AI 基礎プログラミング
第 2 回　効果測定　問題

問題 1　配点 18 点

以下の説明に当てはまる用語を解答群から選択して記号を答えなさい。

データの分布の形状も考慮しながらロジカルに外れ値を除去するために、【①】を用いることがある。
また、1 次元データの外れ値を判定する閾値を決める方法として、データの第 3 四分位数と第 1 四分位数の差分である【②】を用いることがある。このとき、値が大きい外れ値の閾値は、「第 3 四分位数【③】1.5 × 【②】」のように決めることができる。

訓練データから復元抽出によってデータをランダムに抽出し、訓練用のデータセットをいくつか作り、それぞれをモデルに学習させる手法を【④】と呼ぶ。
【④】で学習させるモデルをすべて決定木とし、さらに特徴量の列もランダムに選択するのが【⑤】である。

複数のモデルに順番に学習させ、モデルの学習結果を次に学習するモデルと共有する手法を【⑥】と呼ぶ。
【⑥】の手法の1つであるアダブーストでは、データの法則性を学習するうえで難しい訓練データの情報を共有する。

解答群

ア　四分位範囲	イ　勾配ブースティング	ウ　ブースティング
エ　×	オ　＋	カ　－
キ　バギング	ク　ブートストラップサンプリング	ケ　復元抽出
コ　ランダムフォレスト	サ　逐次学習	シ　ロジスティック回帰
ス　直線距離	セ　ユークリッド距離	ソ　マハラノビス距離

AI 基礎プログラミング
第 2 回　効果測定　問題

問題 2　配点 12 点

踏切の異常を検知するモデルを作成し、予測と実際の件数を調べたところ、下記の結果であった。
以下の問いに答えなさい。(数値を答える際、割り切れないものは小数第 3 位を四捨五入する)

		予測	
		正常	異常
実際	正常	30	10
	異常	20	40

(1) 正解率を計算せよ。
(2) 正常の適合率(正常と予測したところ、実際に正常であった)を計算せよ。
(3) 正常の再現率(実際が正常なとき、予測も正常であった)を計算せよ。
(4) このシステムでは踏切の異常を検出したいということを考慮すると、異常であることの適合率と再現率のどちらを重視すべきか答えよ。(解答欄には「適合率」または「再現率」を記載する)

AI 基礎プログラミング
第 2 回　効果測定　問題

問題 3　配点 40 点

配布データ「Boston2_kouka2_1.tsv」「Boston2_kouka2_2.tsv」を用い、「価格」列のデータを予測するプログラムを作成する。

下記の機能要求を満たすようにプログラムを実装してください。

① 2 つの TSV ファイルを読み込み、読み込んだデータフレームの先頭 5 行をそれぞれ表示する。
② 2 つの TSV ファイルを縦方向に連結し、連結後のデータフレームの行数、列数を表示する。
③ 読み込んだデータの数値以外のデータをダミー変数化する。
④ 読み込んだデータのうちのどれか 1 つの列のヒストグラムを表示する。
⑤ 正解データとそれ以外の全ての列の組み合わせの散布図を表示する。
　正解データ同士の散布図は作成しない。
⑥ ヒストグラム、散布図から外れ値を調べて削除し、削除されたことを確認する。
⑦ 欠損値を補完し、補完されたことを確認する。
　補完方式は自身の考えるベストな方法を選んでください。
⑧ 正解データと他の列の相関係数の絶対値を降順で表示する。
⑨ ヒストグラム、散布図、相関係数を参考にして、既存の列から特徴量を選択する。特徴量は 2 個以上選択してください。
⑩ 上記で選択した特徴量に加えて、多項式特徴量(列を 2 乗したものなど)を 1 項目以上加える。
⑪ 重回帰・多項式回帰、ロジスティクス回帰、決定木(分類)、ランダムフォレスト(分類)から学習に最適なアルゴリズムを選択する。
⑫ K 分割交差検証を行う。分割数は 3 とする。
⑬ K 分割交差検証の決定係数の平均を表示する。
⑭ 学習済みモデルで、テストデータを用いて 2 乗平均平方根誤差 RMSE(平均 2 乗誤差 MSE の平方根値)を表示する。
⑮ 学習済みモデルで、テストデータを用いて決定係数を表示する。

実装においては、下記を考慮してください。

- コメントを適切に記載する。
- 変数名、関数名などはプログラムに適した名称とする。

AI 基礎プログラミング
第 2 回　効果測定　問題

問題 4（配点 30 点）

配布データ「titanic_kouka2_1.csv」、「titanic_kouka2_1.csv」を用い、「生存」列のデータを予測するプログラムを作成する。

下記の機能要求を満たすようにプログラムを実装してください。

① 2 つの CSV ファイルを読み込み、読み込んだデータフレームの先頭 5 行をそれぞれ表示する。
② 2 つの CSV ファイルを外部結合する。
③ 正解データに偏りが有るかどうか(データごとの出現回数)を確認する。
④ 特徴量の欠損値を補完し、補完されたことを確認する。
　補完方式は自身の考えるベストな方法を選んでください。
⑤ 特徴量の数値以外のデータをダミー変数化する。
⑥ 訓練＆検証データとテストデータに分割する
⑦ アダブーストを使用してモデルを作成する。
⑧ ベースモデルとしては、重回帰・多項式回帰、ロジスティクス回帰、決定木(分類)、ランダムフォレスト(分類)から学習に最適なものを選択する。
⑨ テストデータでの正解率を表示する。
⑩ 以下のデータをもとに、モデルで予測した結果を表示する。

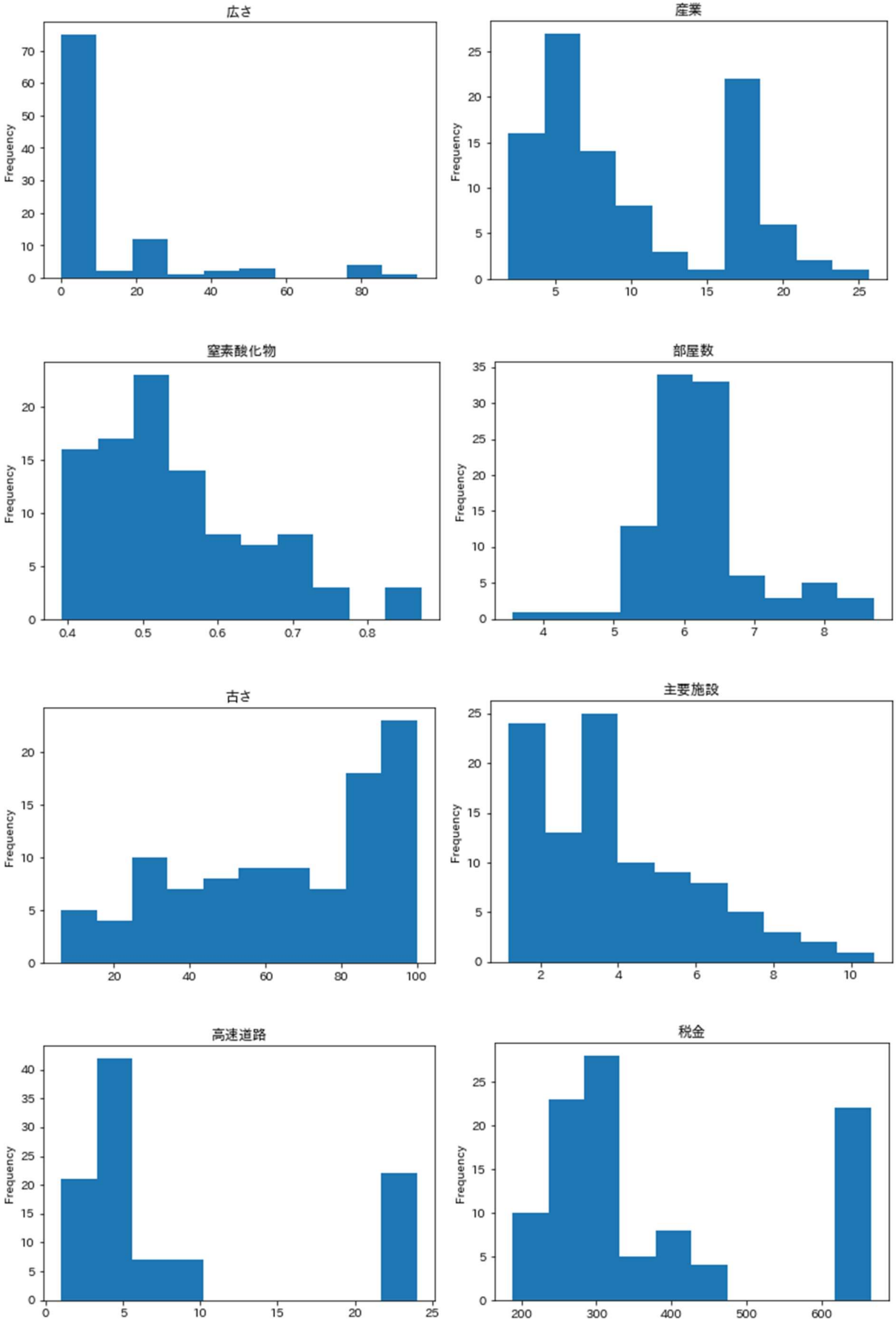
乗客クラス	名前	性別	年齢	兄弟と配偶者	親と子供	チケット番号	運賃	客室番号	出港地
3	nameA	female	13	1	0	2651	11.2417	C99	C
2	nameB	female	13	1	0	2651	11.2417	C99	C
2	nameC	male	13	1	0	2651	11.2417	C99	C

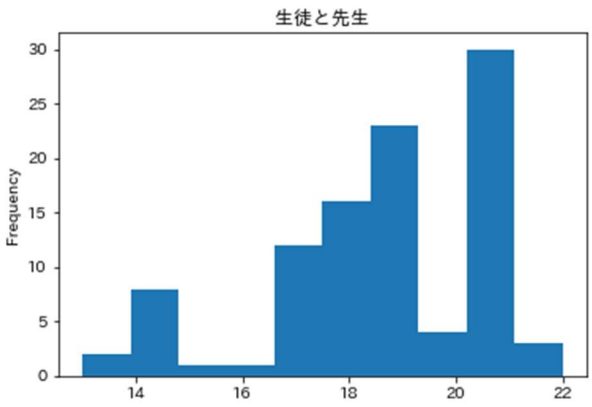
実装においては、下記を考慮してください。

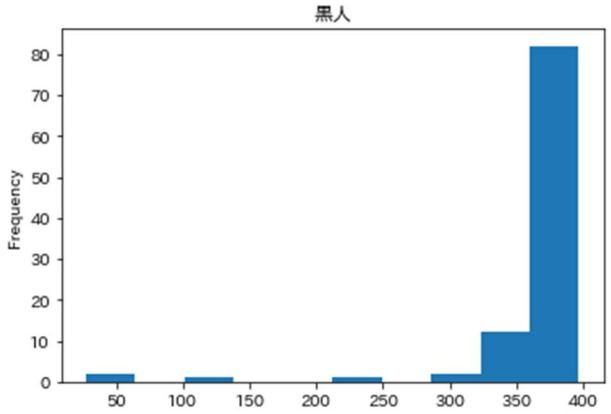
- コメントを適切に記載する。
- 変数名、関数名などはプログラムに適した名称とする。

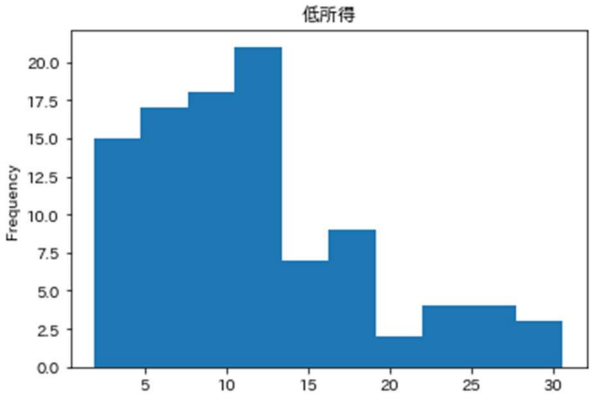
（補足）

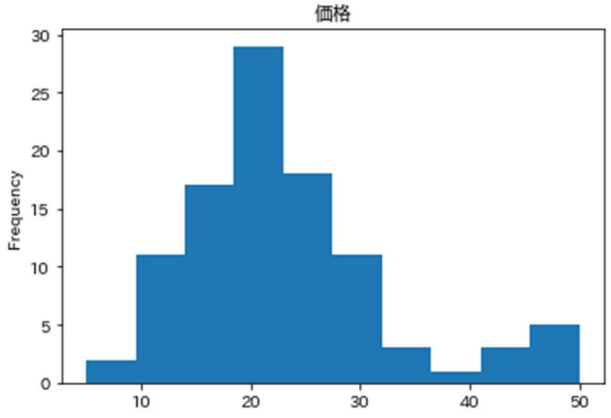
問題 3 ④ 各カラムのヒストグラムは下記のようになる。

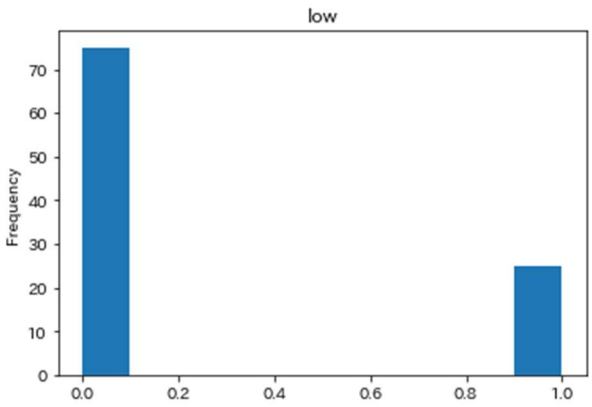


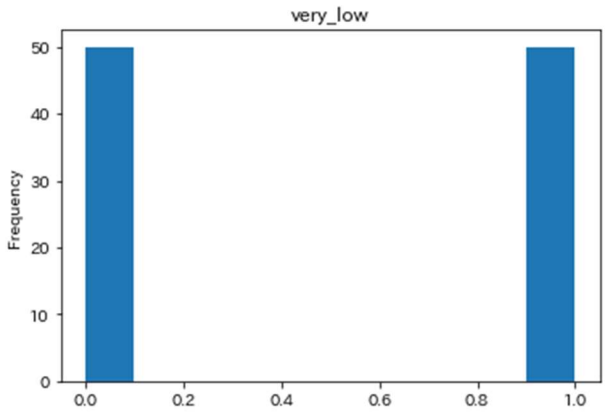

生徒と先生
Frequency

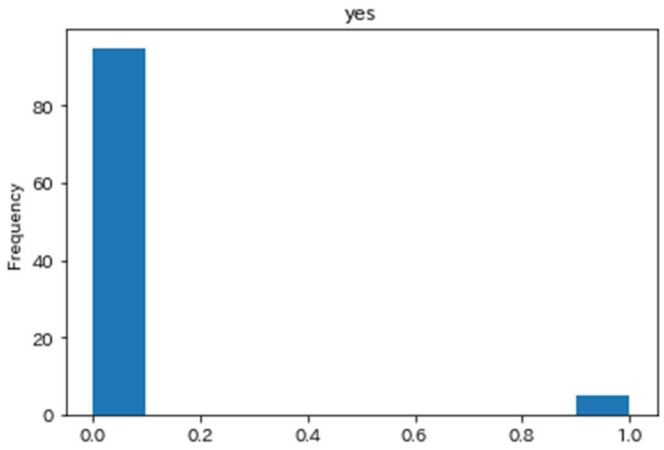

黒人
Frequency


低所得
Frequency


価格
Frequency

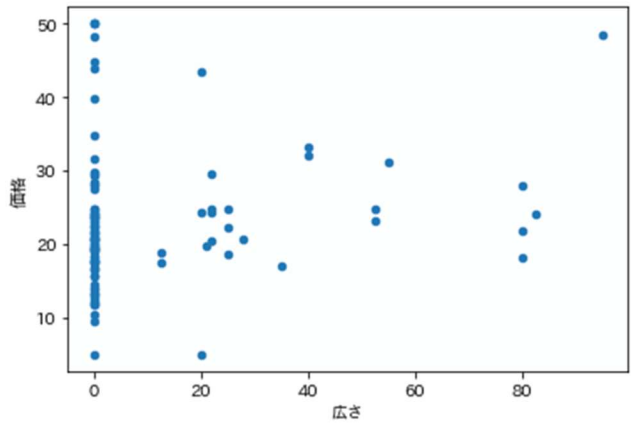

low
Frequency

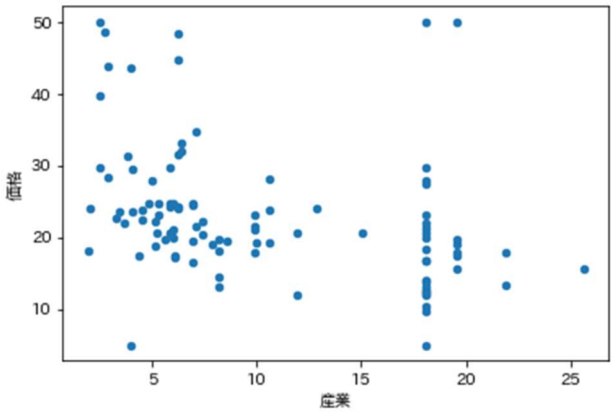

very_low
Frequency

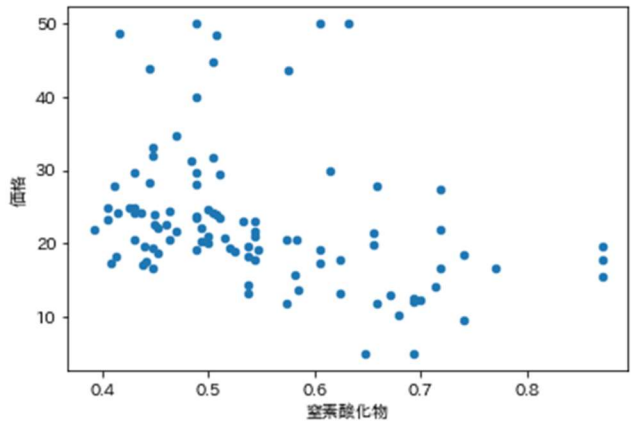

yes
Frequency

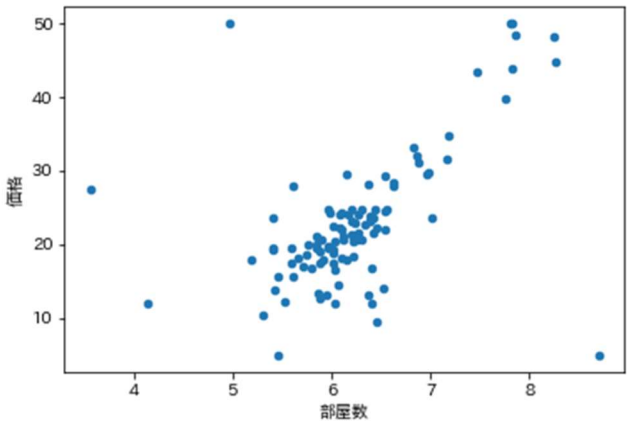
（補足）

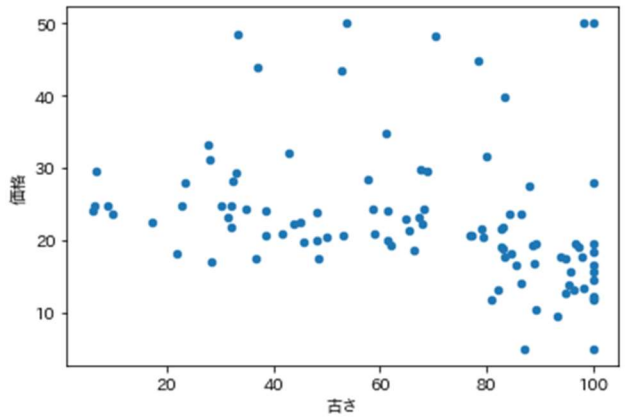
問題 3 ⑤ 各カラムの散布図は下記のようなものになる。

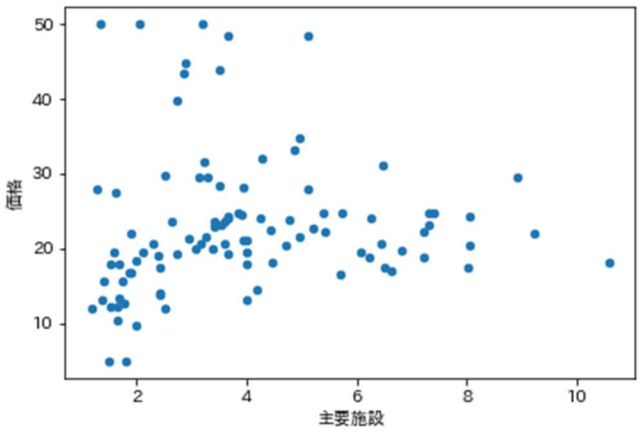


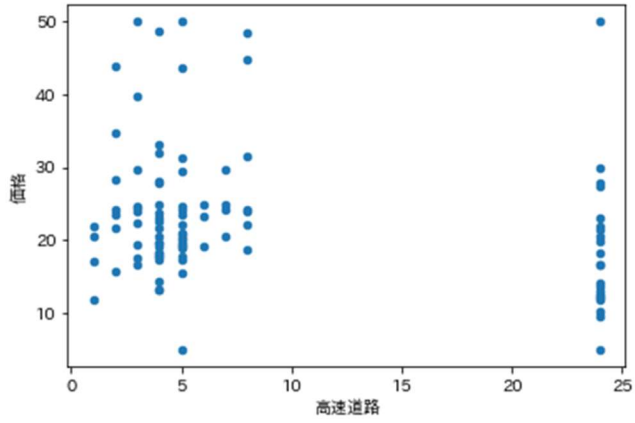


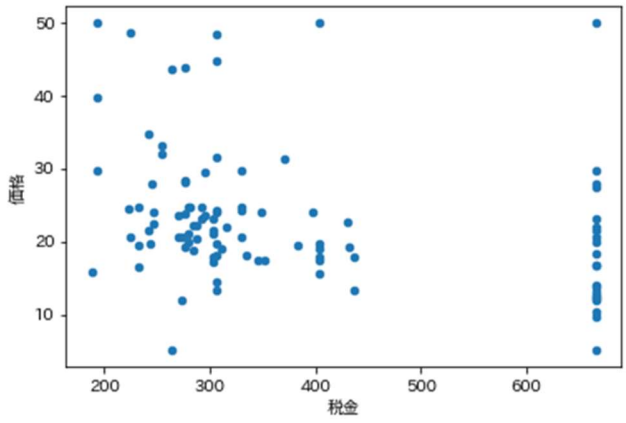


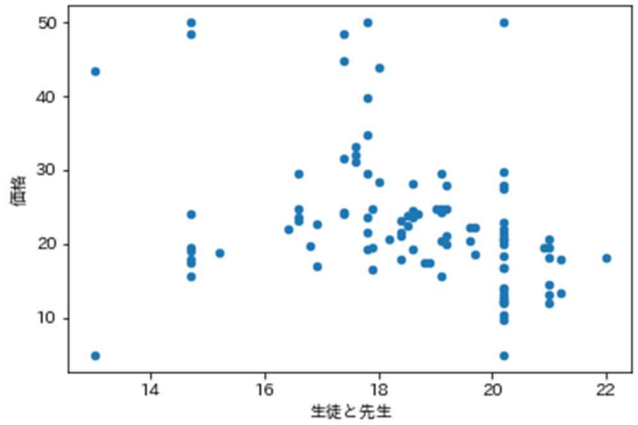


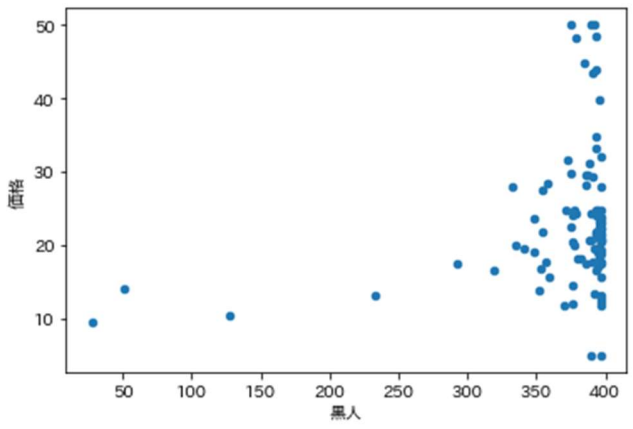


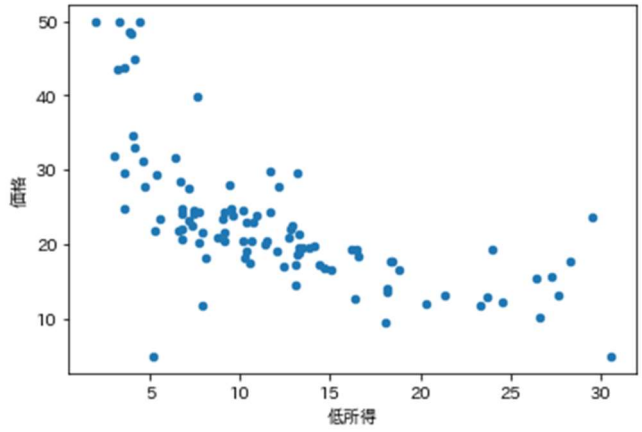


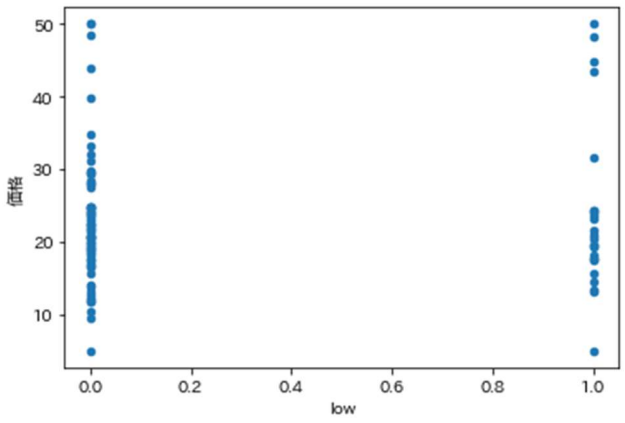


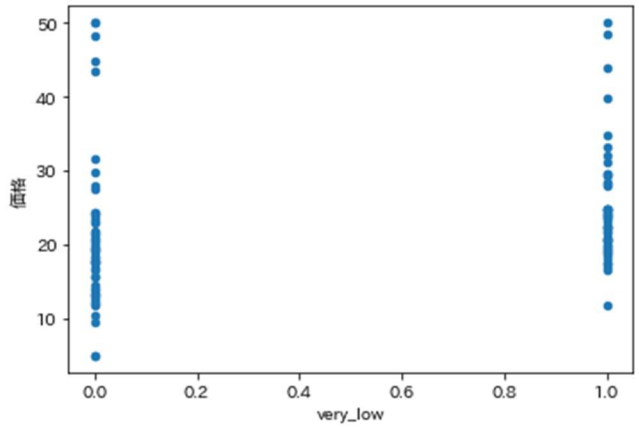


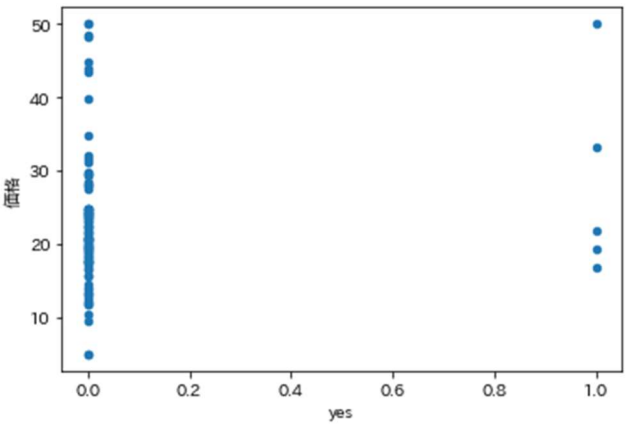

価格
生徒と先生


価格
黒人


価格
低所得


価格
low


価格
very_low


価格
yes