



【情報Ⅱ】 第4回全国指導力向上研修会
”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”のご紹介
重回帰分析の実務応用事例

一般社団法人 データサイエンティスト協会 学生委員会

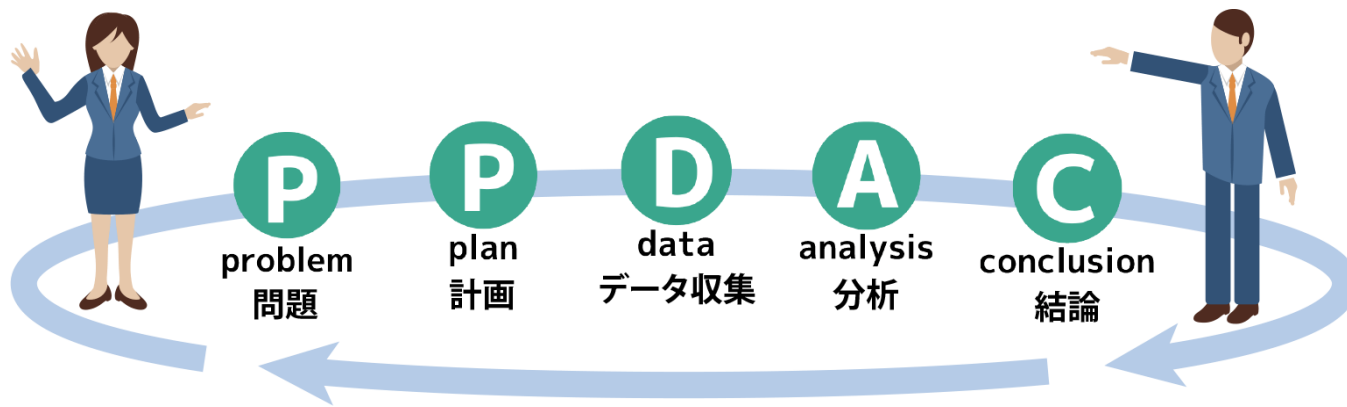
2024年10月29日

1. データサイエンスの実社会での使われ方 | PPDACサイクル
2. Problem | 問題の設定
3. Plan
 1. As-IsとTo-Beの設定、分析方針策定
 2. 要因分析
4. Data | データの取得
5. Analysis
 1. データ可視化
 2. 予測モデル構築
6. Conclusion | 結論
7. Appendix

1. 実社会での使われ方 | PPDACサイクル

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は、
教育現場でも使用されるPPDACサイクルに近いプロセスを進めることが多いため、
今回はPPDACサイクルに則った小売業（コンビニ）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

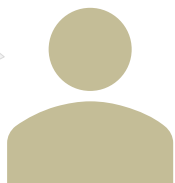
<https://www.datascientist.or.jp/news/news/post-2890/>

2. Problem | 問題の設定

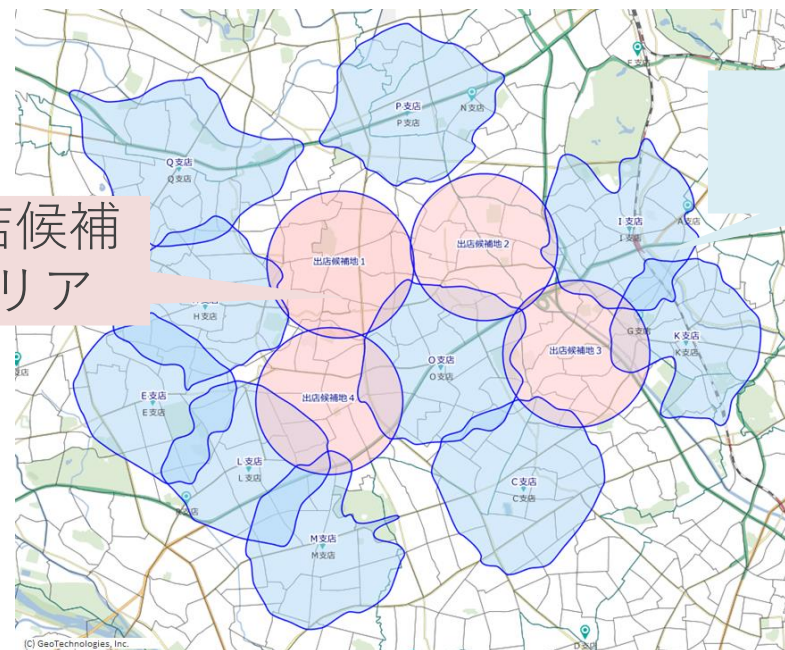
- 問題：**現在の新規店舗出店計画ではデータが活用されておらず、経験や勘に頼っている。正確な売上予測が難しく、出店前売上予測値と実績値の乖離や撤退コストの問題が生じている。

経験的にこのエリアだと、一定の売上が見込めそうですね。

ベテラン担当



出店候補
エリア



既出店
エリア

出店判断
責任者



経験は信じていますが、前回店舗では予測と実績誤差が大きくなっていました。根拠をもう少し具体的に教えてください。

新人担当

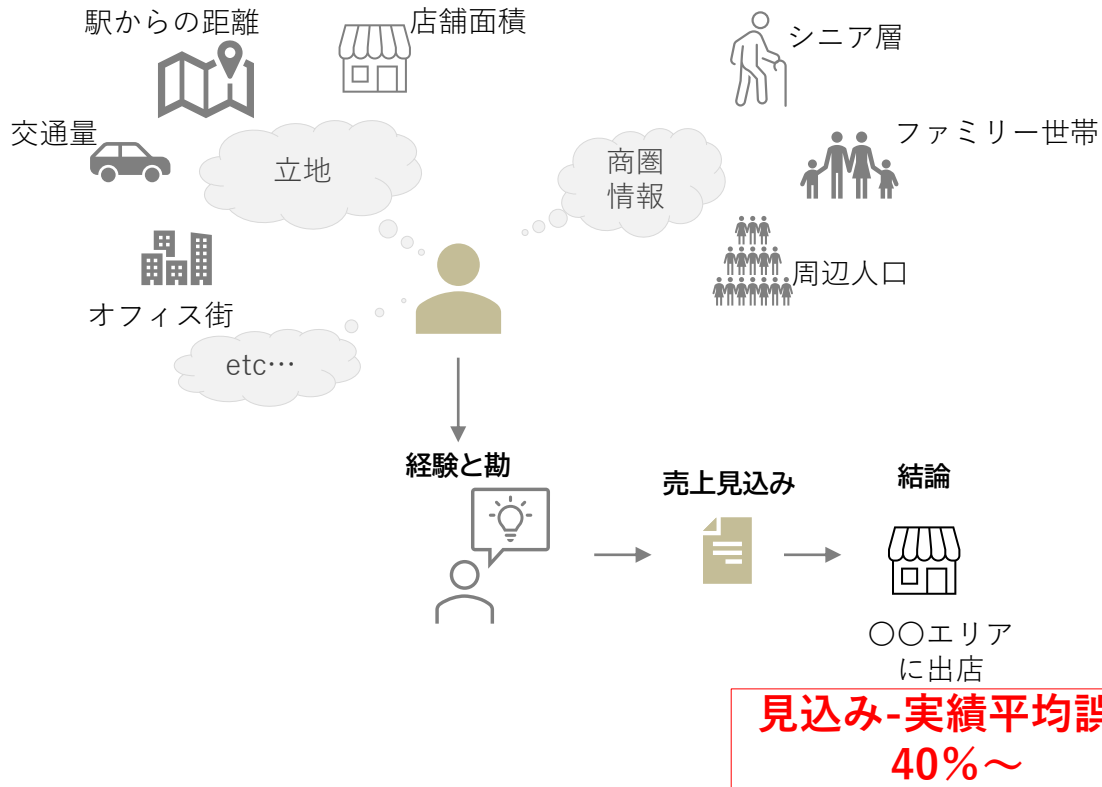


先輩が転職したら困るので、具体的にどのような要因が売上に影響するのかデータで分析しませんか？

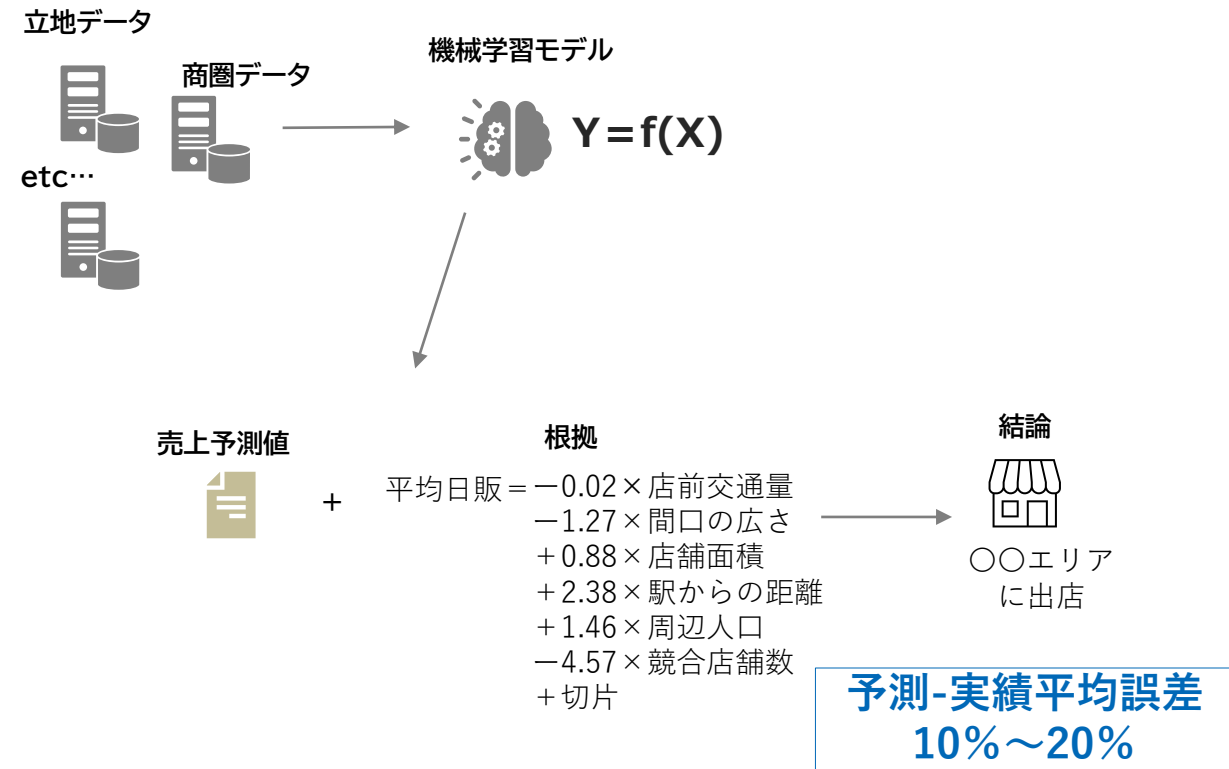
3-1. Plan | As-IsとTo-Beの設定、分析方針の策定

- **As-Is**：経験や勘に頼り正確な売上予測が難しく、出店前の予測値と実績値の乖離や撤退コストの問題が発生
- **To-Be**：データに基づいた店舗の売上予測で予測精度を向上、売上に影響する要因を把握し、出店計画を改善

As-Is（現状）

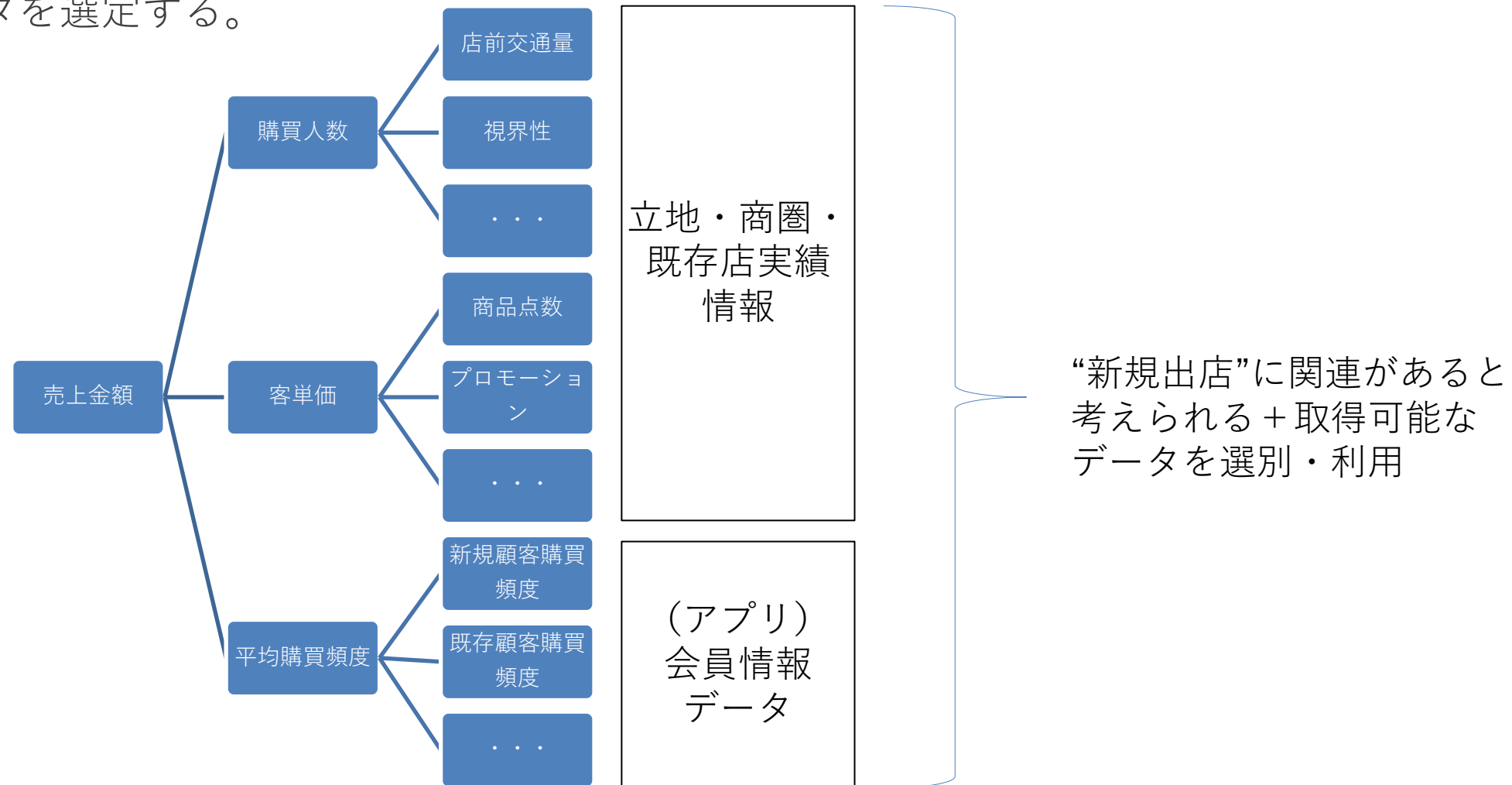


To-Be（あるべき姿）



3-2. Plan | 要因分析

- 売上 に寄与する要因を分解して洗い出す。そのうち、新規出店の候補地の情報として、取得可能なデータを選定する。



4. Data | データの収集

- 新規出店候補地の情報として調査可能な情報と既存店舗の平均日販（売上）を含む、以下のようなデータを利用。
- 既存店舗の取扱商品数やプロモーション、会員情報データなども売上に効く要因だが、新規出店候補地の情報としては取得できないため、今回はデータからは除外。

店舗番号	説明変数								目的変数	
	店前交通量(台/h)	視界性評価 (1:見通せない~ 10:完全に開けている)	間口の広さ	店舗面積(坪)	店長の年齢(歳)	駅からの距離(分)	駐車場台数(台)	周辺人口(千人/km ²)	競合店舗数(店)	平均日販(万円)
1	171.2	3	14	83.98	58	10	5	11	2	107.6
2	165.5	10	14.4	78.53	37	7	4	9.3	2	107
3	169	1	13.4	82.47	51	8	8	6.6	2	106.6
4	168.6	1	13.4	76.86	32	9	8	6.6	2	103.4
5	165.8	3	14.2	76.45	43	7	5	9.1	2	102

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5. 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する。

全体

演習① データ全体の俯瞰

- 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。

演習② 目的変数の分布の確認

- 平均日販の度数分布を確認し、全店舗での平均的な売り上げを確認しよう。

演習③ 目的変数と説明変数の関係の確認

- 平均日販と各説明変数にはどのような関係があるか？平均日販をy軸、各説明変数をx軸とした散布図と、相関係数で確認しよう。

演習④ モデル作成と売上に影響を与える要因の把握

- 重回帰モデルを作成し、各説明変数がどの程度平均日販に影響を与えているか確認しよう。

演習Advanced 出店候補地の検討とモデルの精度確認

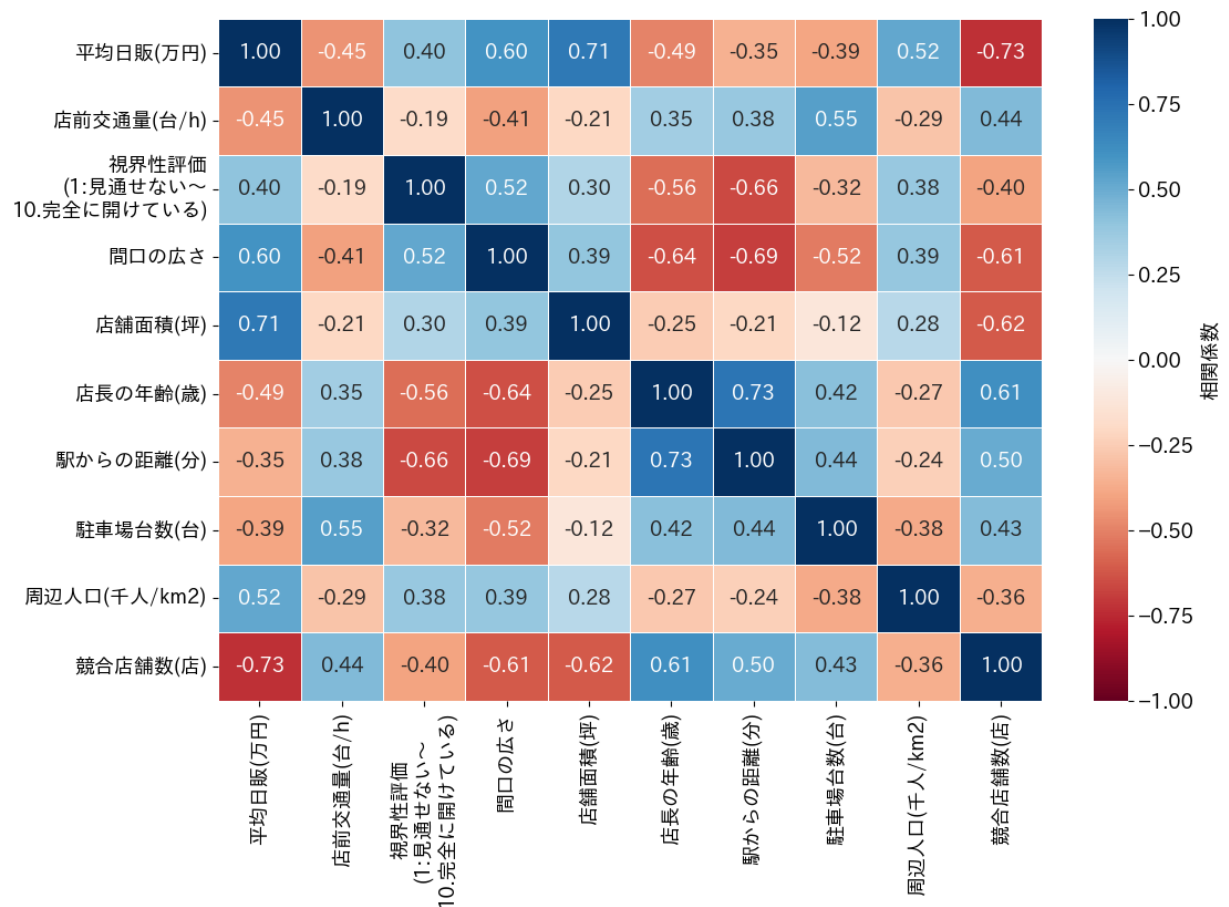
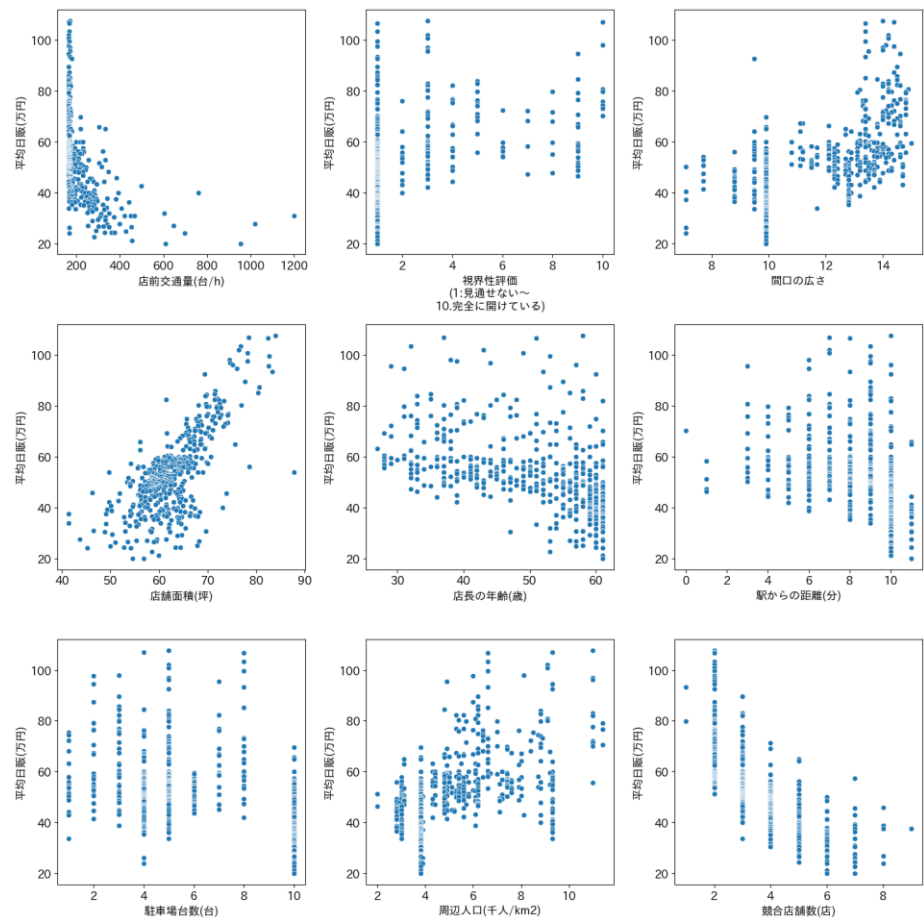
- 新規出店候補地（テストデータ）に対して、予測モデルから平均日販の予測を行い、どの候補地に出店するのがよいか、を考えよう。
- 新規出店候補地（テストデータ）の実績値と作成したモデルの予測値の誤差を確認し、モデルの精度を確認しよう。

詳細



5-1. Analysis | データ可視化

- 平均日販と各データの関係性を散布図・相関係数で確認。目的変数に影響が無さそうな説明変数は除外を検討。



※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5-2. Analysis | 予測モデル構築・要因探索

- 既存店の情報を元に重回帰モデルを構築。特に売り上げに影響を与えている要因を把握。

決定係数は0.724で、ある程度モデルが適合していると判断できる

OLS Regression Results

Dep. Variable:平均日販(万円)R-squared:0.724

Model:OLSAdj. R-squared:0.717

Method:Least SquaresF-statistic:109.6

Date:Tue, 22 Oct 2024Prob (F-statistic):1.25e-88

Time:11:33:28Log-Likelihood:-1191.4

No. Observations:343AIC:2401.

Df Residuals:334BIC:2435.

Df Model:8

Covariance Type:nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	52.8437	0.427	123.751	0.000	52.004	53.684
x1	-2.5334	0.544	-4.655	0.000	-3.616	-1.463
x2	0.7306	0.603	1.211	0.227	-0.456	1.918
x3	3.1456	0.702	4.478	0.000	1.764	4.527
x4	5.6073	0.574	9.767	0.000	4.478	6.737
x5	1.9898	0.712	2.795	0.005	0.589	3.390
x6	-0.2527	0.583	-0.434	0.665	-1.399	0.893
x7	2.7276	0.500	5.459	0.000	1.745	3.711
x8	-4.0382	0.689	-5.859	0.000	-5.394	-2.682

係数でソートして可視化

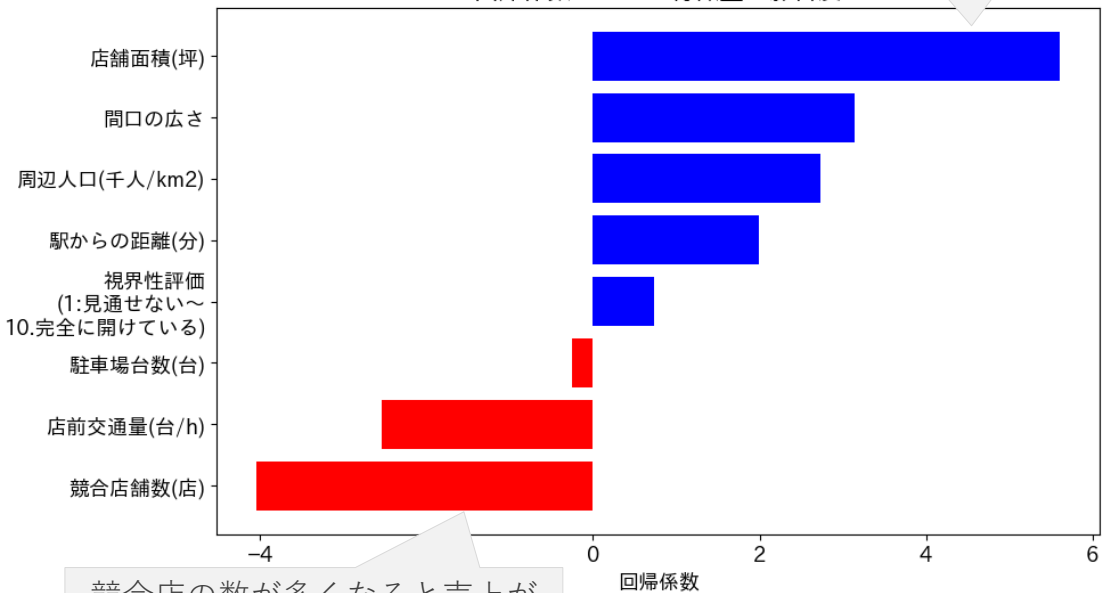
Omnibus:26.901Durbin-Watson:1.892

...

係数でソートして
可視化

店舗面積が大きくなると売上が
上がる方向に影響する

回帰係数に基づく特徴量の影響度



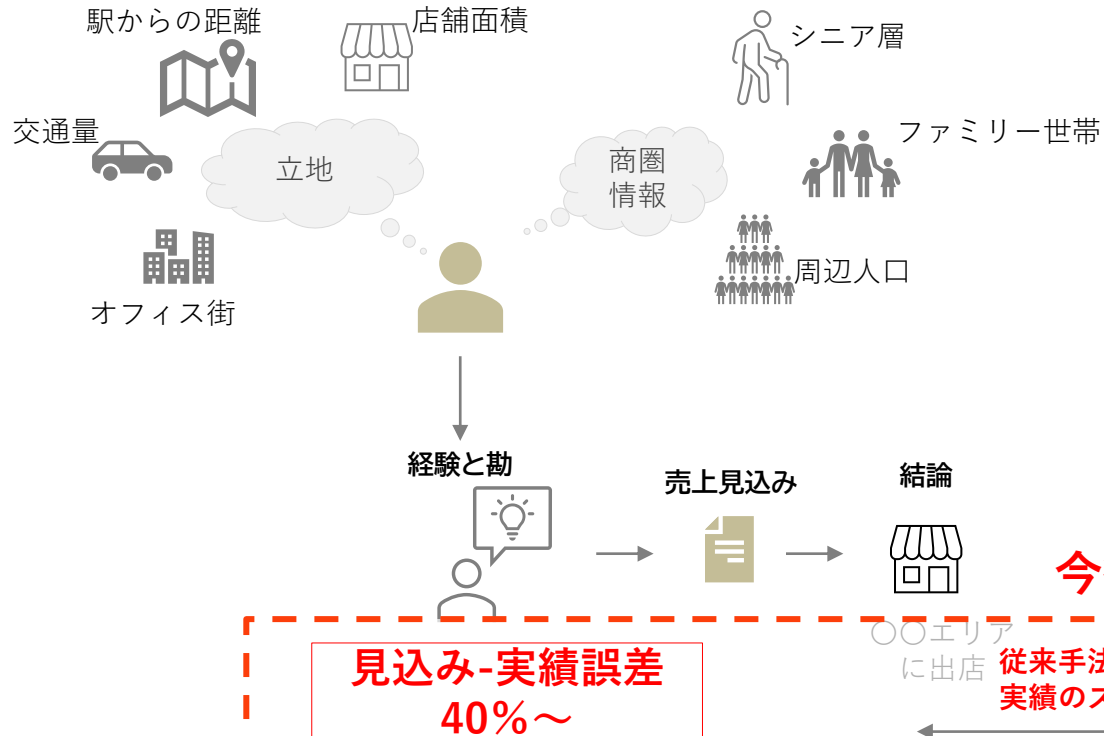
競合店の数が多くなると売上が
下がる方向に影響する

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

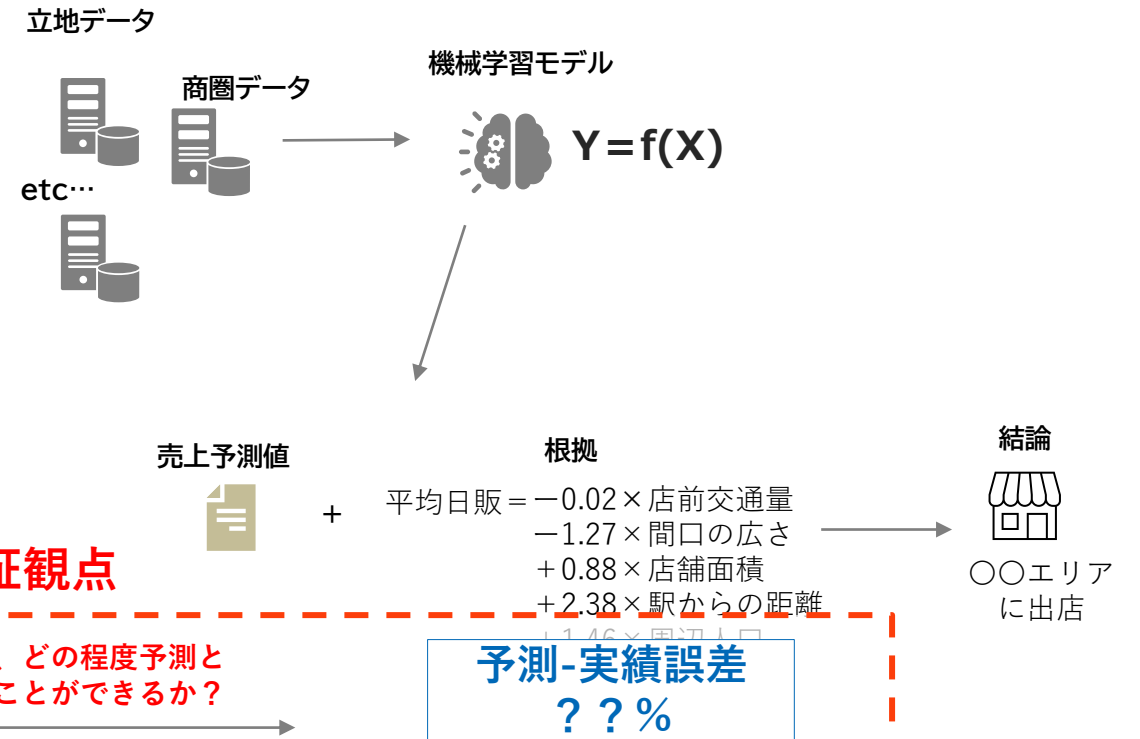
6. Conclusion | 結論

- データに基づく平均日販予測モデルを作成。売上が見込める候補地を根拠を持って選定する見込みが立った。
- 今後は、モデルのブラッシュアップや従来の選定方法とデータに基づいた手法による予測精度差などを検証し、導入提案。

As-Is (現状)



To-Be (あるべき姿)



【情報Ⅱ】第4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。

7. Appendix



演習①

- データは何行×何列あるか確認しよう。

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

```
(490, 11)
```

490行×11列

490店舗分のデータがあることがわかる

- 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)
```

Python

	店舗番号	店前交通量 (台/h)	視界性評価\n(1:見通せない〜\n10.完全に開けている)	間口の広さ	店舗面積 (坪)	店長の年齢 (歳)	駅からの距離 (分)	駐車場台数 (台)	周辺人口(千人/km2)	競合店舗数 (店)	平均日販(万円)
0	1	171.2	3	14.0	83.98	58	10	5	11.0	2	107.6
1	2	165.5	10	14.4	78.53	37	7	4	9.3	2	107.0
2	3	169.0	1	13.4	82.47	51	8	8	6.6	2	106.6
3	4	168.6	1	13.4	76.86	32	9	8	6.6	2	103.4
4	5	165.8	3	14.2	76.45	43	7	5	9.1	2	102.0

演習①

- ・ 欠損値が無いかを確認しよう。
- ・ 要約統計量を確認しよう。

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

店舗番号	0
店前交通量(台/h)	0
視界性評価\n(1:見通せない〜\n10.完全に開けている)	0
間口の広さ	0
店舗面積(坪)	0
店長の年齢(歳)	0
駅からの距離(分)	0
駐車場台数(台)	0
周辺人口(千人/km2)	0
競合店舗数(店)	0
平均日販(万円)	0

dtype: int64

各列で欠損値は無し。

df.describe()

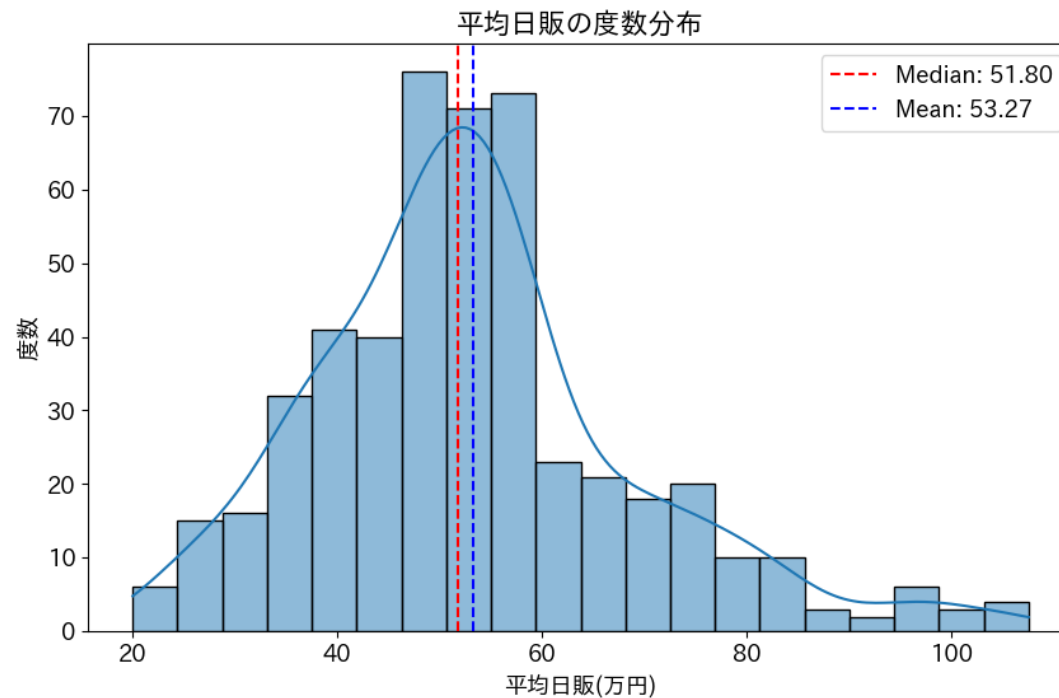
Python

	店舗番号	店前交通量 (台/h)	視界性評価\n(1:見通せない〜\n10.完全に開 けている)	間口の広さ	店舗面積 (坪)	店長の年齢 (歳)	駅からの距離 (分)	駐車場台数 (台)	周辺人口(千 人/km2)	競合店舗数 (店)	平均日販(万 円)
count	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000	490.000000
mean	245.500000	207.481224	2.053061	11.956531	62.617980	49.895918	8.185714	5.885714	5.480000	3.600000	53.271837
std	141.59508	101.402470	2.211986	1.999327	6.394366	10.010801	2.106616	2.812330	2.110478	1.446807	15.730602
min	1.000000	165.200000	1.000000	7.100000	41.380000	27.000000	0.000000	1.000000	2.000000	1.000000	20.000000
25%	123.250000	166.100000	1.000000	9.900000	58.855000	41.250000	7.000000	4.000000	3.800000	2.000000	43.400000
50%	245.500000	168.000000	1.000000	12.400000	61.930000	53.000000	9.000000	5.000000	4.900000	3.000000	51.800000
75%	367.750000	207.500000	2.000000	13.700000	65.922500	59.000000	10.000000	10.000000	6.600000	4.000000	59.350000
max	490.000000	1199.500000	10.000000	15.000000	87.800000	61.000000	11.000000	10.000000	11.400000	9.000000	107.600000

min,maxで明らかにおかしい値が無い事を確認。

演習②

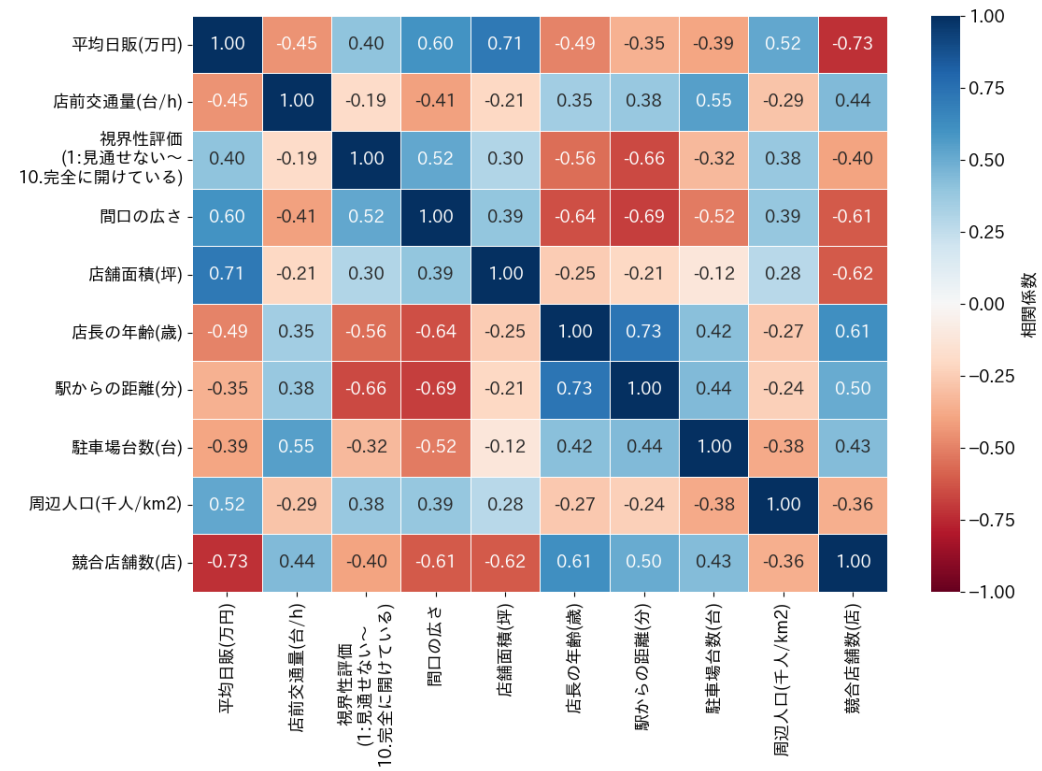
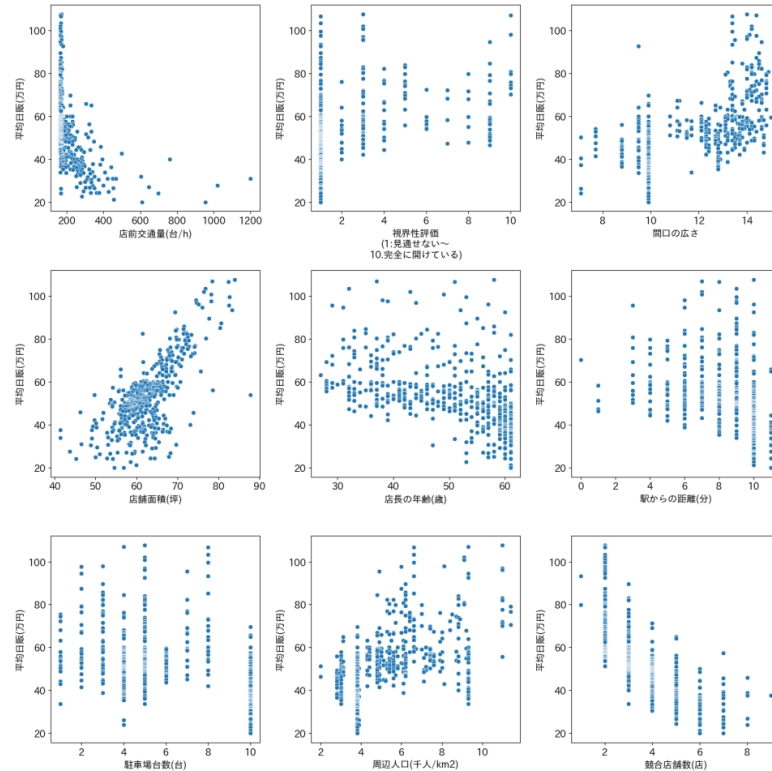
- 平均日販の分布はどのようなになっているか？ヒストグラムで確認しよう。



やや右に裾野が長いデータで、平均値が右側に引っ張られているが、中央値と大きく乖離してはならず、外れ値は含まれていない。

演習③

- 目的変数と各説明変数の関係を散布図と相関係数で確認しよう。



平均日販は、店舗面積、間口の広さ、との間にの正の相関があり、重要な変数と言えそうだ。

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

演習④

- 重回帰モデルを作成し、各説明変数がどの程度平均日販に影響を与えているか確認しよう。

OLS Regression Results

Dep. Variable:平均日販(万円)

R-squared:0.724

Model:OLS

Adj. R-squared:0.717

Method:Least Squares

F-statistic:109.6

Date:Tue, 22 Oct 2024

Prob (F-statistic):1.25e-88

Time:11:33:28

Log-Likelihood:-1191.4

No. Observations:343

AIC:2401.

Df Residuals:334

BIC:2435.

Df Model:8

Covariance Type:nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	52.8437	0.427	123.751	0.000	52.004	53.684
x1	-2.5334	0.544	-4.655	0.000	-3.621	-1.463
x2	0.7306	0.603	1.211	0.227	-0.456	1.918
x3	3.1456	0.702	4.478	0.000	1.764	4.527
x4	5.6073	0.574	9.767	0.000	4.478	6.737
x5	1.9898	0.712	2.795	0.005	0.589	3.390
x6	-0.2527	0.583	-0.434	0.665	-1.399	0.893
x7	2.7276	0.500	5.459	0.000	1.745	3.711
x8	-4.0382	0.689	-5.859	0.000	-5.394	-2.682

係数でソートして可視化

Omnibus:26.901

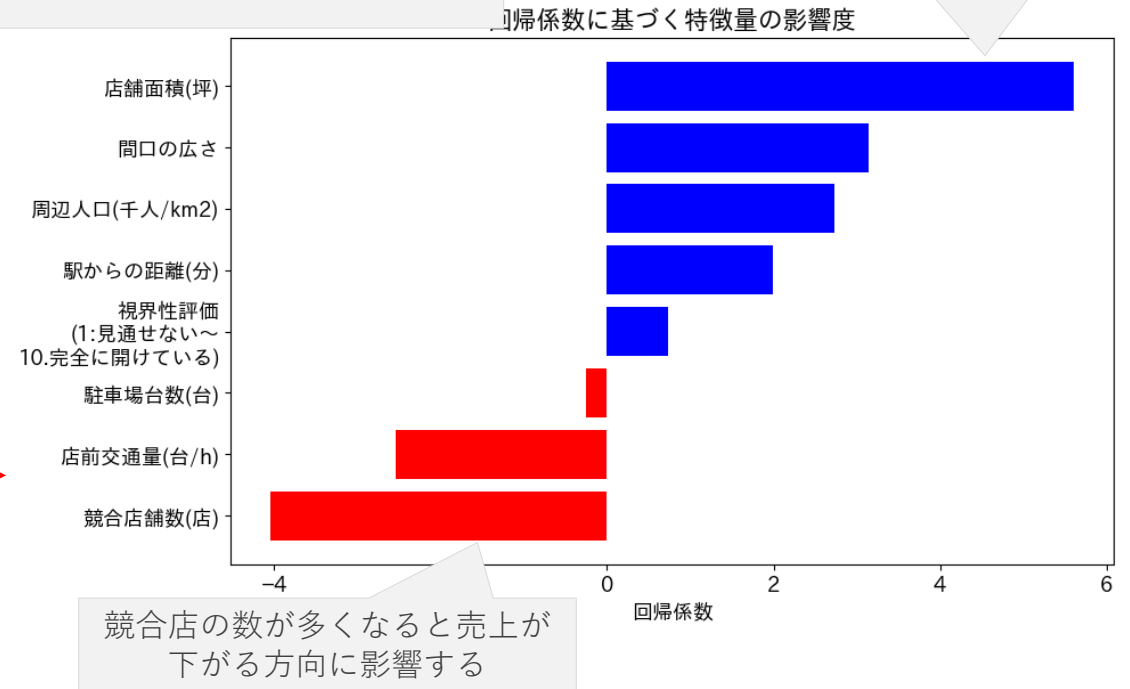
Durbin-Watson:1.892

...

係数でソートして
可視化

決定係数は0.724である程度はモデルが適合していると判断できる

店舗面積が大きくなると売上が上がる方向に影響する



※あくまでダミーデータのため、現実の傾向を反映するものではありません。

演習Advanced

・新規出店候補地に対して、予測モデルから平均日販の予測を行い、どの候補地に出店するのがよいか、候補を考えよう。

```
# 平均日販予測値が大きい順にデータフレームをソートして表示
X_test_df.sort_values('平均日販予測値 (万円)', ascending=False)
```

Python

	店前交通量 (台/h)	視界性評価\n(1:見通せない〜\n10:完全に開けている)	間口の広さ	店舗面積 (坪)	駅からの距離 (分)	駐車場台数 (台)	周辺人口(千人/km2)	競合店舗数 (店)	店舗番号	平均日販予測値 (万円)
54	165.5	10	14.4	78.53	7	4	9.3	2	2	83.145942
47	171.3	3	14.0	75.20	10	5	11.0	2	11	82.013524
35	165.4	9	14.6	76.10	6	2	9.3	2	13	80.247420
90	168.8	1	13.4	82.66	9	8	6.6	2	7	79.980884
137	165.5	3	14.2	78.20	7	5	9.1	2	6	79.785754
...	...	...	...	...	...	...	...	...	...	...
81	331.8	1	9.9	48.80	10	10	3.8	7	465	23.556023
26	427.8	1	9.9	50.00	10	10	3.8	7	481	22.263321
9	448.7	1	9.9	46.52	11	10	3.8	7	460	19.667600
129	954.8	1	9.9	56.83	11	10	3.8	6	490	19.201639
53	383.8	1	9.9	46.28	10	10	3.8	8	343	17.216061

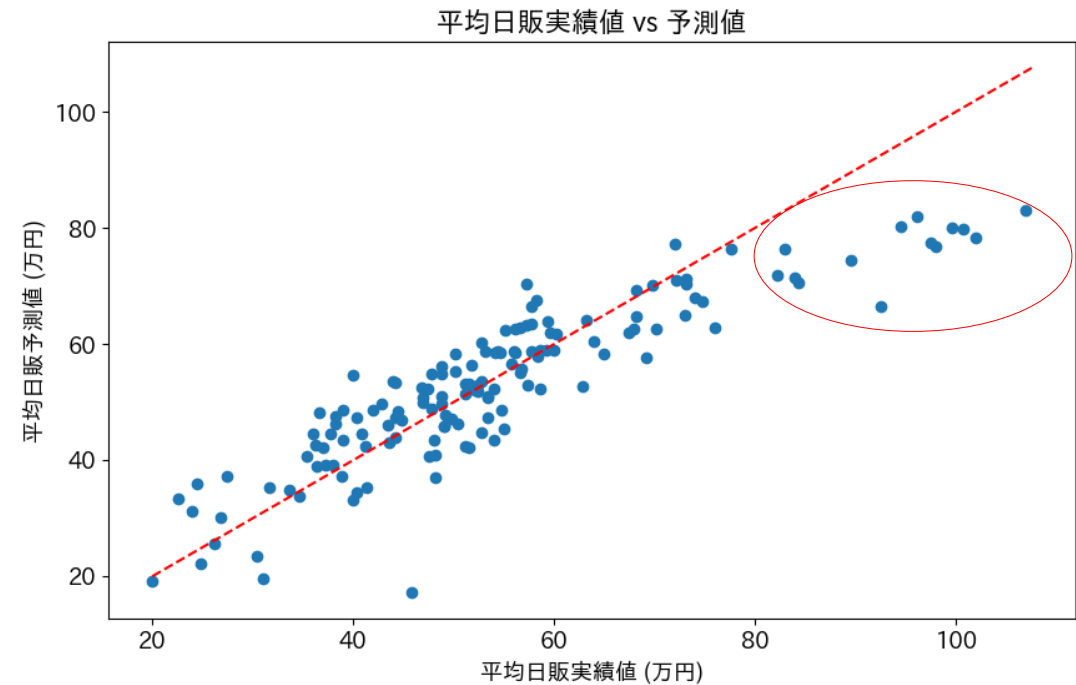
店舗番号2,11,13の店舗では、平均日販予測値が80万円以上と、全店の平均と比較して高い売り上げが見込めそうだ。

演習Advanced

- ・新規出店候補地（テストデータ）の実績値と作成したモデルの予測値の誤差を確認し、モデルの精度を確認しよう。

```
# 平均予測誤差の算出
print("平均予測誤差：" + str((abs(y_test - y_pred)/y_test).mean().round(4)*100) + "%")
```

平均予測誤差：11.76%



- ・平均予測誤差は11.76%とまずまずな精度となっていそうだ。
- ・平均日販の実績値と予測値の散布図を見ると、実績値が80万円以上の店舗で予測値が低めに出ている。その原因は何だろうか？データの追加や説明変数の入れ替えで、精度が向上できるだろうか？