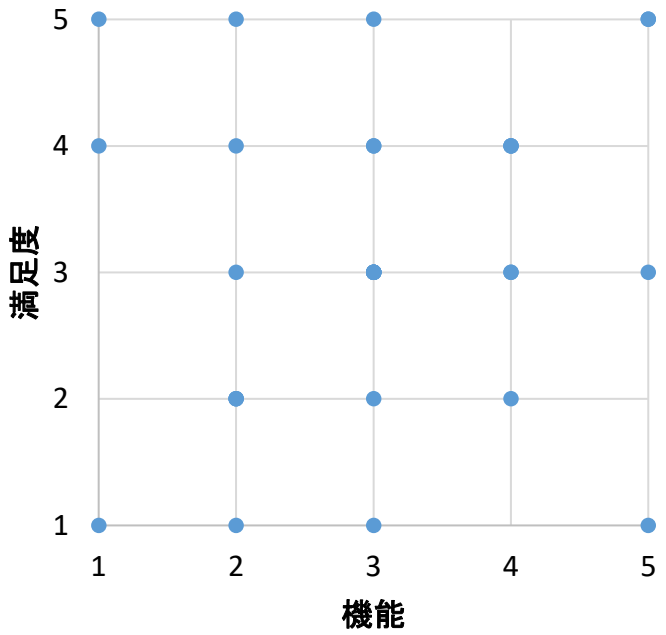
スマホの満足度調査

【調査項目】
- 満足度
- 機能
- デザイン
- 性別

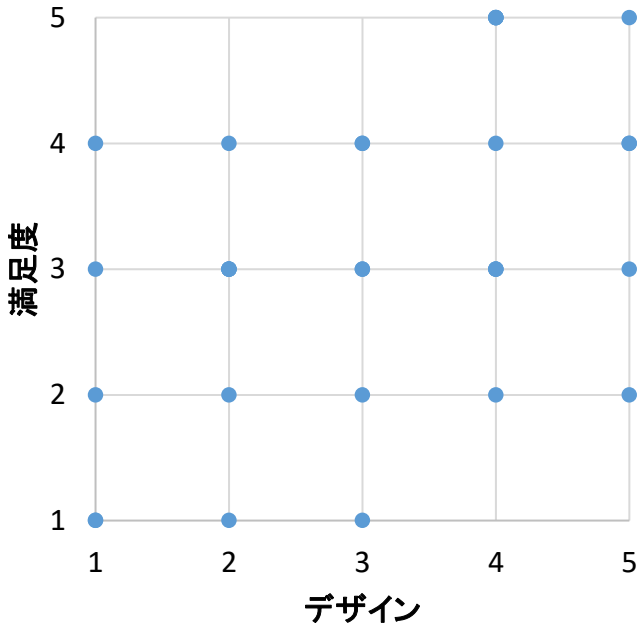
満足度	機能	デザイン	性別	満足度	機能	デザイン	性別
5	5	4	男性	5	3	5	女性
4	4	2	男性	5	2	4	女性
4	4	3	男性	5	5	4	女性
4	3	1	男性	5	1	4	女性
3	4	2	男性	4	4	5	女性
3	3	4	男性	4	2	4	女性
3	3	1	男性	4	1	3	女性
3	3	5	男性	4	3	5	女性
3	3	3	男性	3	2	4	女性
3	3	4	男性	3	5	2	女性
3	3	2	男性	3	3	3	女性
2	2	4	男性	3	4	2	女性
2	2	5	男性	2	3	3	女性
2	2	1	男性	2	4	2	女性
1	2	3	男性	1	5	2	女性
1	1	1	男性	1	3	1	女性

© Data Science Institute

機能と満足度の関係



デザインと満足度の関係



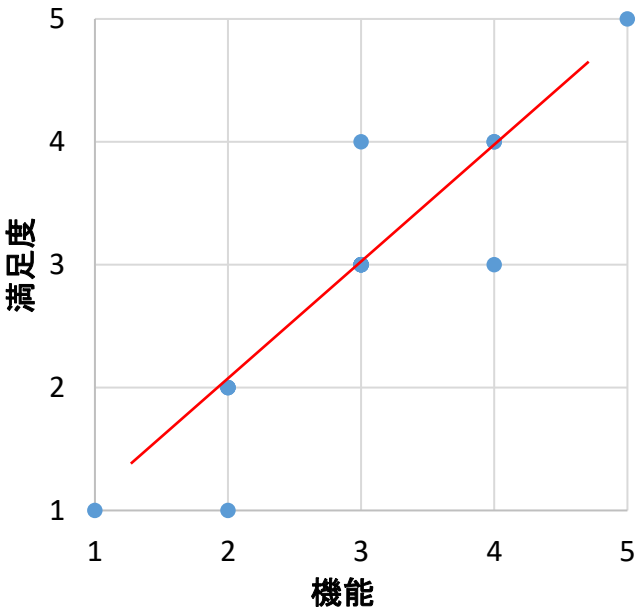
機能、デザインは満足度と全く関係ない!?

© Data Science Institute

交互作用の検討

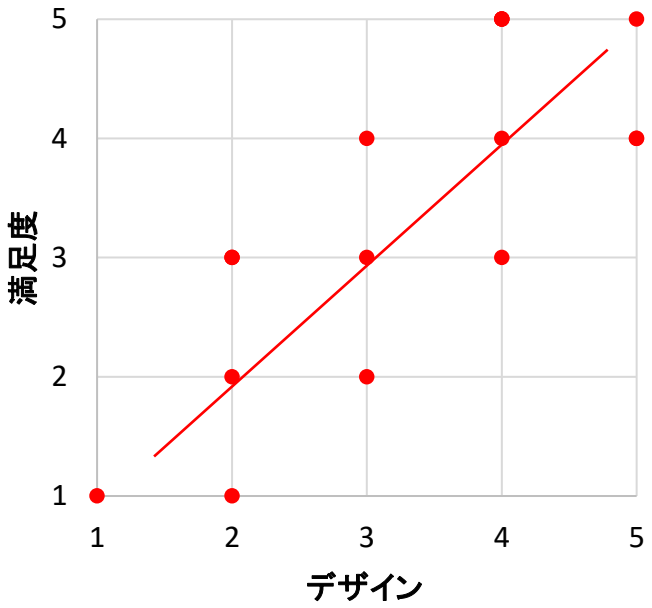
性別に分析した場合

男性



男性は機能を重視

女性



女性はデザインを重視

性別によって影響度合いが異なる（交互作用）
交互作用に注目することで有用な結果が得られる

© Data Science Institute

相関の発見と利用例　Association Rule

{インスタントラーメン} ⇒ {チャーシュー}

インスタントラーメン売り場にチャーシューを並べると
同時購入が増えると予想される。

{牛乳} ⇒ {パン}

牛乳の特売をすると牛乳の販売数が増え、パンの販売数も
増えると予想されるのでパンの仕入を増やす。

© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 回帰分析を知ることで、売上高と売上高に影響を与える要因との関係が調べられる

売上高⇐広告宣伝費、人口、セールスパーソン、・・・

従属変数 ⇐ 説明変数（独立変数）

説明変数が1つ：単回帰
説明変数が2つ以上：重回帰

目的1：回帰式を求め、予測する。

回帰式（ｙ＝a+b ｘ +…）を求める。

a,b,・・・　　偏回帰係数

© Data Science Institute

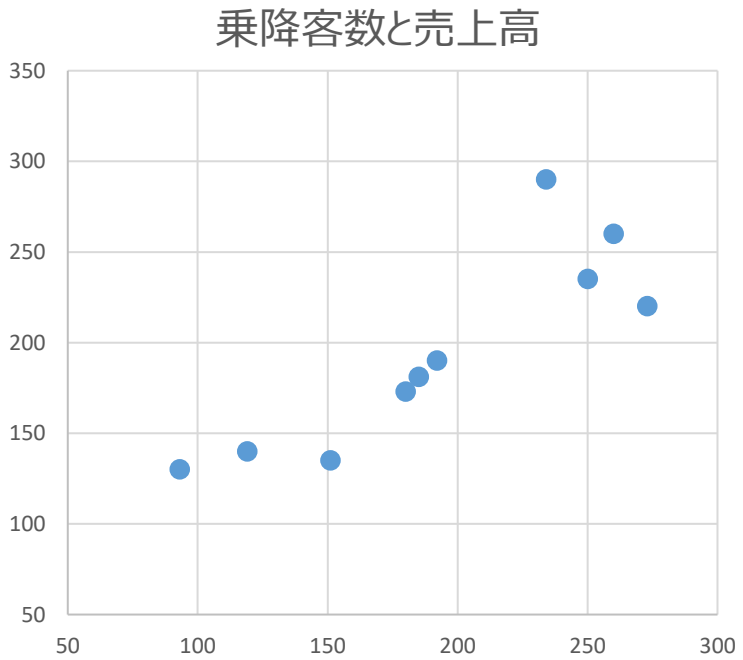
▪Excelでの可視化とデータ観察方法

■ 回帰分析によって、今後を予測することができる

◇駅前コンビニの売上高と乗降客数

売上高⇐乗降客数

	売上高（百万円/月）	乗降客数（百人/日）
1	130	93
2	290	234
3	235	250
4	260	260
5	140	119
6	173	180
7	135	151
8	190	192
9	220	273
10	181	185

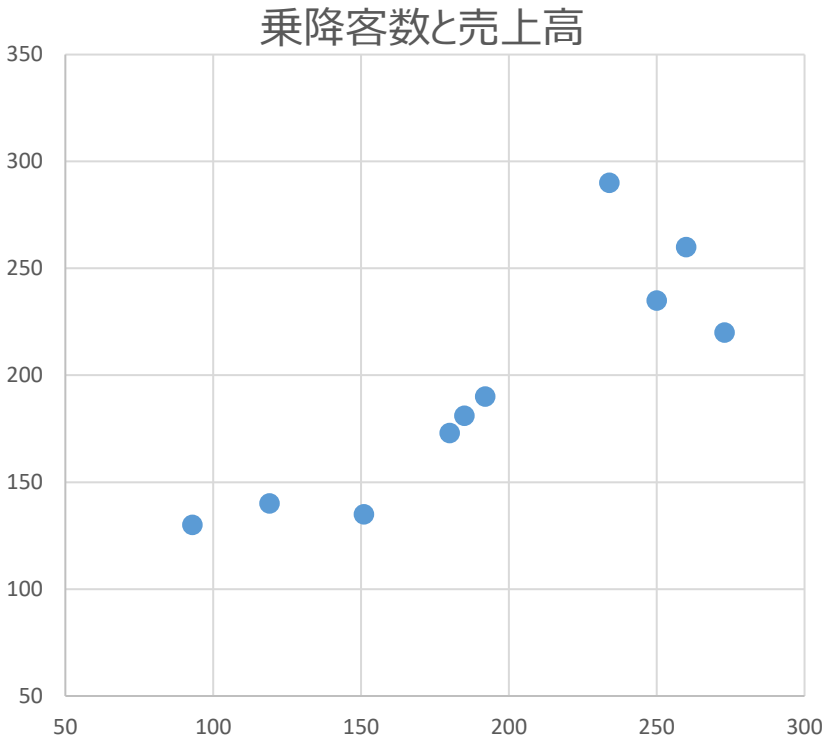


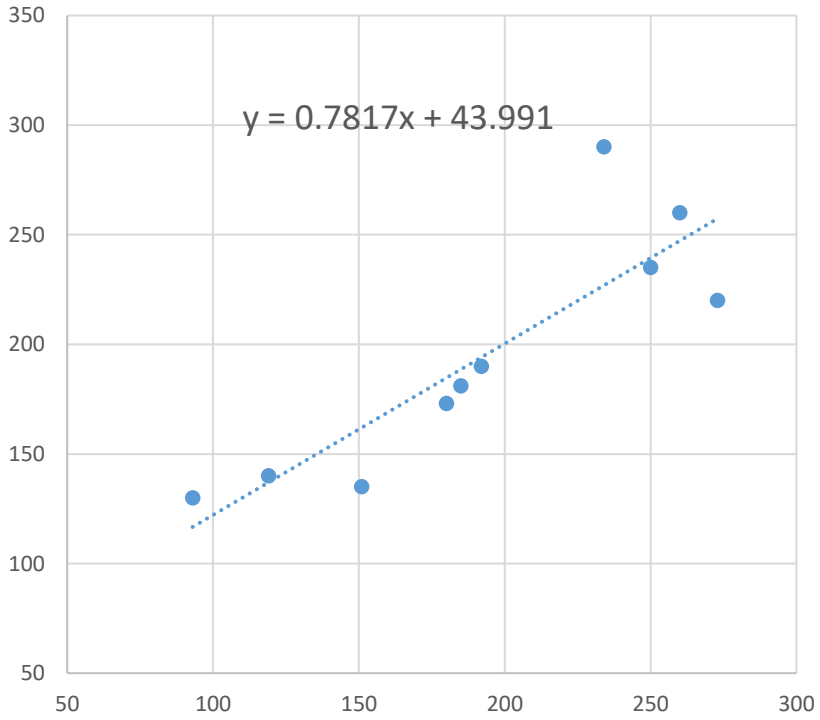
r ＝0.867

© Data Science Institute

▪Excelでの可視化とデータ観察方法

■ 回帰分析によって、今後を予測することができる



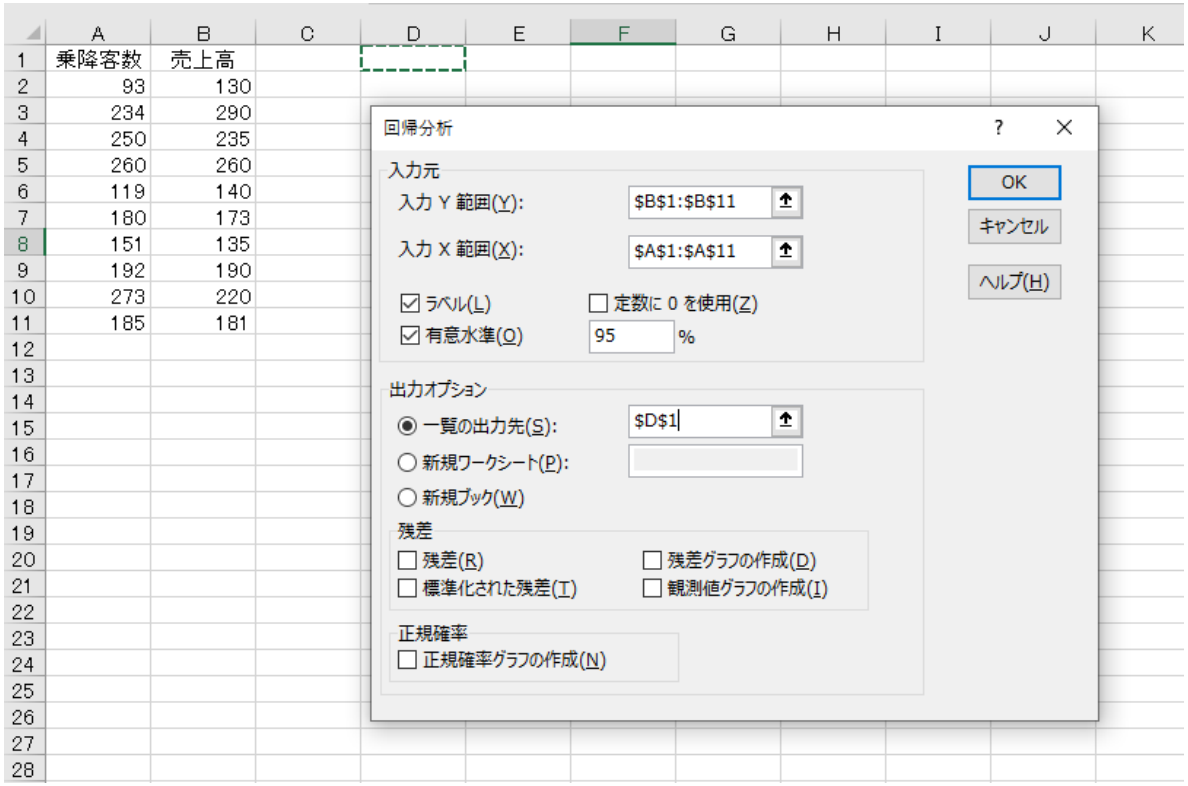


回帰式（ｙ＝a+b ｘ）を求める

今後、乗降客数（ｘ）に値を入れることで、予測売上高（ｙ）が求められるようになる

© Data Science Institute

1）分析ツールメニューから「回帰分析」を選択し、「OK」をクリックする。
2）従属変数、説明変数の範囲を入力する。
3）「ラベル」をチェック、一覧の出力先を指定し「OK」をクリックする。



	A	B
1	乗降客数	売上高
2	93	130
3	234	290
4	250	235
5	260	260
6	119	140
7	180	173
8	151	135
9	192	190
10	273	220
11	185	181

© Data Science Institute

回帰統計	
重相関 R	0.8675
重決定 R2	0.7525
補正 R2	0.7216
標準誤差	28.916
観測数	10

決定係数の大きさ（補正R2）
自由度調整済み決定係数＝0.722

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	43.992	32.03	1.373	0.2069	-29.88	117.9
乗降客数	0.7817	0.159	4.932	0.0011	0.416	1.147

回帰式：売上高＝43.99＋0.782×乗降客数
⇒　回帰式により約72.2%説明できる

© Data Science Institute

演習1

データの理解

演習内容	DM発送数と売上数量との関係を図にしてみましょう。 送り先の性別データがありますので、それぞれで調べてみてください。 もし、コストカットするならば、どのようにしたら良いでしょう
データ	演習データ.xlsx
ツール	Excel
時間	3分

© Data Science Institute