Market Dynamics of Refurbished iPhones: A Web Scraping and Data Analysis Approach

Nico Swart

Layout: typeset by the author using LaTeX.
Cover illustration: Nico Swart

Market Dynamics of Refurbished iPhones: A Web Scraping and Data Analysis Approach

Nico Swart
14034034

Bachelor thesis
Credits: 18 EC

Bachelor *Informatiekunde*

University of Amsterdam
Faculty of Science
Science Park 900
1098 XH Amsterdam

Supervisor
Dr. Robin Langerak

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 900
1098 XH Amsterdam

Semester 2, 2023-2024

Abstract

The refurbished iPhone segment operates in a volatile market, where prices are influenced by factors such as technological advances, new product releases, and changing consumer preferences. These fluctuations can pose challenges to sellers for creating an effective pricing strategy. The goal of this thesis is to gain market trend insights for refurbished iPhones by analysing and predicting prices of pre-owned devices on the Marktplaats platform. The research employs a comprehensive dataset collected from scraping Marktplaats by utilizing Selenium. This research employs moving averages and linear regression for price evaluation and trend analysis. An ARIMA model is used to make price predictions. A grid search method is utilized to identify the optimal model parameters (p, d, q) for each iPhone model. The model's accuracy is evaluated using the Mean Absolute Error (MAE) and Standardized Mean Absolute Error (sMAE). The results indicate that the moving average method effectively provides a stable price valuation for iPhone models, with a standard deviation for all models ranging between €1.45 and €7.69. A growth ratio is provided to indicate the trend of depreciation. The effectiveness of the ARIMA model is questionable due to its variability in performance across different iPhone models, making it less reliable for accurate market trend predictions.

Table of Contents

1. Introduction

In recent years, the market for refurbished electronics has seen a significant increase in size. Refurbished Smartphones in particular seem in the interest of consumers. In the year 2021 only, the refurbished smartphone market in particular saw an increase of 10% compared to the previous year (Counterpoint, 2022). There are multiple factors that make the refurbished smartphone market interesting for the consumer, the primary reason for choosing a refurbished device is the affordability. By thoroughly testing and cleaning used or returned items, refurbishing businesses can present high-quality, performance-guaranteed products that align well with consumer budgets. Besides these economic benefits, the second biggest driver towards purchasing refurbished smartphones are environmental benefits buying pre-owned has over purchasing a brand new device (Ademe, 2022).

Although it is clear the prices of refurbished devices lay lower than brand new devices, the exact price of these devices fluctuate over time. Not only do the devices differ on their own, with for example the different models, state of the phone and capacity, which affect prices. But also technological advances, new product releases, and changing consumer preferences have an effect on the price of such devices. This volatility poses a challenge for sellers who must navigate these fluctuating market to set competitive yet profitable pricing strategies (Hapuli, 2023). Dynamic pricing may offer a solution, as a dynamic pricing strategy has proven to be effective in volatile markets (Jager & Janssen, 2009). Dynamic pricing enables prices to adjust based on market conditions such as demand, inventory levels, and competitors' actions. Studies indicate that a 1% improvement in pricing can lead to a 12.5% rise in profits (Zuora, 2024).

To gain insights on the market dynamics of refurbished iPhones this thesis will study the prices of pre-owned devices being sold on the marketplace Marktplaats, which is the largest online marketplace of The Netherlands. With more than 10 million active monthly users this marketplace is often used by consumers to sell their used iPhone (Bright, 2023). Marktplaats operates as a free market where prices naturally follow market trends, making it an ideal platform to study the price dynamics of pre-owned devices. By focusing on data related to iPhone models, specifically storage capacity and pricing, extracted through web scraping techniques, this research aims to analyze and interpret market trends. This analysis will serve as the foundation for proposing a dynamic pricing model that could help sellers of refurbished smartphones optimize their pricing strategies.

By enabling sellers to price refurbished iPhones more effectively the project supports economic sustainability within the secondary market. It allows technology access at more affordable prices. This broadens the demographic that can afford these devices. Therefore it promotes technological inclusivity. Environmentally, better pricing models and enhanced market efficiency encourage the reuse of electronic devices. This contributes to reduction in electronic waste. This reuse is increasingly important as the global push towards sustainability intensifies (Bruyninckx, 2021). Electronic waste reduction is a critical component of environmental strategies (Hischier & Böni, 2021). Improved pricing strategies not only benefit consumers and sellers. They also support broader environmental goals by promoting the circular economy.

This research is positioned at the intersection of economic theory and data science, utilizing advanced analytical techniques to address a pressing commercial and environmental issue. The outcomes are expected to contribute valuable insights to the field of applied economics and provide a practical tool for businesses in the refurbished electronics sector.

1.1 Research Question

The study is structured around one main research question and a supporting sub-question. The main research question is:

"Can a dataset, systematically created through web scraping techniques on the Marktplaats platform, be utilized to analyze and predict market trends in refurbished iPhones?"

This question focuses on the potential of the dataset to inform pricing strategies for sellers based on identified market dynamics. The sub-question addresses the extraction and cleansing of data:

"How can web scraping be employed to create a comprehensive dataset of pre-owned iPhones on the Marktplaats platform?"

1.2 Relevant Work

1.2.1 Dynamic Pricing

Dynamic pricing is the study of determining optimal selling prices for products or services, in environments where prices can be easily and frequently adjusted. Dynamic pricing techniques are now commonly used across various industries, and in some instances, they are considered an essential part of pricing strategies (Boer, 2014).

Different scientific communities have explored the characteristics of pricing policies. According to Den Boer (2014) the research on pricing policies can be categorized under the operations research/management science literature, the economics literature and the computer science literature. This study mainly will fall under the latter. Notable studies within this category are Raju, Narahari, & Ravikumar (2006) and Chinthalapati (2006) which both apply machine learning to create dynamic pricing strategies for the electronic retail market. Chinthalapati (2006) mentions that dynamic pricing strategies require that a company know not only its own operating costs and supply availability, but also how much the customer values the product and what the future demand will be. The creation of a full pricing strategy lay outside of the scope of this study. It's goal is to give insight into how much the customer values the product, which could be implemented by companies in their own pricing strategy.

No prior research has been identified on the web scraping of a marketplace to predict and analyse price trends for refurbished iPhones, making this an interest field to discover.

1.2.2 Data mining

Data mining is a broad term that refers to generating value from data. A relevant study in which data mining was deployed is the study of Van Nuygen et al (2018). They conducted a study on the prediction of customer demand for remanufactured products. The research has resemblance with this study as it deploys web scraping in order to obtain insights into the market dynamics of pre-owned devices, in particular remanufactured products. Remanufactured products and refurbished products are similar as they both are pre-owned products which are thoroughly tested and often come with warranties. The key difference however is that remanufactured devices are rebuilt to meet the original specifications of new products. In contrast, refurbished devices are repaired to function like new, and may even appear as if they just came out of the factory (Elalj, 2023).

The study of Van Nuygen et al (2018) follows the theoretical Cross Industry Standard Process and Data Mining (CRISP-DM) framework, one of the most popular methodologies for data analytics (Sevim, Oztekin, Bali, Gumus, & Guresen, 2016). This framework consists of 6 fundamental steps: (1) Business understanding; (2) Data understanding involves identifying the data source and obtaining the variables relevant to the problem; (3) Data preparation employs various data cleaning and transformation

techniques to create a well-structured dataset for analysis; (4) Predictive modelling includes variable selection, model development, hyperparameter tuning, and validation; (5) Model evaluation measures and compares the predictive performance of the models using different predefined error measurements; (6) Model deployment generates insights to support managerial decision-making. This study will follow the same theoretical framework to ensure a systematic approach is employed for extracting valuable insights from the data collected.

1.2.3 Web Scraping

Web scraping can enable data mining by providing a dataset. It can ensemble rich datasets by collecting all the text and image content of many websites (Dilmegani, `2023). A study on the legality and ethics of web scraping was conducted by Korov & Johnsen (2020) in which they state that the first step with web scraping should be assessing if the website has any restrictions in place regarding the use of web crawlers, bots, or scraping tools. Restrictions in place are mentioned on the robots.txt file, which can freely be accessed on every website. Kasareka (2020) agrees by stating that ensuring the permission for web scraping is one of the most important, but overlooked step in the web scraping process. Furthermore this study mentions that the internet is dynamic where every website has a particular structure. The right web scraping method should be chosen accordingly. The website of Marktplaats utilizes dynamic elements that are loaded via Javascript. To support web scraping of such dynamic websites, this research employs Selenium, a tool specifically designed for scraping dynamic content (Muthukadan).

2. Methodology

The methodology of this study was divided into six primary phases according to the CRISP-DM framework: (1) understanding of the business, which is covered in chapter 1 through a review of literature, (2) data collection and understanding, (3) data preparation, (4) model development and validation, (5) model evaluation and (6) model deployment. Phases 2 through 5 are detailed in the following subsections, along with the Results section. The final stage, model deployment was not practically implemented within the scope of this thesis due to time constraints. This limitation and its implications are further discussed in the Discussion and Conclusion chapters.

Throughout the stages, several libraries and tools were utilized including Selenium (version 4.21.0 (Muthukadan)), Matplotlib (version 3.9.0 (Matplotlib)), Pandas (version 2.2.2 (Pandas)), Numpy (version 1.26.4 (NumPy)), Chromedriver (version 125.0.6422.76 (Chrome)) and Task Scheduler (version 1.0 (Microsoft)).

Several features were processed using regular expressions. The patterns utilized for this purpose are listed in Appendix A. A GitHub repository with all scripts and data used can be found at: https://github.com/nicoswart/Refurbished-iPhone-Market-Analysis-and-Price-Prediction.git

2.1 Data Collection and Understanding

2.1.1 Scraping Regulations

The first step involved checking the robots.txt file of Marktplaats to determine if there are any rules or restrictions on scraping. The robots.txt file was integrated into every script developed for this research to ensure compliance with ethical scraping practices.

2.1.2 Scraping Tool

After scraping regulations were handled a fitting tool to perform scraping with if needed. The Marktplaats webpage relies on dynamic content for proper functionality. Selenium enables handling dynamic content effectively by interacting with web pages in the same manner a user would (Kumari, 2024), which is why Selenium with Python is chosen to effectively scrape the website.

2.1.3 Structure

To scrape all iPhone models a scripts is created for every series, from the iPhone 8 up to the iPhone 15. The scripts follow a similar structure to each other but have information catered according to the specific series which is scraped. Each script was run daily at the end of the day, capturing listings which have been added that day. Running of the script will be done locally on a laptop, the HP Zbook Studio 4G. The web scraping was started on the third of June up to June 24th.

2.1.4 Target URL

To obtain the data a target uniform resource locator (URL) is provided in each script, this is the first page the web scraper will visit. Conveniently the search term and certain filters can be manipulated within the URL. The intital URL provided in the scripts takes the form of: https://www.marktplaats.nl/l/telecommunicatie/mobiele-telefoons-apple-iphone/#q:iphone+{Series}|offeredSince:Vandaag|searchInTitleAndDescription:true, in which the search term for the iPhone series is specified.

By manipulating the URL three filters are activated:

1. The product category is set to match for iPhones
2. The ‘offerdSince:Vandaag’ filter ensures only listings submitted on the day of scraping are included
3. The ‘searchInTitleAndDescription:true’ filter allows the search term to match either the title or description rather than just the title

2.1.5 Features

The information which is being obtained from the listings are the date of publication, model of the iPhone, the capacity of the device, the listing price and the URL corresponding to the listing page.

Feature 'price'

The listing pricing can take on seven different forms, the only choice that is of a numerical value is the option 'asking price'. The other options are not relevant for this study, in the case where the pricing option of the listing is non numerical the listing is discarded.

Features 'model' and 'capacity'

The model of the device in the listing is extracted from the title with regular expressions. The regular expression searches for the string 'iphone' followed by the corresponding series and other terms associated with the model such as 'pro', 'max', 'plus', dependent on which series is being scraped. The values for the capacity are obtained in a similar manner with strings containing '64', '128', '256' and '516' followed by 'gb' or '1tb'. The capacity is first searched for in the title, if it is not found also the description is checked.

Feature 'date'

The date of publication is extracted using the datetime library in python. As all the listings scraped are filtered to be uploaded on that same day, the datetime tool extracts the current date.

Feature 'URL'

The URL to the corresponding listing is captured. The captured URL serves as a reference point for verifying the data's authenticity and integrity.

2.1.6 Automation

In total there are nine scripts that ensure every model from the iPhone 8-Series up to the iPhone 15-Series can be scraped. To automate the process of running the scripts at the end of the day a .bat script was created that executes each script one by one. The .bat file was executed by Task Schedular daily at 23:40pm. Ensuring it is at the end of the day so it captures all new listings for that day but still leave enough time for the scripts to run before the end of the day is reached.

2.2 Data Preparation

Within the code additional filters are applied to obtain a relevant dataset. Specifically, the data is filtered to exclude irrelevant listings and to focus on the desired attributes of the iPhone models.

2.2.1 Model and Capacity

In the case of a model not being found at all in the title with the regular expressions the listing is discarded. However if the capacity is not identified 'None' is returned leaving the value blank but not discarding the listing.

The regular expressions search for the iPhone model starting with the most specific model first. For example, the regular expression for 'iPhone 15 Pro Max' searches for 'iphone' followed by optional whitespace, then '15' followed by optional whitespace, then 'pro' followed by optional whitespace, and finally 'max'. The order of these searches is crucial. This sequence ensures that more specific models like 'Pro Max' and 'Pro' are correctly identified and not mistakenly categorized as the regular 'iPhone 15'.

For capacities, the regular expressions search for one of the specified capacities numbers (64, 128, 256, GB) followed by optional whitespace and 'gb'. This ensures that the capacity is accurately identified and associated with the correct iPhone model. For both the model and capacity the searches were done in a case-insensitive manner.

2.2.2 Price

To determine whether the price is a value that is assumed to be real several actions have been taken. Firstly a regular expression has been implemented in the script, which ensures the price starts with a Euro symbol (€), followed by optional whitespace, digits, optionally a period with more digits, a comma, and exactly two decimal places. If the price matches this pattern, it is considered valid. Prices which are €0 or have a minimal of 3 digits and are in ascending order are seen as invalid and skips scraping of the listing.

2.2.3 Duplicate Listings

Duplicate listings are removed in the script by checking if the listing which is being scraped has a description which has already been scraped before, in this case the listing is discarded.

2.2.4 Business Listings

The method proposes capturing listings only from individuals, excluding those from businesses. This approach ensures that the free market of individual sellers is represented, as business listings could skew the data with higher prices. To ensure that the gathered data does not contain data from listings that are from businesses several measures have been taken. Marktplaats offers subscriptions to businesses on the platform offering several benefits, one of these benefits is the feature of getting two extra photos shown with their listings on the search page (Marktplaats). The web scraper identifies these listings and discards them. However not all businesses on Marktplaats opt for a subscription, so this does not identify all of the business listings. Another measure taken was removing listings mentioning warranty in their description.

2.2.5 Outliers

Most of the data cleaning was integrated into the scripts, where listings that are not suitable for the data set are skipped by the scraper. This method prevents scraping data which is not relevant, subsequently scraping less data causes the time for the scraping process to decrease. Besides that it eliminates the need of performing the data cleaning of these aspects in further stages, simplifying the process.

The only data cleaning operation that could not be done during the scraping of the listings was the removal of outliers for the price feature. This was done after the dataset was collected using the interquartile range method where the first quartile (Q1) and third quartile (Q3) are calculated. The IQR is the difference between Q3 and Q1. Outliers are defined as data points that lie below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. The outlier removal step was done over each iPhone model separately, this way outliers are removed based on prices in their same price range. This method resulted in 501 listings being removed from the dataset.

2.3 Model Development, Validation and Evaluation

The model development process involves both analysing market trends and predicting prices of pre-owned devices.

2.3.1 Price Stabilisation through Moving Average

With the aim of gaining insights in prices of pre-owned devices, a stable prediction is necessary. Fluctuating predictions of price could affect dynamic pricing models in a negative way. To obtain a prediction of prices first of all, the average of all prices of listings on a day, for a specific model are determined. By taking these outcomes, and averaging them with historic values on the same metric the moving average is calculated. By using a moving average, short-term volatility can be smoothed out, creating a more stable indication of the price valuation and also revealing underlying price trends.

The number of historic data points used to calculate the moving average, known as the window size, influences the degree of smoothing. A larger window size results in smoother averages and a decrease in the standard deviation between points. The downside however of using a larger window size is that price changes are not reflected in the moving average as quickly. While for a dynamic model it is crucial that prices reflect real time market trends. Therefore the moving window should be chosen accordingly. This

study proposes examining the effect of window size on the standard deviation of the moving averages to find an optimal balance.

2.3.2 Market Trend Analysis

The primary goal of performing market trend analysis is to understand how the prices of pre-owned iPhones evolve over time and identify trends. The main statistical tool used to for the trend analysis, is linear regression. Linear regression provides a clear and interpretable measure of the rate at which prices are increasing or decreasing, which is crucial for understanding market dynamics.

The linear regression model is computed over the moving averages, leveraging the strength of both methods. The moving average firstly smooths over the averages, making the underlying trend more apparent. Linear regression then quantifies this smoothed trend providing a clearer and more accurate representation of the underlying relationships in the data. For instance, a study on the detrending-moving-average-based bivariate regression method demonstrates how this combination can significantly enhance analysis accuracy (Fan & Wang, 2020).

The resulting trendline will be used in order to calculate the growth ratio. Which is done by dividing the last point on the trend line by the first. This provides a quantifiable measure that can be implemented in a dynamic pricing model.

2.3.3 Predictive Model

To predict future prices of refurbished iPhones an ARIMA (Autoregressive Integrated Moving Average) model was developed. The ARIMA model is suitable for this study due to it's robustness in handling non-stationary time series data. The model consist out of three main components.

1. An Autoregression (AR) component, which implies there is a changing variable that regresses on its own lagged or prior values.
2. An Integrated (I) component, which incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.
3. Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The model has three parameters that each, respectively, influence one of the main components:

1. *q*: the number of lag observations in the model
2. *d*: the number of times the raw observations are differenced
3. *q*: the size of the moving average window.

Model Configuration and Training

The ARIMA model was fit by using the complete dataset. It was divided into a training and test set. The split ratio was 60% to 40% respectively. The training set consisted of data from the first 14 days. The test set included data from the subsequent 10 days. A grid search was conducted to optimize the ARIMA model parameters *q*, *d* and *p*. The ranges for these parameters were chosen based on preliminary analyses. These analyses suggested they offered the best compromise between model complexity and forecasting accuracy.

Model Validation

The validation of the ARIMA model involved assessing its performance. To achieve this the Mean Absolute Error (MAE) metric was used. This metric is crucial for evaluating accuracy. It measures the average magnitude of the errors between predicted and actual values. This provides a clear indicator of prediction accuracy. In order to be able to compare performance between models a standardized form of the MAE is introduced.

To further examine the evaluation of the model, the residuals, being the difference between the predicted and observed datapoints, will be analysed. In order to do so the residual density of all iPhone

models combined will be plotted. This plot will give insights into the variation of the total residuals. To understand over-time how this residuals are divided the mean residuals will be plotted along the span of the dates in the test set.

Model Evaluation

After the model is configured, and validated, the model will be fitted with future datapoints. This will result in predictions on future datapoints. An appropriate time span will be used to be able to draw conclusions on the predicting capabilities of the model on both long-term and short-term insights.

3. Results

The complete dataset after cleaning and preprocessing consists of 8170 listings. The data is gathered over the period of 20 days, between the third of June up to June 25th. The amount of listings per series together with several descriptives such as the standard deviation and the coëfficiënt of variation are shown in table 1, giving an overview of the dataset.

In some cases displaying all iPhone models in graphs would result in a cluttered presentation. Therefore, one iPhone-series, has been selected for detailed illustration to provide a clearer view of the trends. Additional figures and data on all iPhone models are explained in detail in the text and are referred to in Appendix B-D.

3.1 Descriptive Statistics

The descriptive statistics show us that the average mean value progressively increases with each successive iPhone model. The amount of listings are relatively high for the iPhone series 11, 12 and 13. The oldest iPhone model in the dataset, the iPhone 8 has the least amount of listings. The newer models. such as the iPhone 13, 14 and 15-series have a higher absolute standard deviance than the older models such as the iPhone 8. However, the iPhone 8 and iPhone X-series have the largest coefficient of variation, indicating a higher relative variability in their listing prices.

Series	**Listings**	**Mean (€)**	**Mininum price (€)**	**Maximum price (€)**	**std (€)**	**CV**
iPhone 8	435	98.43	10	200	33.44	0.340
iPhone 10	908	146.04	35	275	41.55	0.285
iPhone 11	1762	220.05	90	419	56.81	0.258
iPhone 12	1535	317.66	125	600	82.25	0.259
iPhone 13	2142	458.12	240	825	104.77	0.229
iPhone 14	898	671.26	300	1000	136.94	0.204
iPhone 15	613	810.54	500	1150	110.34	0.136

Table 1: Summary Descriptive Statistics of iPhone Listings

3.2 Price Stabilization through Moving Average

The average daily prices for the iPhone 12 models are shown in figure 1. The corresponding standard deviation is as follows: iPhone 12 (std = 7.76), iPhone 12 Pro (std = 12.73), iPhone 12 Mini(std = 14.61) and iPhone 12 Pro Max (std = 18.8).

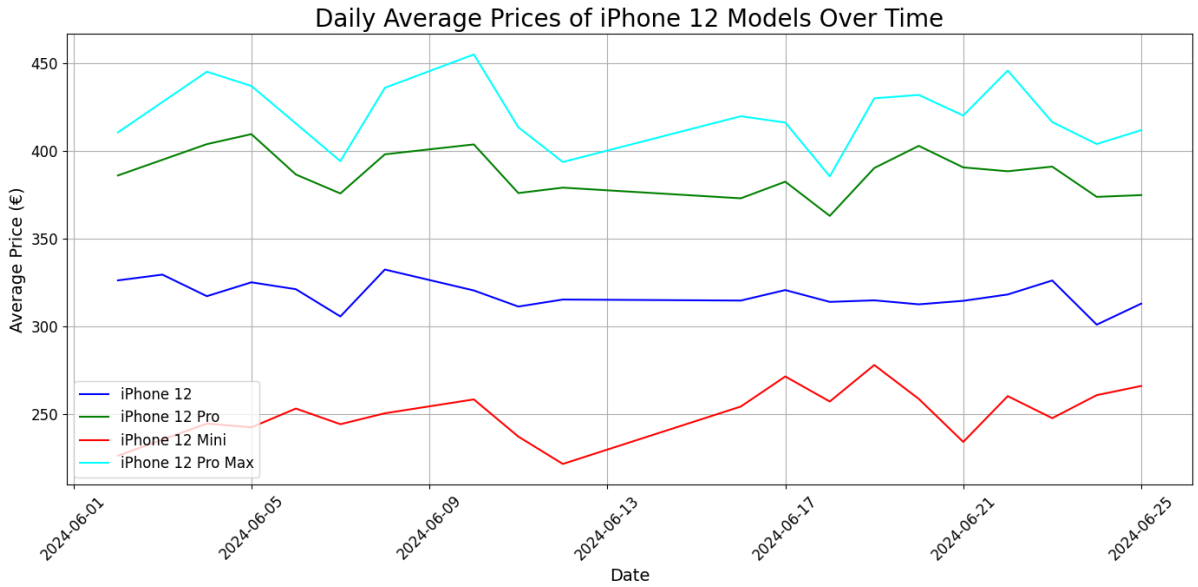

Figure 1: Daily average daily prices of the iPhone 12-series

To further explore the variability in pricing across all iPhone models, Figure 2 presents boxplots of the standard deviations of rolled prices for different moving day windows. In this plot, a moving day window of 1 represents values without any moving average applied, similar to the data shown in Figure 1. The plot considers the standard deviation of prices for all models, ranging from the iPhone 8 to the iPhone 15 Pro Max. The moving day window ranges from 1 to 14 days. The plot highlights the decrease in variability as the window increases. This effect illustrates the smoothing technique on the data. Shorter moving windows show greater variability in standard deviations. Extending the window length consistently reduces this variability. This confirms the effectiveness of the moving average in stabilizing price trends.

This visual analysis was crucial in identifying the optimal window for smoothing the pricing data without dampening the inherent trends. The boxplots clearly show a reduction of the standard deviation with increasing window sizes, particularly up to a 7-day window. Beyond this point the reduction of the standard deviation plateaus. Indicating that an increase of the window size would bring minimal benefit in terms of reducing price variability, while potentially lead to over-smoothing, where market trends might be obscured as the data lags further behind.

According to this information a 7-day moving window was selected as the optimal balance between smoothing the data and maintaining enough granularity to observe meaningful market trends.

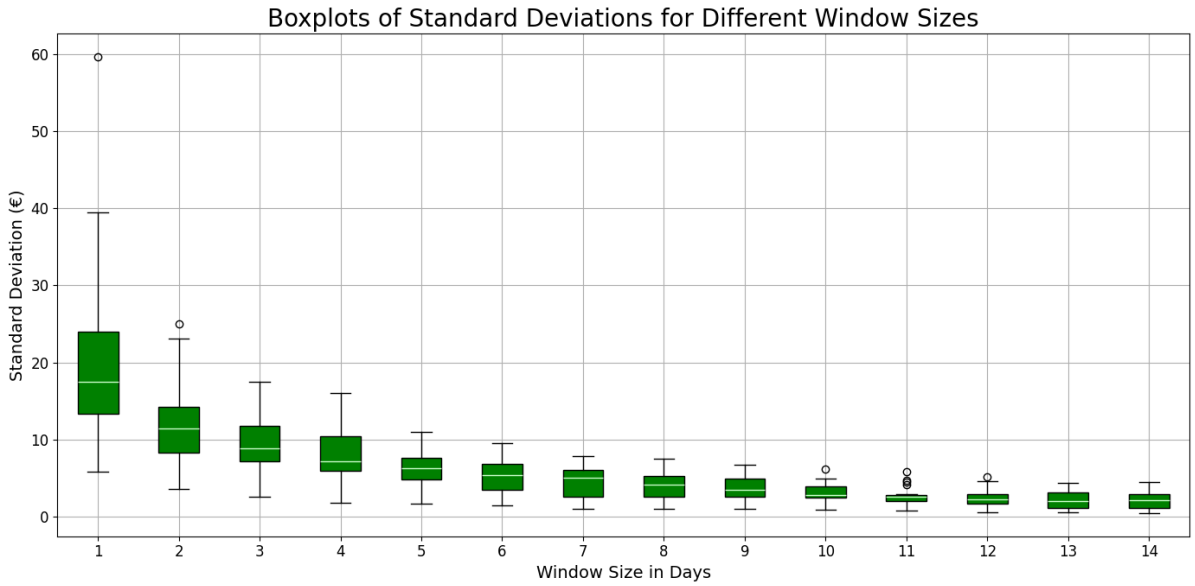

Figure 2: Boxplots on standard deviation per model, for different amount of

To illustrate the effect of the moving window the average prices for the iPhone 12 models are plotted. This time implementing a moving average of the period of the 7 previous days. The result is a smoother line, giving a more stable price indication. The standard deviations of the iPhone 12 models, before and after the implementation of the moving average are shown in table 2.

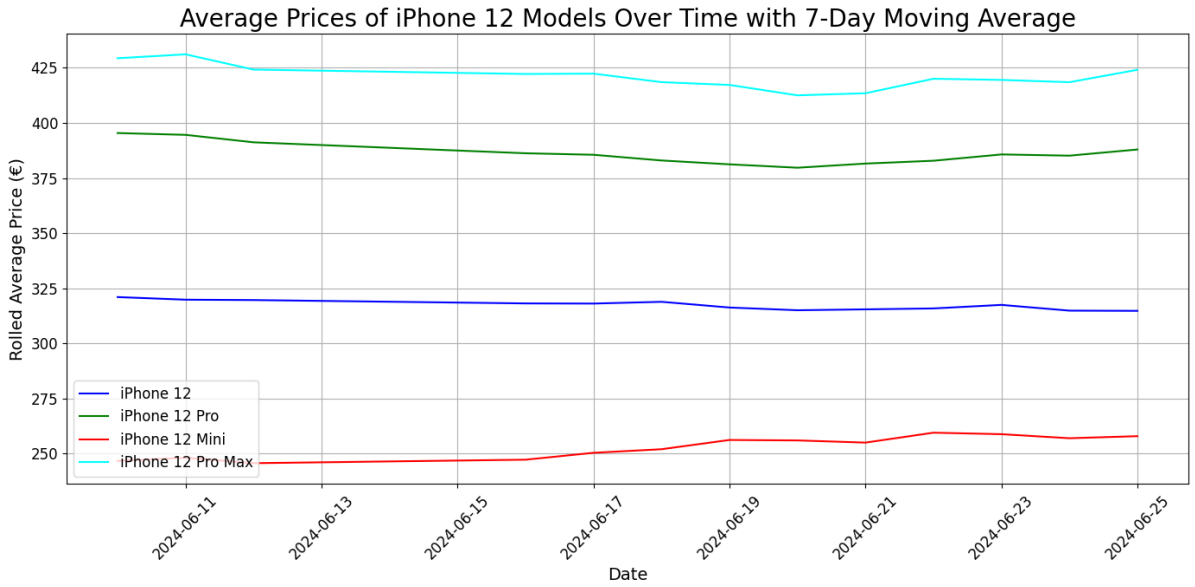

Figure 3: Average prices for the iPhone 12-series over time with a 7-day moving average.

Before applying the moving average, the highest standard deviation was observed in the iPhone 15 Pro Max, with a standard deviation of €59.68. After implementing the moving average, the highest standard deviation remained with the iPhone 15 Pro Max, but it reduced to €7.69. For a detailed table containing the standard deviation values for all the iPhone models refer to Appendix B.

Model	Std prior to implementation of the 7-day moving average (€)	Std prior after implementation of the 7-day moving average (€)
iPhone 12	7,76	2,39
iPhone 12 Pro	12,73	4,96
iPhone 12 Mini	14,61	4,98
iPhone 12 Pro Max	18,80	5,43

Table 2: Summary of iPhone 12's standard deviations, before and after applying moving average of 7 days.

3.3 Average Price Trends

The time period over which the trend lines are calculated spans from June 10th to June 25th. The average price trends for the iPhone 12 models indicate, with exception of the iPhone 12 Mini, that the average price value tends to decrease over time. The iPhone 12 mini however, shows an upward trendline.

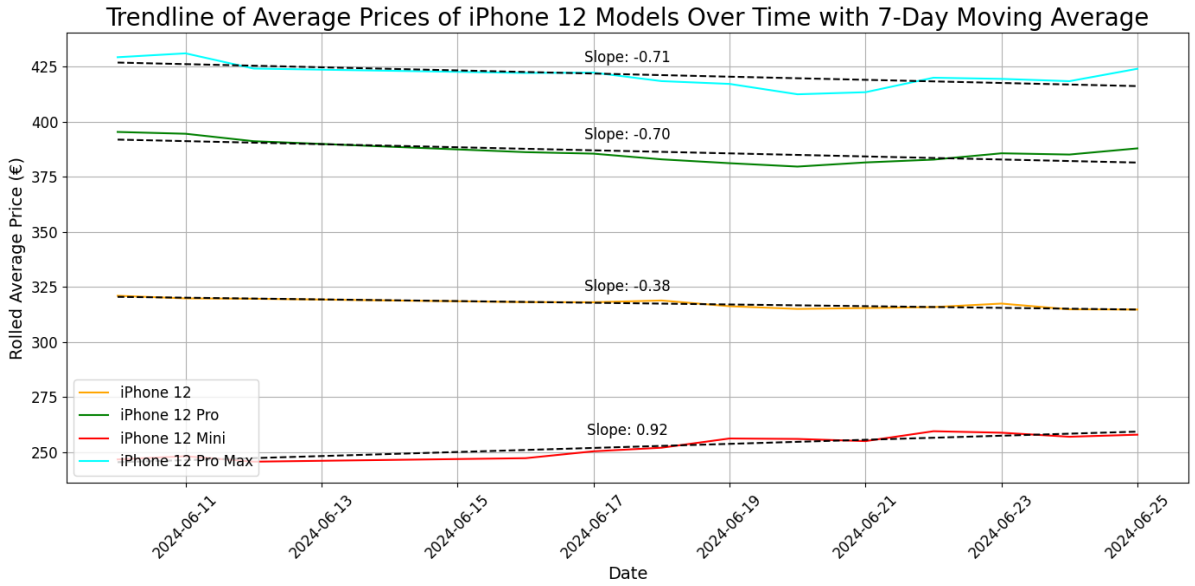

Figure 4: Trendline of the average prices of the iPhone 12 models with a 7-day moving average window

The corresponding growth ratios for these models are shown in table 3. Refer to appendix D for the full table with growth ratios for each model. The average growth ratio across all models is 1.001, indicating that overall the models show a stable trend.

Model	Growth Ratio
iPhone 12	0,979
iPhone 12 Pro	0,973
iPhone 12 Mini	1,057
iPhone 12 Pro Max	0,975

Table 3: Summary of iPhone 12-series growth ratio

3.4 Predictive Model

3.4.1 Model Development and Evaluation

The grid search iterated over the values $q \in \{1 . 2\}$. $d \in \{1 . 2\}$. and $p \in \{7 . 9\}$ to obtain the optimal configuration, the optimal configuration of each model can be found in Appendix C.

The average MAE across all models is €19.5. In standardized form (sMAE) this value is 0.31. The predictions, from the test set, of the best-performing model, the iPhone 12, with an sMAE of 0.176, are shown in Figure 5. It can be seen that the estimations run roughly through the middle of the actual prices.

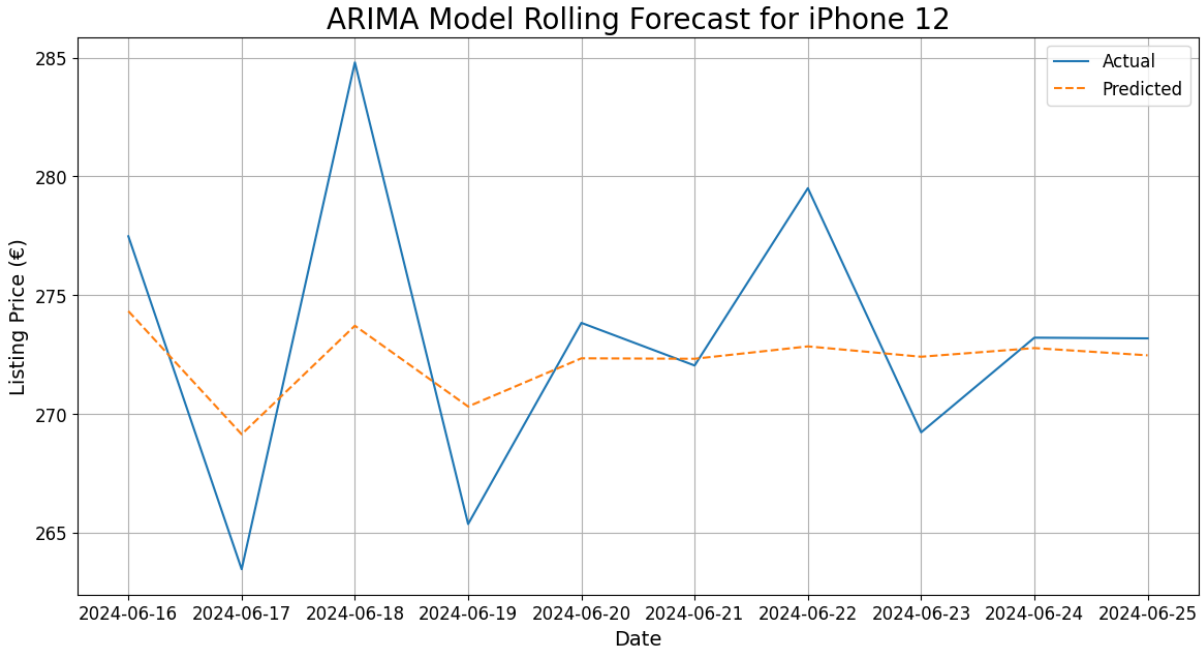

Figure 5: Predictions on the test set for the iPhone 12 vs actual values

For the worst-performing model. the iPhone Xs Max. the prices are consistently overestimated. The corresponding sMAE for this model is 0.927.

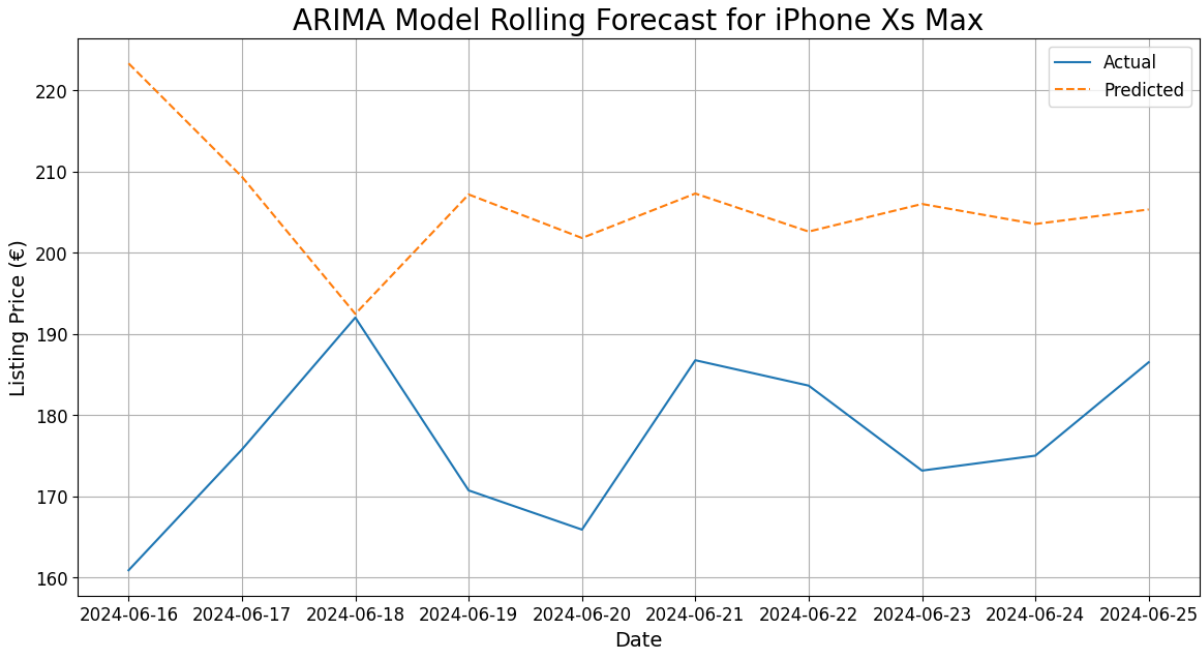

Figure 6: Predictions on the test set for the iPhone Xs Max vs actual values

Model	Parameters (respectively q.d.p)	MAE (€)	sMAE
iPhone 12	1,1,7	3.76	0.176
iPhone Xs Max	1,1,7	28.83	0.927

Table 4: Summary of optimal parameters and model performance on the best and worst performing model, refer to Appendix C for the full table

The residual density plot for all iPhone models shows the distribution of residuals, which are the differences between the actual and predicted listing prices. The plot is centered around zero, indicating that the residuals are approximately normally distributed. The distribution, however, does skew to the right. This suggests that the model tends to underestimate the prices more often than it overestimates. The peak which is slightly to the left of the zero reinforces this observation.

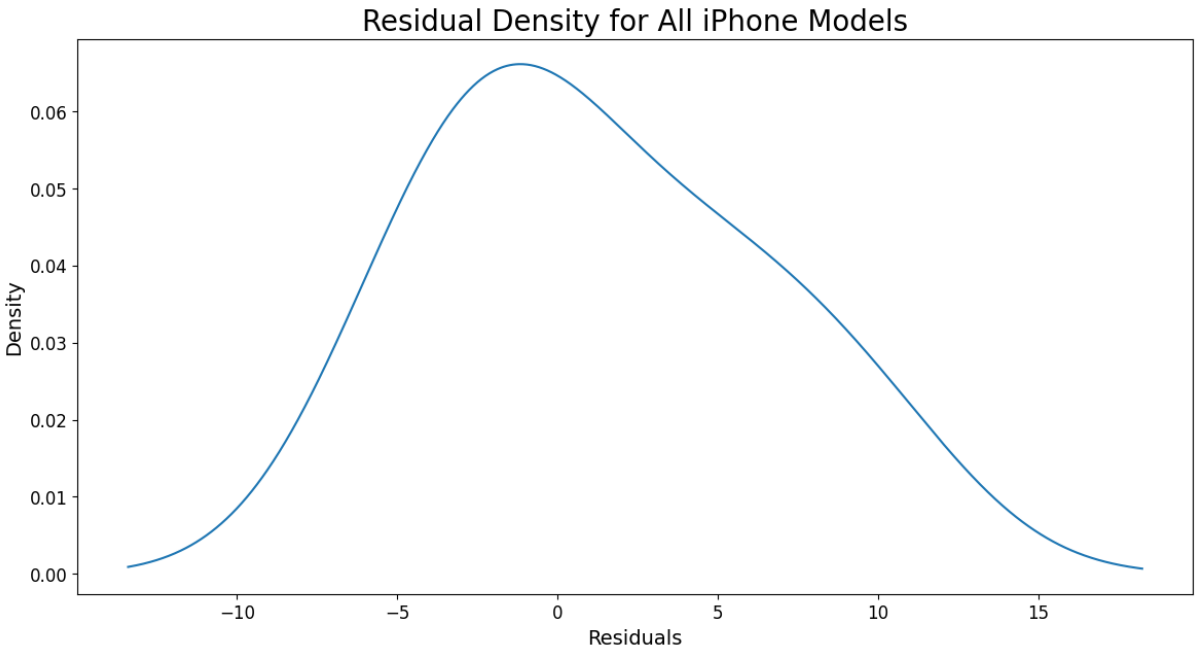

Figure 7: Distribution of the residual density of all models combined

The mean residuals plot for all iPhone models over the test period provides further insight into the model's performance. There are noticeable periods where the residuals are consistently positive or negative. In the first three days and the last day the model overestimates more often, in between these periods the prices are underestimated more often.

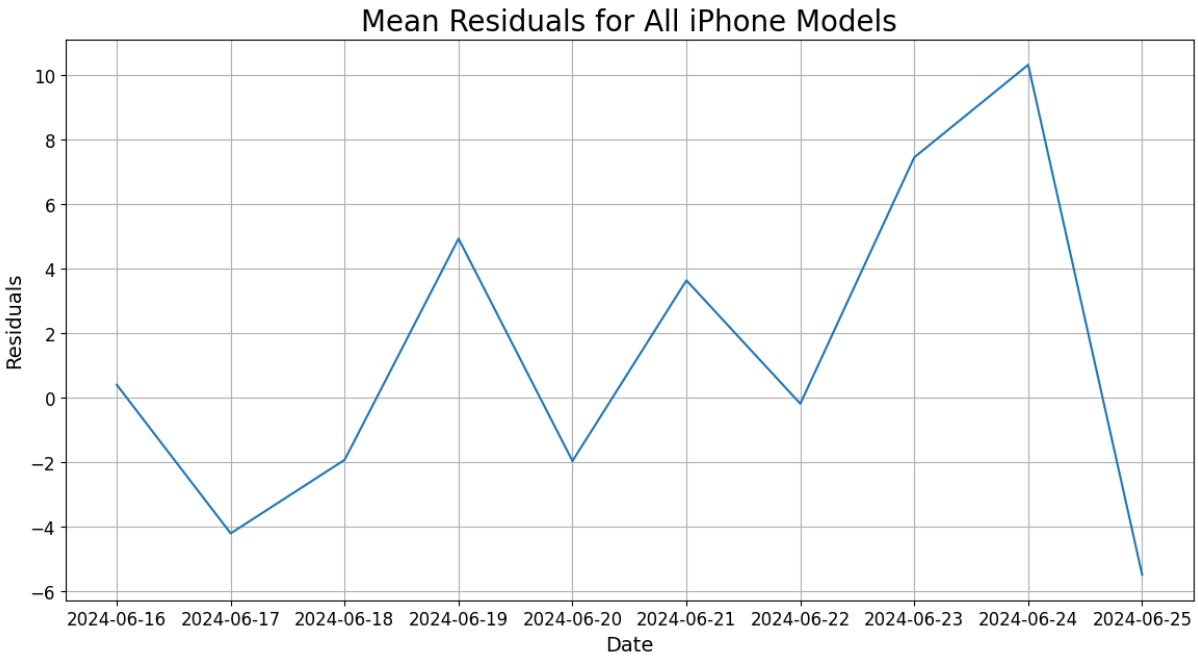

Figure 8: Mean residuals for all iPhone models over time.

The forecast on unseen data of the best performing model, that of the iPhone 12, are plotted in figure 9. There can be observed that the actual listing prices are variable up until the start of the forecast period. The forecasted values indicate a stabilization in the listing prices over the next two months, with the values after six days becoming nearly stationary. The iPhone 12 has a growth ratio of 0.979, indicating in downward trend in the value of the iPhone. In the predictions of the model this downward trend is not identified.

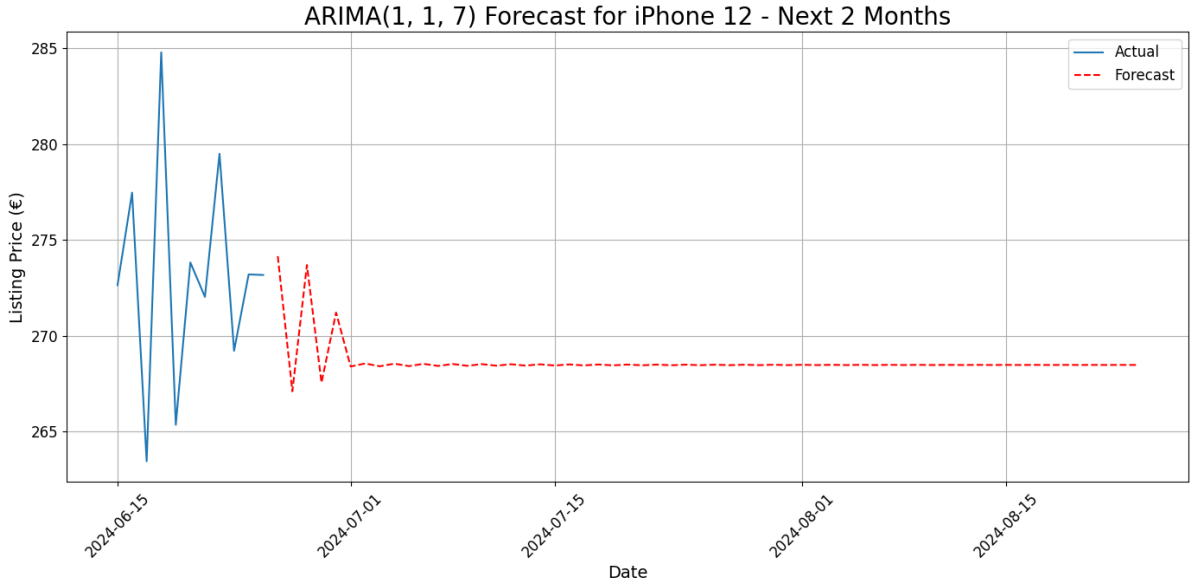

Figure 9: The forecast of the model for the iPhone 12.

4. Discussion

4.1 Data Gathering

Data was gathered at the end of the day, to capture as many listings as possible. Listings that were uploaded and removed on the same day, because for example someone has sold their item, were not scraped. The more often a day the listings would be scraped, the more data would have been gathered.

Within the span of data gathering there were four days on which the web scraping was not run, resulting in four days within the data set without any datapoints. Every time the scraping was not performed this was due to the laptop, on which the script was run, not being powered on. Using for instance a cloud to run it could have proven more useful in running the web scraper program more often, and more consistent. A dataset with listings for every day, and more listings per day could have resulted in better market predictions.

4.2 Web Scraper

The web scraper searches for the model by searching for the mentioning of the model in the title, occasionally though some people use the regular iPhone model as an umbrella term. Some users also don't know that they have for instance a 'pro' version. This can cause the regular iPhone models to have listings included of other models.

The web scraper does not always run smoothly to the last page of the listing pages during scraping. Sometimes it accidentally clicks an ad that has just been loaded. This happens very rarely and in the case of this study it has not affected the dataset as the script, in case of error, was re-run manually. But in the case of running the script in the cloud or without being present this should first be altered by for example using an ad blocker (Agenty, 2022).

4.3 Cleaning Data

The cleaning of the data mostly was handled in the scraping scripts themselves. As a result, the documentation on the cleaning process is less extensive than it would have been if the data cleaning had been performed after storing the data in a file. For instance, the amount of dropped listings because of not having a valid price is not known. Conducting the cleaning process after storing the data would have provided more transparency.

4.4 Model Deployment

The final stage in the CRISP-DM framework, model deployment, lay outside the scope of this research. The absence of a deployment phase means that while the study provides theoretical insights into market trends of pre-owned devices, the practical applicability of these insights for businesses in the refurbished market remains speculative. Without real-world testing and implementation, it is challenging to fully assess the effectiveness of the predictive models in operational settings.

4.5 Future Research

For future research the field of dynamic pricing offers endless opportunities, as models can always be tweaked and improved up on. However, within the realm of refurbished products the existing research is relatively limited. To advance this area, further investigation into key factors influencing dynamic pricing strategies is essential. For example, understanding how demand fluctuates (Chinthalapati, 2006), particularly for refurbished iPhones, could provide valuable insights and significantly enhance the effectiveness of dynamic pricing models.

This study itself also presents several opportunities for future research. Once the web scraper has been run for a longer period of time, perhaps the ARIMA model could be extended with the SARIMA model,

which also takes in seasonality. This extension could potentially yield more accurate predictions by accounting for seasonal patterns in the data. Additionally, analysing the influence of device capacity on prices, as well as examining the supply by periodically assessing the number of listings scraped, could provide valuable insights. Incorporating these factors into the predictive models could further refine their accuracy and usefulness.

5. Conclusion

This thesis systematically explored whether dataset created through web scraping on the Marktplaats platform can be used to analyze and predict market trends in refurbished iPhones. The CRISP-DM framework was employed. This provided a structured approach to the analysis. It supported the process of answering the research questions. Additionally it helped in uncovering pricing dynamics for pre-owned iPhones.

The study successfully addressed the main research question: "Can a dataset, systematically created through web scraping techniques on the Marktplaats platform, be utilized to analyze and predict market trends in refurbished iPhones?" The study finds that the data set created can be utilised to analyse market trends. Over the period of scraping there was a stable trend identified across all models with a growth ratio of 1.001, certain specific models showed an up or downward trend depicted in Appendix D. The ARIMA models performance is not robust along all different iPhone models, with an average MAE of €19.50, the predictive capabilities may not be suitable for dynamic pricing models as they pose variability in effectiveness. Long term trend also are not identified by the model. The effectiveness of utilising identified market trends from pre-owned devices, on the refurbished iPhone market remains speculative as the last step in the framework, the deployment stage was not implemented in this study.

The supporting sub-question "How can web scraping be employed to create a comprehensive dataset of pre-owned iPhones on the Marktplaats platform?" was answered by strategically deploying web scraping techniques to gather targeted data. Making use of Selenium, the study navigated Marktplaats's dynamic website environment to access listings relevant to iPhone models. This was achieved by configuring the starting URL to direct the scraping tool to the appropriate iPhone categories and inputting specific search terms relevant to each model. To ensure the relevance and efficiency of the data collection, the scraper was equipped with filtration steps that selectively captured essential data while excluding the scraping of irrelevant listings. These steps were crucial in assembling a dataset that was not only relevant to answering the research question but also enhances the overall efficiency of the data collection process

Reference List

Ademe. (2022). *Assessment of the environmental impact of a set of refurbished products.* Retrieved from: https://librairie.ademe.fr/dechets-economie-circulaire/5833-assessment-of-the-environmental-impact-of-a-set-of-refurbished-products.html.

Agenty. (2022). Retrieved from: *https://agenty.com/docs/how-to-block-ads-with-puppeteer-to-super-fast-your-web-scraping/309.* https://agenty.com/docs/how-to-block-ads-with-puppeteer-to-super-fast-your-web-scraping/309.

Ahmed, A., Khan, M., & Ishtiaq, A. (2023). *Web Scraping for Scientific Discovery: Strategies for Secure Data.* Retrieved: from https://www.ieeesem.com/researchpaper/Web_Scraping_for_Scientific_Discovery_Strategies_for_Secure_Data_Retrieval_Structured_Transformation_and_Relevant_Content_Selection.pdf.

Boer, A. V. (2014). *Dynamic pricing and learning: Historical origins, current research,.* Retrieved from*:* https://pdf.sciencedirectassets.com/280252/1-s2.0-S1876735415X0002X/1-s2.0-S1876735415000021/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEP%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJIMEYCIQDkRliSYULRLAHYt3ZrIeuVB%2B2kMJ9%2FlRH%2Bcnm8YvzvegIhAP9d.

Bright. (2023). *Techreuzen tonen vanaf vandaag verplicht hun aantallen EU-gebruikers.* Retrieved from: https://www.rtl.nl/tech/artikel/5366538/digital-services-act-dsa-eu-europa-techbedrijven-booking-bol-marktplaats.

Bruyninckx, H. (2021). *Towards global sustainability.* Retrieved from: https://www.eea.europa.eu/articles/towards-global-sustainability.

Chinthalapati, V. L. (2006). *Learning Dynamic Prices in MultiSeller Electronic Retail Markets With Price Sensitive Customers, Stochastic Demands, and Inventory Replenishments.* Retrieved from: https://ieeexplore-ieee-org.proxy.uba.uva.nl/stamp/stamp.jsp?tp=&arnumber=1603740&tag=1.

Chrome. (n.d.). *Downloads.* Retrieved from https://developer.chrome.com/docs/chromedriver/downloads.

Counterpoint. (2022). *Infographic: Global Refurbished Smartphone Market | 2021.* Retrieved from: https://www.counterpointresearch.com/insights/refurbished-smartphone-market-2021-infographic/.

Dilmegani, C. (`2023). *Web Scraping vs Data Mining: Why the Confusion? in 2024.* Retrieved from: https://research.aimultiple.com/web-scraping-vs-data-mining/.

Elalj, S. (2023). *Remanufactured vs. Refurbished: What Are the Differences?* Retrieved from: https://www.refurb.me/blog/remanufactured-vs-refurbished.

Fan, Q., & Wang, F. (2020). *Detrending-moving-average-based bivariate regression estimator.* Retrieved from https://journals.aps.org/pre/abstract/10.1103/PhysRevE.102.012218.

Hapuli, P. (2023). *Circular Economy of Refurbished Smartphones in the European Union.* Retrieved from: https://www.theseus.fi/bitstream/handle/10024/801292/Hapuli_Kumpulainen.pdf?sequence=2#page=56&zoom=100,148,104.

Hischier, R., & Böni, H. (2021). *Combining environmental and economic factors to evaluate the reuse of electrical and electronic equipment – a Swiss case study.* Retrieved from: https://www.sciencedirect.com/science/article/pii/S0921344920306224.

Jager, W., & Janssen, M. (2009). *Dynamic pricing in complex behaving consumer markets: Submitted for the ESSA2009 conference.* Retrieved from :https://www.researchgate.net/publication/288391006_Dynamic_pricing_in_complex_behaving_consumer_markets_Submitted_for_the_ESSA2009_conference.

Kasereka, H. (2020). *Importance of web scraping in e-commerce and e-marketing.* Retrieved from :https://www.researchgate.net/publication/347999311_Importance_of_web_scraping_in_e-commerce_and_e-marketing.

Krotov, V., & Johnsen, L. (2020). *Legality and Ethics of Web Scraping.* Retrieved from :https://www.researchgate.net/publication/352014123_Legality_and_Ethics_of_Web_Scraping.

Kumari, J. (2024). *A Comprehensive Guide to Web Scraping Using Selenium.* Retrieved from :https://www.analyticsvidhya.com/blog/2024/05/a-comprehensive-guide-to-web-scraping-using-selenium/.

Marktplaats. (n.d.). Retrieved from: https://www.marktplaatszakelijk.nl/ondernemers/groei-met-marktplaats/groei-met-standaard/pakketten/.

Matplotlib. (n.d.). *Matplotlib 3.9.0 documentation.* Retrieved from: https://matplotlib.org/stable/index.html.

Microsoft. (2023). *Task Scheduler for developers.* Retrieved from: https://learn.microsoft.com/en-us/windows/win32/taskschd/task-scheduler-start-page.

Muthukadan, B. (n.d.). *Installation.* Retrieved from: https://selenium-python.readthedocs.io/installation.html.

NumPy. (n.d.). *NumPy documentation.* Retrieved from: https://numpy.org/doc/1.26/.

Pandas. (n.d.). *Installation.* Retrieved from: https://pandas.pydata.org/pandas-docs/stable/getting_started/install.html.

Raju, C., Narahari, Y., & Ravikumar , K. (2006). *Learning dynamic prices in electronic retail markets with customer segmentation.* Retrieved from: https://link.springer.com/article/10.1007/s10479-006-7372-3.

Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. (2016). *Developing an early warning system to predict currency crises.* Retrieved from: https://www.sciencedirect.com/science/article/pii/S0377221714001829.

Van Nguyen, T., Zhou, L., Chong, A., Li, B., & Pu, X. (2018). *Predicting customer demand for remanufactured products: A data-mining approach.* Retrieved from: https://www.sciencedirect.com/science/article/pii/S037722171930668X?fr=RR-2&ref=pdf_download&rr=898c58e25f546567.

Zuora. (2024). *Dynamic Pricing: A Comprehensive Guide for Businesses.* Retrieved from: https://www.zuora.com/guides/understanding-dynamic-pricing/.

Appendix A:

The regular expressions for the price, and one model as example, are listed below:

Price:

^\€\s*\d{1,3}(?:\.\d{3})*,\d{2}$

iPhone model:

iphone\s*15\s*pro\s*max

iphone\s*15\s*pro

iphone\s*15\s*plus

iphone\s*15

iPhone capacity:

128\s*GB

256\s*GB

512\s*GB

1TB\s*GB

Appendix B:

Model	Std prior to implementation of the 7-day moving average (€)	Std after implementation of the 7-day moving average (€)
iPhone 8	7,79	1,45
iPhone 8 Plus	13,48	2,52
iPhone X	5,86	2,21
iPhone Xr	7,93	1,18
iPhone Xs Max	20,98	5,02
iPhone 11	11,92	1,03
iPhone 11 Pro	14,10	2,39
iPhone 11 Pro Max	25,48	3,44
iPhone 12	7,76	2,39
iPhone 12 Pro	12,73	4,96
iPhone 12 Mini	14,61	4,98
iPhone 12 Pro Max	18,80	5,43
iPhone 13	11,29	6,19
iPhone 13 Pro	17,40	7,81
iPhone 13 Mini	17,16	4,98
iPhone 13 Pro Max	19,21	6,22
iPhone 14	24,26	7,41
iPhone 14 Plus	39,49	6,78
iPhone 14 Pro	17,59	4,44
iPhone 14 Pro Max	23,88	6,07
iPhone 15	25,83	5,10
iPhone 15 Plus	33,13	5,78
iPhone 15 Pro	14,39	2,56
iPhone 15 Pro Max	59,69	7,69

Table 5: Standard deviations before and after applying moving averages, for all models.

Appendix C:

Model	Parameters (respectively q.d.p)	MAE (€)	sMAE
iPhone 8	2.1.9	9.34	0.369
iPhone 8 Plus	2.1.9	9.44	0.229
iPhone X	1.1.8	9.99	0.288
iPhone Xr	2.1.8	6.09	0.208
iPhone Xs Max	1.1.7	28.83	0.927
iPhone 11	1.1.7	7.91	0.200
iPhone 11 Pro	1.1.7	6.23	0.239
iPhone 11 Pro Max	1.1.7	26.26	0.251
iPhone 12	1.1.7	3.76	0.176
iPhone 12 Mini	2.1.8	29.35	0.669
iPhone 12 Pro	1.2.9	16.04	0.450
iPhone 12 Pro Max	1.1.7	19.06	0.316
iPhone 13	2.1.9	10.95	0.470
iPhone 13 Mini	1.1.9	11.59	0.309
iPhone 13 Pro	1.1.9	17.12	0.485
iPhone 13 Pro Max	2.1.7	12.76	0.235
iPhone 14	2.1.8	37.41	0.384
iPhone 14 Plus	2.2.8	21.36	0.247
iPhone 14 Pro	1.2.7	19.03	0.276
iPhone 14 Pro Max	2.1.7	16.65	0.222
iPhone 15	1.1.8	21.23	0.247
iPhone 15 Plus	1.1.7	39.38	0.337
iPhone 15 Pro	1.1.9	10.48	0.251
iPhone 15 Pro Max	1.2.8	46.65	0.183

Table 6: Optimal parameters and model performance for all models.

Appendix D:

Model	Growth Ratio
iPhone 8	1,032
iPhone 8 Plus	0,999
iPhone X	1,01
iPhone Xr	0,986
iPhone Xs Max	0,943
iPhone 11	0,999
iPhone 11 Pro	0,977
iPhone 11 Pro Max	0,971
iPhone 12	0,979
iPhone 12 Pro	0,973
iPhone 12 Mini	1,057
iPhone 12 Pro Max	0,975
iPhone 13	1,04
iPhone 13 Pro	1,043
iPhone 13 Mini	1,034
iPhone 13 Pro Max	1,031
iPhone 14	0,976
iPhone 14 Plus	0,999
iPhone 14 Pro	0,99
iPhone 14 Pro Max	1,003
iPhone 15	1
iPhone 15 Plus	1,008
iPhone 15 Pro	1,003
iPhone 15 Pro Max	0,992

Table 7: Growth ratios for all models.