# 

**University of Colombo Faculty of Science**

**RESEARCH TRAINING AND INDUSTRIAL TRAINING**

**IT 4009**

**Organization : DATA DISCA (PRIVATE) LIMITED**

**Prepared By : M.U. Abeywickrama**

**Index No.: 14040**

**Reg. No.: 2017s16499**

**DECLARATION**

**I declare that this report does not incorporate, without acknowledgment, any material previously submitted for a Degree in any University, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text and I hereby grant to University of Colombo, Sri Lanka the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or another medium.**

**Candidate Name: M.U. Abeywickrama**

**Reg. No: 2017s16499**

**Signature of Candidate**

**Index No: s14040**

**Date: ………………….**

**Certificate of Approval**

**We hereby declare that this report is from the student’s own work and effort and all other sources of information users have been acknowledged. This report has been submitted with our approval.**

**………………………………………………**

**Signature of Placement Mentor**

**Thilak Fernando**

**BSc(Eng)(Hons)(UoM, SL),**

**MSc(UoM, SL),**

**PhD(Monash, AUS)**

**Managing Director DataDisca Pty Ltd**

**Date: ………………….**

**……………………………………………..**

**Signature of Academic Mentor**

**Dr. S.K.P. Eranga, Deligan SE (PVT) Ltd,**

**ITU 2 Coordinator,**

**Faculty of Science,**

**University of Colombo.**

**Date: ………………….**

**ACKNOWLEDGEMENT**

**I would like to thank my family and**

**all the other ladies and gentlemen who gave**

**the support to complete my training period successfully in Data Disca (Pvt) Ltd.**

**In my training period, Especially Mr.Thilak Fernando (Managing Director Data Disca)**

**who was the mentor for my training period,**

**gave opportunity, knowledge and guided me to gain knowledge.**

**He guided us to learn knowledge by ourselves**

**Further, I could like to thank the all**

**academic and non-academic staff of the University of Colombo.**

**TABLE OF CONTENTS**

Contents

[1 1](#_Toc97467206)

[1. Introduction 6](#_Toc97467207)

[1.1. Scope 6](#_Toc97467208)

[1.2. Sources of the data 6](#_Toc97467209)

[2. Organization 7](#_Toc97467210)

[2.1 Introduction 7](#_Toc97467211)

[2.2 Technology 8](#_Toc97467212)

[2.3 Company Website and LinkedIn profile 9](#_Toc97467213)

[2.4 Progress Monitoring Ethics 10](#_Toc97467214)

[2.5 Acceptance & Validity 10](#_Toc97467215)

[2.6 Intellectual Property Rights 10](#_Toc97467216)

[2.7 Termination 11](#_Toc97467217)

[2.8 Quality Standard of Your Work 11](#_Toc97467218)

[3 Training Experience 12](#_Toc97467219)

[3.1 Introduction 12](#_Toc97467220)

[3.2 Appointment Letter​ 12](#_Toc97467221)

[3.3 Position Description 13](#_Toc97467222)

[3.4 Training Places 14](#_Toc97467223)

[3.5 Meeting 14](#_Toc97467224)

[3.6 Tasks and Activities 14](#_Toc97467225)

[3.6.1 Analysis and visualization basics using Jupyter 14](#_Toc97467226)

[3.6.2 Datasets clustering and algorithm with optimization methods 15](#_Toc97467227)

[3.6.3 Datasets classification algorithm with optimization methods 16](#_Toc97467228)

[3.6.4 Data visualization and dashboard using tableau 17](#_Toc97467229)

[4 Discussion and Conclusion 19](#_Toc97467230)

[5 References 20](#_Toc97467231)

# Introduction

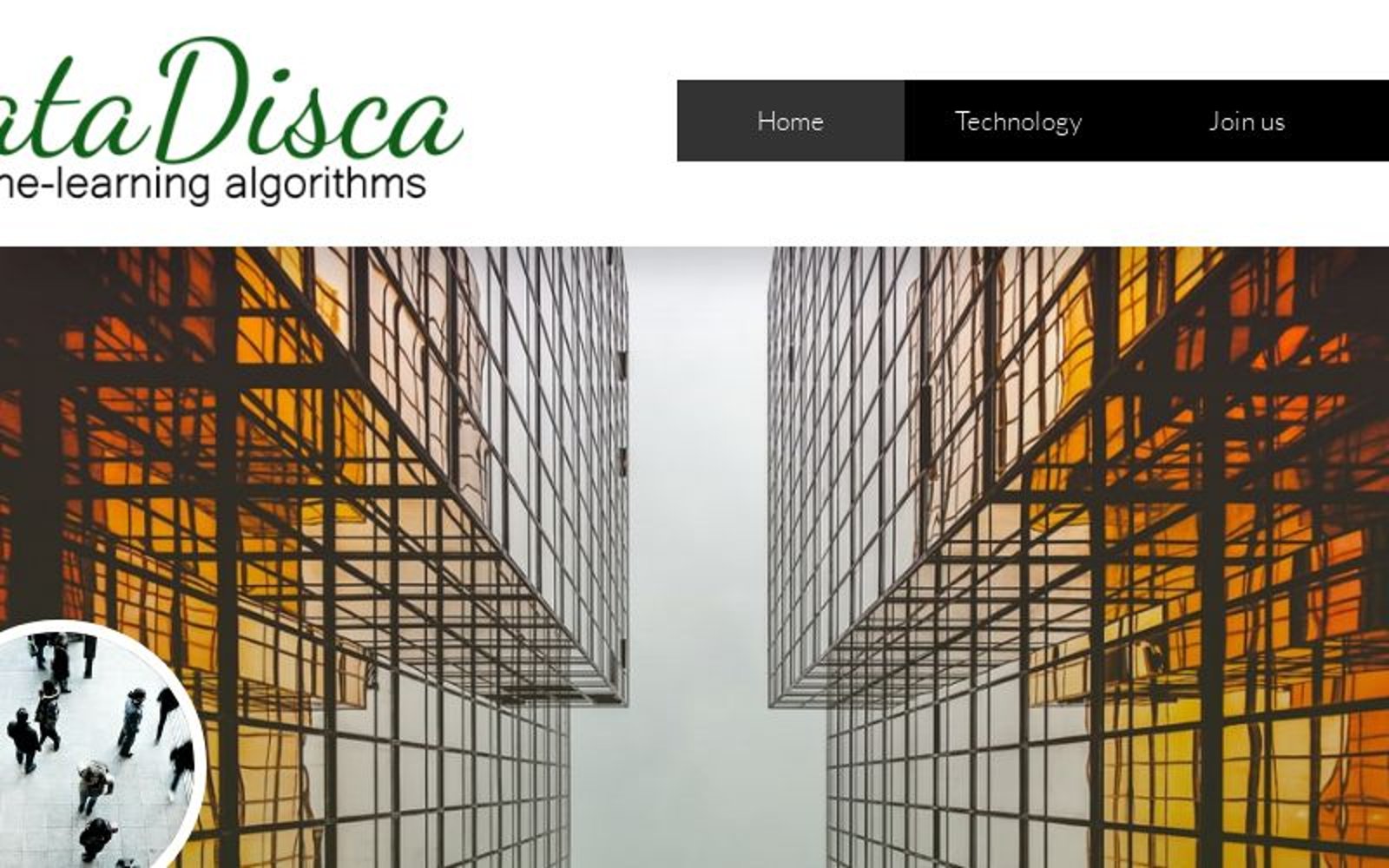
The internship is an opportunity to fill the gap between academics and industry experience. It provides the students an opportunity to use their academic knowledge in practice. Not only that it helps institutions to get in touch with the industry and up to date with their syllabus and another thing in a suitable manner. Industries also get the idea about the research knowledge and get a good chance to identify talented persons. So, Both parties get to benefit from this. Faculty of Science, University of Colombo provides this opportunity to its students within the bachelor program. Every bachelor student of the Faculty of Science has to join the industry as an internship as a degree requirement. Being a bachelor student of the Faculty of Science, I was able to join Data Disca Pvt Ltd as a Trainee data scientist to complete my internship.

## Scope

This report gives an idea about the challenges that I faced during my training period. It also summarizes the outcomes of my internship in Data Disca. A brief discussion about the work that I have done, is also provided. Finally, the most valuable part comparison of my skills before and after the Internship program is included and it will justify my achievements.

## Sources of the data

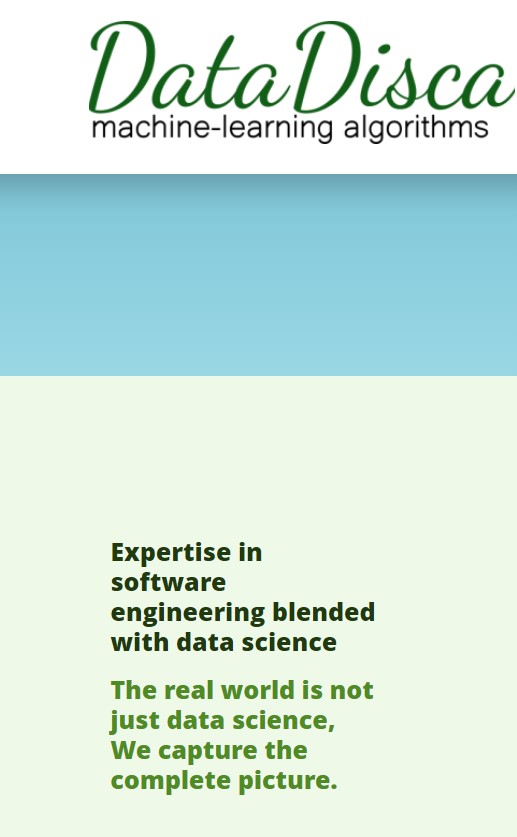
* Primary Data was collected from the practical experience during my internship period.
* Also Using the company website and the internet I collected an extra amount of data



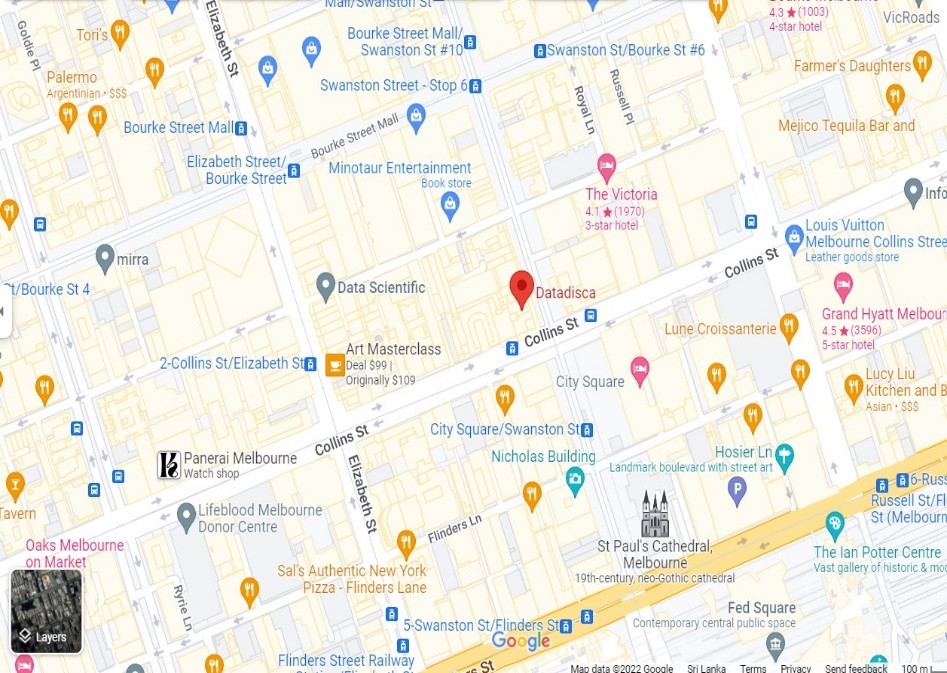
# Organization

## Introduction



A data science company with a research focus. DataDisca is a privately held research-oriented organization. It specialized in machine learning and develops AI for the next generation. Data Disca company is mainly located in Melbourne and Sydney. The headquarters of Data Disca is located in Melbourne, Victoria. Data Disca was founded in 2019 and the company size is 51-200 employees. Now it is a fast-growing and well-reputed company. It specialized in machine learning, Python, TensorFlow, Plotly, AWS, Azure, Maps, Tableau, Visualisations, Object-Oriented Design, Research, and data science. Expertise in software engineering blended with data science

Data Disca logo 1



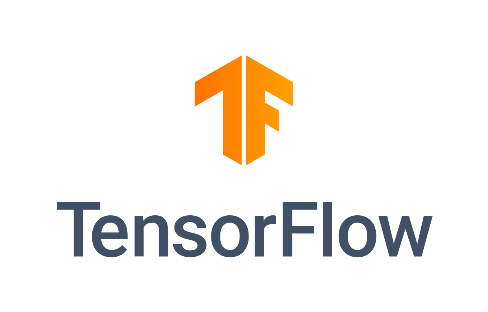
## Technology

Advance data extraction transformation and loading

* Loading data from CSV, Text, Microsoft Excel, Microsoft Word, Facebook, Twitter, Google Search, Wikipedia
* Databases

Classical Machine Learning

* Classification
* Clustering
* Anomaly Detection
* Time Series Analysis
* Market Basket Analysis

Deep Learning

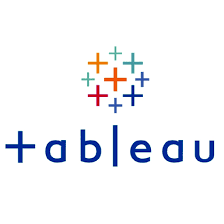
* Computer Vision
* Natural Language Processing
* Time series forecasting
* Classification



Data Visualization

* Tableau
* Power BI
* Plotly
* Interactive Maps
* Interactive HTML



Python Web Development for Data Science

* Fast API
* Django
* Django REST Framework

## Company Website and LinkedIn profile

**LinkedIn Profile**

<https://www.linkedin.com/company/datadisca/mycompany/>

**Website**

<https://datadisca.com/>

## Progress Monitoring Ethics

* Make sure that the Zoom meetings are attended, and the progress calls are answered.
* Person needs to have access to his computer in a calm and settled environment when answering.
* Inform to the company in advance if he is unable to answer or want to reschedule a meeting.
* Be mindful that DataDisca spends the business time to discuss our progress.
* Company does not allow skipping progress calls unless there is a restart date.
* Therefore, person has to discuss with company and agree on a restart date if person want to skip progress calls for any purpose.

## Acceptance & Validity

* person need to be a citizen of Sri Lanka.
* The position is valid only after DataDisca receives and acknowledges this form. On acceptance, DataDisca sends a welcome email with commencement details.
* Terms and conditions given in this agreement apply only to work done under DataDisca mentorship.

## Intellectual Property Rights

* You are free to maintain your copies of work in open-source repositories of your choice under the following conditions.
* DataDisca reserves the right to publish your work that can be of interest to a broader community. In that case, DataDisca may publish selected work in open-source repositories such as github.com or bitbucket.org.
* Open-source repositories published by DataDisca are subjected to the following terms and conditions.

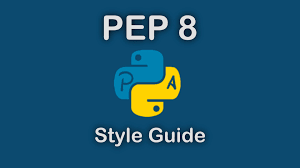
## Termination

* The agreement can be immediately terminated by informing the other party.
* Work done before the termination will abide by the terms and conditions of this agreement.
* If DataDisca does not receive any communication for three weeks, this engagement will be automatically discontinued.
* If you plan not to communicate with DataDisca for more than three weeks, you should agree upon a resumption date with DataDisca.



## Quality Standard of Your Work

* Code should follow PEP8 Standard
* Should host your code on our GitHub in a public or private repository as our prefer.
* If it is a public repository, must send the link to evaluate.
* If it is a private repository, share (view only) with company GitHub usernames.
* Sending a notification to start the evaluation. Then evaluating our code for technical progress.

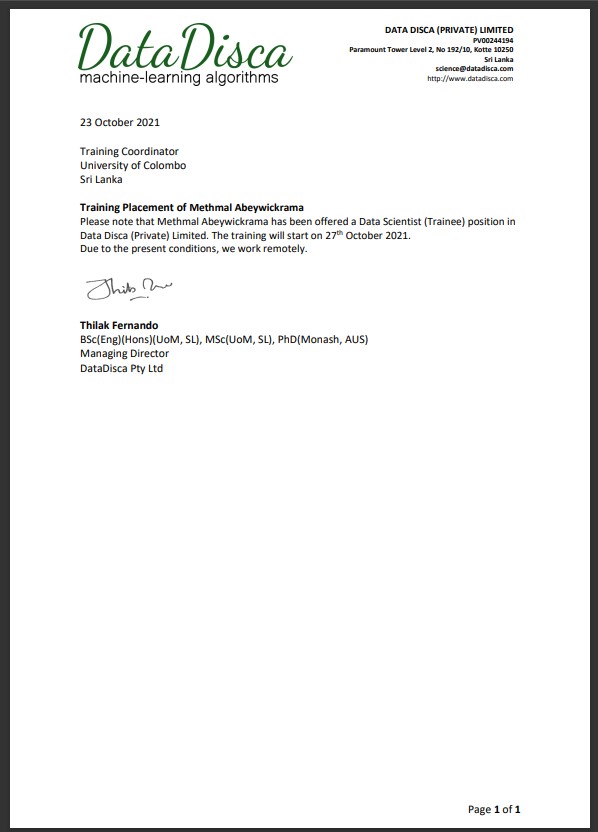


# Training Experience

## Introduction

I started my internship at Data Disca (Pvt.) Ltd. for training period started from 23 October 2021. I started my industrial training as a trainee data scientist. There were many trainers with me. The company provides free tailor-made training to match our skills with industry requirements. Also, we received the necessary knowledge, experience, and due references to be successful in the industry. In their customized training program, we worked with technology that competent data scientists to learn the technology listed on our technology page. ​After identifying our skills during the training program, if there is project availability, the company assigns paid work to us.

## Appointment Letter​



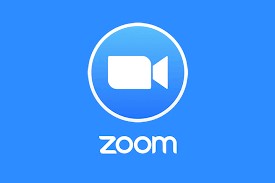
## Position Description

* This is not a paid position.
* The objective of this position is to help me to gain the necessary knowledge.
* DataDisca neither offers nor guarantees you a future employment opportunity in DataDisca or any other organization.
* There is no obligation on my side to follow the instructions or guidelines provided by DataDisca.
* DataDisca does not use unpaid trainees in proprietary projects. If I observe any violation of this term, I should immediately stop doing such assignment(s) and inform the Director of DataDisca.
* All completed work is to be published under appropriate open-source licenses. However, this excludes any proprietary software where your work is built on, proprietary data and their transformed forms.

## Training Places

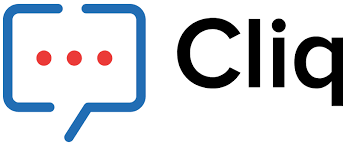
It is completely online and so the working place is work from home for trainers. Also, there isn’t any tough deadline. So trainers can work without stress and freely. Also, we used the GitHub individual branch for saving our works that were ongoing

## Meeting



It weekly meeting through zoom. As a communication tool, we used ZOHO Mail and Cliq tool in ZOHO. Company checks for the weekly progress of assigned works.





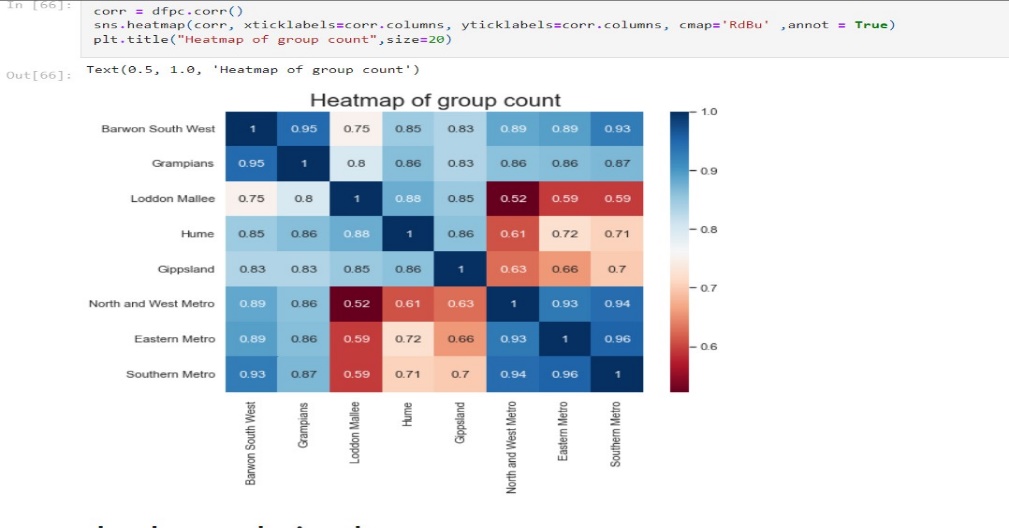
## Tasks and Activities

I was assigned to the following tasks during the training program

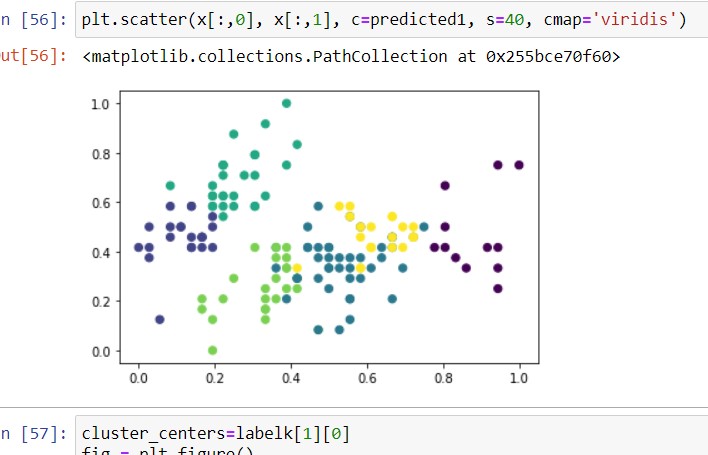
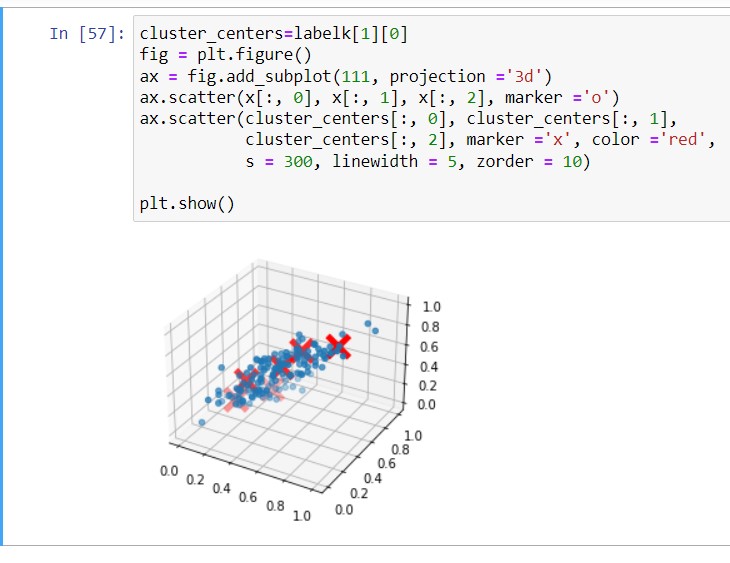
1. Analysis and visualization basics using Jupyter
2. Datasets clustering and algorithm with optimization methods
3. Datasets classification algorithm with optimization methods
4. Data visualization and dashboard using tableau.

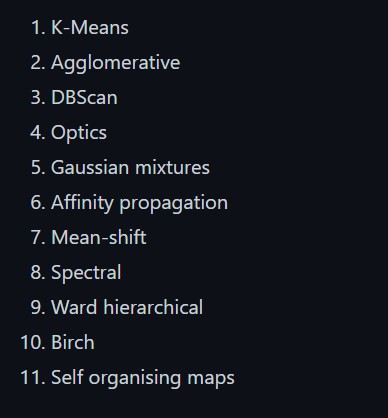
### Analysis and visualization basics using Jupyter

As my first project, I have to do analysis and visualization basics as an open-source project. Because there are numerous examples where the problems are solved with the right statistics and visualizations. So it became interesting to complete that. There I used a dataset provided by the company and after handling the data and Extracting data Using pandas and numpy libraries. Not only that I added the Pearson and Spearman correlation matrices between numerical columns as applicable. And showed the results on annotated heatmaps. Then I visualize the data using the plotly chart (Scatter, Pie, Bubble, Bar, Histogram, and Maps) All the works have been uploaded to GitHub as soon as the end of the day is finished.

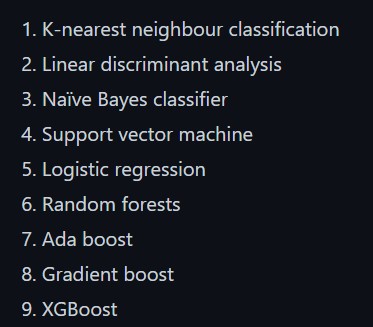


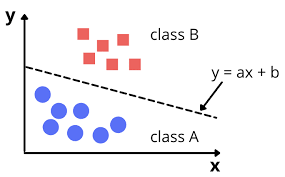
### Datasets clustering and algorithm with optimization methods

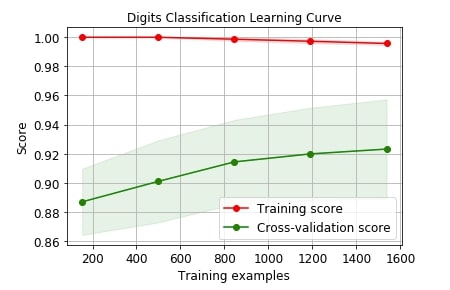
As the second task, using open-source OpenML several datasets, we clustered each dataset by the different skit-learn clustering algorithms. In there we studied python OpenML, data preparation, clustering algorithm, performance measure, Hyper-parameter-tuning, Joblib library and data visualization. Also learned several clustering algorithms. There are K-Means, Agglomerative, DBscan, Optics, Gaussian mixtures, Affinity-propagation, Mean-shift, Spectral, Ward hierarchical, Birch and Self organizing maps. In the code file, download a dataset using the Python OpenML package, Prepare data. For better results added Min-max normalize. Some joblib code to walk through the parameter grid to Record f1\_score, adjusted\_rand\_score, silhouette score and execution time against each parameter combination identified in previously mentioned algorithms. Using the result data, created tableau Dashboards or Plotly visualizations to analyze.



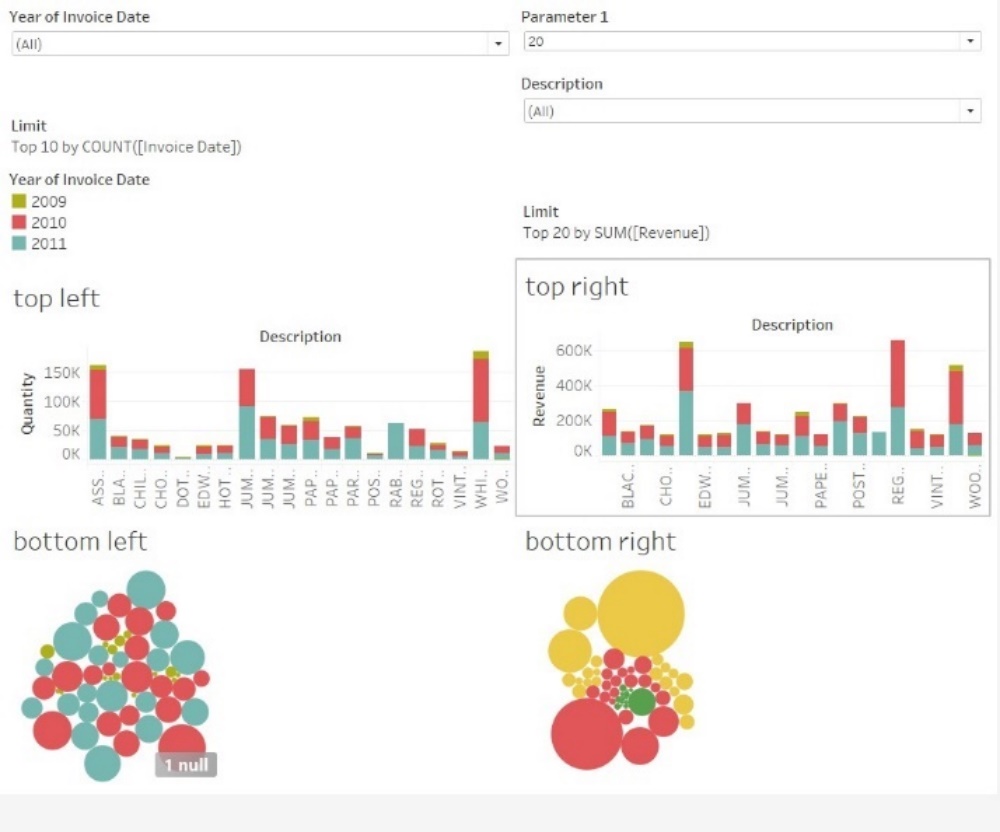
### Datasets classification algorithm with optimization methods

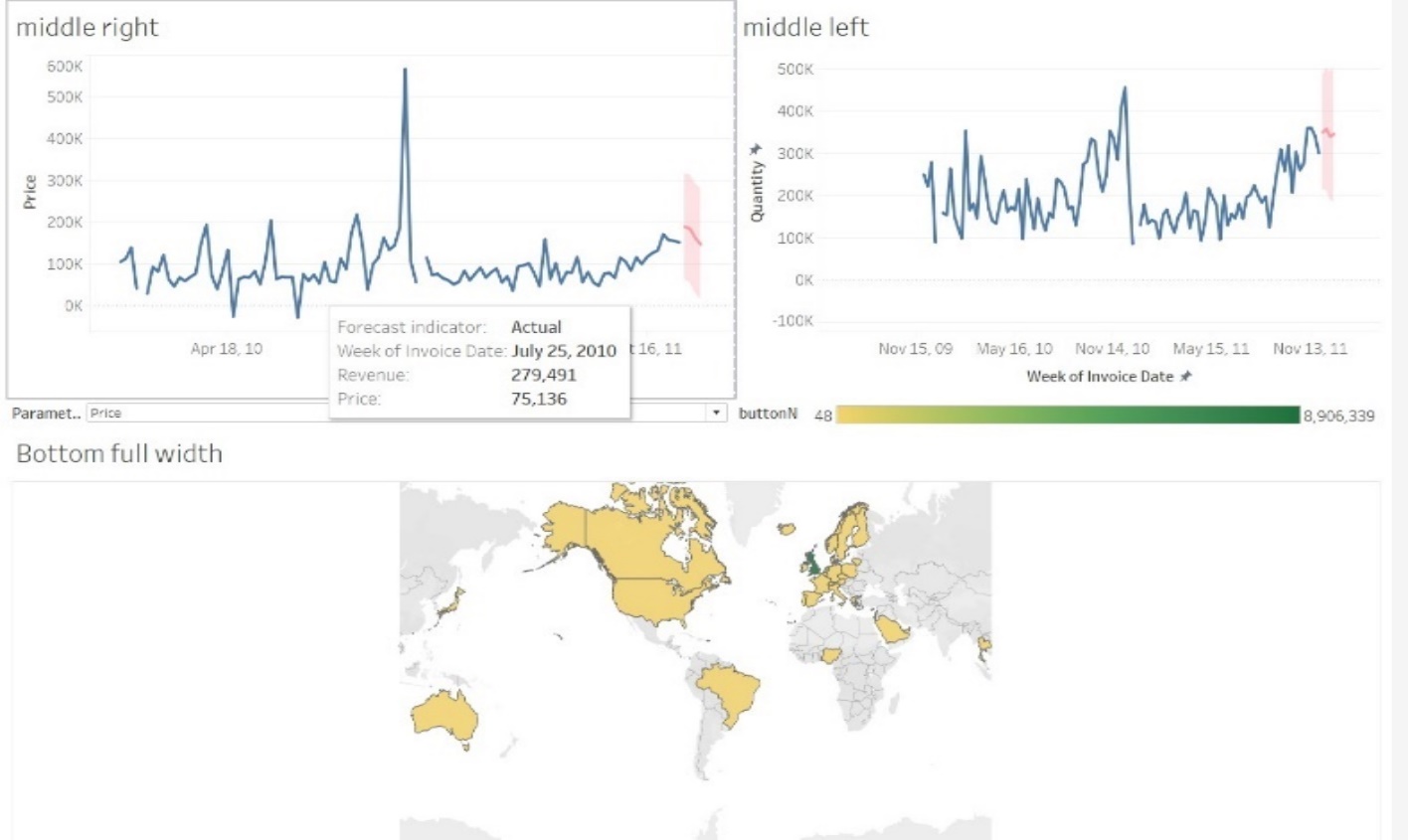




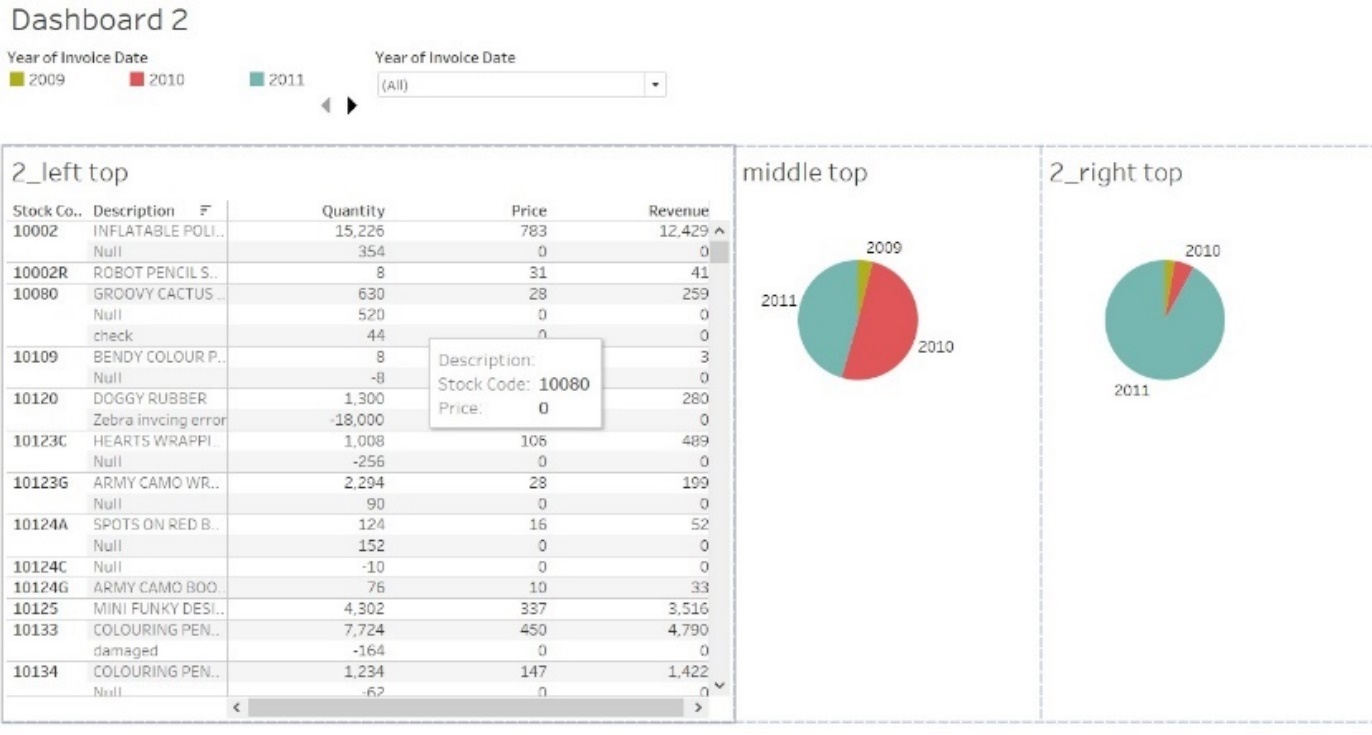


### Data visualization and dashboard using tableau

In the fourth month, to improve tableau dashboard knowledge, Using the online retail open dataset which is “Online Retail II”. using specific guidelines, I have to make the two tableau dashboards. This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers. It has several attributes. There are Invoice No, stock code, Description, Quantity, Unit price, Customer-ID, Country, Quantity and etc. 1 st dashboard has several sections. Selected at the top: The years 2009, 2010, 2011, 2012, or all. All is the default Top N: N = 10 default, 10, 20,50, and 100 are the possible values. Left Top: Top N products based on the quantity bar chart.

Below the plot: included what percentage of the total quantity covered by the top N. Tool-tip: Number, the percentage from the total quantity, percentage from the total price. Right Top: Top N products based on the total revenue bar chart. Also, I added below the plot percentage of the total price (revenue of the item) covered by the top N Tool-tip: Number, the percentage from the total quantity, percentage from the total price. The second dashboard created some tables including stock code, description, total prices order by revenue.

Also add a pie chart for quantities, a pie chart for revenue. Not only that I added a time-series price plot and quantity plot with one-year forecasting finally At the bottom of the second dashboard I created a world map based on price and quantity.



# Discussion and Conclusion

​

The internship program has changed my direction of thinking, approach of working and strategies of problem solving. I feel full utilization of this period. It may provide lots of confidence to a student. The learnings of this period such as regularity, punctuality, analytical ability and ability of judgments will favor me in my future carrier a lot.

From my internship period at Data Disca, I got to learn many things related to the Data Science.

* Knowledge

I have gained lot of new knowledge related to the data science field. Especially it makes me build confidence in this field.

* Self-Learning

After giving training lesson we have to learn things by ourselves to complete. The company provided reference links and material.

* Communication

All the progress meetings were done in English. So, In my training period, I have improved it a lot

# References

* <https://datadisca.com/>
* <https://www.linkedin.com/company/datadisca/mycompany/>