

# 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 | 1 |
| 2 | 課題 2 | 3 |
| 3 | 課題 3 | 4 |

## 1 課題 1

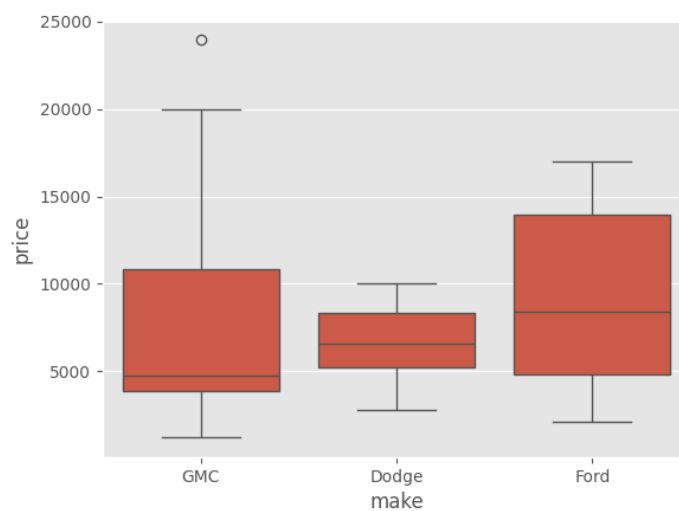


図 1 箱ひげ図

箱ひげ図からメーカーにより価格帯の広がりには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 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

---

\* Email: [hiro-e-tomoya-yp@ynu.jp](mailto:hiro-e-tomoya-yp@ynu.jp)

で、価格の広がり最も小さいのは 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。

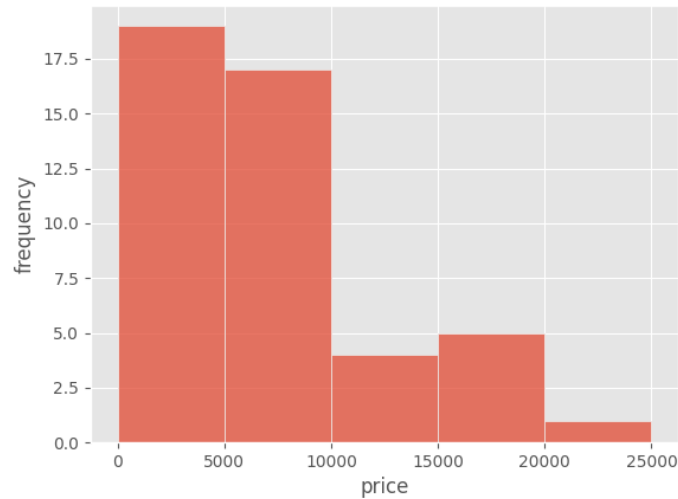


図2 ヒストグラム

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。

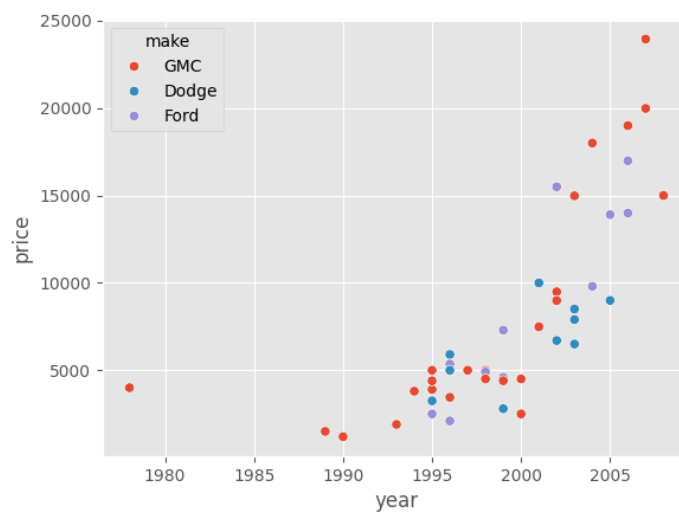


図3 散布図

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

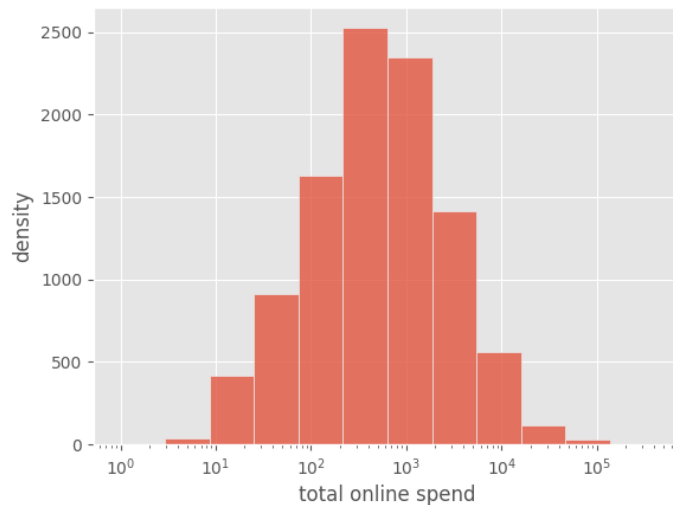


図4 ヒストグラム

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

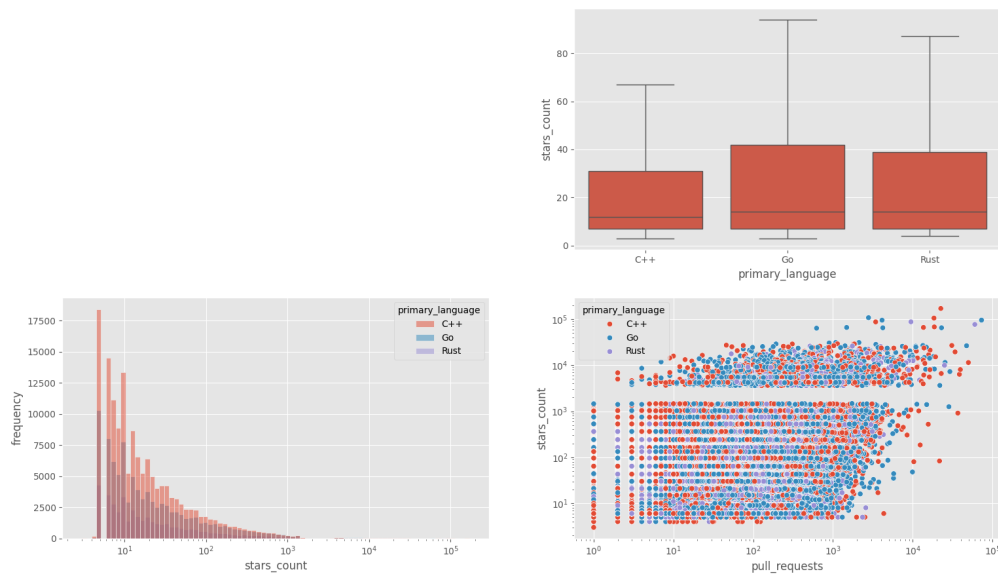


図 5 Kaggle GitHub dataset のグラフ

データは Kaggle 上で公開されていた GitHub Dataset<sup>1)</sup>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<sup>2)</sup>毎のデータが行単位で格納されており、レポジトリ名、プルリクエスト<sup>3)</sup>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<sup>4)</sup>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

1) [https://www.kaggle.com/datasets/nikhil25803/github-dataset?select=repository\\_data.csv](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。

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、レビューヤ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。

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 のレ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い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。