統計・機械学習モデル宿題１ *

廣江友哉 2125178†

2024 年 11 月 4 日

目次

1 課題１

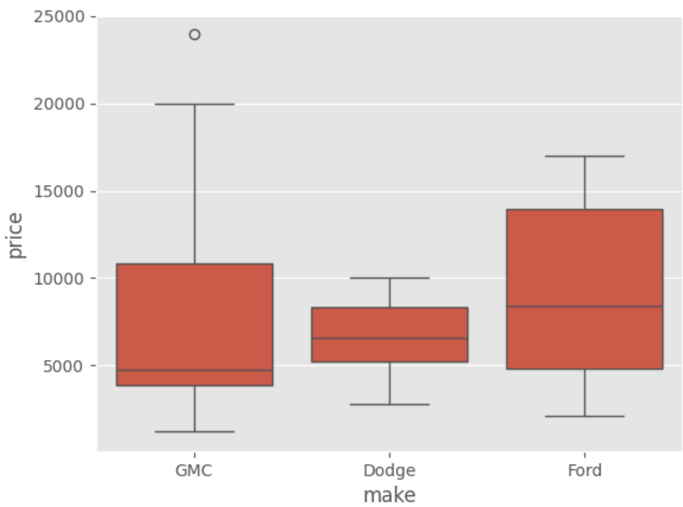

図１ 箱ひげ図

箱ひげ図からメーカーにより価格帯の広がりに違いがあることがわかる．例えば，最低価格と最高価格のトラックのメーカーは共に GMC 社である．また，製品価格の中央値が最も小さいのも

* https://github.com/tomoyahiroe/stats-ml-course
† Email: hiroe-tomoya-yp@ynu.jp

GMC 社である．ここから GMC 社は製品価格の分散が他社と比較して大きいことがわかる．一方で，価格の広がりが最も小さいのは Dodge 社である．四分位範囲も狭く，最高価格は他社と比較して安く，最低価格も他社と比較して高いことがわかる．Ford 社は価格の中央値が 3 社の中で最も高い．

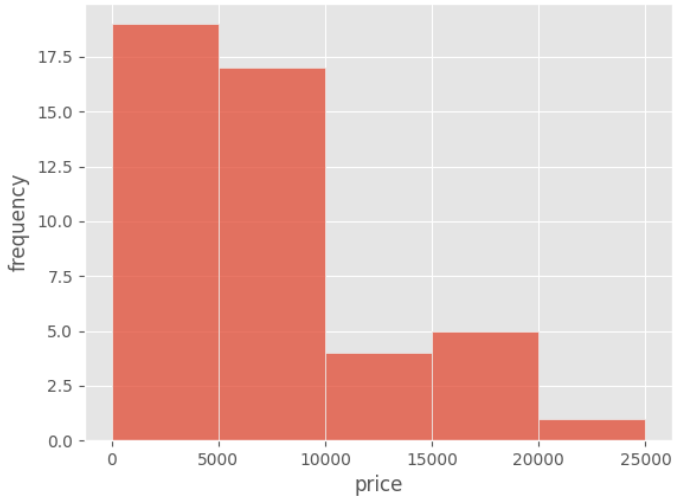

図 2　ヒストグラム

図 2 のヒストグラムは，価格を 5 つの階級に分けて，0 ドル以上 25000 ドル未満の範囲でプロットしている．0 ドル以上 5000 ドル未満の階級の度数が最も高く，19 台である．元のデータは 46 行からなる CSV データであるから，中央値が 5000 ドル以上 10000 ドル未満の階級に属していることがわかる．

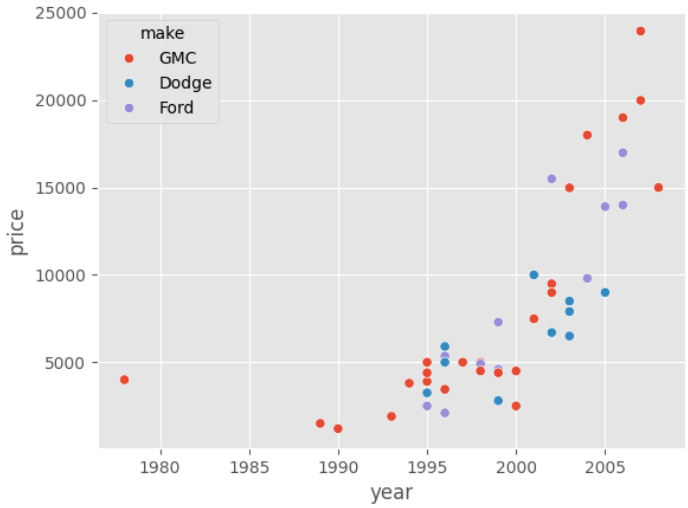

図 3　散布図

図 3 の散布図を確認すると，年々トラックの製品価格が高くなる傾向が 3 社ともに見て取れる．また，図 1 の箱ひげ図でも確認したように，GMC 社の製品には 3 社の中で最低価格のものと最高価格のものが存在する．さらに，年代に注目すると，最も古いトラックは GMC 社のもので，Ford 社と Dodge 社が同時期からデータとして存在することがわかる．

2　課題 2

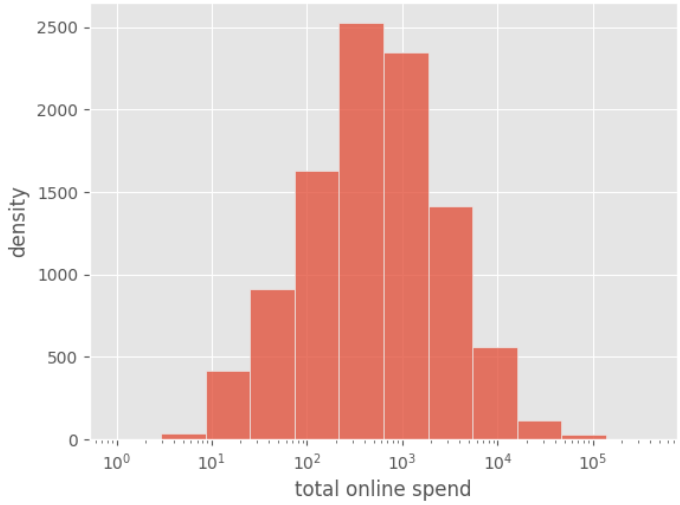

図 4　ヒストグラム

図４のヒストグラムは，オンライン消費額の度数分布を表している．横軸は対数メモリのため，0 ドルから 10 万ドルまでの間に 12 の階級がある．最も度数の高い階級は，だいたい 200 ドルから 600 ドルの間にある階級で 2500 となっている．

Python ファイルは完全に動くものを提出するという規則があるため，データは GitHub に get リクエストを送る形で取得している．また，課題３では Kaggle のデータセットを使用したため，kagglehub というデータセットをインポートするためのライブラリを使用している．また，提出できるファイルにも規則があるため，requirements.txt の代わりに Python ファイルのはじめに必要なライブラリとバージョンをコメントしている．

3　課題３

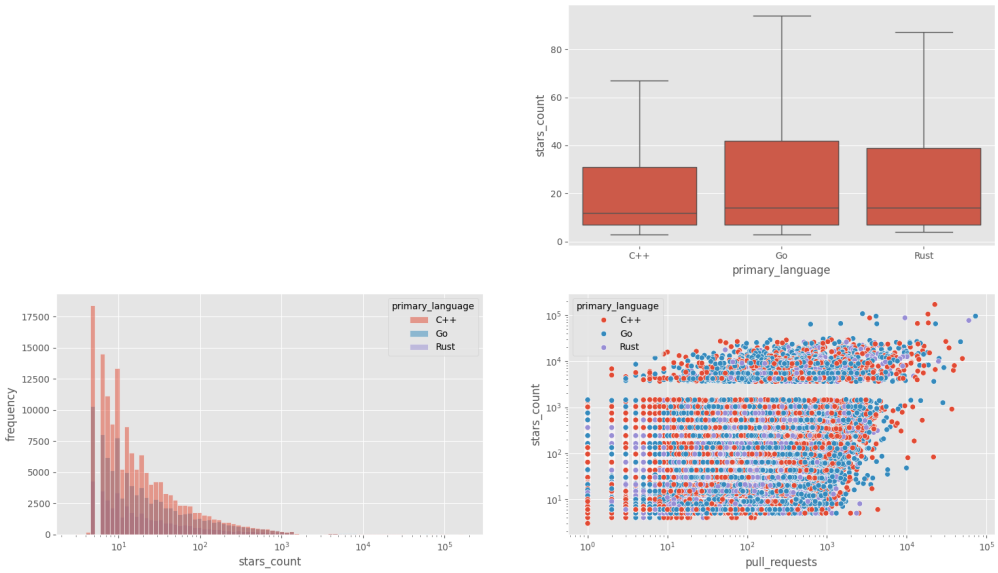

図 5　Kaggle GitHub dataset のグラフ

データは Kaggle 上で公開されていた GitHub Dataset[1)]を使用した．このデータセットを選んだ理由は，近年コミュニティの盛り上がりを見せている Go 言語，Rust 言語，と比較して古典的なプログラミング言語である C++言語の GitHub 上で公開されているリポジトリの中での使用状況を確認するためである．データセットには，レポジトリ[2)]毎のデータが行単位で格納されており，レポジ

1）　https://www.kaggle.com/datasets/nikhil25803/github-dataset?select=repository_data.csv

2）　レポジトリとは GitHub 上でコードを扱う際のフォルダのようなもの．例えば，ライブラリのコードは基本的に一つのレポジトリ内に格納される．

トリ名，プルリクエスト[3)]の回数，そのレポジトリで使用されている主なプログラミング言語，スター数[4)]などのカラムが存在する．

まず，3 言語とも外れ値が大きくグラフから関係が読み取りづらい為，箱ひげ図に関しては外れ値を非表示とし，ヒストグラムは x 軸を対数目盛りにし，散布図に関しては両軸とも対数目盛りでプロットした．

箱ひげ図では言語毎のスター数をプロットした．外れ値は表示していない．四分位範囲は Go 言語が最も広く，中央値も Go 言語が一番高いことがわかる．ヒストグラムもスター数を 80 の階級に分けて表示している．横軸が対数目盛りなのにも関わらず分布が左に寄っている．最も度数が高いのはスター数が 4 から 5 の階級であり，約 18000 のリポジトリがこの階級に属している．したがって，ほとんどのレポジトリのスター数が 1 ケタから 2 ケタの間に収まっていることがわかる．特に 4 ケタを超えるスター数を持つレポジトリはわずかしかないことがわかる．最後に両軸を対数目盛で表示した散布図を確認する．横軸はプルリクエストの回数，縦軸はスター数である．プルリクエストの回数が多いと言うことはそれだけ開発が行われていることを意味するので，開発が行われるだけの需要がありスター数も比例して大きくなると予想していたが，プロットした点は満遍なく広がっており，あまり傾向がみられない．

3） 共同開発のために存在する GitHub の機能．コードに変更を加えたいものがリクエストを作成し，レビュワーがその変更を確認する．プルリクエストの回数はコードの変更回数の目安となる．

4） GitHub のお気に入り機能のようなもの．