DATA MINING WITH PYTHON

TADEJ ROŠKARIČ, SAMO BOBEK
University of Maribor, Faculty of Economics and Business, Maribor, Slovenia
tadej.roskaric@student.um.si, samo.bobek@um.si

Abstract As the amount of data in the world is exponentially on the rise, we need all the tools and knowledge we can get to analyse this data and extract valuable information. This allows important stakeholders to make data-driven decisions, thus providing added value in any organisation. The data mining process can be applied in virtually all kinds of organisations ranging from the public to the private sector. Employees use data in their professional lives and therefore need to be familiar with the knowledge discovery process. The focus of this article is Python as a tool for data mining. The authors concluded that Python is a great option for this task since it is open-source, free and comes with a huge community that develops the packages needed for analytics workloads and it also has lots of documentation. Its capabilities are demonstrated at the end of this paper, where the authors have set up a case study relating to airline passenger satisfaction. The main approach is exploratory data analysis through visualisations with the goal of finding hidden patterns in the data. A decision tree machine learning model was also developed to extract the features that contribute to a higher satisfaction level.

Keywords:
data mining, Python, knowledge discovery process, data mining techniques, machine learning

University of Maribor Press

DOI https://doi.org/10.18690/um.epf.5.2022.32
ISBN 978-961-286-600-6

1 Introduction

In the digital era, large amounts of data are being produced. Due to its huge volume, high velocity and variety, the term Big Data came into play. In order to find interesting insights and information in these huge piles of data, it is necessary to analyse them. This process is called data mining. It is the process of extracting information from datasets and providing end-users with added value. It provides insights into potential trends, correlations, patterns and outliers in the given data (Prasdika & Sugiantoro, 2018).

Data mining tends to be relevant for all sectors of the economy. It is usually mentioned in the context of business applications, such as identifying potential customers. There are also a lot of use cases in healthcare, energy production and other industries (Sumiran, 2018).

Its usage is also well known in manufacturing since new technologies, such as the Internet of Things and Cyber-physical Production Systems, make real-time data acquisition possible (Huber, Wiemer, Schneider, & Ihlenfeldt, 2019).

There is a vast array of tools and software available for data mining. Some of them are open-source and free, such as Python, R, and Orange, while others are offered by big corporations under a commercial licence. Some of the most commonly mentioned commercial tools are MATLAB, SAS Enterprise Miner, and IBM SPSS Modeler (Mikut & Reischl, 2011).

The focus of this article is the Python open-source programming language and specifically its applications and usage in data mining. The capacities of Python will be demonstrated on a use case analysing airline passenger satisfaction data.

2 Data Mining

The data mining process makes information accessible to users by extracting knowledge from the given datasets. The term 'knowledge' in this context means the acquisition of important correlations, patterns and relations in the data. The term itself is deeply connected with disciplines such as statistics and the newest technologies such as artificial intelligence and machine learning. As such, data mining uses statistical methods and subdomains such as exploratory data analysis at

its core, while at the same time enabling users to analyse data that was not intentionally designed for statistical analysis, since formal procedures or hypothesis testing, which are seen in the field of statistics, are not required (Schuh, et al., 2019).

2.1 The knowledge discovery process

Data mining usually consists of several standard phases that form the so-called knowledge discovery process. This is a guideline to get from raw data to valuable insights (Sumiran, 2018).

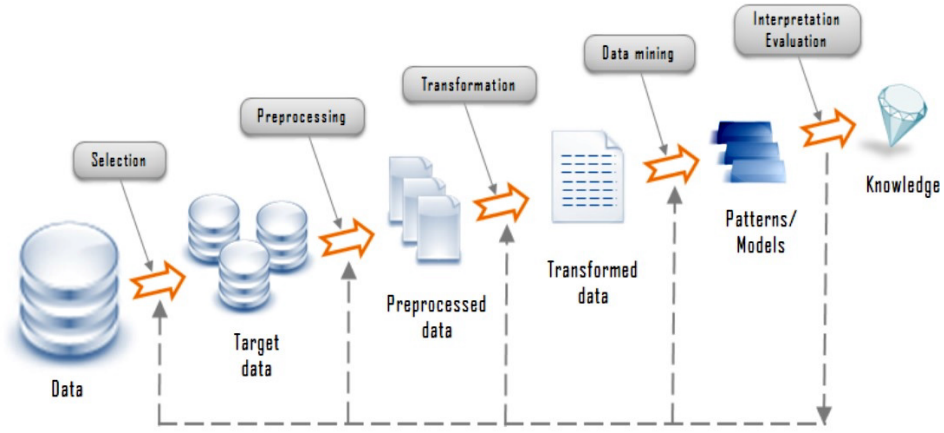

Figure 1: The knowledge discovery process
Source: (Sumiran, 2018)

The phases of the knowledge discovery process are shown in Figure 1 above. The first stage is the collection of raw data. Once this has been collected, certain parts of the data need to be selected that could be useful for solving the initial data mining problem. This data must then be preprocessed and cleaned to get rid of errors, inconsistencies, and missing values. Phase three is the transformation part, where the data is brought in an appropriate format to solve the data mining task. The next phase is the data mining itself. One or more techniques are applied with the main goal of discovering new patterns. Thereafter, it is a good practice to visualise and evaluate the findings in order to reach useful conclusions and complete the knowledge discovery process successfully. These insights can now be used to make data-driven decisions.

2.2 Common data mining techniques

Classification is a very common approach. The general idea is to classify each item into a predefined class or group (Sumiran, 2018).

In this method, a dataset of correctly pre-classified examples is used to develop a model for further classification of unknown examples. This technique is useful for problems such as fraud detection (Ramageri, 2010).

Another fundamental technique in data mining is association. This approach allows patterns to be derived from the frequent item sets in the data. This tends to be useful for discovering cross-marketing opportunities and analysing customer behaviour. So-called association rules are the output of this technique, however, there can be a lot for just one dataset, therefore it can be hard to filter out the ones that do not provide added value (Ramageri, 2010).

An additional technique is called regression, in which the relationship is modelled between a dependent variable, which is explained with the help of independent variables (Ramageri, 2010).

In a regression problem, the value trying to be predicted is numeric. An example of this is to predict house prices in the future (Sumiran, 2018).

Clustering is another method that focuses on identifying similar objects. It allows correlations and distribution patterns in data to be determined. A use case for clustering is to form groups of customers based on their purchasing history and look at the attributes they have in common (Ramageri, 2010).

3 Python

Python is a very useful and powerful programming language that is beginner-friendly due to its simple and easy-to-read syntax. The interpreter needed to develop code in this language and all the standard libraries is open-source and available free of charge. Python has a lot of different use cases ranging from software engineering, internet protocols and operating system interfaces. However, its usefulness does not stop at this point. The language has a vast variety of third-party modules that enable the use of Python in a lot of additional scenarios. These modules, which are also

called packages or libraries, can be found in the Python Package Index (PyPI) repository (Python Sofware Foundation, 2022a).

NumPy is a commonly used library that is the base for a lot of other data analysis modules. It provides the tools required to work with numerical data and is thus the core of scientific computing in Python since it also powers many other modules such as pandas, SciPy, scikit-learn and others. The package contains features such as multidimensional arrays, matrix operations and mathematical functions that can be used on these objects (NumPy Developers, 2022).

More information about NumPy is available in the paper 'Array programming with NumPy (Harris, et al., 2020).

Pandas is a very important package for this article. This library provides a huge data analysis toolkit. Its data structures enable users to work with relational and labelled data. Pandas is good at dealing with errors and missing data and also offers the possibility to split or merge datasets. Pandas has aggregation features such as the 'group by' functionality. The package allows users to load and read data very intuitively since it supports flat files, Excel files and databases. It also offers tools specifically needed for time series analysis (Python Software Foundation, 2022).

Further information about the pandas library is available in the article 'Data structures for statistical computing in Python' (McKinney, 2010).

An important thing to note when using packages are dependencies. A lot of the features in modules depend on other packages. Pandas, for example, has a direct connection with NumPy. This means that it is also necessary to install these packages in order to use the preferred module. Recommended dependencies can sometimes also be encountered that are complementary to the toolkit of a library. Pandas, for example, does not have any features that enable plotting and visualisations, therefore Matplotlib is an optional dependency to solve this (The pandas development team, 2022).

In the next chapter, the authors of this paper also used the scikit-learn library. This allows predictive data analysis with the help of pre-made machine learning algorithms. Scikit-learn includes functionalities such as classification, clustering, regression, dimensionality reduction (reducing the number of variables), model

selection (comparing models), and preprocessing (transformation of data) (Scikit-learn developers, 2022).

For an overview of the scikit-learn package and its features, you can read the paper 'Scikit-learn: Machine Learning in Python' (Pedregosa, et al., 2011).

Some papers also mention the subdomain of data mining called web mining. This functionality is enabled by the Beautiful Soup module. Its main purpose is to pull data from HTML and XML files and as such is a great tool for web scraping (Richardson, 2020).

4 Data mining with Python

This chapter focuses on the exploratory data analysis and development of a machine learning model for an airline. The data is based on passenger satisfaction surveys and can be found in the appendix of the paper titled 'Investigating airline passenger satisfaction: Data mining method', although the authors mention its origins as being the Kaggle machine learning and data science community (Noviantoro & Huang, 2021).

Apart from the libraries mentioned in the previous chapter, we will also be using Seaborn and Matplotlib for drawing visualisations.

For further reference on the Seaborn package, please refer to the paper 'Seaborn: Statistical data visualization' (Waskom, 2021).

Since visualising data relies heavily on Matplotlib, you should also give the publication 'Matplotlib: A 2D graphics environment' a read (Hunter, 2007).

Let's now put ourselves in the role of an airline company. Of course, the company's goal is to have satisfied customers, therefore it collects feedback from them in the form of satisfaction surveys. They are asked multiple questions mostly based on their flight experience and a general satisfaction evaluation which is either 'Satisfied' or 'Neutral/Dissatisfied'. The next stage of the process is to obtain insights into these people and see any potentially interesting patterns. This process is done with the help of some Seaborn visualisations in Python. The main focus is to look at the satisfaction ratios within different attributes. For example, 'How do customers rate

the flight based on the class in which they travelled?' The age factor can then be taken into account and younger and older passengers and their opinions can be compared.

Since the dataset contains 24 columns and around 130,000 rows, it is obviously not possible to extract all the available knowledge in this paper, since that would involve writing an entire paper on its own.

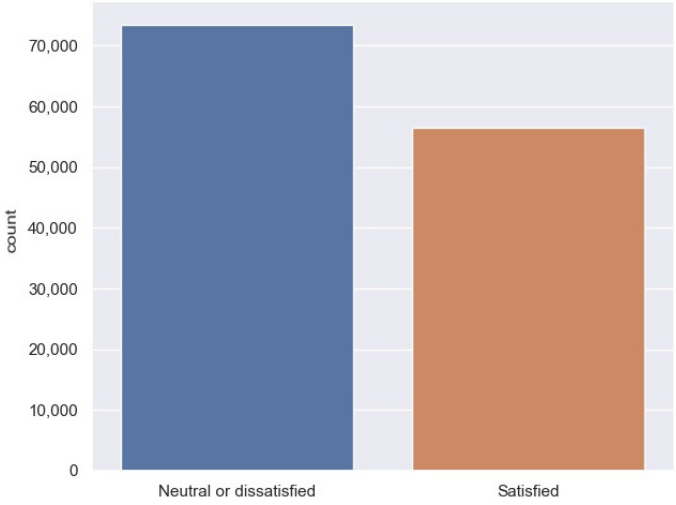

Figure 2: The count of Neutral/Dissatisfied and Satisfied passengers

Figure 2 shows the number of satisfied and neutral/dissatisfied customers. The number of satisfied customers is lower, albeit not significantly so, therefore the balance of the dataset is completely acceptable for future work.

The dataset is divided into male and female genders. There is a slightly higher ratio of females, but the difference is minimal. The satisfaction levels of both groups are practically identical, therefore there is no difference in the general rating of the flight between men and women. Passenger satisfaction can also be measured based on the class in which they travelled. There is a noticeable negative perception when it comes to economy and economy plus travellers since the service is probably not as high quality as in business class.

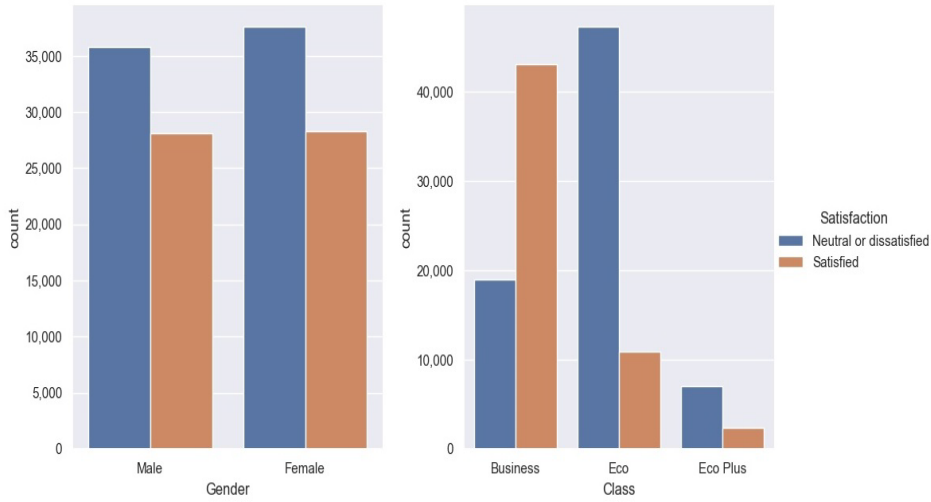

Figure 3: Satisfaction rating counts based on gender and travel class

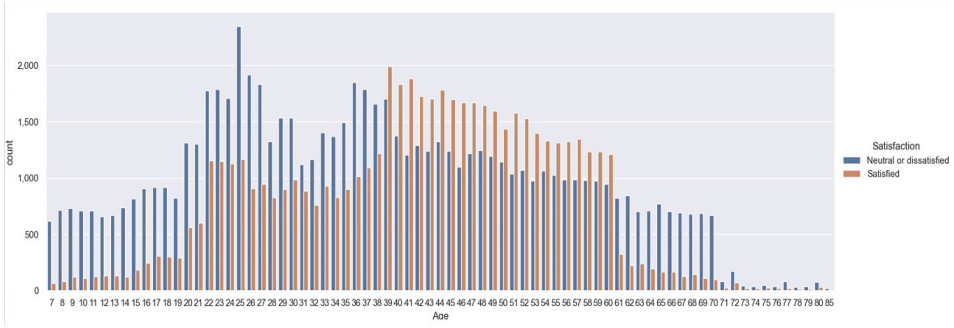

Figure 4: Satisfaction counts based on age

The figure above shows the age distribution of customers as well as their general opinion on the flight. The results show that middle-aged people tend to be less critical compared to their younger and older counterparts, since the satisfaction levels rise significantly at age 40 and also decrease quickly from age 60 onwards.

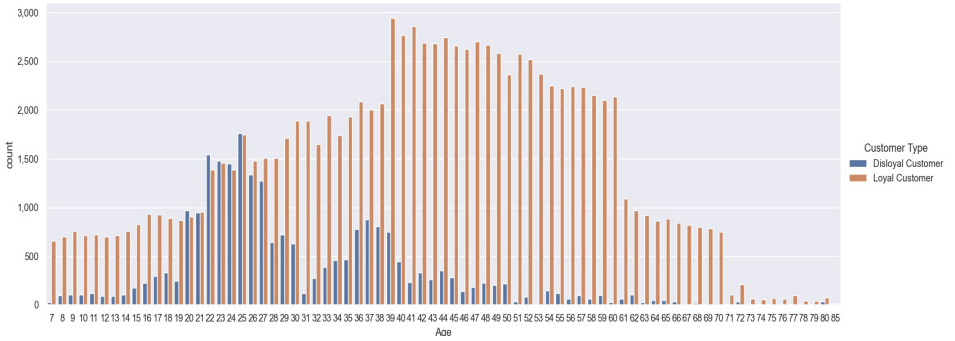

Figure 5: Number of loyal and disloyal customers and their age

In terms of age, it is also possible to check how loyal customers are. The loyal customers group tends to be larger. An interesting pattern that emerges is that younger passengers are on average less loyal than older ones.

After the visualisation part, further examination of the data can be carried out to find the importance of the attributes using a correlation matrix. For the purposes of this process, it is necessary to convert the variable types of some attributes from text to categorical/ordinal values. An example of this would be type of travel which has the possibility of 'Personal Travel' and 'Business Travel'. These values can be encoded as a 0 and a 1.

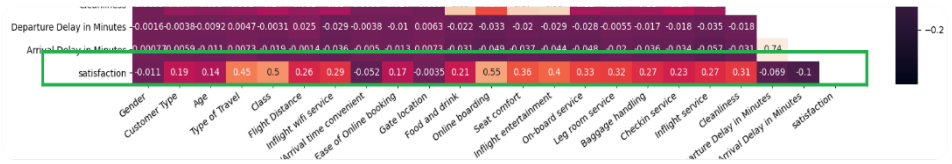

Figure 6: Lower part of the correlation matrix with the satisfaction attribute highlighted

Once the data transformation part has been completed, the lower part of the correlation matrix can be displayed. The Spearman correlation was used for this calculation. According to the visualisation, the most important attributes for satisfaction tend to be class, type of travel, online boarding and inflight entertainment.

However, one issue arises. Since some variables (such as satisfaction where 1 = 'Satisfied' and 0 = 'Neutral/Dissatisfied') are categorical while others are ordinal (for example the Wi-Fi service is labelled as a score from 1-5), the results of the correlation matrix might not be the most reliable source of information. In order to address this issue, the authors created a decision tree machine learning model to classify passengers into satisfied and dissatisfied groups and to then extract the attributes (in this case features) that have the biggest impact on the evaluation.
All that remained was to create the model with the help of the scikit-learn library. The dataset was split into training and test data. The x and y values of the training data were input into the model (with x being the independent variables and y the target variable – customer satisfaction level). Once the model had been created, it was easy to use the .predict() method and to assign the independent variables of the test set as the parameter to predict the y values in the test set. Once complete, the output was documented in the form of a confusion matrix to evaluate the results.

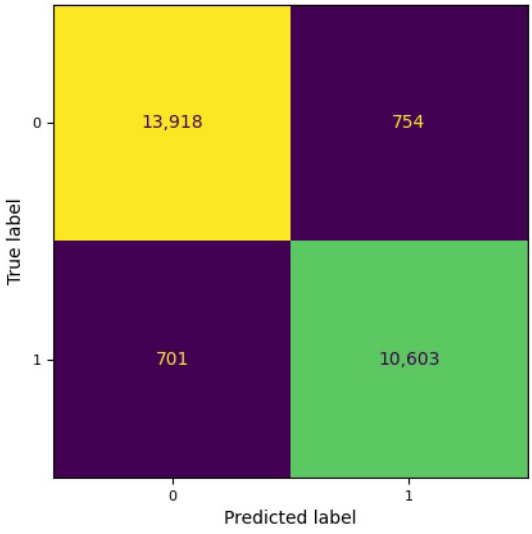

Figure 7: Confusion matrix of decision tree

The confusion matrix displays the output of the machine learning model. The top left quadrant are the true negatives while the true positives are shown in the lower right quadrant. The violet quadrants are misclassifications. There is a relatively small number of misclassifications compared to the whole test dataset, therefore the accuracy, recall and precision of the model are high.

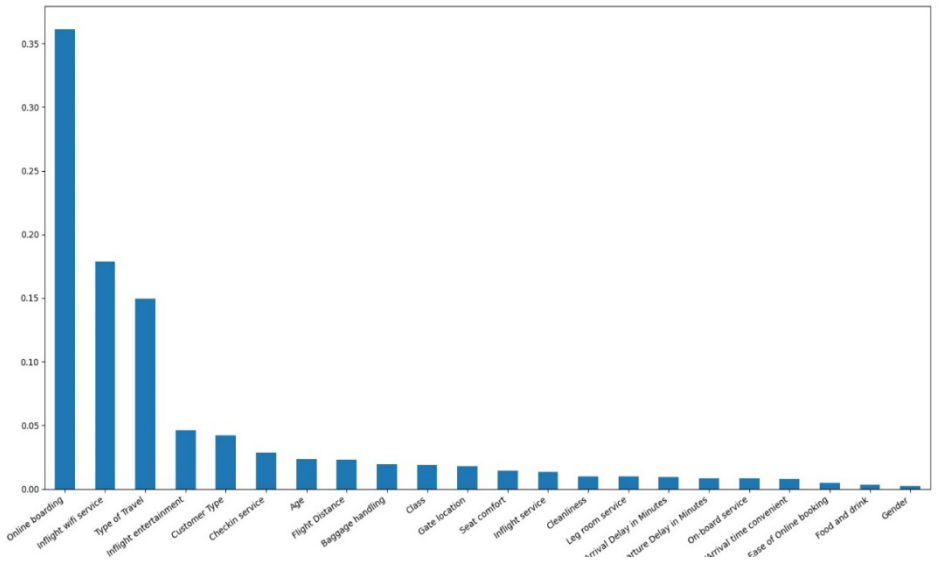

Figure 8: Feature importance of the decision tree model

The final task is to display the features that the model uses to classify the objects into groups and place those attributes on a scale to find the most significant ones. According to the model, the most important attributes are online boarding, inflight Wi-Fi service, and type of travel (business or private travel). The task of the airline at this point would be to improve the aspects on which they have an influence. For example, they cannot control whether or not a passenger flies business class, but they can improve the online boarding experience if it increases satisfaction.

5 Conclusion

This paper focuses on data mining and the knowledge discovery process. The authors defined all the necessary terminology and looked at some common techniques that are used by data mining experts. There is a wide range of tools that can be used to accomplish the same data mining task, but for the purposes of this research the authors concentrated on the Python programming language. The main advantage of Python is that it is free and it has a huge community of developers with a lot of documentation to make the learning process easier. While doing data mining with Python is a 'hard coding' approach, the syntax tends to be easy to read and, with the help of libraries, a lot of the work is already done for the user.

The second part of this paper is empirical. Based on the theoretical concepts of the knowledge discovery process, the classification data mining technique, and the Python packages that were described, the authors set up a case study to gain new insights into the customers of an airline. It was found that younger passengers tend to be less loyal and more critical of the services, which is also the case for economy and economy plus travellers. This makes sense, since business class usually offers the best service, therefore customers are more satisfied as a consequence. The results show that online boarding, inflight wi-fi service and type of travel tend to be the most important factors when it comes to the overall opinion of the flight. Therefore, the output of this research is that some patterns were discovered, and a machine learning model was used to predict traveller satisfaction and also help discover important areas that can be improved to make the customer experience as enjoyable as possible.

An idea for further research would be to do an expanded exploratory data analysis to discover more relationships in the data. Additional machine learning models could also be produced in order to compare the results, which would provide greater reliability and accurately predict the most important features of a satisfied passenger.

References

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 357-362. doi:10.1038/s41586-020-2649-2

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic module to the CRISP-DM model. *Procedia CIRP*, 403-408. doi:https://doi.org/10.1016/j.procir.2019.02.106

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 90-95. doi:10.1109/MCSE.2007.55

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, (pp. 56-61). doi:10.25080/Majora-92bf1922-00a

Mikut, R., & Reischl, M. (2011). Data mining tools. *WIREs data mining and knowledge discovery*, 431-443. doi:10.1002/widm.24

Noviantoro, T., & Huang, J.-P. (2021). Investigating airline passenger satisfaction: Data mining method. *Reasearch in Transportation Business & Management*. doi:https://doi.org/10.1016/j.rtbm.2021.100726

NumPy Developers. (2022). *NumPy: the absolute basics for beginners*. Retrieved from NumPy documentation: https://numpy.org/doc/stable/user/absolute_beginners.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Prasdika, F. B., & Sugiantoro, B. (2018). A Review Paper on Big Data and Data Mining Concepts and Techniques. *International Journal on Informatics for Development*, 33-35. doi:10.14421/ijid.2018.07107

Python Software Foundation. (2022). *pandas*. Retrieved from Python Packages Index: https://pypi.org/project/pandas/

Python Sofware Foundation. (2022a, March 20). *General Python FAQ*. Retrieved from Python documentation: https://docs.python.org/3/faq/general.html

Ramageri, B. M. (2010). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering*, 301-305. Retrieved from https://www.researchgate.net/publication/49616224_Data_mining_techniques_and_applications

Richardson, L. (2020). *Beatiful Soup Documentation*. Retrieved from Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Schuh, G., Reinhart, G., Prote, J.-P., Sauermann, F., Horsthofer, J., Oppolzer, F., & Knoll, D. (2019). Data Mining Definitions and Applications for the Management of Production Complexity. *Procedia CIRP*, 874-879. doi:https://doi.org/10.1016/j.procir.2019.03.217

Scikit-learn developers. (2022). *Home*. Retrieved from scikit-learn: https://scikit-learn.org/stable/index.html

Sumiran, K. (2018). An Overview of Data Mining Techniques and Their Application in Industrial Engineering. *Asian Journal of Applied Science and Technology*, 947-953. Retrieved from https://www.researchgate.net/publication/326098515_An_Overview_of_Data_Mining_Techniques_and_Their_Application_in_Industrial_Engineering

The pandas development team. (2022). *Installation*. Retrieved from pandas documentation: https://pandas.pydata.org/pandas-docs/stable/getting_started/install.html

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 3021. doi:10.21105/joss.03021