Rick van Ravestein (Student)

HAN University of Applied Sciences

19-12-2023



**MDD Individual Assignment  
A Portfolio showcasing Data Science Analytics Skills**

**Title:**  MDD Individual Assignment

**Sub-title:** A Portfolio showcasing Data Science Analytics Skills

**Student:**  Rick van Ravestein   
 1639258   
 E: [DJ.vanRavestein@student.han.nl](mailto:DJ.vanRavestein@student.han.nl)

**Study Background:** Business Administration

**Class:** MDD-01

**Educational Institution**:  HAN University of Applied Sciences

 Ruitenberglaan 26

 6826 CC Arnhem   
 T: 024 353 0500

**Lecturer:**  John Smits

 E: [John.Smits@han.nl](mailto:John.Smits@han.nl)

**Date:** 19-12-2023

**Place:** Arnhem

**Version:**  1.0

**Preface**

Welcome to my Data Science Analytics Portfolio, the culmination of my journey through the MDD Minor Programme. This collection reflects the application of acquired skills in real-life projects, demonstrating proficiency in the CRISP-DM process and data science tools. The projects showcase my ability to solve business problems independently, apply diverse data mining techniques, and communicate effectively.

**Summary**

This project focused on analyzing a Nike shoe sales dataset using Python, with the aim of providing valuable insights for Nike and other stakeholders. The objectives included identifying popular Nike shoes, understanding customer preferences, and offering information for informed decision-making by consumers and retailers.

The dataset, sourced from various platforms, offered comprehensive details on shoe models, prices, reviews, and ratings. The methodology followed the CRISP-DM model, encompassing key phases like Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Key findings from the analysis included the identification of consumer-favorite shoes, insights into features valued by customers through spaCy-based Natural Language Processing (NLP) analysis and showcasing the top-rated Nike shoes based on customer reviews.

The regression analysis examining the relationship between sale price and customer ratings revealed a minimal association, with a low R-squared value (0.008) and a small positive coefficient for sale price.

Deployment recommendations encompassed the integration of spaCy for continuous textual analysis, strategic promotion of popular and highly rated shoes in marketing campaigns and emphasizing customer-valued features in product descriptions.

In conclusion, these insights position Nike to make informed decisions, enhance customer satisfaction, and optimize marketing strategies. The deployment of these recommendations ensures Nike's adaptability and competitiveness in a dynamic market.

Table of Contents

[Chapter 1: Introduction 5](#_Toc153832666)

[1.1: Nike Shoes Sales Python Analysis 5](#_Toc153832667)

[1.2: CRISP-DM 5](#_Toc153832668)

[Chapter 2: Project Phases 6](#_Toc153832669)

[2.1: Business Understanding 6](#_Toc153832670)

[2.2: Data Understanding 6](#_Toc153832671)

[2.3: Data Preparation 7](#_Toc153832672)

[2.4: Modelling 8](#_Toc153832673)

[2.5: Evaluation 12](#_Toc153832674)

[2.6: Deployment 13](#_Toc153832675)

[Bibliography 14](#_Toc153832676)

# Chapter 1: Introduction

## 1.1: Nike Shoes Sales Python Analysis

Throughout this minor, I acquired diverse skills in analyzing datasets with Python. The methodologies implemented in this project were assimilated through the courses FMSR-Foundation Mathematics & Statistics in Python and FIML-Introduction to Modelling. The objective of this project is to derive valuable insights for Nike and other footwear brands, as well as for consumers, shoe enthusiasts, and retailers.

**Objectives**

* By analyzing this dataset, the researcher wants to gain insight into which Nike shoes are most popular among consumers, as well as the features and characteristics that customers value most in a shoe. This information could be used to inform future product design and marketing strategies for Nike and other shoe brands.
* Furthermore, the researcher wants to inform consumers and shoe enthusiasts so that they can benefit from this dataset to make more informed decisions about purchasing Nike shoes based on the experiences of other customers.
* The dataset will also be used to advice retailers on how to optimize their inventory and marketing strategies based on popular shoe models.

**About Dataset**

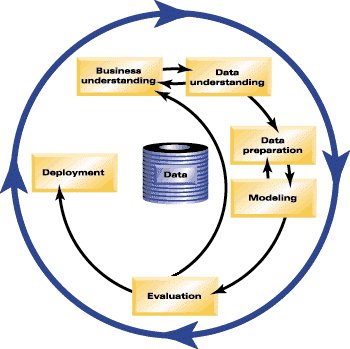
Link: <https://www.kaggle.com/datasets/ulrikthygepedersen/nike-shoes-sales>

Nike is one of the largest and most recognizable athletic shoe brands in the world, known for their high-quality products and innovative designs. To gain insight into how Nike's shoes are received by consumers, a dataset has been compiled that contains information on Nike shoe sales, including customer reviews and ratings (Pedersen, 2023).

The dataset includes information on Nike shoes sold across a variety of platforms, including Nike's own website, Amazon, and other retailers. It contains data on the shoe model, colorway, size, price, and the number of reviews and ratings. Additionally, the dataset contains text data on customer reviews, allowing for a detailed analysis of consumer opinions and feedback (Pedersen, 2023).

## 1.2: CRISP-DM

During this project the different phases of the CRISP-DM model will be carried out. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide data mining efforts (IBM, n.d.).



*Figure 1: CRISP-DM Model (IBM, n.d.)*

# Chapter 2: Project Phases

## 2.1: Business Understanding

This phase of the CRISP-DM model is about understanding the business and exploring what the organization can gain from data analysis. Getting to know the business reasons for data analysis efforts helps to ensure that everyone is on the same page before expending valuable resources (IBM, n.d.)

Nike, Inc. is a renowned American multinational corporation specializing in the design, development, manufacturing, and global marketing of apparel, footwear, accessories, equipment, and services. Nike is a leading producer of sports equipment, ranking among the world's largest suppliers of athletic shoes and apparel. With a workforce exceeding 44,000 people globally, Nike boasts a brand value of $19 billion as of 2014, making it one of the most valuable names in the sports industry. The company markets its products under various brands, including Nike Pro, Nike+, Nike Golf, and subsidiaries such as Jordan, Hurley Int., and Converse. Known for iconic trademarks like "Just Do It" and the Swoosh logo, Nike sponsors high-profile athletes and sports teams worldwide (NIKE COMPANY, 2015).

Analyzing the data allows Nike to identify the most popular shoe models and understand customer preferences, informing future design and marketing strategies. This data-driven approach benefits not only Nike but also assists retailers in optimizing inventory and marketing based on popular models. Additionally, sharing insights with consumers empowers them to make informed purchasing decisions, strengthening the brand's connection with its customer base.

## 2.2: Data Understanding

The data understanding phase of CRISP-DM model involves taking a closer look at the data available for analysis. This step is critical in avoiding unexpected problems during the next phase, data preparation, which is typically the longest part of a project (IBM, n.d.).

Columns in Dataset Explained:

**Product Name:** The name of the product (shoe model).

**Product ID:** Each shoe has its own unique product ID, which is displayed in this column.

**Listing Price:** A shoe’s listing price is what it is listed for on the market. Most of the records in this column contain the value ‘0’, which is ‘no value’. In that case the is no difference between the listing price and the sales price. The variable listing price is expressed in US Dollars.

**Sale Price:** A shoe’s sale price is what the shoe sells for on the market, also expressed in US Dollars.

**Discount:** The amount of discount that is been given on a shoe, expressed in US Dollars. This variable only contains the value ‘0’, which means that there are no discounts in this data.

**Brand:** The brand of the shoe. The only value in this column is ‘Nike’.

**Description:** The description of the shoe.

**Rating:** What customers rated the shoe from one to five. If rating is ‘0’, that means that specific customer did not rate the shoe after purchasing it.

**Reviews:** This column should contain text data of customer reviews, but unfortunately it does not. Therefore, this variable will not be used for analysis.

**Images:** This column contains URLs of the images of the products.

## 2.3: Data Preparation

Data preparation is one of the most important and often time-consuming aspects of data analysis. Depending on the organization and its goals, data preparation typically involves the following tasks (IBM, n.d.):

* Merging data sets and/or records
* Selecting a sample subset of data
* Aggregating records
* Deriving new attributes
* Sorting the data for modelling
* Removing or replacing blank or missing values
* Splitting into training and test data sets

For this research, the following data preparation tasks were carried out:

**Checking for and Deleting Missing Data**

To check for missing data and to delete observations with missing data, the functions isna() and drop.na() from the Pandas package were used.

Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijving

*Figure 2: The count of missing values per column (Own Work, 2023)*

After the observations with missing values were dropped, the shape of the cleaned Dataframe was printed:

Shape of the cleaned DataFrame: (569, 10)

**Dropping the 'images' Column**

Since the ‘images’ variable was not going to be used for analysis, the researcher decided to drop it.

**Identifying and Removing Outliers Using the IQR Method**

Some observations within a set of data may fall outside the general scope of the other observations. Such observations are called outliers. The IQR method of identifying outliers is used to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence 1.5 times the IQR is taken and then subtracted from Q1, then this value is added to Q3. This gives the minimum and maximum fence posts that each observation is compared to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers (Pennsylvania State University, n.d.).

With the use of the IQR method, 107 outliers were identified:

Afbeelding met tekst, Lettertype, lijn, schermopname

Automatisch gegenereerde beschrijving

*Figure 3: Identified outlier (Own Work, 2023)*

After the outliers were identified, they were removed from the Dataframe.

## 2.4: Modelling

At this point the prepared data is brought into the analysis tools, and the results begin to shed some light on the business problem (IBM, n.d.).

**Most Popular Nike Shoes Among Consumers**

To make a visualization of the most popular Nike shoes among consumers, a bar plot of the top 10 most popular Nike shoes was made:

Afbeelding met tekst, schermopname, Lettertype, diagram

Automatisch gegenereerde beschrijving

*Figure 4: Most popular Nike shoes bar plot (Own Work, 2023)*

As can be seen in the figure above, the Nike Air Max 90 is the most popular Nike shoe, followed by the Air Max 270 React and the Air Max 97.

**Features and Characteristics that Customers Value Most in a (Nike) Shoe**

The Column “description” contains textual data that describes the shoe models. The researcher analyzed this data to discover which features and characteristics customers value most in a shoe. This information will be used to inform future product design and marketing strategies for Nike and other shoe brands. To analyze the data of the “description” column the researcher applied a natural language processing (NLP) technique, using the NLP library spaCy.

Natural Language Processing (NLP)

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written, referred to as natural language. It is a component of artificial intelligence (AI) (Lutkevich & Burns, 2023).

spaCy

spaCy is an open-source natural language processing (NLP) library for Python that provides efficient and fast tools for processing and understanding human language. It offers pre-trained models for tasks such as tokenization, part-of-speech tagging, named entity recognition, and dependency parsing (spaCy, n.d.).

The following result came forth from the text analysis:

Top Key Phrases:

you: 225

that: 219

it: 116

your foot: 77

It: 59

comfort: 50

traction: 35

style: 33

the Nike Air Max: 30

the court: 30

The word "comfort" appears 50 times. This indicates that customers value the comfort aspect of Nike footwear.

The word "traction" appears 35 times. This suggests that customers value traction in Nike sports shoes.

The word "style" appears 33 times. This indicates that customers value the aesthetic or design aspects of Nike footwear.

To visually represent the frequency of words in the “description” variable, the researcher generated a word cloud in Python. This word cloud provides insights into the most common terms used in product descriptions.

Afbeelding met tekst, Lettertype, schermopname, grafische vormgeving

Automatisch gegenereerde beschrijving

*Figure 5: Word cloud for product descriptions (Own Work, 2023)*

**Customer Ratings on Nike Shoes**

To make a visualization of the best Nike shoes based on customer ratings, a bar plot of the top 10 Nike shoes based on the average of customer ratings was made:

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

*Figure 6: Customer ratings on Nike shoes bar plot (Own Work, 2023)*

As can be seen in the figure above, the Nike Air Force 1 ’07 LV8 is the best Nike shoe based on customer ratings, followed by the Air Max 2090 and the Metcon 5 AMP.

**Relationship Between Sale Price and Customer Ratings**

To examine the relationship between "sale\_price" and "rating" the researcher used the OLS regression model:

Afbeelding met tekst, schermopname, menu, nummer

Automatisch gegenereerde beschrijving

*Figure 7: OLS Regression sale price and rating (Own Work, 2023)*

The following can be concluded based on the results in figure 7:

* The R-squared value is 0.008, indicating that only 0.8% of the variability in "rating" can be explained by the variability in "sale\_price."
* The coefficient for "sale\_price" is 1.741e-05, suggesting a very small positive association with "rating."
* The p-value ia 0.064, indicating that the overall model may not be statistically significant at the conventional significance level.

In summary, the model explains a very small portion of the variability in "rating," and the coefficient for "sale\_price" is not statistically significant at the 0.05 level.

## 2.5: Evaluation

At this point, the results of the data mining efforts are evaluated. This is the key to ensuring that organizations can make use of the obtained results. Two types of results are produced by data mining (IBM, n.d.):

* The final **models** selected in the previous phase of CRISP-DM model.
* Any conclusions or inferences drawn from the models themselves as well as from the data mining process. These are known as **findings**.

**Models**

Nike can benefit from continuing to use the natural language processing (NLP) library spaCy for several reasons:

* spaCy allows Nike to effectively analyze and extract valuable information from textual data, such as customer reviews and product descriptions in the dataset. This can provide insights into customer sentiments, preferences, and key features that influence purchasing decisions.
* By leveraging spaCy, Nike can gain a deeper understanding of the textual content in the "description" column. This analysis can inform product development strategies by identifying features and characteristics that customers value most in a shoe. Additionally, marketing strategies can be optimized based on customer sentiments and language used in reviews.
* Natural language processing allows Nike to compare its product descriptions, customer reviews, and overall language use against competitors.

The researcher recommends not to continue using the regression model for sale price and rating. The model explains a very small portion of the variability in "rating," and the coefficient for "sale\_price" is not statistically significant at the 0.05 level.

**Findings**

Most Popular Nike Shoes Among Consumers

Utilizing a bar plot of the top 10 most popular Nike shoes, the analysis revealed that the Nike Air Max 90, Air Max 270 React, and Air Max 97 are the most favored among consumers.

Features and Characteristics Valued by Customers

Leveraging natural language processing (NLP) through spaCy, the researcher analyzed the "description" column to identify key phrases. Customers highly value aspects such as comfort, traction in sports shoes, and the aesthetic or design elements of Nike footwear.

Customer Ratings on Nike Shoes

A bar plot of the top 10 Nike shoes based on average customer ratings showcased the Nike Air Force 1 ’07 LV8 as the highest-rated, followed by the Air Max 2090 and Metcon 5 AMP.

Relationship Between Sale Price and Customer Ratings

Using the OLS regression model, the analysis examined the relationship between "sale\_price" and "rating." The results indicated a very small portion of variability in "rating" explained by "sale\_price," with the coefficient not statistically significant at the conventional level.

## 2.6: Deployment

Deployment is the process of using new insights to make improvements within organizations. This can mean a formal integration such as the implementation of a model producing churn scores that are then read into a data warehouse. Alternatively, deployment can mean that the insights gained from data mining are used to elicit change in organizations (IBM, n.d.).

In general, the deployment phase of CRISP-DM model includes two types of activities (IBM, n.d.):

* Planning and monitoring the deployment of results.
* Completing wrap-up tasks such as producing a final report and conducting a project review.

Deployment recommendations for Nike:

Integration of spaCy for Textual Analysis

**Implementation:** Nike should integrate spaCy into its analytics infrastructure for ongoing analysis of textual data, including customer reviews and product descriptions.

**Benefits:** By integrating spaCy in its analytics infrastructure Nike gains continuous insights into customer sentiments, preferences, and key features influencing purchasing decisions. This information should be used for agile adjustments in product development and marketing strategies.

Promotion of Most Popular Shoes

**Marketing Strategy:** Nike should act on the insight that the Nike Air Max 90, Air Max 270 React, and Air Max 97 are the most popular shoes among consumers.

**Implementation:** Nike should feature these shoes prominently in marketing campaigns and optimize its inventory based on their popularity to meet customer demand effectively.

Product Feature Emphasis

**Feature Highlighting:** Nike should emphasize key features identified through spaCy analysis, such as comfort, traction, and aesthetic design.

**Implementation:** Nike should integrate these features into marketing materials and product descriptions to communicate value propositions more effectively to customers.

Customer-Rated Shoes Promotion

**Marketing Strategy:** Nike should capitalize on customer ratings to promote the highest-rated shoes, including the Nike Air Force 1 ’07 LV8, Air Max 2090, and Metcon 5 AMP.

**Implementation:** Nike should highlight positive reviews and ratings in marketing materials, potentially offering promotions or exclusives for these top-rated shoes.

# Bibliography

IBM. (n.d.). *IBM SPSS Modeler*. Retrieved December 6, 2023, from <https://www.ibm.com/docs/zh/spss-modeler/18.0.0?topic=dm-crisp-help-overview>

Lutkevich, B., & Burns, E. (2023, January 20). *natural language processing (NLP)*. TechTarget | Enterprise AI. Retrieved December 18, 2023, from <https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP#:~:text=Natural%20language%20processing%20(NLP)%20is,of%20artificial%20intelligence%20(AI)>.

NIKE COMPANY. (2015, June 2). *ABOUT US*. Retrieved December 11, 2023, from <https://nikecompanyblog.wordpress.com/about-us/>

Pedersen, U. T. (2023, March 1). *Nike Shoes Sales*. Kaggle. Retrieved November 29, 2023, from <https://www.kaggle.com/datasets/ulrikthygepedersen/nike-shoes-sales>

Pennsylvania State University. (n.d.). *3.2 - Identifying outliers: IQR Method | STAT 200*. PennState: Statistics Online Courses. Retrieved December 12, 2023, from <https://online.stat.psu.edu/stat200/lesson/3/3.2>

spaCy. (n.d.). *SPACY · Industrial-strength Natural language processing in Python*. Retrieved December 18, 2023, from <https://spacy.io/>