Pythonic vs Refactorable Pythonic: On the Relationship between Pythonic Idioms and Code Quality in Machine Learning Projects

Gerardo Festa,[1] Giammaria Giordano,[1] Valeria Pontillo,[2]
Massimiliano Di Penta,[3] Damian A. Tamburri[3,4], Fabio Palomba[1]

[1] *Software Engineering (SeSa) Lab - Department of Computer Science, University of Salerno, Italy* — [2] *Gran Sasso Science Institute (GSSI), L'Aquila, Italy* — [3] *University of Sannio, Italy* — [4] *JADS/NXP Semiconductors, Netherlands*
g.festa22@studenti.unisa.it, giagiordano@unisa.it, valeria.pontillo@gssi.it,
dipenta@unisannio.it, datamburri@unisannio.it, fpalomba@unisa.it

Abstract

Context: Python is increasingly becoming the *lingua franca* for developing Machine Learning (ML) systems, thanks to a rich ecosystem of libraries and an emphasis on readability. In this context, *Pythonic* idioms are seen as stylistic conventions that support maintainable and efficient code. Conversely, *Refactorable-Pythonic* idioms refer to patterns that can be refactored into more idiomatic Python, improving code quality in terms of maintainability, performance, and clarity.

Objective: While the assumptions about idiomaticity are widely accepted in practice, the extent to which *Pythonic* or *Refactorable-Pythonic* idioms relate to software quality in ML projects has not been systematically validated. To address this lack of empirical evidence, this paper conducts a large-scale study to assess how *Pythonic* and *Refactorable-Pythonic* idioms are related to code quality in ML systems.

Method: We analyze 303 open-source Python projects from the NICHE dataset, distinguishing between "well-engineered" (i.e., projects that adopt structured development practices such as testing, CI, documentation, and packaging) and "non-engineered" (i.e., projects that lack such characteristics). Our analysis proceeds in two main phases: (i) idiom detection, where we extract *Pythonic* and *Refactorable-Pythonic* code patterns using a combination of existing and custom detectors; and (ii) quality assessment, where we detect Python-specific smells and relate them to code metrics and other quality indicators.

Result: Truth Value Test and *Assign Multiple Targets* are the most common *Pythonic* and *Refactorable-Pythonic* idioms, respectively. In "well-engineered" projects, both idiom types positively correlate with Python-specific code smells, suggesting that idiomatic usage does not always align with higher code quality. In contrast, in "non-engineered" projects, the presence of smells is more strongly influenced by structural factors such as the number of lines of code, complexity, and commit activity.

Conclusion: We conclude by distilling lessons learned, implications, and future research directions.

Keywords: Software Quality, Software Engineering for Artificial Intelligence, Software Maintenance and Evolution, Empirical Software Engineering.

1. Introduction

As Machine Learning (ML) becomes increasingly widespread, so too does the textitasis on producing code that is readable, maintainable, and scalable [1]. Python—currently the dominant language in ML development—encourages using *Pythonic* idioms: coding practices that leverage the language's expressive syntax to promote simplicity, clarity, and efficiency [2]. For example, a typical *Pythonic* idiom, the list comprehension, replaces a manual loop for computing the squares of a list of numbers with the more concise function `[x**2 for x in range(10)]`. These idioms are widely promoted within the Python community and development guidelines as a means to write cleaner, more maintainable code [3, 4].

However, while *Pythonic* idioms are among the practices most widely acknowledged by developers as contributing to more maintainable code, their relationship with software quality has not been empirically validated. Some idioms, although concise, may inadvertently foster patterns that lead to maintainability issues. Consider, for instance, the *Assign Multiple Targets* idiom, where multiple variables are assigned in a single line (e.g., `a = b = 0`). While this syntax is compact and valid, the idiom may inadvertently promote a coding style in which shared mutable state is accessed across extended blocks of code, making it more challenging to isolate responsibilities and encapsulate logic. Moreover, the idiom can encourage the propagation of multiple interdependent variables into downstream functions, increasing the complexity of function interfaces and making the code harder to maintain and evolve.

These risks are particularly relevant in ML systems, where functions are frequently used to configure models or compose data processing pipelines. In such settings, the need for rapid prototyping and experimentation often leads to function signatures expanding over time, especially when variables assigned together are passed through multiple layers of the codebase. These concerns are further amplified by the fact that contributors to ML systems often come from non-software engineering backgrounds and may lack formal training in principles such as code design and maintainability [5, 6]. So, these arguments raise a fundamental question:

> **Main Question**
>
> *Does the use of Pythonic idioms correlate with higher or lower code quality in real-world ML codebases?*

Answering this question has important implications for both research and practice. For practitioners, empirical evidence can validate or challenge current recommendations on idiomatic coding, helping teams make more informed decisions about code conventions and training practices in ML development environments. For researchers, such findings contribute to a better understanding of how high-level coding patterns influence maintainability and may inform the design of future tools, linters, or refactoring strategies that are sensitive to the specific needs of ML projects.

This paper addresses this question through an empirical investigation into whether the use of *Pythonic*

idioms correlates with software quality in ML projects. As a comparison baseline, we consider *Refactorable-Pythonic* idioms, namely code patterns that are syntactically correct but deviate from idiomatic best practices and can be mechanically transformed into more *Pythonic* alternatives. We adopt this category as a baseline because it naturally contrasts with idiomatic usage, representing functionally correct but stylistically suboptimal code. At the same time, these idioms can be reliably and automatically detected through publicly available tools. As for code quality, we operationalize it through the presence of *code smells*, i.e., recurring design deficiencies that hinder maintainability and may signal deeper structural issues [7]. Specifically, we consider *Pythonic* smells detected using DPy [8]. Our study analyzes 303 open-source Python repositories for ML development, drawn from the NICHE dataset [9], and classified as either "well-engineered" or "non-engineered" based on established criteria, such as the presence of tests, continuous integration, or packaging metadata. We focus on nine *Pythonic* idioms and their *Refactorable-Pythonic* counterparts, examining their statistical association with code smells and how these patterns vary across the two groups. Importantly, the maintainability of real-world ML systems depends on multiple interacting factors, such as developer experience, team practices, and project domain. As such, our study investigates correlations rather than causal relationships between idiomatic usage and quality indicators such as code smells. Furthermore, by leveraging the NICHE dataset, Giordano et al. [10] provided an evidence-based quantitative assessment of the relationship between software engineering practices in Python projects and code smells. Their findings indicate that "well-engineered" projects are statistically associated with a significantly lower overall prevalence of code smells across various project sizes. The empirically demonstrated relationship between engineering maturity, idiomatic usage, and code quality motivates us to select the NICHE dataset for investigating the interplay between Pythonic idioms and code quality in ML projects.

The results of our study indicate that *Truth Value Test* and *Assign Multiple Targets* are the most common idioms in *Pythonic* and *Refactorable-Pythonic* idioms, respectively. In addition, we find a statistical relationship between idioms and the presence of Python-specific code smells, especially in "well-engineered" projects, suggesting that the adoption of these idioms is not in all cases the best solution to improve code quality in ML systems. To summarize, our paper makes the following contributions:

1. A large-scale empirical study investigating the adoption of nine *(Refactorable-)Pythonic* idioms across 303 ML projects and how such idioms vary with codebase size and density;
2. A statistical analysis exploring the relationship between *(Refactorable-)Pythonic* idioms and Python-specific code smells, highlighting the idioms that most strongly correlate with quality issues;
3. A publicly available replication package including all scripts, datasets, and results used in our study, to ensure transparency and foster future research [11].

Structure of the paper. Section 2 summarizes the state-of-the-art and discusses the most closely

related work. Section 3 details the research questions of our study and the research method applied to address them. Section 4 summarizes the results obtained. Section 5 discusses a further analysis to improve the generalizability of the study, as well as the implications of this work. Section 6 discusses the potential threats to validity and the mitigation strategies applied. Section 7 concludes the paper.

2. Background and Related Work

This section provides the necessary information to understand the rest of the paper and summarizes the state-of-the-art in the context of code smells in Python projects.

2.1. Background and Motivation

Pythonic code refers to a style of Python programming that adheres to the idiomatic principles of the Python language, textitizing readability, simplicity, and conciseness. These principles are encapsulated in Python's unofficial mantra, *"The Zen of Python"* [2], which includes precepts such as *"Beautiful is better than ugly"* and *"Readability counts"*. Writing *Pythonic* code involves leveraging Python's built-in features and capabilities in a way that maximizes the language's expressiveness and minimizes code complexity. Additionally, the use of *Pythonic* is among the most widely acknowledged practices by developers for contributing to more maintainable code.

Refactorable-Pythonic code, instead, refers to Python code that does not use these idiomatic principles and often resembles patterns that might be more common in other programming languages. This code can typically be refactored into a *Pythonic* style, which not only enhances readability but also aligns with Python's philosophy of simplicity and efficiency. The refactoring into *Pythonic* code often involves replacing cumbersome and verbose constructs with more streamlined and effective idioms.

For example, a common *Refactorable-Pythonic* approach might involve using a loop to filter active users whose age is greater than 18 and whose name starts with an uppercase "A", and collecting their names into a list. Listing 1 provides an example.

Listing 1: Filter active users with age over 18 whose names start with 'A', and collect their names.

```
filtered_names = []
for user in users_list:
    if user.is_active and user.age > 18 and user.name.startswith("A"):
        filtered_names.append(user.name)
```

A *Pythonic* refactor of the above code would use a *list comprehension* function, which is more concise and transparent, as shown in Listing 2.

Listing 2: Pythonic filter active users with over 18 whose names start with 'A', and collect their names.

```
filtered_names = [user.name for user in users_list if user.is_active and user.
    age > 18 and user.name.startswith("A")]
```

In the context of ML projects, writing *Pythonic* code can have a particularly significant impact. For instance, ML workflows often involve dynamic configuration of models, where parameters may be generated programmatically or loaded from external sources. In such cases, idioms such as *Star in Function Call* can enhance both readability and flexibility. In the snippet provided in Listing 3, the unpacking operator * is used to pass a list of hyperparameters directly to the `LogisticRegression` constructor in a clean and maintainable way:

Listing 3: Star in Function Call example.

```
# Load hyperparameters from external config file
with open('config.json') as f:
    config = json.load(f)

# Unpack the dictionary into the model constructor
model = LogisticRegression(**config)
```

This *Pythonic* idiom not only makes the code more concise but also improves execution efficiency by optimizing parameter handling internally, representing an important advantage for data-intensive tasks typical in machine learning workflows.

In other terms, *Pythonic* idioms provide not only aesthetic and stylistic benefits but also practical advantages that may affect multiple properties of ML projects [12]. The Python's dominance in the ML domain, along with the potential benefits of *Pythonic* code, motivates our work: we aim to *empirically validate these benefits and provide a data-driven understanding of how Pythonic coding influences ML projects.*

Given the importance of using *Pythonic* code, recent research has focused on the automatic detection of refactorable-idiomatic constructs. One such tool is RIdiom [13], a framework specifically designed to identify code snippets that can be rewritten in a more idiomatic style. RIdiom addresses the challenge of promoting idiomatic usage by analyzing source code through its Abstract Syntax Tree (AST) representation. It systematically detects non-idiomatic code, i.e., *Refactorable-Pythonic*. As elaborated in Section 3.2, we employed this tool in the scope of our study, as it provides a reliable and systematic way to identify *Pythonic* idioms in real-world code. In our investigation, we focus on analyzing code smells, which we adopt as a proxy metric to evaluate the potential impact of *Pythonic* code on software quality.

In this context, several static analyzers have been proposed to detect code smells in Python codebases. Among the most prominent are PySmell [14] and DPy [8], both capable of detecting 11 Python-specific code smells with high precision and recall. While PySmell was the first tool designed for this purpose,

it is no longer maintained and is currently unavailable [8]. In contrast, DPy is actively supported and accessible—which explains why we opted for this tool in the context of our paper.

2.2. Related Work

Fowler and Beck [7] originally defined *code smells* as indicators of sub-optimal design decisions that could exacerbate code complexity, particularly during maintenance and evolution tasks. A significant amount of research has been conducted on code smells in traditional software systems, *e.g.,* systems developed in Java, investigating their origins, persistence, and mitigation strategies [15, 16, 17, 18].

Further studies have correlated code smells to reusability mechanisms, such as inheritance and delegation, noting their benefits and drawbacks [17]. Giordano *et al.* [19] found that while certain design patterns negatively correlated with code smells, others positively correlated with their presence. Although the majority of this research has concentrated on Java systems [20, 21, 22, 23], studies by Vavrová *et al.* [24] have started to explore the prevalence and characteristics of code smells in Python, discovering that smells like *Long Method* are statistically more prevalent in Python systems than in Java.

In the context of ML projects, Chen *et al.* [25] analyzed 106 Python ML projects by introducing PySmell, a Python code smell detection tool. Their findings reinforced previous work and highlighted that the *Long Method* smell is the most prevalent one. Van Oort *et al.* [26] investigated Python-specific code smells by analyzing 74 ML projects, finding that *Duplicate Code* is one of the most widespread smells in ML projects. Jebnoun *et al.* [27] used PySmell to explore deep learning projects, discovering that *Long Lambda Expression*, *Long Ternary Conditional Expression*, and *Complex Container Comprehension* smells are more frequent within deep learning code than in traditional software code. More recently, Giordano et al. [10] conducted an evidence-based investigation into how CI mechanisms affect the emergence of code smells in ML systems, finding that they may actually lead to a significant reduction of code quality issues.

When comparing our work to those discussed above, we differentiate ourselves by explicitly focusing on the influence of *Pythonic* coding practices on the prevalence of code smells in ML projects. Unlike previous studies that predominantly examine code smells within the context of traditional software systems or general Python applications, our research directly targets ML environments where *Pythonic* practices are hypothesized to have a distinct impact on software quality. In this respect, our study contributes to the field by systematically assessing whether adopting *Pythonic* idioms leads to a measurable improvement in code maintainability and a reduction in code smells, potentially providing recommendations for ML practitioners who heavily rely on Python for implementing complex algorithms and data processing tasks.

Also, our research contributes to advancing the current body of knowledge on *Pythonic* code and its practical impact. In this respect, Alexandru *et al.* [3] explored the adoption and benefits of *Pythonic* idioms, such as *list comprehensions* and *decorators*. Their study, informed by interviews with developers and an analysis of 1,000 Python repositories, highlights that these idioms are favored for improving code

maintainability, aligning with the principles outlined in *"The Zen of Python"* [2].

Zid et al. [4] conducted a controlled experiment involving 209 developers to evaluate the understandability of *Pythonic* functional constructs such as *lambdas*, *comprehensions*, and *map/reduce/filter* functions, compared to their procedural counterparts. The study revealed that procedural code was generally found to be more readable than functional alternatives.

Zampetti *et al.* [28, 29] also studied whether *Pythonic* functional constructs are more prone to induced fixes than others. They found that, in general, changes to such constructs have higher odds of inducing fixes than other changes, and that this is more likely the case for lambdas and to some extent for comprehensions.

Sakulniwat *et al.* [30] investigated *when* and *why* developers adopt *Pythonic* idioms in open-source projects. They discovered that practitioners tend to use *Pythonic* idioms during evolutionary activities and, in a non-negligible number of cases, developers improve their source code refactoring using *Pythonic* idioms.

Phan-udom *et al.* [31] released Teddy, a tool that provides *Pythonic* idioms examples starting from *Refactorable-Pythonic* idioms. The main limitation of this tool is the absence of refactoring operations; it only provides examples starting from a predefined knowledge base. Zhang *et al.* [32] introduced a tool for refactoring *non-Pythonic* code into *Pythonic* one, demonstrating its efficacy in real-world applications.

Some studies investigated the performance improvement achieved when refactoring code towards *Pythonic* idioms. Leelaprute *et al.* [33] provided evidence that *Pythonic* idioms could significantly enhance both memory usage and execution times, suggesting areas for further practical research. Zhang *et al.* [34] leveraged their refactoring tool and showed that *non-Pythonic* code refactored into *Pythonic* one led to performance improvements. However, Zid *et al.* [35] found that performance improvements are usually negligible for real-world source code elements, and become tangible only when the refactoring action is amplified.

While the work above discusses the relationship between *Pythonic* constructs and software quality, we further analyze this relationship for ML programs.

Similarly, some work investigates the use of Python in ML-intensive projects. Nagpal and Gabrani [36] investigate the suitability of Python for scientific applications, highlighting its advantages because of a simpler syntax and portability. Instead, we focus more on *Pythonic* idioms, and the extent to which they relate (or not) to smells in ML code.

3. Research Method

The *goal* of this study is to understand whether the adoption of *Pythonic* idioms in ML projects correlates with better software quality, measured through the presence of code smells. The focus on quality lies in assessing how idiomatic coding practices may influence maintainability-related decisions in ML systems.

The *perspective* is for both developers and researchers; the former are interested in understanding whether their coding practices are beneficial or detrimental to the quality of their software, while the latter aim to

deepen their knowledge of both the adoption of *Pythonic* idioms in ML projects and the relationship between idioms and code smells. Starting from the overarching research question proposed in Section 1, we formulate the following research questions that guide our analysis.

> **🔍 RQ$_1$. On the diffusion of *(Refactorable-)Pythonic* idioms.**
> *To what extent do (Refactorable-)Pythonic idioms occur in ML systems?*

RQ$_1$ serves as a preliminary investigation and provides a descriptive perspective to characterize how *Pythonic* and *Refactorable-Pythonic* idioms are used in practice across ML projects. This question is instrumental for the subsequent analysis of the potential relationship between *Pythonic* idioms and code quality. In particular, understanding the frequency and context of idiom usage helps interpret any correlation with code smells and supports a more detailed view of their role in real-world systems. To explore idiom adoption in more detail, we split **RQ$_1$** into three sub-questions:

RQ$_{1.1}$ What are the most frequent *(Refactorable-)Pythonic* idioms in ML projects?

RQ$_{1.2}$ How does the adoption of *(Refactorable-)Pythonic* idioms vary as the project grows in size?

RQ$_{1.3}$ What is the density of *(Refactorable-)Pythonic* idioms in source code?

> **🔍 RQ$_2$. On the relationship between *Pythonic* idioms and code smells.**
> *What is the relationship between the usage of Pythonic and Refactorable-Pythonic idioms and the presence of code smells in ML systems?*

The goal of **RQ$_2$** is to investigate whether the use of *(Refactorable-)Pythonic* idioms is statistically correlated with the presence of code smells. We focused on this dimension because code smells, representing poor design and implementation choices, can degrade the overall system quality [7, 10, 18]. As such, they serve as effective proxy metrics for understanding and evaluating software quality. By adopting code smells as a proxy, our study aims to provide an objective and reproducible method for measuring the potential impact of *Pythonic* idioms on software quality.

Building on the descriptive findings from **RQ$_1$**, this analysis aims to determine whether certain idioms tend to co-occur with higher or lower smell incidence. This question lies at the core of our study, as it empirically tests the assumption that *Pythonic* idioms inherently lead to cleaner, higher-quality code. Considering the two categories of projects *i.e.,* "well-engineered" and "non-engineered", we decided to split the **RQ$_2$** into two sub-questions. Indeed, we asked:

RQ$_{2.1}$ What is the relationship between the usage of *Pythonic* and *Refactorable-Pythonic* idioms in "well-engineered" projects and the presence of code smells in ML systems?

$RQ_{2.2}$ What is the relationship between the usage of *Pythonic* and *Refactorable-Pythonic* idioms in "non-engineered" projects and the presence of code smells in ML systems?

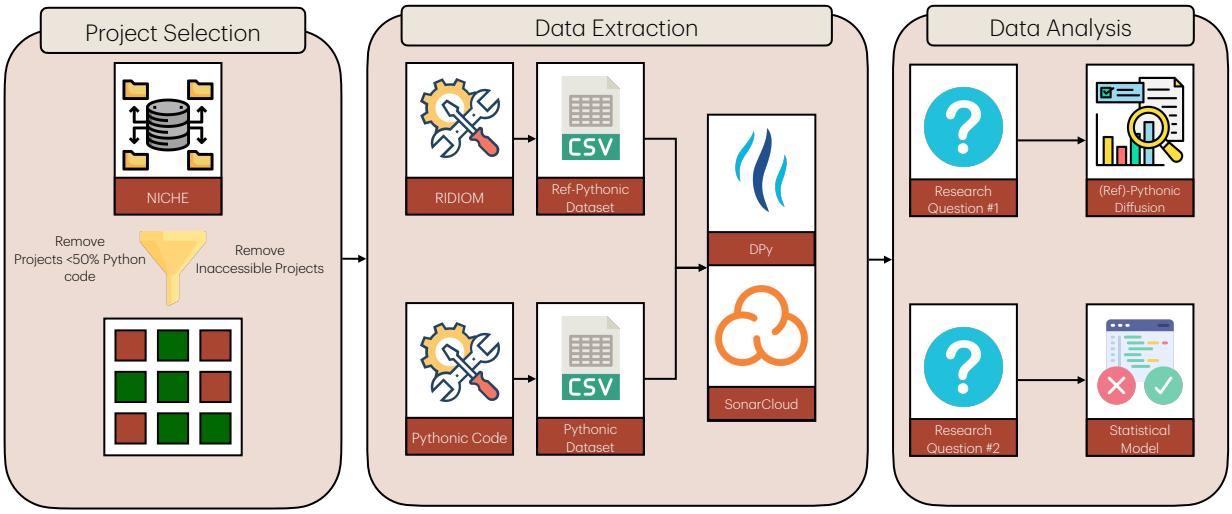

Figure 1: Research Method Overview.

Our investigation has a statistical connotation. We adopted the guidelines of Wohlin *et al.* [37] in terms of reporting and the *ACM/SIGSOFT Empirical Standards*;[1] - we leveraged the "General Standard", "Data Science", and "Repository Mining" guidelines. As overviewed in Figure 1, we initially mined ML projects from GitHub. Next, we simultaneously ran RIdiom to extract code snippets that could be refactored into a *Pythonic* way (*i.e., Refactorable-Pythonic* idioms) and built and tested a homemade tool to detect *Pythonic* snippets in repositories based on AST. After obtaining *Pythonic* and *Refactorable-Pythonic* idioms, to address RQ_1, we computed the' frequency, variation, and density of the idioms. Finally, we ran SonarCloud [38] to extract information regarding project metrics *i.e.,* lines of code (LOC), file complexity, and number of commits, DPy [8] to extract Python-specific code smells from repositories, and applied statistical tests to analyze the relationship between *(Refactorable-)Pythonic* idioms and the frequency of code smells.

3.1. Dataset Description and Project Selection

The *context* of our analysis is the NICHE dataset [9], which includes 572 ML projects categorized as either "well-engineered" or "non-engineered", based on eight-dimensional metrics (see Table 1). We selected this dataset for three main reasons. First, NICHE comprises only popular, active projects with a substantial development history—specifically, projects with at least 100 GitHub stars, a minimum of 100 commits, and a most recent commit dated after May 1st, 2020. These criteria ensure the exclusion of inactive or personal

[1] Available at:`https://github.com/acmsigsoft/EmpiricalStandards`

Table 1: Description of the Eight Dimension Metrics that Characterize Well-Engineered Projects.

Dimension	Description
Unit testing	The project clearly shows the presence of unit tests.
Architecture	The project architecture is well-defined
Documentation	The project demonstrates sufficient documentation to enable developers to perform evolutionary and maintainability activities.
Issues	The community uses the GitHub issue management tools.
Continuous Integration (CI)	Instruments to permit CI, such as Jenkins or GitHub Actions, are regularly used in software projects.
History	The project shows a long history, demonstrating the community's maturity.
Community	The project has a large number of collaborators, showing their capabilities as a working team.
License	The project is hosted with an appropriate license, clearly defined and documented.

repositories. Second, each project in the dataset has been manually labeled by the dataset authors as either "well-engineered" or "non-engineered", based on whether it satisfies at least four of the eight defined dimensions. This labeling facilitates a robust comparative statistical analysis between the two groups: 441 "well-engineered" and 131 "non-engineered" projects.

Finally, the dataset encompasses a diverse range of project domains, thereby enhancing the generalizability and relevance of our findings. To ensure that only projects useful for our analysis are considered, we applied two additional filters *i.e.,* we removed projects with less than 50% of Python code to consider only those mainly written in Python and projects not yet accessible on GitHub (*e.g.,* projects migrated to other system versioning). As a result of these steps, we ended up with 303 projects, of which 76 were "not engineered" and 227 were "well-engineered". Table 2 provides the descriptive statistics for the variables *Stars*, *Commits*, *NLOC*, and *Pythonic Percentage* divided according to the column *Engineered*.

It is important to clarify that our study did not perform any classification or labeling of the projects. We relied on the manual categorization provided by Widyasari et al. [9], who, following the framework by Munaiah et al. [39], assessed each project according to the eight dimensions above. Therefore, our analysis builds upon an already validated dataset, using this binary label as a means to compare patterns of idiom adoption and quality indicators across two empirically derived groups. Regarding the rationale behind the dataset's choice, we selected this dataset because we assume that these dimensions are indicators not only of process maturity but also of developer expertise and adherence to best practices. Teams capable of maintaining testing pipelines, documentation, and CI workflows are typically composed of contributors with software engineering backgrounds and a greater tendency to follow community-endorsed conventions. In the Python ecosystem, such conventions explicitly include the use of *Pythonic* idioms, as promoted by "*The Zen of Python*", to enhance readability and maintainability. Therefore, we expect "well-engineered" projects to exhibit a more idiomatic coding style, reflecting the habits of experienced developers who value clean and readable code. Conversely, "non-engineered" projects, often exploratory or research-oriented, tend to rely

on ad-hoc solutions and show less consistent idiomatic adoption. This rationale underpins our analysis of idiomatic usage and its relationship with code quality.

Table 2: Descriptive Statistics for Variables *Stars*, *Commits*, *NLOC*, and *Python Percentage* divided according to the *Engineered* Column.

		Stars	Commits	NLOC	Python %
Well-Engineered	Min	100	102	234	52%
	Mean	1,843	1,088	20,055	91%
	Median	650	465	11,130	98%
	Max	31,057	47,094	232,930	100%
Non-Engineered	Min	111	100	396	51%
	Mean	2,033	422	15,158	89%
	Median	405	223	4,303	97%
	Max	64,439	4,914	268,628	100%

3.2. Selection of Pythonic Idioms and Code Smells

This study focuses on the relationship between *Pythonic* coding practices and the presence of code smells in ML systems. To this aim, we selected a representative set of idioms and code smells.

The *Pythonic* idioms were chosen based on their expressiveness, frequency of use, and alignment with the principles outlined in the *"The Zen of Python"* book [2]. We rely on RIdiom [13], a state-of-the-art tool for the detection of candidate idioms in Python. The tool operates on the AST and supports a catalog of idioms, including *List Comprehension*, *Dict Comprehension*, *Set Comprehension*, *Chain Comparison*, *Truth Value Test*, *Loop Else*, *Assign Multiple Targets*, *Star in Function Call*, and *For Multiple Targets*. These idioms improve code conciseness, readability, and efficiency. Table 3 presents the idioms used in this study.

As for code smells, we leverage DPy [8], a Python static analyzer capable of identifying 11 Python-specific implementation smells, such as *Long Method*, *Complex Conditional*, *Long Statement*, and *Empty Catch Block*. These smells are commonly associated with reduced readability and maintainability in codebases. As explained in Section 2, we preferred DPy over earlier tools (e.g., PySmell) due to its active maintenance and superior detection performance. Table 4 lists the smells considered in the scope of the study.

Starting from the full set of implementation smells detected by the DPy tool, we retained only those for which a mitigation strategy could plausibly involve using a *Pythonic* idiom. This filtering step ensured that our analysis focused on smells with a *direct* and *interpretable* connection to idiomatic constructs, rather than broader structural or semantic issues. The resulting mapping is presented in Table 5, organized by row, each row describing a single code smell instance. For each smell, the Table shows a concrete code example,

Table 3: List of *Pythonic* idioms recognized by RIdiom.

Idiom	Description
List Comprehension	A concise way to create lists in a single line by applying an expression to each item in an iterable.
Set Comprehension	A method for creating sets by directly iterating over items and applying transformations, eliminating the need for loops and 'add' operations.
Dict Comprehension	Allows building dictionaries in a single line by iterating over items and dynamically defining the keys and values.
Chain Comparison	Permits combining multiple comparison expressions in a single statement, reducing redundancy by avoiding separate if conditions.
Truth Value Test	Leverages Python's inherent truthiness rules to simplify conditions. Instead of explicitly comparing a variable to values like zero or None, it directly evaluates its truthy or falsy state.
Loop Else	Adds an 'else' clause to a loop, which runs only if the loop completes normally (i.e., without encountering a 'break').
Assign Multiple Targets	Allow assigning multiple variables in one line.
Star in Func Call	Uses the unpacking operator '*' to pass a list or tuple as multiple arguments to a function call, improving readability and flexibility when handling dynamic data.
For Multiple Targets	Enhances readability and efficiency in loops by unpacking multiple items from an iterable (e.g., tuples or lists) directly into separate variables, avoiding index-based access.

a refactored version, the *Pythonic* idioms applied during refactoring, and whether the transformation was implemented through idiomatic constructs (✅) or not (❌). Additional columns provide a brief textual description of the problem and the rationale behind the mitigation. This layout enables the reader to assess both the nature of the smells and the plausibility of using idiomatic code as a remedy. Accordingly, we included in our analysis only those smells for which the application of *Pythonic* idioms represents a plausible mitigation strategy. This choice aligns with the core objective of our study, i.e., to empirically assess whether the use of idiomatic Python correlates with measurable improvements in software quality. By focusing on smells that can be explicitly addressed through idiomatic constructs, we ensure that any observed associations (or lack thereof) can be meaningfully interpreted in terms of the role that *Pythonic* practices play in mitigating common quality issues in ML codebases.

Looking more closely at the Table, an initial hypothesis emerges: *Pythonic* idioms appear to offer natural mitigation strategies for smells that involve verbose or repetitive syntactic patterns (e.g., *Long Statement*, *Complex Method*, *Empty Catch Block*). In contrast, smells rooted in naming (e.g., *Long Identifier*,

Table 4: Code Smells Detectable Using DPy.

Implementation Smell	Description
Long statement	A line of code that is excessively long, making it harder to read and maintain.
Long parameter list	A function definition that includes too many input parameters, which can reduce clarity and increase complexity.
Long method	A method or function that performs too many tasks, resulting in low readability and poor modularity.
Long identifier	An overly verbose name for a function, class, variable, or field that impacts code readability.
Empty catch block	A catch or except block that lacks handling logic, potentially hiding errors during execution.
Complex method	A method with intricate logic or too many responsibilities, making it difficult to understand or test.
Complex conditional	A condition expression containing numerous logical operators, which can obscure intent and cause errors.
Missing default	A `match-case` statement that lacks a default case, risking unhandled inputs.
Long lambda function	A lambda expression that is too lengthy or complicated, reducing code brevity and clarity.
Long message chain	A deep chain of method calls that hinders understanding and makes debugging more difficult.
Magic number	A numeric literal used without explanation, making the code harder to interpret and maintain.

Magic Number) or structural design (e.g., *Missing Default*) do not lend themselves to idiomatic refactoring and instead require interventions beyond what idioms alone can express. This observation motivates our empirical investigation: whether the presence of *Pythonic* idioms in real-world ML codebases correlates with a reduction in detectable smells, or whether idioms serve primarily stylistic purposes with limited practical effect. In this sense, Table 5 serves not only as a reference for the kinds of issues idioms may address, but also as a conceptual bridge to our experimental design.

3.3. $\boldsymbol{RQ}_1$: On the Diffusion of (Refactorable-)Pythonic Idioms

To address $\mathbf{RQ}_1$, we proceeded from two sides. On the one hand, we needed to count the instances of snippets of code refactorable in *Pythonic* idioms. To achieve this, we employed RIdiom [13], a tool designed to identify *Refactorable-Pythonic* code by analyzing the Abstract Syntax Tree (AST). The tool was evaluated on over 7,000 Python projects and achieved 100% accuracy for all nine idioms. Table 3 describes the idioms refactorable by the tool.

To address $\mathbf{RQ}_{1.1}$, first, we instrumented RIdiom in order to count the instances of *Refactorable-Pythonic*. Then, to extract and quantify occurrences of *Pythonic* idioms in software projects, we developed a custom analysis tool. The analysis tool is composed of two main modules: PythonicExtractor and PythonicVisitor. The PythonicExtractor module is responsible for cloning the target repository locally and generating an AST for each Python file in the project. Once the AST is created, the PythonicVisitor module processes it by traversing each node. This module extends the NodeVisitor class and overrides the visit methods corresponding to the nine idioms under analysis, *i.e., Assign Multi Targets*, *Call*

Table 5: List of Code Smells Detectable Using DPy, the Mitigation Strategy, and Whether a Pythonic Idiom Is Applied.

Code Smell	Example	Refactored Code	Idioms Used	Implemented?	Description	Mitigation Rationale
Long Statement	`sum([x**2 for x in range(1, 101) if x % 2 == 0 and x > 10 and x < 90])`	`sum([x**2 for x in range(1, 101) if x % 2 == 0 and 10 < x < 90])`	List Comprehension, Chain Comparison	✅	Computes the sum of squares of even numbers in a range, but the verbose bounds make the expression hard to scan.	Chain comparison shortens the numeric test, keeping the list comprehension readable and reducing visual noise.
Long Parameter List	`def f(a, b, c, d, e, f): pass` `f(1, 2, 3, 4, 5, 6)`	`args = (1, 2, 3, 4, 5, 6)` `f(*args)`	Star in Function Call	✅	Defines and calls a function with six positional parameters, forcing every caller to remember order and count.	Packs values into one tuple and unpacks with *, making the call site shorter and signalling that the arguments travel as a bundle.

Long Method	`def process_data(data):`
	`  result = {}`
	`  for x in data:`
	`  if x:`
	`  x = x.strip().upper()`
	`  result[x] = len(x)`
	`  return result`
		`def process_data(data):`
		`  cleaned = [x.strip().upper() for x in data if x]`
		`  return {x: len(x) for x in cleaned}`
			List Comprehension, Dict Comprehension	✅	Cleans a list of strings and maps each to its length, but interleaves cleansing and accumulation in one loop.	Splits the concerns: a list comprehension does filtering/normalising, a dict comprehension builds the mapping, yielding shorter, testable code.

Complex Conditional	`if x > 10 and x < 50 and is_valid and not is_expired:` `  ...`	`if 10 < x < 50 and is_valid and not is_expired:` `  ...`	Chain Comparison, Truth-Value Test	✅	Evaluates bounds and flags in a verbose Boolean chain.	Chain comparison condenses the numeric test; relying on truthiness eliminates needless `== True/False` clutter.
Long Identifier	`total_number_of_successful_login = 12`	`total = 12`	Not Refactorable using Pythonic idioms	❌	Uses an overly wordy name for a simple scalar.	Replaces with a concise but meaningful identifier, improving scanability and maintenance.

Empty Catch Block	`try:`
	`  risky_operation()`
	`except:`
	`  pass`
		`try:`
		`  risky_operation()`
		`except Exception as e:`
		`  if e:`
		`  print(f"Error: {e}")`
			Truth-Value Test	✅	Silently ignores every exception, hiding failures.	Catches the specific base class, truth-tests the exception object, and logs it, surfacing problems without crashing.

Missing Default (match–case)	`match val:`
	`  case "a": ...`
	`  case "b": ...`
		`match val:`
		`  case "a": ...`
		`  case "b": ...`
		`  case _: handle_unknown()`
			Not Refactorable using Pythonic idioms	❌	Handles only two explicit cases, ignoring others.	Adds a wildcard branch so unexpected values are processed safely instead of falling through.

Long Lambda Function	`filter(`
	`  lambda user: len(user.posts) != 0,`
	`  users`
	`)`
		`[u for u in users if u.posts]`	Truth-Value Test, List Comprehension	✅	Anonymous lambda hides branching logic and forces a length check.	Uses truthiness (`if u.posts`) and a list comprehension to state intent directly, making the expression shorter and testable.

Long Message Chain	`user.get_profile().get_settings()`
	`.get_theme().get_color()`
		`profile, settings, theme = (`
		`  p := user.get_profile(),`
		`  s := p.get_settings(),`
		`  t := s.get_theme()`
		`)`
		`color = t.get_color()`
			Assign Multiple Targets	✅	Traverses four layers in one expression; a `None` anywhere raises late and obscures where it failed.	Binds each hop to a name in a single unpacking statement, surfaces errors sooner, and lets debuggers inspect each intermediate object.

Magic Number	`if age > 65:`	`SENIOR_AGE = 65` `if age > SENIOR_AGE:`	Not Refactorable using Pythonic idioms	❌	Uses a raw numeric literal with no context.	Extracts the constant into a named variable, documenting meaning and centralising future changes.

Complex Method	`def analyze(data):`
	`  result = {}`
	`  for item in data:`
	`  if isinstance(item, str):`
	`  if item:`
	`  cleaned = item.strip()`
	`  result[cleaned] = len(cleaned)`
	`  return result`
		`def analyze(data):`
		`  cleaned = [x.strip() for x in data if isinstance(x, str) and x]`
		`  return {x: len(x) for x in cleaned}`
			List Comprehension, Dict Comprehension	✅	Filters non-empty strings, strips them, and records their lengths with nested loops.	Comprehensions express filtering and mapping declaratively, producing shorter, clearer, and easily testable code.

Star, *List Comprehension*, *Dict Comprehension*, *Set Comprehension*, *Truth Value Test*, *Chain Compare*, *For Multi Targets*, and *For Else*. We selected these idioms because they are detectable using RIdiom. When a specific node is encountered, the module checks whether it conforms to the definition or identification rule associated with that idiom. For idioms with a dedicated AST node (e.g., *Dict Comprehension*), the module directly increments the corresponding counter. In contrast, for idioms that do not have a dedicated node (e.g., *Assign Multi Targets*, which uses the Assign node), the module verifies whether the code matches the definition provided in the *"The Zen of Python"* book [2]. Table 6 shows the Idioms Extraction based on AST Nodes and Conditions.

Table 6: Idioms Extraction based on AST Nodes and Conditions.

Idiom	AST Node	Condition
Assign Multi Targets	`ast.Assign`	`len(node.targets) > 1`
Call Star	`ast.Call`	`isinstance(expr, ast.Starred)`
List Comprehension	`ast.ListComp`	Presence
Dict Comprehension	`ast.DictComp`	Presence
Set Comprehension	`ast.SetComp`	Presence
Truth Value Test	`ast.If`	`not isinstance(node.test, ast.Compare)`
Chain Compare	`ast.Compare`	`len(node.ops) > 1`
For Multiple Targets	`ast.For`	`isinstance(node.target, ast.List)`
Loop Else	`ast.For`	`node.orelse Presence`

Examining the Table 6, we can observe that for specific nodes, such as *Assign Multi Targets*, detection can be performed if the current node is an `ast.Assign` node, i.e., the node corresponding to the assignment operation, and the condition is that the number of targets involved in the assignment is greater than one `(len(node.targets) > 1)`. This indicates that the statement assigns the same value to multiple variables simultaneously. Similarly, the detection of *List Comprehension* can be achieved by checking for the presence of an `ast.ListComp` node in the Abstract Syntax Tree. The presence of this node directly implies the use of a list comprehension construct, which syntactically represents the creation of a new list through an iterative expression within square brackets. The detection process implemented in the tool is deterministic by design, as it is based on the AST representation of the source code. Since the AST provides a canonical and unambiguous structural interpretation of syntactically correct Python code, each node type and its associated attributes can be mapped precisely to the definition of a given idiom. As a result, the mapping between idioms and their detection rules is one-to-one and rule-based, ensuring that the same code fragment

yields the same analytical outcome.

Before executing RIdiom against the considered software projects, we evaluated its actual detection capabilities. While the tool implements a set of heuristics grounded in the original *Pythonic* principles [2], we sought to ensure that these heuristics were correctly and consistently applied in practice. To this end, the first two authors jointly assessed a statistically significant random sample of the idiom instances detected by the tool across the dataset. Specifically, the validation was conducted using a sample size designed to achieve a 5% margin of error at a 95% confidence level. The goal of this manual inspection was to estimate the precision of the tool, that is, the proportion of correctly identified idioms among those flagged, thereby verifying whether the tool accurately implements the intended detection rules. We focused exclusively on precision, as estimating recall would have required a complete ground truth of all *Pythonic* idioms in the analyzed projects, which is not available. In fact, the lack of such a ground truth was a primary motivation for developing the custom detection tool in the first place. The results of this validation confirmed that the tool faithfully captures the idiomatic patterns it was designed to detect, corroborating the results obtained by the original authors [13] (accuracy = 100%) and supporting its use in ML systems.

To address the $\mathbf{RQ}_{1.2}$ and $\mathbf{RQ}_{1.3}$, we calculated both the percentage of *Pythonic* idioms and the *Pythonic* density in the source code as follows:

$$\text{Pythonic Perc.} = \frac{\#PythonicIdioms}{\#PythonicIdioms + \#Refactorable - PythonicIdioms} \tag{1}$$

Equation 1 calculates the *Pythonic* Percentage, which measures the share of idiomatic (*Pythonic*) code patterns relative to all idioms (both *Pythonic* and *Refactorable-Pythonic*) in the codebase. A higher value indicates a more idiomatic use of Python.

Finally, we computed the density as follows:

$$\text{(Refactorable-)Pythonic Density} = \frac{\#(Refactorable-)PythonicIdioms}{LOC} \tag{2}$$

Equation 2 computes the *(Refactorable-)Pythonic* Density, i.e., how often *Pythonic* or *Refactorable-Pythonic* appear per line of code. This helps us to assess how densely idiomatic patterns are used throughout the project.

3.4. $\mathbf{RQ}_2$: On the Relationship between Pythonic Idioms and Code Smells

To answer $\mathbf{RQ}_2$, we first checked the normality of the data to determine the most appropriate statistical test for our experiments. To assess the normality, we employed the *Shapiro-Wilk test* [40], which is particularly useful for small sample sizes and helps us understand whether our data is normally distributed. For each distribution, we split the data according to the NICHE dataset's "engineered" column, which distinguishes "well-engineered" from "non-engineered" projects - this distinction allowed us to address $\mathbf{RQ}_{2.1}$ and $\mathbf{RQ}_{2.2}$ separately using the same data analysis methods described in the following. Since the variables did not satisfy the normality assumption, we selected the Spearman test [41]—the non-parametric counterpart of Pearson [42]. We normalize data by dividing by NLOC. Furthermore, to mitigate the risk of false positives arising from multiple comparisons, we applied the Bonferroni correction to the resulting p-values [43]. We considered the p-value $\leqslant 0.05$ as statistically significant.

To perform our analysis, we considered the following dependent, independent, and control variables:

Dependent Variables: Since our objective is to evaluate the relationship between *(Refactorable)-Pythonic* idioms and code smells, we considered, as dependent variables, code smells detectable by DPy with a plausible mitigation strategy through *Pythonic* idioms. In particular, we thought the following Python-specific code smells: *Long Statement*, *Long Parameter List*, *Long Method*, *empty Catch Block*, *Complex Method*, *Complex Conditional*, *Long Lambda Function*, and *Long Message Chain*.

Independent Variable: As independent variables we selected all the nine *(Refactorable)-Pythonic* idioms detectable by RIdiom, *i.e., Assign Multiple Targets*, *Call Star*, *List Comprehension*, *Dict Comprehension*, *Set Comprehension*, *Truth Value Test*, *Chain Compare*, *For Multi Targets*, and *Loop Else*.

Control Variables: We selected the number of lines of code without comments (NLOC), the cyclomatic complexity, and the number of commits. The NLOC serves as an indicator of the codebase's size, a factor that naturally correlates with the likelihood of encountering code smells [7]. Cyclomatic complexity reflects the structural intricacy of the code and can influence the presence of code smells independently of the proportion of Python code. Finally, the number of commits captures the overall development and maintenance activity within a project's history, offering insight into how actively the codebase has evolved. We extracted these metrics using SonarQube [44].

It is important to note that, on the one hand, we performed the same analysis considering *Pythonic* and *Refactorable-Pythonic* idioms; on the other hand, we consider only these metrics as control variables due to the lack of useful tools for extracting other candidate control variables.

4. Analysis of the Results

Before reporting and discussing our results, we provide a preliminary statistical description of the NICHE dataset of code smell diffusion.

Table 7: Descriptive statistics of code smells in well-engineered and not well-engineered projects.

Group	Mean	Std. Dev.	Median	Min	Max
Well-Engineered	785.790	587.650	616	9	2,782
Non-engineered	459.270	499.090	284	5	2,576

Table 7 presents descriptive statistics on Python-specific code smells, comparing "well-engineered" and "non-engineered" projects. On average, "well-engineered" projects exhibit a higher number of Python-specific smells (mean = 785.790) compared to "non-engineered" projects (mean = 459.270). The standard deviation values are also substantial in both groups, indicating considerable variability in smell counts across projects. These findings may suggest that "well-engineered" projects tend to contain more code overall—or that smell detection is more thorough in such projects—resulting in higher raw counts of Python-specific code smells.

To understand whether the categorization of projects into "well-engineered" and "not-engineered" groups can be representative for our investigations, we analyzed the distributions of *Pythonic* percentage and density using the

Mann–Whitney U test and Cliff's δ to assess effect size. The Mann–Whitney test was chosen because the data did not follow a normal distribution. Regarding the *Pythonic* percentage, we found a statistically significant difference between "well-engineered" and "non-engineered" projects (p-value = 0.0195). When considering a one-sided test, we observed that the distribution of *Pythonic* idioms in "well-engineered" projects is significantly higher than in "non-engineered" projects (p-value = 0.0098). However, the effect size was relatively low (Cliff's δ 0.179). When we considered the density, we discovered that, similarly, the distribution of idioms in "well-engineered" projects is statistically significantly greater than in "non-engineered" projects (p-value = 1.93e-05) and exhibits a moderate effect size (Cliff's δ = 0.315). Overall, these findings support the suitability of the selected dataset for our analysis. The observed statistical differences in idiom adoption between "well-engineered" and "non-engineered" projects demonstrate that the underlying categorization captures meaningful distinctions in development maturity and coding practices. As a consequence, the dataset provides an appropriate and empirically grounded foundation for investigating the relationship between idiomatic usage and code quality.

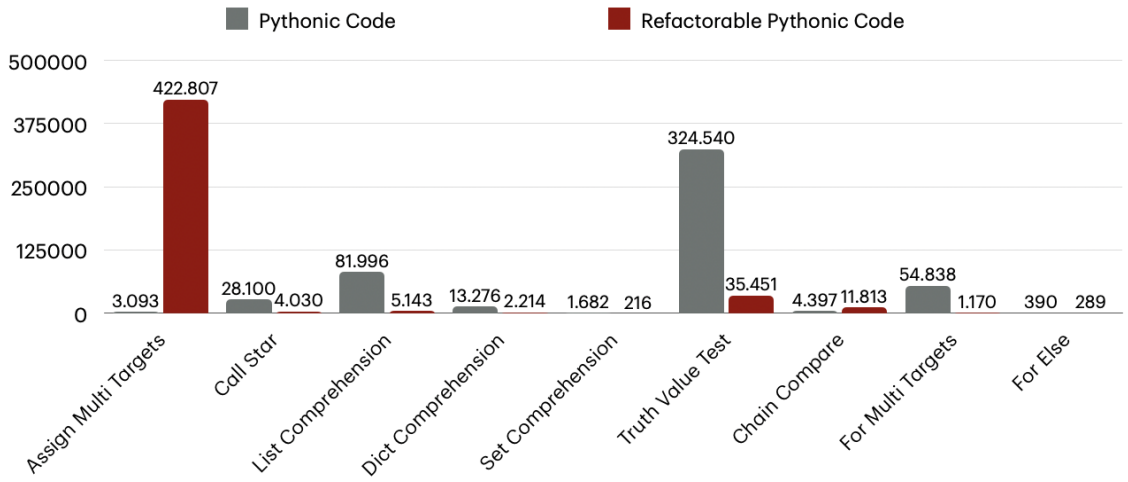

Figure 2: Frequencies of Pythonic and Refactorable-Pythonic Idioms.

4.1. $\boldsymbol{RQ}_{1.1}$*: On the Frequencies of Idioms in ML Projects*

Figure 2 displays the frequencies of *Pythonic* (gray bar) and *Refactorable-Pythonic* (red bar) idioms. Among *Pythonic* idioms, the most frequent is the *Truth Value Test*, followed by *List Comprehension*, and *For Multiple Targets*. These idioms enable concise and readable code, particularly when dealing with loops or complex data manipulations. In addition, among the "comprehension" constructs, *List Comprehension* is the most widely used. This result corroborates previous ones [3], identifying *List Comprehension* as one of the most adopted idioms in the Python codebase.

Moving on to *Refactorable-Pythonic* idioms, the most frequent idiom observed is *Assign Multiple Targets*, followed by *Truth Value Test* and *Chain Compare*. These results could suggest that practitioners may often write multiple assignment statements separately, rather than utilizing Python's ability to assign multiple variables in a single

statement. In both *Pythonic* and *Refactorable-Pythonic* idioms, the less frequent idiom is *For Else*, suggesting the unpopularity of this construct in ML projects.

RQ$_{1.1}$ Summary

The most frequent idioms are *Truth Value Test* and *Assign Multi Targets* for *Pythonic* and *Refactorable-Pythonic* idioms, respectively, while the least frequent is the *For Else* idiom.

4.2. ***RQ***$_{1.2}$*: On the Variation of Project Size*

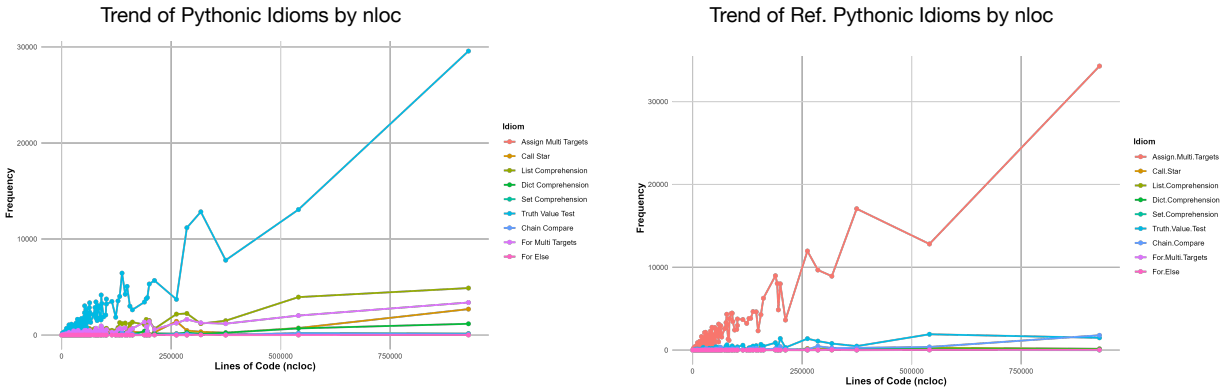

Figure 3: Variation of Pythonic Idioms and Non-Pythonic Idioms with Respect to Project Size.

Figure 3 illustrates the variation in the usage of *Pythonic* idioms (left side) and *Refactorable-Pythonic* idioms (right side) with respect to the project size (in NLOC). As shown in the figure, the most frequently adopted *Pythonic* idiom is the *Truth Value Test*. In contrast, the most prevalent *Refactorable-Pythonic*, refactorable idiom is *Assign Multiple Targets*. In both cases, it is possible to observe that the growth of the *Truth Value Test* and *Assign Multiple Targets* idioms accelerates dramatically after 250,000 lines. We observed that as the project size increases, both *Pythonic* and *Refactorable-Pythonic* idioms increase. The substantial rise in *Truth Value Test* indicates that *Pythonic* best practices become increasingly essential as the complexity of the codebase grows. At the same time, the persistent use of *Assign Multi Targets* highlights the presence of Refactorable-idiomatic patterns prevalent in larger projects, underscoring the need for targeted refactoring efforts.

RQ$_{1.2}$ Summary

If we normalize by project size, the most prevalent *Pythonic* is *Truth Value Test*, and the most prevalent *Refactorable-Pythonic* idiom is *Assign Multiple Targets*.

4.3. ***RQ***1_3*: On the Density of Idioms*

Figure 4 presents the density distribution of idioms relative to NLOC, separately for *Pythonic* idioms (left) and *Refactorable-Pythonic* idioms (right). Among the idioms considered, *Truth Value Test* emerges as the most frequently used in the Pythonic category, while *Assign Multiple Targets* is the most prevalent among the refactorable subset. These findings align with the observations made in RQ$_{1.2}$, where both idioms also appeared as prominent in terms of frequency.

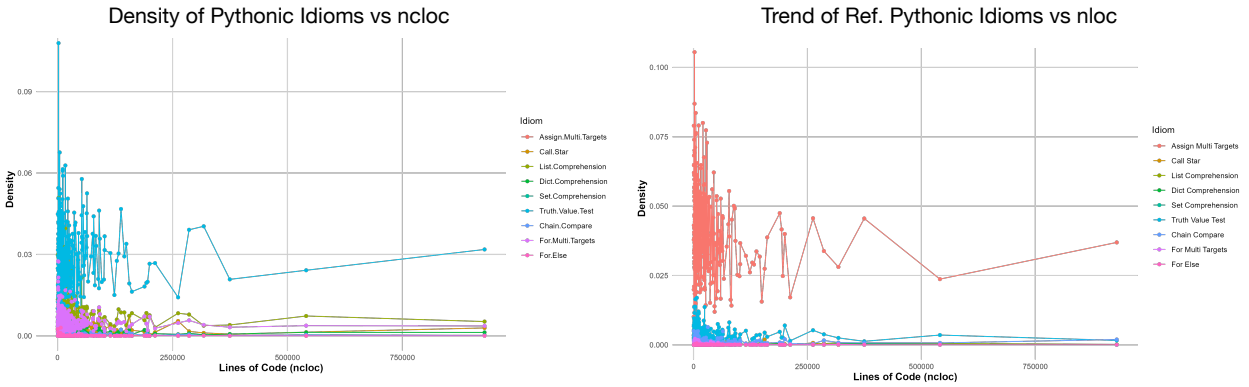

Figure 4: Density of Pythonic and Refactoraizable-Pythonic Idioms.

The high density of *Truth Value Test* likely reflects its widespread use as a concise and idiomatic pattern for evaluating conditions—an operation that is both common and semantically lightweight. Its idiomaticity and simplicity may explain its frequent appearance even in projects with varying levels of engineering discipline. Similarly, *Assign Multiple Targets* is syntactically compact and often used to streamline code, which may contribute to its diffusion across projects regardless of quality.

$RQ_{1.3}$ Summary

The most adopted *Pythonic* idiom in terms of density is *Truth Value Test*, while the most Refactorable idiom is *Assign Multi Targets*.

4.4. RQ_2: On the Relationship Between Idioms and Smells

Before analyzing the differences observed between "well-engineered" and "non-engineered" projects, let us provide an overall overview of the correlation between the presence of idioms and code smells. Table 8 shows the results of our statistical test, with p-values adjusted with Bonferroni correction [43]. We can observe that, even after normalizing, certain *Refactorable-Pythonic* idioms (*e.g., Call Star*, *Chain Compare*) maintain moderate positive correlations ($\rho \approx 0.30$–0.36) with Python-specific code smells (such as *Complex Conditional*, and *Long Parameter List*). *Pythonic* constructs like *Truth Value Test* also appear, but with smaller effect sizes ($\rho \approx 0.23$–0.26). In contrast, control variables, especially *NLOC*, emerge as dominant factors. For instance, $\rho(\text{NLOC}, \text{Long Statement}) = 0.73$ and $\rho(\text{NLOC}, \text{Long Method}) = 0.69$, indicating that file size is the strongest predictor of code smell incidence. Additionally, *File Complexity* correlates with *Complex Method* at $\rho = 0.37$ ($p < 10^{-11}$, ***), and *Commits* also shows moderate correlations with several smells. Overall, these results indicate that while idiom usage, particularly of *Refactorable-Pythonic* constructs, is related to certain code smells, structural factors such as file size and complexity are more strongly associated with the incidence of these smells. This provides a baseline understanding for interpreting the role of engineering practices in the patterns observed. In the remainder of this section, we turn to the analysis of the results that directly addresses $\mathbf{RQ}_{2.1}$ and $\mathbf{RQ}_{2.2}$.

Table 8: Spearman Correlation between *(Refactorable-)Pythonic* Idioms and Control Variables versus Code Smells (normalized by NLOC), with Bonferroni-adjusted significance levels

Idiom	Smell	Spearman$_\rho$	p-value	Significance
Refactorable-Pythonic Chain Compare	Complex Conditional	0.35	2.92e-10	***
Refactorable-Pythonic Call Star	Long Parameter List	0.34	1.86e-09	***
Assign Multi Targets	Long Parameter List	0.30	7.30e-08	***
Refactorable-Pythonic Chain Compare	Complex Method	0.30	1.23e-07	***
Refactorable-Pythonic Call Star	Complex Method	0.30	1.35e-07	***
Refactorable-Pythonic Call Star	Long Method	0.28	9.90e-07	***
Refactorable-Pythonic Call Star	Complex Conditional	0.26	3.92e-06	**
Refactorable-Pythonic List Compreh.	Complex Method	0.26	4.79e-06	**
Truth Value Test	Complex Method	0.25	1.17e-05	**
Refactorable-Pythonic Chain Compare	Long Parameter List	0.25	1.47e-05	**
Set Comprehension	Empty Catch Block	0.24	2.50e-05	**
Refactorable-Pythonic Chain Compare	Long Method	0.24	2.73e-05	**
Assign Multi Targets	Complex Method	0.23	5.20e-05	*
Refactorable-Pythonic Call Star	Long Statement	0.23	5.66e-05	*
Truth Value Test	Complex Conditional	0.23	5.85e-05	*
Truth Value Test	Empty Catch Block	0.23	6.54e-05	*
Chain Compare	Complex Method	0.22	8.08e-05	*
Assign Multi Targets	Long Method	0.22	9.70e-05	*
Control Variable	**Smell**	**Spearman$_\rho$**	**p-value**	**Significance**
NLOC	Long Statement	0.73	1.64e-51	***
NLOC	Long Method	0.69	1.62e-44	***
NLOC	Complex Method	0.65	1.75e-38	***
NLOC	Long Parameter List	0.61	1.16e-31	***
NLOC	Complex Conditional	0.41	7.72e-14	***
NLOC	Long Lambda Function	0.31	2.31e-08	***
NLOC	Long Message Chain	0.25	1.36e-05	*
File Complexity	Complex Method	0.37	1.65e-11	***
File Complexity	Complex Conditional	0.32	1.09e-07	***
File Complexity	Long Method	0.31	2.54e-08	***
File Complexity	Long Parameter List	0.27	2.27e-06	***
Commits	Long Statement	0.29	2.49e-07	***

Continued on next page

Idiom	Smell	Spearman$_\rho$	p-value	Significance
Commits	Complex Method	0.27	1.44e-06	***
Commits	Empty Catch Block	0.24	1.98e-05	*
Commits	Long Method	0.22	8.43e-05	*

$^*\, p < 0.1;\ ^{**}\, p < 0.05;\ ^{***}\, p < 0.01$

4.4.1. $\boldsymbol{RQ}_{2.1}$: *On the Correlation between Idioms and Code Smells in Well-Engineered Projects*

Table 9: Spearman Correlation between *(Refactorable-)Pythonic* Idioms and Control Variables versus Code Smells for "Well-Engineered" Projects (normalized by NLOC), with Bonferroni-corrected significance levels

Idiom	Smell	Spearman$_\rho$	p-value	Significance
Chain Compare	Complex Conditional	0.35	4.48e-08	***
Call Star	Long Parameter List	0.33	5.01e-07	***
Chain Compare	Complex Method	0.31	1.78e-06	***
Call Star	Complex Method	0.29	6.26e-06	**
Call Star	Complex Conditional	0.27	2.94e-05	**
Chain Compare	Long Parameter List	0.26	7.31e-05	*
Truth Value Test	Complex Method	0.26	6.54e-05	*
Call Star	Long Method	0.26	9.46e-05	*
Control Variable	**Smell**	**Spearman$_\rho$**	**p-value**	**Significance**
NLOC	Long Statement	0.71	4.49e-13	***
NLOC	Long Method	0.69	2.91e-11	***
NLOC	Complex Method	0.64	9.38e-10	***
NLOC	Long Parameter List	0.61	5.39e-08	***
NLOC	Complex Conditional	0.39	1.21e-04	***
File Complexity	Complex Method	0.48	1.19e-05	**
File Complexity	Long Method	0.31	2.50e-06	***
File Complexity	Long Parameter List	0.25	1.06e-04	*

$^*\, p < 0.1;\ ^{**}\, p < 0.05;\ ^{***}\, p < 0.01$

Restricting the analysis to "well-engineered" projects (Table 9), we observe that the same idioms, i.e., particularly *Call Star* and *Chain Compare*, still correlate moderately with Python-specific smells such as *Complex Conditional* and *Long Parameter List* ($\rho \approx 0.30$–0.36). This suggests that the mere presence of engineering practices (e.g., tests, CI, structured packaging) does not eliminate the co-occurrence of idioms and smells, especially for constructs that inherently increase syntactic or semantic complexity. Interestingly, *Pythonic* idioms show weaker associations overall, reinforcing the intuition that idioms aligned with Python's idiomatic style may be less likely to contribute

to maintainability issues, even in well-structured repositories. Such a correlation analysis seems to reinforce the value of idiomatic style, while also highlighting that some idioms, when overused or applied without sufficient caution, may correlate with specific smells, regardless of the engineering discipline being adopted. However, the strongest correlations remain associated with control variables, particularly structural complexity indicators. *NLOC* continues to be the most dominant factor, with $\rho(\text{NLOC} = 0.71)$ for *Long Statement* and $\rho(\text{NLOC} = 0.69)$ for *Long Method*, showing that larger files are more prone to certain smells, even in projects with sound engineering processes. *File Complexity* and *Commits* also show non-negligible correlations, emphasizing the multifaceted nature of smell emergence. These findings suggest that while engineering practices improve overall quality, they do not fully mitigate the stylistic or structural risks associated with specific idiom usages. Therefore, effective engineering should combine process-oriented practices with style-aware and complexity-aware development guidelines.

$RQ_{2.1}$ Summary

Considering "well-engineered" projects, we observe that some idioms (*e.g., Call Star* and *Chain Compare*) positively correlate with the presence of smells. Our results also confirm the correlation between NLOC and file complexity with some smells.

4.4.2. ***$RQ_{2.2}$****: On the Correlation between Idioms and Code Smells in non-engineered Projects*

Table 10: Spearman Correlation between *(Refactorable-)Pythonic* Idioms and Control Variables versus Code Smells for "Not-Engineered" Projects (normalized by NLOC), with Bonferroni-corrected significance levels

Control Variable	Smell	Spearman_ρ	p-value	Significance
NLOC	Long Statement	0.71	4.49e-13	***
NLOC	Long Method	0.67	2.91e-11	***
NLOC	Long Parameter List	0.64	4.53e-10	***
NLOC	Complex Method	0.63	8.29e-10	***
File Complexity	Complex Method	0.48	1.19e-05	**

$^*\, p < 0.1$; $^{**}\, p < 0.05$; $^{***}\, p < 0.01$

The results for "non-engineered" projects (Table 10) show a markedly different profile compared to their well-engineered counterparts. Here, control variables overwhelmingly dominate the top correlations. In particular, we noticed: $\rho(NLOC, \textit{Long Statement}) = 0.71$, $\rho(NLOC, \textit{Long Method}) = 0.67$, and $\rho(NLOC, \textit{Long Parameter List}) = 0.64$. On the one hand, these strong associations seem to confirm that, in the absence of structured engineering practices, code size alone becomes the primary driver of code quality degradation. Moreover, *File Complexity* also shows a substantial correlation with *Complex Method* ($\rho = 0.48$, **), reinforcing the idea that cyclomatic or structural complexity further compounds the likelihood of smell incidence. Interestingly, no statistically significant correlations were found between idioms, either *Pythonic* or *Refactorable-Pythonic*, and code smells in this group. This absence suggests that, in poorly governed codebases, the stylistic layer provided by idioms is largely irrelevant in the face of overwhelming structural issues. This stark contrast with the "well-engineered" subset highlights an important

insight: idiom–smell associations may only become visible or meaningful when basic quality controls are in place. In other words, engineering practices appear to serve as a kind of baseline hygiene, allowing finer-grained influences, such as idiomatic usage, to emerge.

On the other hand, the relatively low number of "non-engineered" projects (76)—and therefore a low statistical power (e.g., at least 80 data points would be necessary to observe a correlation of 0.3, and over 200 for a correlation of 0.2)—may be a reason for failing to find statistically significant correlations with idiom-related variables. Therefore, such results must be interpreted cautiously and deserve further investigations on larger datasets.

$RQ_{2.2}$ Summary

In "non-engineered" projects, code smells are most strongly associated with structural control variables rather than idiomatic usage. *NLOC* shows the highest correlations, particularly with *Long Statement* ($\rho = 0.71$), *Long Method* ($\rho = 0.67$), and *Long Parameter List* ($\rho = 0.64$). *File Complexity* also correlates with *Complex Method* ($\rho = 0.48$). No significant correlations involving idioms were found, yet this may be due to the relatively small number of "non-engineering" projects.

5. Discussion and Implications

This section first provides a synthesis of the key findings from the analysis performed in $\mathbf{RQ}_1$ and $\mathbf{RQ}_2$, to make more transparent the logical progression from the individual analyses to the main outcomes. Then, we enhance the generalizability of our results in a less controlled environment by performing further investigation on a different sample of ML projects. Finally, we provide implications and discussion for both practitioners and researchers.

5.1. Summarizing the Findings

From $\mathbf{RQ}_1$, we observe that idioms such as *Truth Value Test* and *Assign Multiple Targets* are among the most frequently used across ML projects, suggesting that developers systematically rely on these constructs when implementing control-flow structures or variable assignments. At the same time, idioms like *For Else* remain rarely employed, confirming their marginal role in day-to-day development and suggesting that certain *Pythonic* constructs, although expressive, may not align well with the needs of ML workflows.

When considering how idiom adoption varies with project size, the results show a growth in both *Pythonic* and *Refactorable-Pythonic* idioms as systems scale. In particular, the adoption of the *Truth Value Test* increases significantly in large-scale projects, indicating that idioms become increasingly central as practitioners attempt to achieve readability and maintainability in complex ML systems. Conversely, the widespread use of *Assign Multiple Targets* in these systems highlights that some developers continue to prefer compact but less idiomatic constructs that, while valid, may compromise long-term clarity.

The analysis of idiom density confirms these tendencies. The high density of *Truth Value Test* instances across both small and large projects suggests that this idiom is deeply embedded in Python, while the dense presence of *Assign Multiple Targets* reveals a recurring stylistic pattern that might represent an intermediate step toward fully

idiomatic code. These observations suggest a development landscape in which *Pythonic* idioms are not only pervasive but also evolve alongside a project's growth and engineering maturity.

RQ$_1$ Summary

The results highlight a clear diffusion of *(Refactorable-)Pythonic* idioms in ML systems, with their frequency and density increasing as projects become larger and more complex. The prevalence of idioms like *Truth Value Test* reflects a well-established idiomatic culture among developers, whereas the continued use of patterns such as *Assign Multiple Targets* indicates areas where idiomatic practices do not appear to be mature yet.

Moving to **RQ$_2$**, we observe that in "well-engineered" projects, the adoption of both *Pythonic* and *Refactorable-Pythonic* idioms shows statistically significant correlations with several code smells, suggesting that idiomatic usage does not always align with higher code quality. This finding challenges the common assumption that idiomatic code inherently promotes maintainability.

In contrast, for "non-engineered" projects, the emergence of code smells appears to be connected to structural factors such as code size, complexity, and development activity rather than idiomatic usage. In these projects, we observe that idioms play a comparatively minor role in influencing code quality, indicating that the lack of disciplined engineering practices overshadows any potential benefit of writing idiomatic code. This result emphasizes that the effectiveness of idiomatic constructs depends not only on their structure but also on the surrounding engineering context in which they are applied.

RQ$_2$ Summary

(Refactorable-)Pythonic idioms do not consistently correlate with higher code quality. In "well-engineered" projects, idioms are statistically associated with code smells, while in "non-engineered" projects, quality issues are more influenced by external factors like size and complexity.

On the basis of the findings of the study, we may finally address the overarching question of the study asked in Section 1, i.e., whether the use of *Pythonic* idioms correlate with higher or lower code quality in real-world ML codebases: applying idioms or refactoring to *Pythonic* style does not guarantee cleaner or more maintainable code. While idioms promote readability and conciseness, their misapplication may introduce or obscure structural issues. The effectiveness of idioms may be influenced by the developers' background, project size, and architectural practices. Therefore, *Pythonic* idioms should be considered complementary tools rather than guarantees of quality, requiring careful and informed use supported by context-sensitive development tools.

Main Question Summary

The use of *Pythonic* idioms does not inherently lead to higher code quality in ML projects. In well-structured projects, idioms can complement good engineering practices and contribute to maintainability; however, when applied inconsistently or without architectural awareness, they may coexist with or even mask design issues. Hence, idioms should be regarded as supportive practices rather than indicators of quality, whose real benefits emerge only within disciplined and context-aware development environments.

5.2. Further Analysis

To enhance the generalizability of our results, we replicated our experiments on an additional set of 30 small- to medium-scale ML projects, representing approximately 10% of those analyzed in NICHE [9]. To identify these projects, we relaxed two of the original inclusion criteria, *History* and *Population*, previously adopted as proxies for project popularity and development maturity. The main motivation behind this additional analysis was to assess whether our findings could be generalized to a broader set of projects selected under less stringent conditions. After cloning projects, we replicated our experiments without applying the "well-engineered" and "non-engineered" categorization used in NICHE: Table 11 reports the statistical description of the sample.

Table 11: Descriptive statistics of project characteristics.

Statistic	Stars	Commit	NLOC	Python %
Min	50	56	1,239	19.5%
Mean	72.70	81.27	4,221	44.77%
Median	71	83	2,189	38.62%
Max	96	100	46,403	95.7%

As shown in the table, the projects exhibit similar central tendencies for both *stars* and *commits*, with a mean of 72.70 and 81.27, respectively and a median of 71 and 83. This suggests a relatively balanced distribution for these two metrics. Conversely, the *NLOC* and *Python* percentage values display a much wider range, indicating higher variability across projects. In particular, *NLOC* ranges from 1,239 to 46,403 (mean 4,221; median 2,189), while the *Pythonic* Percentage spans from 19.5% to 95.7% (mean 44.77%; median 38.62%). These results reveal a substantial dispersion in project size and code "Pythonicity".

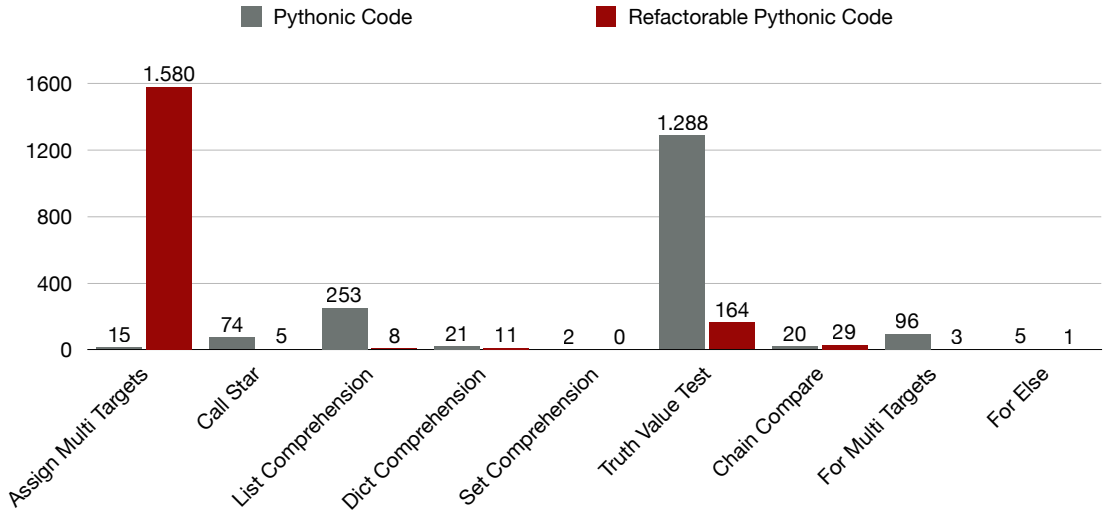

Figure 5: Frequencies of Pythonic and Refactorable-Pythonic Idioms.

RQ*$_{1.1}$*: What are the most frequent (Refactorable-)Pythonic idioms in ML projects?. Figure 5 illustrates the frequencies of the most commonly adopted *Pythonic* idioms as well as those that are most prone to refactoring. As shown in the figure, *Assign Multi Targets* emerges as the idiom most frequently refactored, whereas *Truth Value Test* appears as the most widely adopted Pythonic construct. In both cases, *For Else* represents the least frequent idiom. Taken together, these observations align with the results reported in **RQ**$_{1.1}$, thereby providing additional support for the earlier findings and pointing toward a comparable distribution trend.

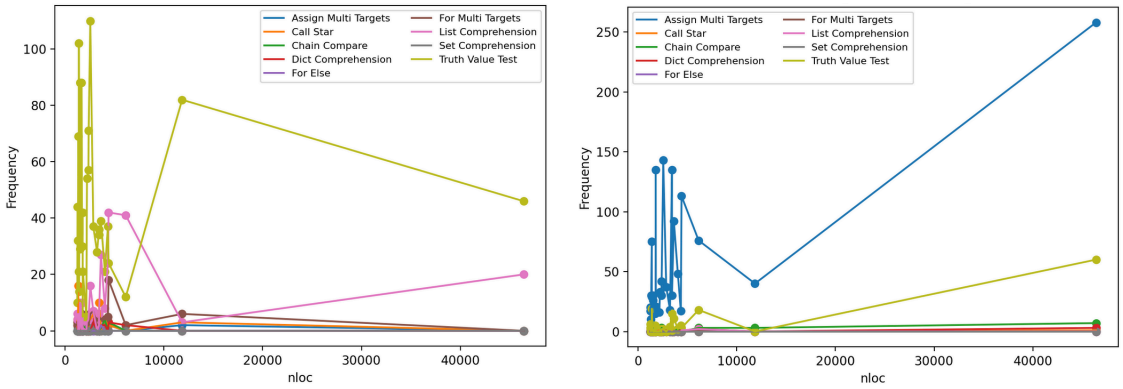

Figure 6: Results of the frequencies of *(Refactorable-)Pythonic* idiom at the growth of LOC.

RQ*$_{1.2}$*: How does the adoption of (Refactorable-)Pythonic idioms vary as the project grows in size?. Figure 6 shows the results of the frequencies of *(Refactorable-)Pythonic* idiom at the growth of *LOC*. From the *Pythonic* idioms chart (left chart), three distinct patterns can be observed: (1) idioms that show initial instability, then fall to a minimum peak, then rise again and then fall again; (2) idioms that show initial instability, then fall to a maximum peak, then fall and then rise again; and (3) idioms whose density remains roughly constant as the number of lines of code increases. The most frequent idiom belongs to the first category. Specifically, *Truth Value Test* idiom exhibits a substantially higher adoption rate than other idioms, particularly during the initial development phases. However, despite its early prominence, its usage tends to decrease after approximately 10,000 lines of code. The *List Comprehension* idiom ranks second in frequency and falls into the second category (increase–decrease–increase). Its adoption grows up to around 5,000 lines of code, declines thereafter, and rises again beyond 10,000 lines, suggesting that this idiom is more prevalent in medium-to-large projects. Finally, we observe idioms whose frequencies remain relatively stable as project size increases (e.g., *For Else*). The variations in their usage are less pronounced compared to idioms such as *Truth Value Test* or *List Comprehension*, suggesting that their changes only marginally affect the overall distribution of idiom frequencies. Turning to the *Refactorable-Pythonic* idioms (right chart), two patterns emerge: (1) idioms that show initial instability, then reach a minimum peak and then rise; and (2) idioms whose value remains roughly constant as the number of lines of code increases. The most frequent idiom, i.e., *Assign Multi Targets*, falls in the first category. In the early stages of a project, its adoption appears unstable, showing frequent upward and downward fluctuations. However, its usage becomes highly relevant after 10,000 lines of code. The *Truth Value Test* displays a similar pattern, though its overall frequency is significantly lower than that of *Assign Multi Targets*. In both cases, we observe that most oscillations in idiom frequencies are concentrated within the range of

0–5,000 lines of code. This concentration can be explained by considering the distribution of project sizes; Indeed, the mean value is around 4,000 lines of code, while the median is approximately 2,000. These values indicate that the majority of projects fall within this interval, which could explain why the *Refactorable-Pythonic* idioms occur with the highest variability in this specific range. Overall, when we compare these results with the main findings of our study, we can observe a consistent pattern. In fact, as shown in Figure 2, the most adopted *Pythonic* idiom is the *Truth Value Test*, while among *Refactorable-Pythonic* idioms, the most frequent is *Assign Multi Targets.*

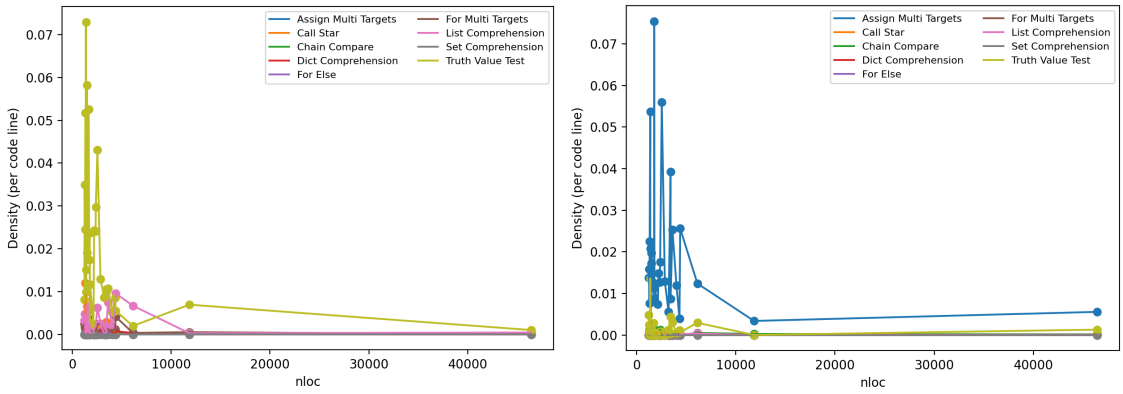

Figure 7: Results of the densities of (refactorable-) Pythonic idiom at the growth of LOC.

RQ1.3: What is the density of (Refactorable-)Pythonic idioms in source code?. Figure 7 shows the results of the densities of (Refactorable-) Pythonic idiom at the growth of LOC. Starting from the *Pythonic* idioms (left chart), three patterns arise: (1) idioms that show initial instability then go up and down; (2) idioms that show initial instability, then rise to a peak, then fall to a steady level; and (3) idioms that remain constant at the growth of lines of code. The *Truth Value Test* remains the most prominent idiom (which falls under the first category), showing a remarkably high density in small projects. Its sharp decline as the codebase grows reinforces the idea that developers tend to use this idiom more frequently in compact or exploratory code. At the same time, its presence becomes diluted as systems grow in size. The *List Comprehension* idiom (second category) follows a similar initial pattern but with smoother variations, until it becomes constant, indicating sustained, though less intense, adoption across project sizes. Its density suggests that while frequently used, it represents a stylistic choice distributed more evenly across different development stages. Finally, as also noted in $\mathbf{RQ}_{1.2}$, we can see idioms that remain fairly constant as the number of lines of code increases (e.g., *Call Star*). Turning to the *Refactorable-Pythonic* idioms (right chart), two patterns emerge: (1) idioms that show initial instability, then decline, and then remain constant as the number of lines of code increases, and 2) idioms whose density remains constant as the number of lines of code increases. Looking at the figure, we can see that the most dense idiom is the *Assign Multi Targets* idiom (first category), which is the most dominant. Its density peaks within the first few thousand lines of code, showing irregular yet concentrated use, before gradually stabilizing as projects grow. This indicates that refactorable idioms tend to appear more frequently in smaller codebases, possibly due to early design experimentation or less-structured implementation phases. The *Truth Value Test* in this context displays a similar shape, although with a considerably lower overall density, confirming its secondary role among the *Refactorable-Pythonic* idioms. The remaining idioms,

such as *Call Star*, *Chain Compare*, and *Dict Comprehension*, show minimal variation, maintaining consistently low density across different project sizes. Overall, when we compare the results with the main findings of $\mathbf{RQ}_{1.3}$, we can notice that the density trends for both *(Refactorable-)Pythonic* idioms exhibit a similar overall behavior.

Table 12: Spearman correlation between *(Refactorable-)Pythonic* Idioms and code smells.

Idiom	Smell	Spearman	P-value	Sig.
Truth Value Test	Long Method	0.657	0.009684	***
Truth Value Test	Empty catch block	0.609	0.032123	**
File Complexity	Complex Method	0.678	0.004533	***

RQ2: On the relationship between (Refactorable-)Pythonic idioms and code smells. Table 12 shows the results of the comparison between *(Refactorable-)Pythonic* idioms and code smells. The results show that *Truth Value Test* is the only Pythonic idiom showing a statistically significant association with code smells. Specifically, it shows a strong positive correlation (***) with the *Long Method* smell, suggesting that projects with a higher prevalence of this idiom tend to include longer methods. We also observed a moderate correlation (**) with the *Empty Catch Block* smell. The control variable *File Complexity* shows a strong correlation with the *Complex Method* smell. This result is not surprising, as the concepts are conceptually linked to the notion of code complexity. This result refines and contextualizes our main findings: in this extended sample, the associations between *(Refactorable-)Pythonic* idioms and code smells appear weaker and more localized than in the main dataset, suggesting that the relationships observed earlier are not universal but depend on project context and engineering maturity.

In summary, the results of this extended analysis partially confirmed the main findings observed in the study. In particular, for $\mathbf{RQ}_1$, we observed that *Truth Value Test* still remains the most widely adopted *Pythonic* idiom, while *Assign Multi Targets* is the most prevalent among *(Refactorable-)Pythonic* idioms. In contrast, when we considered $\mathbf{RQ}_2$, we found that a limited subset of idioms is statistically correlated with smells, suggesting that this outcome may depend on external factors such as project size or the inclusion of specific filter criteria.

5.3. Implications for Practitioners and Researchers

While the findings of our study are correlational rather than causal, they nonetheless provide actionable insights for both practitioners and researchers seeking to understand how the use of *Pythonic* idioms affects code quality in ML projects. From $\mathbf{RQ}_1$, we observed that idioms such as *Truth Value Test* and *Assign Multiple Targets* are among the most frequently used across ML projects—both in absolute frequency and density. However, the results of $\mathbf{RQ}_2$ revealed that these same idioms exhibit moderate to weak positive correlations with various code smells. Specifically, the *Truth Value Test* correlates with *Complex Method*, while *Assign Multi Targets* shows moderate correlation with *Long Parameter List* and weaker correlations with *Long Method* and *Complex Method*. These patterns indicate that even idioms typically introduced to improve readability and maintainability may coexist with structural or semantic issues, potentially reflecting deeper quality concerns within the software. Such ignishts are evident in "well-engineered" ML projects, where idioms are more consistently adopted but still appear alongside a high prevalence

of Python-specific code smells. This finding underscores that *Pythonic* idioms may obscure structural complexity if applied without architectural awareness.

For practitioners, these observations reinforce that idioms should be treated as stylistic instruments to enhance clarity, rather than as automatic indicators of good design. Developers should combine their use with engineering practices such as modularization, systematic refactoring, and code review. Incorporating idiom-awareness into code review processes can help teams identify potentially misleading uses of idiomatic constructs, especially within large conditional blocks or exception-handling routines. Automated tools could further support this process by detecting recurring co-occurrences between idioms and code smells, providing early feedback before complexity escalates.

Implication for Practitioners. Practitioners should treat *Pythonic* idioms as stylistic tools to improve clarity rather than as indicators of code quality. Their use must be complemented by modularization, systematic refactoring, and code review to avoid introducing or concealing structural complexity.

The interpretation of the *Pythonic Percentage* and *Pythonic Density* metrics introduced in the study also carries practical implications. These indicators can serve as diagnostic measures for evaluating how consistently and intensively idiomatic practices are applied across a project. A higher *Pythonic Percentage* suggests that idioms are more uniformly distributed, while *Pythonic Density* reflects their concentration relative to project size and complexity. These measures can help practitioners identify codebase where idioms may be either under- or overused, guiding targeted refactoring and style alignment efforts. Importantly, these metrics should not be interpreted as direct proxies for quality, but as contextual indicators of structural patterns that can inform maintenance decisions.

Implication for Practitioners. The *Pythonic Percentage* and *Pythonic Density* metrics can be used as diagnostic indicators to assess how idiomatic practices are distributed and concentrated across a project, to guide targeted refactoring and balanced style alignment.

For researchers, our findings emphasize the need for more fine-grained investigations that move beyond syntactic detection. Future work should explore how idioms interact with quality attributes such as maintainability, defect proneness, and testability, and examine the causal mechanisms that link idiomatic usage to design degradation. Expanding current datasets to include version histories and developer-level information could reveal whether idioms emerge as a result of deliberate refactoring or evolve organically through maintenance activities.

Implication for Researchers. Researchers should move beyond syntactic detection and examine how idioms interact with deeper quality attributes such as maintainability, defect proneness, and testability. Longitudinal studies including version histories and developer information could clarify whether idioms emerge through deliberate refactoring or organic code evolution.

Concerning the *Pythonic Percentage* and *Pythonic Density* metrics, these enable normalized comparisons across heterogeneous projects, facilitating empirical analyses of how idiomatic usage scales with system growth and its relationship to maintainability or complexity. Importantly, both measures should be interpreted as contextual indicators rather than direct proxies for code quality: they capture the stylistic and structural footprint of idiomatic coding, which, as our findings suggest, does not always coincide with the absence of quality issues. Building on this

perspective, future research could further investigate how these metrics evolve over time, how they interact with other quality dimensions such as testability and performance, and how they might inform adaptive refactoring tools capable of providing context-aware feedback on idiomatic usage in ML systems.

Implication for Researchers. The *Pythonic Percentage* and *Pythonic Density* metrics enable project comparisons of idiomatic usage. Future research can leverage these measures to explore how idiomaticity evolves with system growth and to design adaptive refactoring tools that provide context-aware feedback on idiom application.

6. Threats to Validity

This section discusses threats that might have affected the findings of our study, and how we mitigated them.

Construct Validity. This category concerns the potential discrepancies between theory and observation. Dataset selection plays a crucial role, and using an established dataset from prior literature helps mitigate bias from uncontrolled variables. Since our study centers on the concept of *Pythonic* code, rooted in the Python programming community, we focused on projects predominantly written in Python. Additionally, tool selection is critical, as different tools yield different metrics.

For code smell extraction, we used an already validated tool *i.e.,* we selected DPy for Pythonic-specific code smells. In addition, we used SonarQube to extract additional metrics, *i.e.,* NLOC, file complexity, and number of commits. To maintain focus on *Pythonic* elements, we analyzed only Python files in projects where more than 50% of the code was in Python. This approach helps avoid noise from *Refactorable-Pythonic* sources, such as code written in other languages. To identify Refactorable-idiomatic Python code, we used the RIdiom tool. We selected these tools, because they are considered the state-of-the-art. RIdiom was evaluated by its authors [13] on 3,215 refactorings from 479 repositories, with a manual review of 900 refactorings from 672 repositories. The tool achieved 100% detection accuracy for six idioms: *List-*, *Set-*, and *Dict-Comprehension*, *Loop-Else*, *For-Multi-Targets*, and *Chain-comparison*. For the remaining three idioms (*Truth-Value-Test*, *Assign-Multi-Targets*, and *Star-In-Func-Call*), the paper does not explicitly report performance metrics. To assess its usefulness, the RIdiom authors submitted 90 pull requests (10 per idiom) to 84 GitHub repositories, receiving 57 responses (63%). Of these, 34 pull requests were accepted (60%), and 28 (50%) were merged. Developers' feedback highlighted improved readability and a "more Pythonic" style. However, some rejections were attributed to readability, performance, or stylistic preferences, particularly for *Chain-Comparison*, *Truth-Value-Test*, and *Star-In-Func-Call* idioms. DPy combines AST-based analysis with a custom scope-based type inference system. The original validation involved four open-source projects, which were manually analyzed by two evaluators, yielding a Cohen's k of 0.87, indicating a strong inter-rater agreement. The tool achieved a precision of 0.96 and a recall of 0.93 across 929 verified instances. Reported limitations include the lack of support for nested entities and reliance on a simplified type inference mechanism.

The *Pythonic* variables in the study (*Density, Percentage*, *Count*) were extracted through AST traversing and validated with manual testing to ensure alignment with the nine idioms defined in the RIdiom paper [13].

While *Pythonic* code is not limited to these idioms, future studies incorporating additional idioms could enrich our findings. Another threat to construct validity concerns our tool for detecting Pythonic idioms. To ensure the reliability of our idiom-detection process, we derived rules for identifying Pythonic idioms by following the guidelines and principles outlined in The Zen of Python. Furthermore, we implemented a comprehensive suite of unit tests to verify the correctness and robustness of our detection tool, ensuring that each rule accurately captures the intended idiomatic construct. Nevertheless, a potential threat to construct validity remains due to the nature of our detection approach. Our tool relies on the construction of an AST, which by design can analyze only syntactically correct files. This limitation may lead to an underestimation of the actual number of idioms if the files under analysis contain syntax errors and are therefore excluded from the AST parsing process. Finally, the authors of NICHE considered the number of stars as a proxy metric for estimating project popularity. However, this metric may not fully capture all aspects related to popularity [45]. To mitigate this threat, we removed this criterion in our further analysis.

Internal Validity. These threats relate to factors that might have influenced the study's results. In $\mathbf{RQ}_2$, we examined the relationship between *Pythonic* idioms and code quality, measured via code smells. While collecting data through DPy, we also gathered metrics on project size, number of files, and cyclomatic complexity. These were used as control variables alongside the *Pythonic* metrics to reduce confounding effects.

We also acknowledge the role of potential confounding factors, including developer experience, project domain, team diversity, and community-related aspects, that may independently influence maintainability regardless of idiomatic usage. It is important to note that our study does not aim to establish causation between the presence (or absence) of *Pythonic* idioms and code quality, but rather to demonstrate correlations.

Conclusion Validity. Conclusion Validity threats pertain to the choice and use of statistical tests. Before applying statistical tests (*e.g.,* t-tests, Wilcoxon tests, or correlation analyses) and models, we ensured that the data met the necessary assumptions. We set a significance level of 0.05, which was consistently used across all tests and models. While our study is quantitative and serves as a preliminary exploration, further qualitative research is needed to understand how *Pythonic* idioms impact code quality.

A threat particularly worth discussing is the one that affects the results of $\mathbf{RQ}_{2.2}$, and particularly the lack of correlation found between smells and idioms for non-engineered projects. As explained in Section 4.4.2, this lack of correlation can be attributed to the low statistical power resulting from the small number of non-engineered projects (76). Therefore, the RQ2.2 results for such projects need to be carefully interpreted.

External Validity. The primary threat to external validity stems from the dataset used. The selection of projects is crucial, so we selected the NICHE dataset, a large collection of 303 real-world ML-enabled systems. These projects vary in context, size, and number of files, supporting the generalization of our findings to open-source ML projects with similar characteristics. Although the results may not directly apply to industrial or private/personal projects, we partially mitigate this threat by simulating private/personal projects by including small- and medium-sized projects from GitHub, relaxing two filters imposed by the authors of NICHE (i.e., the population and history), and replicating our analysis. The results of this additional experiment partially confirmed our findings, suggesting a possible transfer of learning for these project categories as well. Another potential threat is the selection of

Pythonic idioms. Although *Pythonic* code encompasses many idioms and conventions, RIdiom is the only tool that can identify Refactorable-idiomatic Python code. Our focus on nine specific idioms reflects a balance between identifying commonly used idioms and detecting those that are underutilized but potentially beneficial. While this limits our definition of *Pythonic* to these idioms, future studies employing additional idioms or other coding practices could provide a more comprehensive view.

7. Conclusions

This paper proposed an empirical investigation into the relationship between *Pythonic* and *Refactorable-Pythonic* idioms and code smells in ML projects. We analyzed 303 Python projects classified as "well-engineered" or "non-engineered", discovering that the more frequent *Pythonic* idioms are *Truth Value Test* and *Assign Multi Target* for "well-engineered" and "non-engineered", respectively.

Our statistical analyses revealed a correlation between the usage of *(Refactorable-)Pythonic* idioms and code smells, indicating that refactoring *Refactorable-Pythonic* code into *Pythonic* idioms does not consistently correspond with improved code quality in ML systems. Indeed, in a non-negligible number of cases, *Pythonic* idioms correlated with some code smells, especially in "well-engineered" projects. We conclude our study by providing implications and discussions to drive further research on the matter.

As future work, we plan to extend our investigation by incorporating a broader range of *Pythonic* idioms beyond the nine currently studied, potentially capturing a more comprehensive picture of idiomatic usage in ML codebases. Furthermore, we aim to complement our quantitative analysis with qualitative methods to understand better the rationale behind using certain idioms and their perceived impact on maintainability. Finally, another promising direction involves exploring the interplay between idiomatic usage and other quality attributes beyond code smells, such as performance or testability.

Acknowledgment

This work has been partially supported by the *Qual-AI* national research project, funded by the MUR under the PRIN 2022 programs (Code: D53D23008570006), and the project *"FAIR"* (PE0000013).

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

The manuscript includes data as electronic supplementary material. In particular, datasets, detailed results, as well as scripts and additional resources useful for reproducing the study, are available as part of our online appendix on Figshare available at: [11] and at the following GitHub page: `https://giammariagiordano.github.io/pythonic_and_smells/index.html`

Credits

Gerardo Festa: Formal analysis, Investigation, Data Curation, Validation, Writing - Original Draft, Visualization. **Giammaria Giordano:** Conceptualization, Methodology, Validation, Supervision, Resources, Writing - Original Draft & Editing. **Valeria Pontillo:** Conceptualization, Methodology, Validation, Writing - Review & Editing. **Massimiliano Di Penta:** Writing - Review & Editing. **Damian A. Tamburri:** Resources, Writing - Review & Editing. **Fabio Palomba:** Supervision, Resources, Writing - Review & Editing.

References

[1] I. Karamitsos, S. Albarhami, C. Apostolopoulos, Applying devops practices of continuous automation for machine learning, Information 11 (2020) 363.

[2] T. Peters, The zen of python, PEP 20, https://www.python.org/dev/peps/pep-0020/, 2004. Accessed: October 6, 2024.

[3] C. V. Alexandru, J. J. Merchante, S. Panichella, S. Proksch, H. C. Gall, G. Robles, On the usage of pythonic idioms, in: Proceedings of the 2018 ACM SIGPLAN international symposium on new ideas, new paradigms, and reflections on programming and software, 2018.

[4] C. Zid, F. Zampetti, G. Antoniol, M. Di Penta, A study on the pythonic functional constructs' understandability, in: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, 2024, pp. 1–13.

[5] G. Busquim, H. Villamizar, M. J. Lima, M. Kalinowski, On the interaction between software engineers and data scientists when building machine learning-enabled systems, in: International Conference on Software Quality, Springer, 2024, pp. 55–75.

[6] G. Annunziata, S. Lambiase, D. A. Tamburri, W.-J. van den Heuvel, F. Palomba, G. Catolino, F. Ferrucci, A. De Lucia, Uncovering community smells in machine learning-enabled systems: Causes, effects, and mitigation strategies, ACM Transactions on Software Engineering and Methodology (2024).

[7] M. Fowler, Refactoring: improving the design of existing code, Addison-Wesley Professional, 2018.

[8] A. Boloori, T. Sharma, Dpy: Code smells detection tool for python (2025) 826–830.

[9] R. Widyasari, Z. Yang, F. Thung, S. Q. Sim, F. Wee, C. Lok, J. Phan, H. Qi, C. Tan, Q. Tay, et al., Niche: A curated dataset of engineered machine learning projects in python, in: 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR), IEEE, 2023, pp. 62–66.

[10] G. Giordano, A. Della Porta, F. Ferrucci, F. Palomba, An evidence-based study on the relationship of software engineering practices on code smells in python ml projects, in: 2025 51st Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, 2025.

[11] G. Festa, G. Giammaria, P. Valeria, D. P. Max, T. Damian, P. Fabio, Pythonic vs refactorable pythonic: On the relationship between pythonic idioms and code quality in machine learning projects, 2025. URL: https://zenodo.org/records/17557873.

[12] A. Martelli, A. Ravenscroft, S. Holden, Python in a Nutshell: A desktop quick reference, " O'Reilly Media, Inc.", 2017.

[13] Z. Zhang, Z. Xing, X. Xu, L. Zhu, Ridiom: Automatically refactoring non-idiomatic python code with pythonic idioms, in: 2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 2023, pp. 102–106. doi:10.1109/ICSE-Companion58688.2023.00034.

[14] Z. Chen, L. Chen, W. Ma, B. Xu, Detecting code smells in python programs, in: 2016 International Conference on Software Analysis, Testing and Evolution (SATE), 2016, pp. 18–23. doi:10.1109/SATE.2016.10.

[15] F. Khomh, M. D. Penta, Y.-G. Guéhéneuc, G. Antoniol, An exploratory study of the impact of antipatterns on class change-and fault-proneness, Empirical Software Engineering 17 (2012) 243–275.

[16] F. Palomba, G. Bavota, M. Di Penta, F. Fasano, R. Oliveto, A. De Lucia, On the diffuseness and the impact on maintainability of code smells: a large scale empirical investigation, in: Proceedings of the 40th International Conference on Software Engineering, 2018.

[17] G. Giordano, A. Fasulo, G. Catolino, F. Palomba, F. Ferrucci, C. Gravino, On the evolution of inheritance and delegation mechanisms and their impact on code quality, in: 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, 2022, pp. 947–958.

[18] M. Tufano, F. Palomba, G. Bavota, R. Oliveto, M. Di Penta, A. De Lucia, D. Poshyvanyk, When and why your code starts to smell bad (and whether the smells go away), IEEE Transactions on Software Engineering 43 (2017) 1063–1088.

[19] G. Giordano, G. Sellitto, A. Sepe, F. Palomba, F. Ferrucci, The yin and yang of software quality: On the relationship between design patterns and code smells, in: 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, 2023.

[20] F. Pecorelli, F. Palomba, D. Di Nucci, A. De Lucia, Comparing heuristic and machine learning approaches for metric-based code smell detection, in: 2019 IEEE/ACM 27th international conference on program comprehension (ICPC), IEEE, 2019, pp. 93–104.

[21] F. Pecorelli, F. Palomba, F. Khomh, A. De Lucia, Developer-driven code smell prioritization, in: Proceedings of the 17th International Conference on Mining Software Repositories, 2020, pp. 220–231.

[22] M. De Stefano, F. Pecorelli, F. Palomba, A. De Lucia, Comparing within-and cross-project machine learning algorithms for code smell detection, in: Proceedings of the 5th international workshop on machine learning techniques for software quality evolution, 2021, pp. 1–6.

[23] F. Palomba, G. Bavota, M. Di Penta, R. Oliveto, A. De Lucia, D. Poshyvanyk, Detecting bad smells in source code using change history information, in: 2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE), IEEE, 2013, pp. 268–278.

[24] N. Vavrová, V. Zaytsev, Does python smell like java? tool support for design defect discovery in python, arXiv preprint arXiv:1703.10882 (2017).

[25] Z. Chen, L. Chen, W. Ma, X. Zhou, Y. Zhou, B. Xu, Understanding metric-based detectable smells in python software: A comparative study, Information and Software Technology 94 (2018) 14–29.

[26] B. Van Oort, L. Cruz, M. Aniche, A. Van Deursen, The prevalence of code smells in machine learning projects, in: 2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN), IEEE, 2021, pp. 1–8.

[27] H. Jebnoun, H. Ben Braiek, M. M. Rahman, F. Khomh, The scent of deep learning code: An empirical study, in: Proceedings of the 17th International Conference on Mining Software Repositories, 2020.

[28] F. Zampetti, F. Belias, C. Zid, G. Antoniol, M. Di Penta, An empirical study on the fault-inducing effect of functional constructs in python, in: IEEE International Conference on Software Maintenance and Evolution, ICSME 2022, Limassol, Cyprus, October 3-7, 2022, 2022, pp. 47–58. URL: https://doi.org/10.1109/ICSME55016.2022.00013. doi:10.1109/ICSME55016.2022.00013.

[29] F. Zampetti, C. Zid, G. Antoniol, M. Di Penta, The downside of functional constructs: a quantitative and qualitative analysis of their fix-inducing effects, Empir. Softw. Eng. 30 (2025) 9. URL: https://doi.org/10.1007/s10664-024-10568-z. doi:10.1007/S10664-024-10568-Z.

[30] T. Sakulniwat, R. G. Kula, C. Ragkhitwetsagul, M. Choetkiertikul, T. Sunetnanta, D. Wang, T. Ishio, K. Matsumoto, Visualizing the usage of pythonic idioms over time: A case study of the with open idiom, in: 2019 10th International Workshop on Empirical Software Engineering in Practice (IWESEP), 2019, pp. 43–435. doi:10.1109/IWESEP49350.2019.00016.

[31] P. Phan-Udom, N. Wattanakul, T. Sakulniwat, C. Ragkhitwetsagul, T. Sunetnanta, M. Choetkiertikul, R. G. Kula, Teddy: automatic recommendation of pythonic idiom usage for pull-based software projects, in: 2020 IEEE International

Conference on Software Maintenance and Evolution (ICSME), IEEE, 2020, pp. 806–809.

[32] Z. Zhang, Z. Xing, X. Xia, X. Xu, L. Zhu, Making Python code idiomatic by automatic refactoring non-idiomatic python code with Pythonic idioms, in: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 696–708.

[33] P. Leelaprute, B. Chinthanet, S. Wattanakriengkrai, R. G. Kula, P. Jaisri, T. Ishio, Does coding in pythonic zen peak performance? preliminary experiments of nine pythonic idioms at scale, in: Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension, 2022, pp. 575–579.

[34] Z. Zhang, Z. Xing, X. Xia, X. Xu, L. Zhu, Q. Lu, Faster or slower? performance mystery of python idioms unveiled with empirical evidence, in: 45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023, IEEE, 2023, pp. 1495–1507.

[35] C. Zid, F. Belias, M. Di Penta, F. Khomh, G. Antoniol, List comprehension versus for loops performance in real python projects: Should we care?, in: IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2024, Rovaniemi, Finland, March 12-15, 2024, 2024, pp. 592–601. URL: https://doi.org/10.1109/SANER60148.2024.00066. doi:10.1109/SANER60148.2024.00066.

[36] A. Nagpal, G. Gabrani, Python for data analytics, scientific and technical applications, in: 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 140–145. doi:10.1109/AICAI.2019.8701341.

[37] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, et al., Experimentation in software engineering, Springer, 2012.

[38] J. Tan, M. Lungu, P. Avgeriou, Towards studying the evolution of technical debt in the python projects from the apache software ecosystem., in: BENEVOL, 2018, pp. 43–45.

[39] N. Munaiah, S. Kroh, C. Cabrey, M. Nagappan, Curating github for engineered software projects, Empirical Software Engineering 22 (2017) 3219–3253.

[40] Z. Hanusz, J. Tarasinska, W. Zielinski, Shapiro–wilk test with known mean, REVSTAT-Statistical Journal 14 (2016) 89–100.

[41] J. C. De Winter, S. D. Gosling, J. Potter, Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data., Psychological methods 21 (2016) 273.

[42] P. Sedgwick, Pearson's correlation coefficient, Bmj 345 (2012).

[43] E. W. Weisstein, Bonferroni correction, https://mathworld. wolfram. com/ (2004).

[44] G. A. Campbell, P. P. Papapetrou, SonarQube in action, Manning Publications Co., 2013.

[45] H. Borges, M. T. Valente, What's in a github star? understanding repository starring practices in a social coding platform, J. Syst. Softw. 146 (2018) 112–129.