COMP 551: Mini Project 1 - Linear Classification Techniques

GROUP-97 Nikhil Podila[1], Shantanil Bagchi[1] and Surya Penmetsa[1]

Abstract— **We investigate the performance of two linear classification techniques–logistic regression and linear discriminant analysis–on the red wine quality [1] and breast cancer [2] datasets. We preprocess the data, analyse the features, before implementing the linear models and comparing their performance.**

We found that LDA is computationally more intensive than logistic regression, but that logistic regression needs careful selection of the hyper parameters such as the learning rate and stopping criteria. We also inferred that selection of appropriate features and transformations during data preprocessing are crucial for linear classification techniques. Additionally, we tested different learning rates for logistic regression, plotted individual feature histograms, perform correlation analysis on the features and compare the accuracies of the models.

I. INTRODUCTION

In this project, our goal is to implement two Machine Learning algorithms – logistic regression (LR) and Linear Discriminant Analysis (LDA) – and compare their performance on two classification problems – Breast Cancer classification and Red wine recognition.

A. Task Description

The aim of the first task is to predict if a particular breast cancer is benign or malignant from a standard dataset obtained from Dr. William H. Wolberg of University of Wisconsin Hospitals, Madison[3]. A study was conducted over a period of 3 years where data from 8 groups were taken to form a database of 699 patients. The analysis on the dataset determined 10 features of importance.

The second task is to predict quality of a wine given quantitative descriptions of its constituents, from the wine recognition dataset[1]. The dataset contains results of a chemical analysis of wines grown in a particular region in Italy, but derived from three different cultivars. The analysis on the wines determined quantities of 13 constituents found in each of the three types of wines.

[1]McGill University

B. Related work

On the breast cancer dataset, Panahi et al. [9] compares LDA with other classifiers. Lee et al. [10] implements efficient computation method for LR and tests on the Breast Cancer dataset. In this paper, we replicate the results from most of the available literature on this dataset and further analyze the input data of algorithm to derive important conclusions.

In the wine dataset, Bredensteiner et al. [6] performs classification to predict the cultivar. Appalasamy et al. [7] and Yesim et al. [8] have classified wine quality on 11 classes for each quality value. As auditors would prefer binary (good or bad) indication of wine quality, we propose a novel approach – binary classification setting wine qualities $6, 7, 8, 9, 10$ to positive and the remaining to negative.

C. Algorithm and approach

A supervised Machine Learning (ML) algorithm is one where a descriptions (the inputs) relevant to the problem are used to predict an unknown qualitative value (the output) in classification task [5]. The training phase of the algorithm updates numerical parameters using a data where the outputs are known. In testing phase, the parameters are used for prediction. LR is a discriminative linear model of ML, in which the data is summarized to w vector during training. LDA is a generative linear model of ML assuming Gaussian Distribution in input data. The distribution parameters μ_1, μ_2 and Σ are learned in training phase.

D. Important findings from results

- Logistic regression is computationally efficient compared to LDA.
- Logistic regression doesn't converge when the learning rate is too low, or when it's too high.
- LDA shows minor performance improvement by adding interaction features compared to LR.

The rest of the paper is organized as follows. Section II covers analysis of datasets. Section III discusses implementation and results obtained. Section IV contains discussions and conclusions made from the project. Section V highlights each members' contribution.

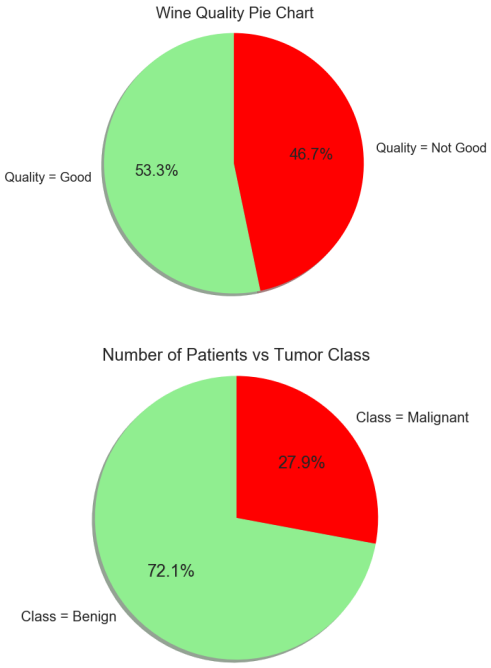

Fig. 1. Split of positive and negative class records in wine dataset (top) and breast cancer dataset (bottom)

II. ANALYSIS OF DATA

A. Dataset and features

The red wine dataset contains 1599 wine samples across 11 features with wine quality as the target variable. The Breast cancer dataset contains 699 patients' information (16 entries incomplete) across 9 features with the class (benign/malignant) as target variable. The split of positive and negative classes are shown in Fig. 1. We split both the datasets, with 80% data for training and rest for testing.

Initially, we noted that the breast cancer dataset consists of 16 entries with missing features and removed those entries. We also converted the dataset's quality variable to a categorical 0 or 1 as described in Section I. For Wine dataset, we converted the class variable values 2 and 4 to 0 and 1 respectively. We explored the individual features and their distributions for both the datasets. We discussed interesting questions about the data and described our observations. One such analysis is shown in Fig. 2. We also performed cross correlation analysis on the datasets to visualise the correlation between input features. Further, we explored interaction features specifically for the wine dataset.

B. Features in the dataset, their distributions and correlations

1) Red Wine Dataset:

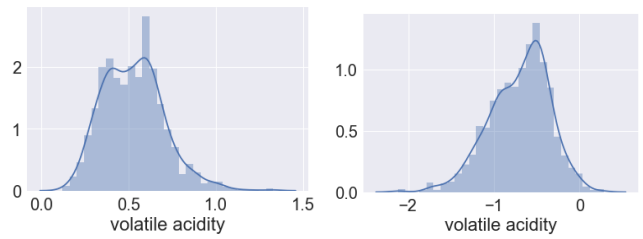

Fig. 2. Available volatile acidity histogram (left) and log of volatile acidity histogram (right). This analysis shows that taking log of feature transforms plot closer to normal distribution.

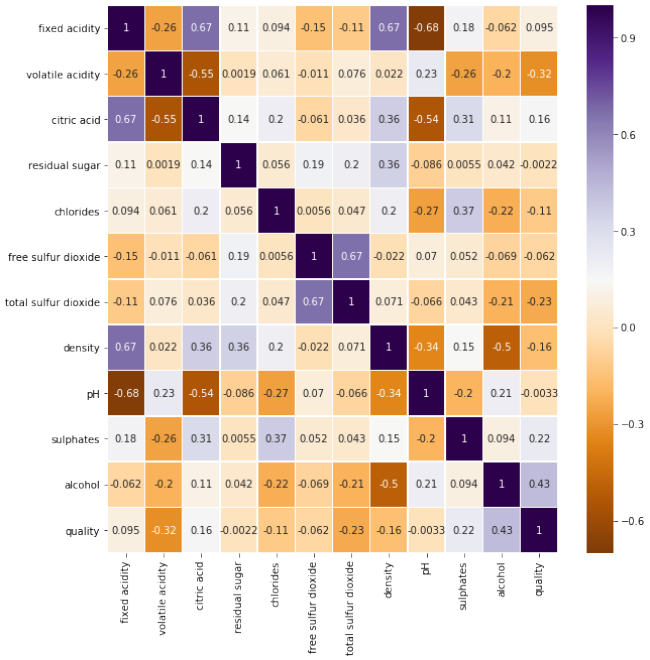

Fig. 3. Cross correlation analysis using seaborn [14] for wine dataset

- Fixed Acidity- Normal distribution with major samples exhibiting values between 6.5 g/dm3 to 9.2 g/dm3.
- Volatile Acidity- Majority between 0.25g/dm3 to 0.79g/dm3. Taking log of Volatile Acidity makes the plot normally distributed.
- Residual Sugar- Highly left skewed histogram, with very few samples having residual sugar more than 8.
- Chlorides- Highly skewed histogram with most values between 0 and 0.1.
- Free sulfur dioxide- Most of the free sulfur dioxide values are between 1 and 40.
- Total sulfur dioxide- Most of the total sulfur dioxide have a spread between 0 and 150.
- Density- Normally distributed with values between

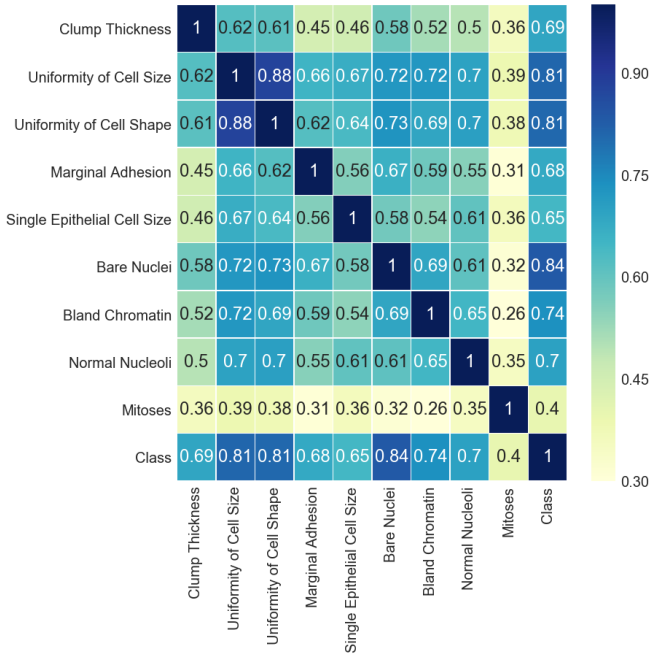

Fig. 4. Cross correlation analysis using seaborn for breast cancer dataset

0.99 g/cm3 and 1.004 g/cm3.

- pH- Normally distributed with major samples exhibiting values between 3.0 and 3.5.
- Alcohol- Varies from 8 to 14 with most of data around 9-10.
- Correlations-
  - pH is negatively correlated with fixed acidity and citric acid which is natural since $pH = -log(H+)$. See Fig. 5
  - Fixed acidity is positively correlated with citric acid as citric acid is the main acidic component of grapes and wine is made from grapes.
  - Alcohol content is positively correlated to quality.See Fig.6

2) Breast cancer Dataset: For the breast cancer dataset, we observe that almost all features are positively correlated with each other.

C. Derived Features

Most of the following features are developed after exhaustive data analysis. We attempt taking each of log, $square$, $square\,root$ of data and note their effect on accuracy. We also try combining two features separately and note their effect on correlation. The features are selected based on higher correlation to quality.

The following derived features, mainly logarithmic and interaction features, have been used in the classification task. See table I

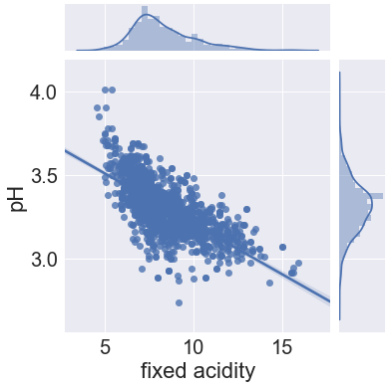

Fig. 5. Correlation between fixed acidity and pH

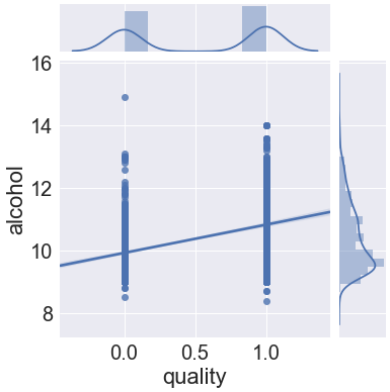

Fig. 6. Correlation between fixed alcohol and quality

TABLE I

DERIVED FEATURES AND THEIR CONCENTRATIONS

Feature	Correlation
density-alcohol	0.43
pH-alcohol	0.36
sulphates-alcohol	0.35
totalsulfurdioxide-volatileacidity	-0.31
density-volatileacidity	-0.31
pH-volatileacidity	-0.3

D. Outlier Detection

We used Boxplot method [12] to perform Outlier detection on wine dataset (specifically $quality$ feature) using interquartile range. It is equal to the difference between 75th and 25th percentiles, IQR = Q3 − Q1. We calculated data spread beyond 1.5 times the interquartile range and removed the data for further processing.

Breast cancer dataset is already classified into class 2 and 4 so it was not intuitive to utilise Boxplot

method on the $class$ so we used it on $bare\ nuclei$ and $Uniformity\ of\ cell\ size$ because of their higher correlation value. Number of outliers detected were 0 and 67 respectively. For bare nuclei, the data was highed skewed and centered on a single value, which was the reason for 0 outliers.

E. Ethical concerns

Datasets like the breast cancer are private to an individual. They must be anonymized before sharing in public as they can be misused in a variety of ways. One example would be insurance companies exploiting this information to hike the premiums of customers who are sick. During data collection, consent of all the patients must be taken before their data is used or released for any research purposes. Careful precautions need to be taken to avoid intentional or unintentional leaks of these data.

III. RESULTS

A. Model implementation

We implement[1] both the algorithms, namely logistic regression and linear discriminant analysis, as classes with the following methods:

- Constructor - Setup important parameters for the classification.
- fit method - Perform training of weights on the Training dataset.
- $predict$ method - Predict the class labels for new/test dataset.
- $normalize\_data$ method - Normalize the dataset over Training mean and standard deviation

The $evaluate\_acc$ function takes the predicted labels (y_pred) and target labels (y_test) as inputs, and returns the accuracy values of the predictions.

The loss function plots for both the datasets when trained with logistic regression is shown in Fig. 7.

We also implement a stopping criteria, where norm of the gradient of loss function is considered [13]. Below a certain threshold, this norm indicates the there will be no further improvements in the weights. This threshold is chosen as 0.0001 and 0.001 in this paper.

B. Task: Run experiments

K-fold cross validation is implemented from scratch and used to evaluate various parameters and inputs for the tasks.

[1]Results and implementation can be found in the link: https://github.com/nikhilpodila/Classification-UCI-Datasets

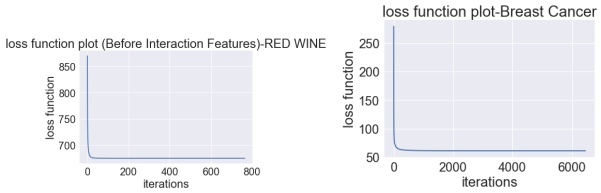

Fig. 7. Loss function plot for wine dataset (left) and breast cancer dataset (right)

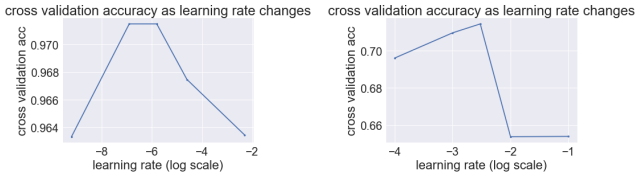

Fig. 8. Cross-validation accuracy % vs. Learning rate for Breast Cancer classification dataset (left) and Wine dataset (right)

We analyse how logistic regression performs for varying learning rates. The plots for the same have been shown in Fig. 8. We notice peak in values at specific values of the learning rate. This explains that the algorithm doesn't converge for a low learning rate and oscillates around the optima for a high learning rate.

The algorithms are timed using the time library. On averaging over multiple runs, we observe that LDA takes more time which is inline with the results observed in the literature[4].

TABLE II
PERFORMANCE AND ACCURACY FOR LOGISTIC REGRESSION ON WINE DATASET

Detail	**Interaction terms**	**Run time**	**k-fold accuracy**	**Test set accuracy**
Full Data	No	0.21	64.94	69.37
Full Data	Yes	0.20	71.21	71.56
Data without outliers	No	0.22	63.13	70
Data without outliers	Yes	0.23	69.3	69.7

We can see major improvement in the accuracy with the addition of interaction terms for logistic regression both when full dataset was taken as well as when outliers data were removed. See table II.

LDA shows only slight improvement with the addition of interaction terms. Overall LR has quicker execution time but is less accurate compared to LDA which is taking more time. See table III.

TABLE III
PERFORMANCE AND ACCURACY FOR LDA ON WINE DATASET

Detail	Interaction terms	Run time	k-fold accuracy	Test set accuracy
Full Data	No	1.33	74.35	74.06
Full Data	Yes	1.23	74.58	75.88
Data without outliers	No	1.26	73.33	74.2
Data without outliers	Yes	1.3	73.5	75.13

TABLE IV
PERFORMANCE AND ACCURACY OF LR AND LDA ON BREAST CANCER DATASET

Description	Run time	k-fold accuracy	Test set accuracy
LR - Full Data	0.09	97.88	96.82
LDA - Full Data	0.44	96.51	94.16
LR - Without outlier	0.09	96.74	95.35
LDA - Without outlier	0.44	95.12	95.43

IV. DISCUSSION AND CONCLUSIONS

We implemented logistic regression and LDA and tested them on the breast cancer and wine datasets. We analysed these algorithms further in terms of convergence, optimal learning rates, etc.

Here are some important findings from our study:

- Logistic regression is efficient compared to LDA (with appropriate input features). This is consistent with what was discussed in the class, that LDA is much more computationally expensive because of the matrix operations and inverses that it needs to calculate.
- We have low accuracy on the logistic regression when the learning rate is too low (because it doesn't converge fast) and when it's too high (it fails to hit the optima). Parameter search is shown in Fig. 8.
- We observed an interesting phenomena that LDA shows minor performance improvement by adding interaction features compared to LR.
- LDA did not perform well on the Breast Cancer dataset when compared to LR, since most features in the dataset are not normally distributed. This aligns with results from literature [11].

We also used vectorized implementations in the code for efficiency. We completed all the requirements for the project and additionally worked on the understanding what these algorithms do in detail.

This work has helped us understand these algorithms more deeply. This project provided us a great hands-on experience of building a model from beginning and improving it until it can perform well on prediction.

Now, we are able to sketch out a procedure for working on a machine learning project based on what we have done. This involves: Data preprocessing, model training, model validation and model testing. Data preprocessing focuses on the features. We may also encounter non-numeric features in the future, which require more advanced encoding operations, But the basic idea here is simply converting data of any kind into numerical values that can be taken as input to our models.

Model training and validation can be seen as a whole separate task. The main goal is to select out the good features and tune for the better combinations of hyperparameters in the model. Examples would be choosing among the various interaction features that can be formulated.

Finally, once we have our best model, we run it on the test set. This step makes sure that our model doesn't over-fit. When given another unseen dataset, our model should still make good prediction. Also, if we have multiple models and want to pick out the one that performs the best on a specific problem, the final step will provide us valuable information.

V. STATEMENT OF CONTRIBUTIONS

The collaborative work of this team involved everyone contributing to the report, Shantanil contributing to data processing, analysis and visualisation, Nikhil contributing to cross validation, logistic regression implementations and Surya working on implementing LDA and determining timing of the algorithms.

REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[2] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[3] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 18

[4] Pohar, Maja, Mateja Blas, and Sandra Turk. "Comparison of logistic regression and linear discriminant analysis: a simulation study." Metodoloski zvezki 1, no. 1 (2004): 143.

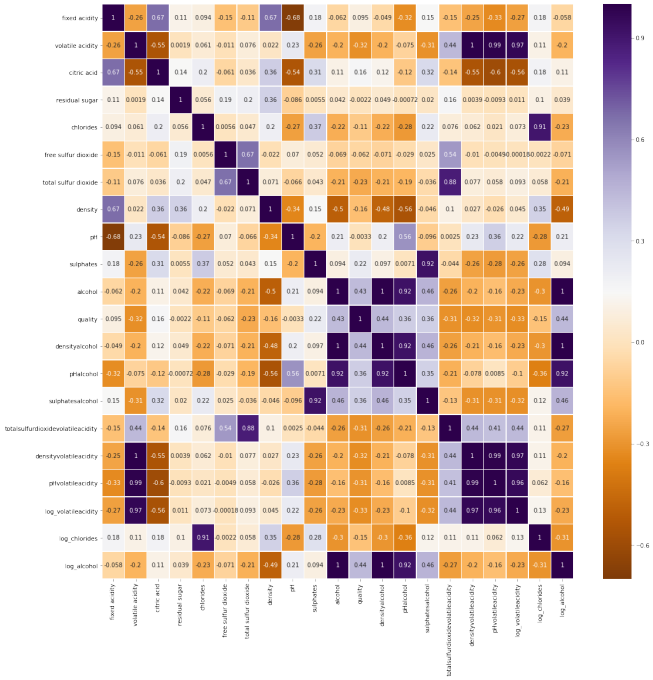

Fig. 9. Plots of correlation when new interaction features are included for the wine data set.

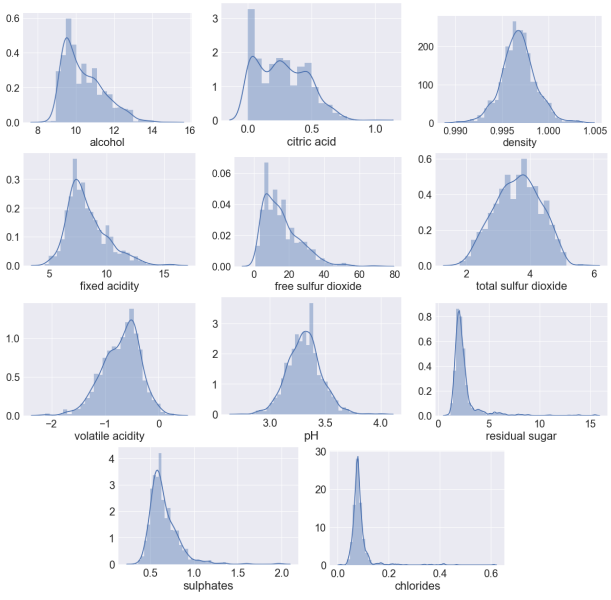

Fig. 10. Plots of individual data features for the wine data set.

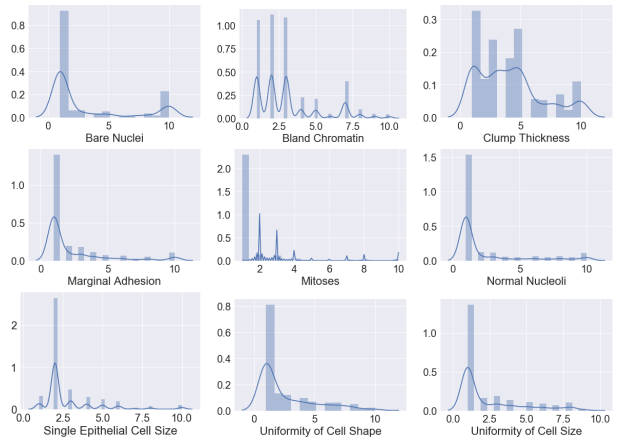

Fig. 11. Plots of individual data features for the breast cancer data set.

[5] Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The elements of statistical learning : data mining, inference, and prediction. New York: Springer, 2001. Print.

[6] Bredensteiner, Erin J., and Kristin P. Bennett. "Multicategory classification by support vector machines." In Computational Optimization, pp. 53-79. Springer, Boston, MA, 1999.

[7] Appalasamy, P., A. Mustapha, N. D. Rizal, F. Johari, and A. F. Mansor. "Classication-based Data Mining Approach for Quality Control in Wine Production." Journal of Applied Sciences 12, no. 6 (2012): 598-601.

[8] Er, Yeşim, and Ayten Atasoy. "The classification of white wine and red wine according to their physicochemical qualities." International Journal of Intelligent Systems and Applications in Engineering (2016): 23-26.

[9] Panahi, Nazila, Mahrokh G. Shayesteh, Sara Mihandoost, and Behrooz Zali Varghahan. "Recognition of different datasets using PCA, LDA, and various classifiers." In 2011 5th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-5. IEEE, 2011.

[10] Lee, Su-In, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng. "Efficient l 1 regularized logistic regression." In AAAI, vol. 6, pp. 401-408. 2006.

[11] Pohar, Maja, Mateja Blas, and Sandra Turk. "Comparison of logistic regression and linear discriminant analysis: a simulation study." Metodoloski zvezki 1, no. 1 (2004): 143.

[12] Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley

[13] Nocedal, Jorge, and Stephen Wright. Numerical optimization. Springer Science Business Media, 2006.

[14] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, … Adel Qalieh. (2017, September 3). mwaskom/seaborn: v0.8.1 (September 2017) (Version v0.8.1). Zenodo.