Exploratory Data Analysis Driven Feature Engineering for Modeling and Forecasting Expenditure in Tourism Entrepreneurship

Juan Ignacio Acciardi[1] and Antonio Quintero-Rincón[1,2]

[1] Departamento de Ciencia de Datos, Laboratorio de Ciencia de Datos e Inteligencia Artificial, Universidad Católica Argentina (UCA), Buenos Aires, Argentina
[2] Departamento de Ingeniería Informática , Universidad Católica Argentina (UCA), Buenos Aires, Argentina {juanacciardi,antonioquintero}@uca.edu.ar

Abstract. This pilot study introduces a feature engineering (FE) methodology based on exploratory data analysis (EDA) to augment and enrich an initially limited dataset. The primary objective is to establish a systematic framework for accurately predicting tourism-related expenditures at an operational resort in Esquel, Argentina. Data collection practices in such tourism facilities are often suboptimal: reservations are typically recorded manually or stored in simple spreadsheets, then examined informally to identify apparent trends. Under these conditions, EDA is a valuable tool for investigating relationships among the original variables and for constructing derived variables that more effectively capture their association with the studys target variable. A multivariate linear regression model was specified to predict *Tourist Spending*. The newly engineered predictors derived through FE informed by EDA, namely *Interaction between guests and days*, *Spending per person*, and *Staydays*, enabled a transparent and interpretable decomposition of the variance in the dependent variable. Model performance, assessed using RMSE, MAE, MSE, and the F-test, suggests that the model explains approximately 66.6% of the variability in *Tourist Spending* at the resort. These results are promising and provide further evidence of the critical role of FE and EDA in developing enterprise-level predictive models for tourism-related expenditures.

Keywords: Exploratory data analysis, Multivariate analysis, Regression analysis, Tourist industry, Feature engineering.

Ingeniería de características basada en análisis exploratorio de datos para el modelado y la previsión de gastos en turismo.

Abstract. Este estudio piloto introduce una metodología de ingeniería de características (EC) basada en el análisis exploratorio de datos (AED) para ampliar y enriquecer un conjunto de datos inicialmente limitado. El objetivo principal es establecer un marco sistemático para predecir con precisión los gastos relacionados con el turismo en un resort en funcionamiento en Esquel, Argentina. Las prácticas de recopilación de datos en este tipo de establecimientos turísticos suelen ser deficientes: las reservas generalmente se registran manualmente o se almacenan en hojas de cálculo simples, y luego se examinan informalmente para identificar tendencias aparentes. En estas condiciones, el AED es una herramienta valiosa para investigar las relaciones entre las variables originales y para construir variables derivadas que capturen de manera más efectiva su asociación con la variable objetivo del estudio. Se especificó un modelo de regresión lineal multivariante para predecir el gasto turístico. Los nuevos predictores diseñados mediante EC con base en el AED, a saber, la *Interacción entre huéspedes* y *Días*, el *Gasto por persona* y la *Duración de la estadía*, permitieron una descomposición transparente e interpretable de la varianza en la variable dependiente. El rendimiento del modelo, evaluado mediante RMSE, MAE, MSE y la prueba F, sugiere que explica aproximadamente el 66.6% de la variabilidad del gasto turístico en el complejo. Estos resultados son muy prometedores y refuerzan el papel central de la EC y el AED en la creación y validación de modelos predictivos en el ámbito empresarial turístico.

Keywords: Análisis exploratorio de datos, análisis multivariante, análisis de regresión, industria turística, ingeniería de características.

1 Introducction

In recent years, the problem of effectively exploiting the limited data typically available in small and medium-sized enterprises (SMEs) has emerged as an increasingly critical research and practical challenge. A major constraint in this context is the intrinsic scarcity of data, which frequently renders the straightforward deployment of machine learning (ML) models impracticable. Contemporary ML approaches generally require access to large-scale datasets to autonomously learn complex patterns and perform tasks with minimal human supervision. One strategy to mitigate data scarcity is to apply feature engineering (FE). FE is a human-centered process in which humans and computational tools collaborate to address a problem using only the available data (Egger, 2022). Within an FE workflow, two main types of analysis can be conducted: Exploratory Data Analysis (EDA) and Feature Selection (FS). EDA is a fundamental component of statistical analysis. It involves acquiring a detailed understanding of the variability, distributions, and statistical properties of the input features, as well as their interactions and relationships with the target (or class) variable to be predicted. This procedure enables the systematic identification of potentially informative features, as well as those that are redundant or irrelevant, which can therefore be excluded from subsequent analyses. Furthermore, the outcomes of exploratory analyses enable the visualization of specific feature combinations that may be particularly relevant to the target problem, thereby informing and guiding the design of subsequent hypothesis tests, predictive models, and related analytic procedures (Duboue, 2020). In contrast, feature selection (FS) focuses on eliminating variables that are either non-informative or contribute minimally to the predictive performance for the task at hand. This dimensionality reduction procedure increases the dataset's information content and discriminative capacity by projecting it into a lower-dimensional latent space while preserving the most relevant variance and class-distinguishing structures (Guyon et al., 2006). In numerous contexts, FE encompasses the systematic transformation of existing variables and/or the derivation of novel variables based on domain-specific knowledge, heuristic reasoning, and accumulated experiential evidence, to enhance the effectiveness and efficiency of problem-solving processes in human-computer interaction environments. Specifically, this work aims to develop an EDA-based FE methodology to generate an extended dataset from a limited amount of initial data. This enriched dataset can then be analyzed using ML techniques to predict *Tourist Spending* (TS). In this context, TS is defined as the monetary expenditure incurred by tourists to cover their stay in an accommodation complex. TS is a fundamental variable in tourism economics, as it directly and significantly influences the economic performance of tourist destinations, generates tangible benefits for local communities, and contributes to the broader social and economic development of the surrounding region. In countries where tourism constitutes a substantial proportion of Gross Domestic Product (GDP), TS becomes particularly crucial for maintaining macroeconomic stability.

Esquel is a city located in the northwest of the *Provincia del Chubut*, in *Patagonia Argentina*. One of the numerous tourist resorts in Esquel provided the empirical basis for this pilot study. As is customary in this context, essential reservation data are either recorded manually or stored in spreadsheet software and subsequently subjected to rudimentary analyses to detect trends in the information collected. Consequently, the application of FE is particularly pertinent for deriving new variables from a limited set of original attributes, thereby enhancing the analysis, interpretation, discovery, and optimization of patterns within the dataset. This process generates the variables necessary for training and deploying an ML model, which in turn enables the prediction of new observations for a target variable that, in this study, corresponds to tourist expenditure. This methodological approach offers several advantages:

- Regarding decision-making, it helps reduce reliance on intuition and ad hoc judgments.
- In the prediction stage, it allows anticipation of trends and supports the forecasting of variables of interest.
- It enhances knowledge of customers, thereby supporting the design of proposals, products, and/or services with demonstrable added value.
- The analysis of customer experience makes it possible to identify preferences and behavioral patterns that, once modeled, can be used to strengthen customer loyalty and increase sales.
- It facilitates revenue management by enabling the design and implementation of quantitatively optimized revenue-maximization strategies, commonly referred to as yield management, which provides a methodological framework for setting prices, managing reservations, and closing sales.
- It facilitates enhancements in operational productivity through the implementation of systematic process management methodologies.
- It generates deeper insights into competitors activities and, consequently, increases the capacity to implement or develop efficient competitive strategies and deploy more sophisticated marketing practices.

Studies using ML techniques within the tourism literature are both heterogeneous and extensive. Representative examples include Bravo et al., 2023; Ghosal et al., 2022; Jan et al., 2022; Xiaoling et al., 2021. These works share common methodological challenges, such as the identification and selection of appropriate explanatory features to predict target variables, including *tourism growth* and *tourist demand*. In addition, some contributions focus on the digitalization of data from tourism enterprises Bravo et al., 2023, which is consistent with the strategic orientation and potential objectives of the tourism resort analyzed in the present study. Specifically, in relation to this pilot study, the work of Moreno-Bastante, 2023 proposed a methodology for comparing different ML models to analyze the evolution of tourism expenditure in the post COVID-19 context. Two scenarios, each comprising 13 variables, were evaluated. Scenario 1 corresponded to pre-pandemic conditions, using data from 2019, and included 26,370 observations. Scenario 2 reflected the situation in 2022, one year after the pandemic, and contained 21,497 observations. A preliminary EDA was complemented by correlation analysis and feature selection procedures, including the F-test and mutual information, to identify the most informative explanatory variables (e.g., *country*, *transport*, and *number of overnights*). Based on these results, a Gaussian process regression (GPR) model was implemented to predict the target variable *Total Spending*. The model, however, did not exhibit adequate predictive performance, apparently due to class imbalance introduced during the variable categorization. Despite this limitation, the study successfully identified the *Luxury Tourism* segment, achieving a classification accuracy of 65% with the XGBoost model. This performance can be attributed to the algorithms demonstrated robustness in handling imbalanced datasets.

The current state-of-the-art does not adequately cover studies, research, or projects in small and medium-sized tourism enterprises. Moreover, FE in this domain is neither well defined nor systematically documented. This gap allows the present study to conduct exploratory data analysis and develop predictive models of tourist expenditure in tourism enterprises using FE techniques. Highquality ML models rely on relevant, reliable, and semantically adequate features. According to the principle "Garbage in, garbage out" (GIGO), models trained on highquality data should outperform those trained on data of low relevance, accuracy, or coherence. FE is therefore conceptualized as the systematic procedure of transforming existing features and constructing novel ones to more precisely represent the underlying characteristics and structure of the problem under investigation. These transformations are intrinsically linked to the performance of ML models. For a model to work correctly, its internal feature vectors must accurately reflect the structure and properties of the real-world phenomenon in the data. First, choosing an appropriate feature representation is essential. Second, features can often be preprocessed outside the model, allowing domain-specific knowledge to be incorporated more effectively. Such data-driven operations, which lack a universal rule set, often yield greater performance gains than tuning hyperparameters or changing algorithms alone. Consequently, the FE process is time-consuming and labor-intensive, relying on iterative experimentation, trial and error, and substantial domain expertise to simplify and structure the learning task. External data can also be added to the feature set based on prior experience and domain expertise.. In tourism, incorporating external data sources is particularly promising, as many relevant datasets are proprietary or purchasable only (Egger, 2022). However, this study did not use such data augmentation.

This manuscript is structured as follows. Section 2 describes the adopted methodology, including the dataset and the feature engineering process based on exploratory data analysis (Section 2.1). Section 3 introduces a multivariate linear regression model encompassing three distinct scenarios for predicting the target variable, namely *tourism spending*. Section 4 presents and analyzes the results obtained from the application of the proposed methodology. Finally, Section 5 summarizes the main conclusions of the study and discusses its limitations and strengths, as well as directions for future research.

2 Methodology

2.1 Dataset

The initial dataset for the tourist resort contains 201 booking records (observations) and 5 variables in a spreadsheet table. The limited data were collected between February and November 2023. November observations were reserved for validation, and earlier data were used for training. The 5 initial variables (var) are defined as follows:

- *Name* of the guest who made the booking: qualitative var.
- *Staydays* of the booking (length of stay in days): quantitative var.
- *Number of guests* associated with each booking: quantitative var.
- Total *spending* amount per booking: quantitative var.
- *Month* during which customers visited the establishment: qualitative var.

2.2 Feature engineering informed by exploratory data analysis

This work assesses how the initial variables introduced in Section 2.1 affect the target variable, *Tourist spending*. However, this dataset is not very informative due to the limited number of observations (201) and explanatory variables (5). Therefore, we must increase the feature space dimensionality by combining, selecting, or transforming existing variables to reveal more informative relationships between the new variables and the target. In other words, FE is undertaken, guided by exploratory data analysis (Egger, 2022; Kronthaler and Zollner, 2021). Following the recommendations of a tourism industry expert, various combinations (i.e., cross-products and interactions) of the initial variables were generated to define new derived variables and to establish relationships that may better explain the target variable, *Tourist spending*. The incorporation of domain knowledge from the tourism expert ensures coherence between the EDA and the subsequent confirmatory data analysis. This integration facilitates an iterative research process that begins with a clearly formulated research question, proceeds through the design of an appropriate empirical strategy, the collection of data in accordance with this design, the statistical analysis of the resulting data, and, finally, the derivation of logically consistent, data-driven conclusions. Below, all newly constructed variables are presented:

- *Sex*: Qualitative (categorical) variable inferred from the guest who made the booking.

– *Number of guests*: Quantitative (discrete) variable indicating the total number of guests who stayed at the tourist establishment during the booking period.
– *Staydays*: Quantitative (discrete) variable representing the total number of days associated with each booking.
– *Spending*: Quantitative (continuous) variable measured in U.S. dollars. For this variable, a monthly adjustment was applied to account for the variation in the dollar value at the time of booking. This procedure was implemented to standardize the amounts and obtain values that more accurately reflect real spending.
– *Month*: Qualitative (categorical) variable that records the calendar month during which the guests stay at the establishment took place.
– *Spending per guest*: Derived quantitative variable obtained by dividing *Spending* by *Number of guests*. It provides the average spending per guest for each booking. Formally,

$$\text{Spending per guest} = \frac{\text{Spending}}{\text{Number of guests}} \tag{1}$$

It is important to emphasize that, since the target variable is defined within the same resort and assessed at an identical temporal reference, no data leakage is introduced. In an operational (production or serving) setting, both variables (*Spending* and the *Tourism spending* target) would be simultaneously observable at the conclusion of the billing cycle. Furthermore, the transformation applied to the data is deterministic.
– *Spending per day*: Derived quantitative variable defined as the ratio between *Spending* and *Staydays*. Its purpose is to estimate the average daily spending over the entire duration of each booking. Formally,

$$\text{Spending per day} = \frac{\text{Spending}}{\text{Staydays}} \tag{2}$$

– *Spending per day and guest*: Quantitative variable derived from the variables *Spending per guest* and *Spending per day*. Its purpose is to obtain an estimate interpretable as the average expenditure attributable to each guest per day of stay in the establishment. It is defined as:

$$\begin{array}{c}\text{Spending}\\ \text{day/guest}\end{array} = \frac{\text{Spending per guest}}{\text{Spending per day}} \tag{3}$$

– *Interaction between guests and days*: Composite quantitative variable constructed as the product of the variables *Number of guests* and *Staydays*. It yields a summary measure of total guest-days in the establishment, i.e., the cumulative number of guests weighted by the length of their stay. It is defined as:

$$\begin{array}{c}\text{Interaction}\\ \text{guest/days}\end{array} = \text{Number of guests} \times \text{Staydays} \tag{4}$$

– *Numeric Month*: Discrete numerical (coded) variable whose objective is to evaluate potential associations between *Season* and *Spending per booking*. Specifically, it corresponds to the numerical encoding of the calendar month (e.g., February = 2, March = 3, etc.). This encoding is used to examine whether a statistical relationship exists between *Spending* and *Month*.

- *Season*: Binary categorical variable that partitions the months into two groups. Group 1 corresponds to high season and includes the months of February, July, August, and October. Group 2 comprises the remaining months, classified as low season.

The final dataset comprises 201 observations and the following 11 variables: *Sex, Number of guests, Staydays, Spending, Month, Spending per guest, Spending per day, Spending per guest and day, Interaction between guests and days, Numeric month, Season*.

3 Multivariate linear regression-based ML model

An ML model is a mathematical or computational construct that learns patterns from input data to generate predictions or support decision-making processes. In the context of the present work, a supervised learning model was implemented using labeled data to learn a mapping that can be subsequently applied to new, unlabeled instances (Flach, 2012). A wide variety of models can be developed and deployed within the tourism sector, including neural networks, decision trees, k-nearest neighbors, and linear regression, among others (Bravo et al., 2023; Ghosal et al., 2022; Jan et al., 2022; Xiaoling et al., 2021). In this study, a multivariate linear regression model (MLRM) was developed as the primary machine learning approach (James et al., 2023; Kronthaler and Zollner, 2021). Multivariate linear regression constitutes a fundamental statistical methodology for characterizing and quantifying the relationships between multiple explanatory (predictor) variables and a single response (target) variable. In contrast to univariate regression models, which examine the effect of a single predictor on the response, multivariate regression enables the concurrent evaluation of several predictors, thereby allowing the joint influence of these variables on the response to be systematically assessed. In the context of the present study, a MLRM framework was employed to quantify and characterize the joint effects of multiple predictor variables on the response of interest, thereby enabling the simultaneous estimation of their individual and combined associations with the target outcome.This modeling strategy is particularly well suited to representing the complexity that emerges from interactions and interdependencies among multiple factors, thereby enabling a more comprehensive and fine-grained characterization of the determinants governing the observed response.

The primary objective of employing a MLRM in this work is to efficiently identify and quantify how the set of predictor variables collectively affects the total tourist expenditure associated with each booking. This holistic approach makes it possible to estimate the relative contribution of each predictor while controlling for the presence of the others, thereby providing a more robust and accurate understanding of the behavior of the target variable, *Spending*. The objective is to determine the extent to which multiple independent variables collectively influence a given response variable. Multivariate (multiple) linear regression is employed to model and quantify the relationship between a single dependent variable and a set of predictor (explanatory) variables (Flach, 2012; Heumann et al., 2022; James et al., 2023). The MLRM is formally specified as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{5}$$

where Y denotes the dependent (response) variable, β_0 is the intercept or constant term, $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients associated with the predictor variables $X_1, X_2, \ldots, X_p$, and ε denotes the random error term. The primary objective is to estimate the coefficients β_i to fit the model to the observed data. This is achieved by minimizing the sum of squared residuals. Multivariate linear regression is particularly suitable for quantifying the relative contribution of multiple predictor variables and for examining how these variables jointly influence the dependent variable.

The dataset was partitioned into two subsets for model development and performance assessment: 70% of the observations were used for training and 30% for testing. The MLRM was constructed to examine the associations between the explanatory variables and the dependent (target) variable, and to derive alternative predictive scenarios based on the magnitude and statistical relevance of these associations. To this end, three scenarios under different combinations of variables were defined to evaluate the proposed model and to predict the target variable, *Spending*:

- Scenario 1: *Guest, days* and *Staydays*.
- Scenario 2: *Guest, days* and *Spending per guest*.
- Scenario 3: *Guest, days, Spending per guest*, and *Staydays*.

4 Results and discussion

In an exploratory analysis, the Pearson correlation coefficient was computed to examine associations among the quantitative variables. The results indicated a positive relationship between *Staydays* and the target variable *Spending*. This preliminary finding is a useful starting point for more rigorous analysis. However, the original dataset includes few variables, which significantly limits the ability to identify robust patterns, characterize meaningful relationships, and extract comprehensive information. The original dataset was expanded from 5 to 11 variables using FE techniques (Section 2.2), enhancing the ability to extract more informative and nuanced data characteristics via EDA. Descriptive statistics are a key part of EDA, as they characterize the main properties of the sample (Guyon et al., 2006). This exploratory framework facilitates the detection of potential outliers and observations that warrant further scrutiny. It should be emphasized that, in the presence of NaN or missing values, summary measures such as quartiles, mean, median, and mode are often employed as imputation values to correct or mitigate the impact of incomplete data. In the present study, however, no such missing values were identified because the dataset was manually entered, supervised, validated, and cross-checked against the source.

Table 1 describes each analyzed variable. These variables yield noteworthy results that can be directly compared with their histograms in Fig. 1. For example, the maximum values of the target variable *Spending* (910 dollars) and *Staydays* (15 days) appear in the histograms as isolated observations about four times greater than their respective means (see Fig. 1a for *Spending* and Fig. 1d for *Staydays*).

It should be emphasized that the expert administrator of the tourist resort under investigation was systematically consulted throughout all stages of the research process, to ensure the validity and internal consistency of the data. In particular, values that could be considered outliers or otherwise questionable were carefully examined to clarify any detected inconsistencies. This rigorous validation procedure ensured the correct use of the complete dataset during the development of this study and provided a robust ground truth for the subsequent analyses.

Table 1: *Spending (S) expressed in dollars; Number of guests (NG); Staydays (SD); Spending per guest (SG); Spending per day (SD); Spending per guest and day (SGD); Interaction between guest and days (IGD).*

	S	*NG*	*SD*	*SG*	*SD*	*SGD*	*IGD*
Minimum	37	1	1	10.40	17.10	9.30	2
1st Quartile	101	2	2	35.50	42.20	14.10	6
Median	164	3	3	53.90	51.20	17	10
Mean	203	3.41	3.54	62.65	58.31	18.08	12.12
3rd Quartile	256	4	5	77	67.40	21	16
Maximum	910	10	15	288.50	198.30	66.20	63

The variables are described below:

Spending: Table 1 reports a minimum value of 37 dollars, a maximum value of 910 dollars, and a median and mean that differ, indicating skewness in this variable. The scatter plot in Fig. 1b shows that the distribution of this variable is relatively regular up to approximately 300 dollars, with pronounced peaks above 400 dollars. This distribution does not reveal a clear association between high spending and bookings made during the high season. The histogram in Fig. 1a indicates a left-skewed (negatively skewed) pattern, as both the mean and the median are located in the lower value intervals. Nonetheless, the dispersion of spending values remains substantial, reflecting generally high expenditure levels.

Sex: This variable did not exhibit substantial relevance with respect to the target variable, nor were any statistically significant effects identified. Fig. 1b presents the dispersion of the variable *Spending* as a function of *Sex*. The data set comprises 83 bookings made by men and 118 by women.

Number of guests: As reported in Table 1, the number of guests per booking ranges from a minimum of 1 to a maximum of 10, with a mean of 3.41. The histogram in Fig. 1c depicts the frequency per booking, indicating that bookings with 2, 3, and 4 guests are predominant.

Staydays: According to Table 1, the length of stay ranges from a minimum of 1 day to a maximum of 15 days, with a mean of 3.5 days. The histogram in Fig. 1d illustrates the distribution of stay duration. This distribution is left-skewed, reflecting guests preference for stays between 1 and 5 days, either for short vacations or brief visits to another city.

Month and Season: These variables were analyzed jointly due to their direct relationship, as the *Season* variable is derived from the *Month* variable. Their main contribution lies in providing operationally relevant information for the resort. The histogram in Fig. 1e indicates that the months with the highest booking frequency coincide with the *high season*, with all 25 bookings occurring between July and October, corresponding to the winter and spring periods. Nonetheless, a higher number of bookings does not necessarily translate into higher per-booking spending. It can be inferred that total monthly revenue is higher primarily due to increased booking frequency, rather than elevated spending per booking.

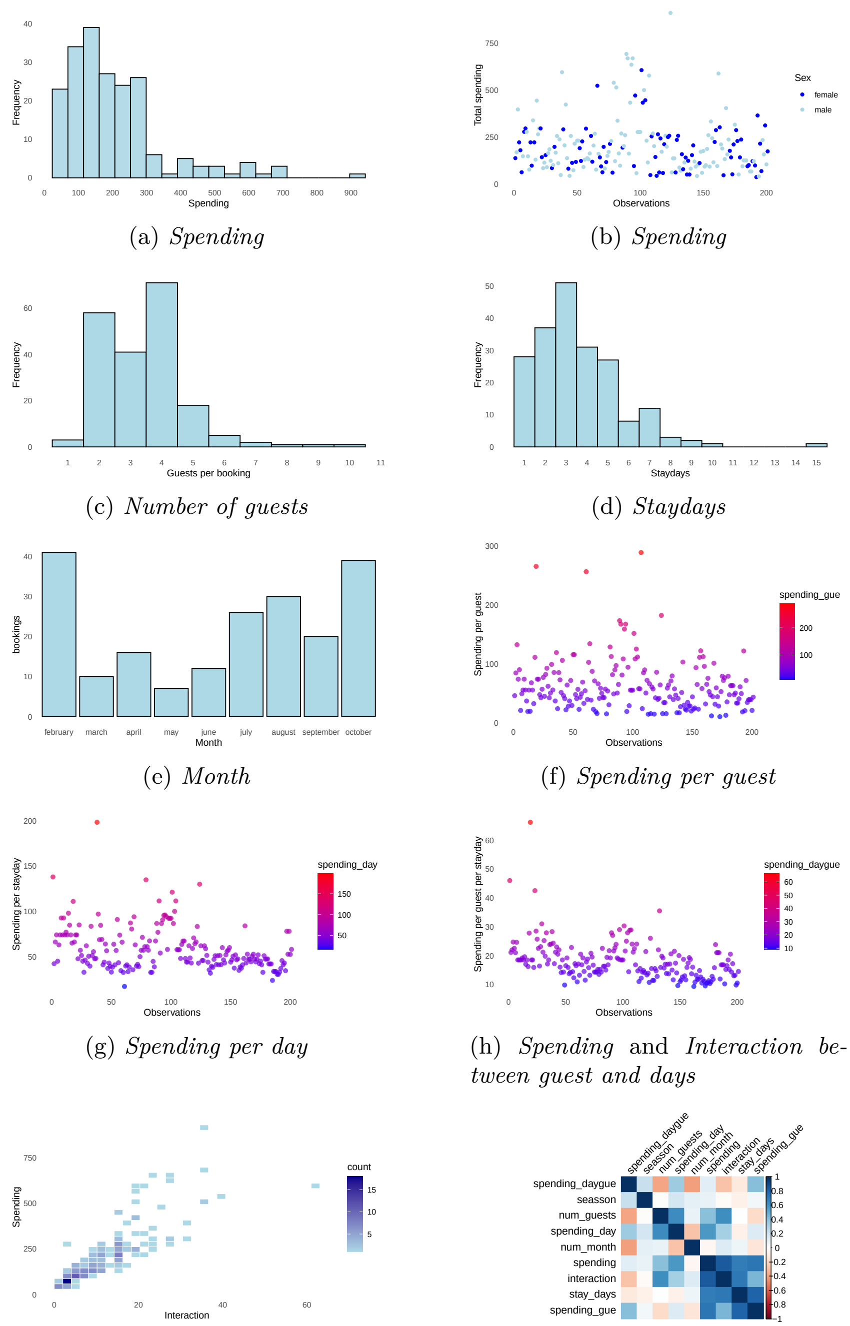
Spending per guest: Table 1 reports a minimum spending per guest of 10 dollars and a maximum of 288 dollars, with a mean of 62 dollars. The scatter plot in Fig. 1f reveals an irregular dispersion pattern, with a slight increase in variability during the winter season around observation 100, which does not appear to be substantively relevant.

Spending per day: Table 1 indicates that spending per day ranges from a minimum of 17 dollars to a maximum of 288 dollars, with a mean of 58 dollars. The scatter plot in Fig. 1g shows that the data are initially highly dispersed but become more homogeneous toward the end of the year (after approximately 130 observations), where a discernible downward trend can be observed.

Spending per guest and day: This variable is derived from the two variables *Spending per guest* and *Spending per day*, see Section 2.1. The scatter plot in Fig. 1h exhibits pronounced peaks during the high-season months (July and August), corresponding to between 80 and 120 observations, and a reduced dispersion of the observations for each booking. This relatively homogeneous distribution with comparatively small values is a direct consequence of the division of the two underlying variables. As a result, this derived variable does not contribute additional information beyond that already contained in its components.

Interaction between guests and days: This variable, representing the interaction between the number of guests and the number of *Staydays* in the resort, emerged as one of the most influential predictors during model development from a statistical standpoint. As reported in Table 1, it takes a minimum value of 2, a maximum of 63, with a mean of 12 and a median of 10. The two-dimensional histogram in Fig. 1i reveals an approximately linear relationship between the underlying variables, characterized by higher frequencies (darker colors) at lower interaction values and a greater dispersion at higher values. These findings provide a quantitative characterization of the interaction and covariation between the two variables, which is essential for the present analysis.

Fig. 1j presents the Pearson correlation matrix heat map for all variables considered in this study. Dark blue tones represent positive correlation coefficients between variables, whereas white indicates near-zero coefficients, corresponding to an absence of linear association. In contrast, the few warm colors denote negative correlations of varying magnitude (Heumann et al., 2022). This visualization permits a detailed assessment of the strength of the relationships between the target variable *Spending* and the explanatory variables. The variables *Interaction between guest and days*, *Staydays*, *Spending per guest*, and *Spending per day* display the strongest positive associations with *Spending*, with *Interaction between guest and days* exhibiting the highest correlation. Furthermore, it is both intuitive and empirically corroborated that *Staydays*, representing the number of nights spent by visitors, is directly and positively associated with the target variable *Spending*.



(a) *Spending* (b) *Spending*

(c) *Number of guests* (d) *Staydays*

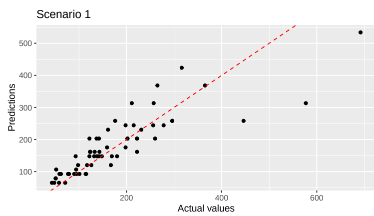
(e) *Month* (f) *Spending per guest*

(g) *Spending per day* (h) *Spending* and *Interaction between guest and days*

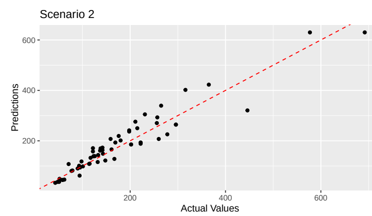
(i) (j) Final correlation matrix

Fig. 1: **(a)** Histogram of *Spending*, grouped in bins of 50 dollars.**(b)** Scatter plot for *Spending* regarding the variable *Sex* of the person who made the booking. The male gender is in blue, and the female gender is in cyan. **(c)** Histogram of *Number of guests*. **(d)** Histogram of *Staydays*. **(e)** Histogram of *Month*. **(f)** Scatter plot for *Spending per guest*. **(g)** Scatter plot for *Spending per day*. **(h)** Scatter plot for *Spending per guest and day*.**(i)** Two-dimensional histogram of *Spending* as a function of *Interaction between guest and days*. **(j)** Correlation matrix of all final variables in the dataset.

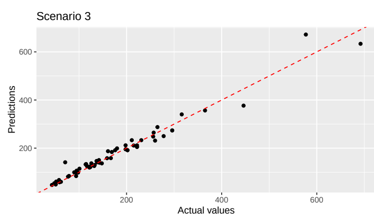
This study aims to identify a modeling framework to predict the target variable *Spending* using ML. Accordingly, EDA-based FE is essential for robust, practical deployment in real-world business contexts, such as tourism enterprises (Flach, 2012; Heumann et al., 2022). For this purpose, the outcomes of the visual inspection results of the Pearson correlation matrix are shown in Fig. 1j.



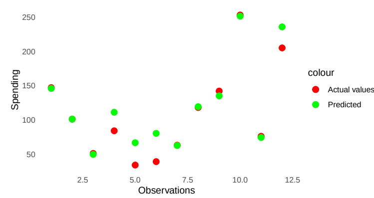
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Predicted values

Fig. 2: **(a)** Scenario 1: MLRM with *Interaction between guest and days* and *Staydays*. **(b)** Scenario 2: MLRM with *Interaction between guest and days* and *Spending per guest*. **(c)** Scenario 3: MLRM with *Interaction between guest and days*, *Spending per guest*, and *Staydays*. **(d)** November scatter plot of predicted values: green for model 3 and red for actual values.

Table 2 and Figures 2a to 2c present the assessment of the three proposed models using multivariate linear regression, evaluated by means of the F-test and error metrics including the root mean square error (RMSE), the mean absolute error (MAE), and the mean squared error (MSE) (Draper and Smith, 1998; Heumann et al., 2022; James et al., 2023). Each figure depicts the prediction plot for the corresponding model, where the points represent the predicted values and the superimposed line corresponds to the fitted multivariate linear regression model. A combined assessment of these figures and the corresponding table is essential to verify the coherence between the graphical representations and the quantitative values of the selected performance metrics. November was selected as the validation period to assess the models predictive performance on previously unseen data. During this validation stage, 12 observations were analyzed using the same dataset structure and the same procedure for deriving the new variables as in the model development stage.

The MLRM was estimated using observations from the preceding months and subsequently validated against the realized November values.
Fig. 2d presents the predicted (green) and actual (red) November values. The

overlap between the two colors indicates an exact correspondence between the predicted and observed points. This overlap demonstrates that, for these observations, the predicted values coincide with the actual reserve expenditure. These results support the conclusion that the model for scenario 3 is suitable for predicting the target variable *Spending*.

As measures of predictive performance, the coefficient of determination (R^2) was 0.89, and the Pearson correlation coefficient was 0.96. Defining 10 dollars as the maximum acceptable prediction error for the proposed model, the model achieved an accuracy of 66.6% (8/12). This model yields reliable and precise forecasts for previously unobserved data points and facilitates prospective analyses of the target variable Spending over forthcoming months and years.

Table 2: Evaluation of the proposed models based on their F-tests and associated error metrics.

	F-Test	MAE	MSE	RMSE
Scenario model 1	257.34	39.82	3753	61.26
Scenario model 2	196.87	28.87	1390	37.29
Scenario model 3	94.70	14.23	514	22.67

5 Conclusions

This study presents a methodology to expand a small dataset by using exploratory data analysis to guide feature engineering on the available variables. The proposed methodology is tailored to an analysis and prediction problem situated in the context of tourism entrepreneurship. By means of the FE design, the number of variables was increased from 5 to 11. This process resulted in the identification of the most informative variable in the new dataset called, termed *Interaction between people and duration of stay*. Three predictive modeling scenarios were evaluated using multivariate linear regression. The most suitable model was the one incorporating the variables *Interaction, spending per person, and Staydays*. This model provides effective and accurate predictions of new observations and supports the prospective analysis of the target variable *Spending* over the coming months and years. The primary limitation of this work is the scarcity of data pertaining to the tourism sector in the region. Efforts to raise awareness of the importance of systematic data collection are ongoing; however, these initiatives remain at an early stage of development.

The main contribution of this study lies in the formulation of a methodological framework that employs exploratory data analysisdriven FE to construct novel variables from a constrained dataset. This strategy facilitated the development of a predictive model based on multivariate linear regression, which can be systematically updated and replicated as additional data become available. Future research will focus on increasing both the volume and heterogeneity of the dataset, by enlarging the number of variables and observations and by incorporating a broader range of establishment types. The objective is to standardize the dataset and to detect robust associations among heterogeneous features that potentially influence the social and economic development of the regions. These features include, for example, categories of tourist establishments, climatic conditions, cultural offerings, natural attractions, historical heritage, lifestyle patterns, and structural characteristics of cities, towns, and provinces. This multidimensional heterogeneity represents a substantial challenge for the quantification, characterization, and integration of all these dimensions within a unified and coherent analytical framework. In addition, the implementation of sentiment analysis techniques Kontogianni et al., 2024 can facilitate the identification of critical challenges within the tourism sector and enable their targeted resolution in cooperation with tourism enterprises. Furthermore, fostering awareness of the significance of systematic data collection by tourism companies is crucial for addressing pertinent research questions, assessing intervention outcomes, and improving the accuracy of forecasts regarding future probabilities and trends, thereby contributing to the economic and social development of the region.

References

Bravo, J., Alarcón, R., Valdivia, C., & Serquén, O. (2023). Application of machine learning techniques to predict visitors to the tourist attractions of the moche route in Perú. *Sustainability*, *15*(11). https://doi.org/0.3390/su15118967

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. Wiley-Interscience.

Duboue, P. (2020). *The art of feature engineering essentials for machine learning*. Cambridge University Press.

Egger, R. (2022). *Applied data science in tourism*. Springer.

Flach, P. (2012). *Machine learning, the art and science of algorithms that make sense of data*. Cambridge.

Ghosal, A., Singh, J., & Singh, A. (2022). Predicting growth in tourism industry using machine learning methods. *2022 COM-IT-CON*, 317–321. https://doi.org/10.1109/COM-IT-CON54601.2022.9850491

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature extraction foundations and applications*. Springer.

Heumann, C., Schomaker, M., & Shalabh. (2022). *Introduction to statistics and data analysis: With exercises, solutions and applications in R*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning with applications in R*. Springer.

Jan, N., Yu, N., & Chen, J. (2022). Design of machine learning algorithm for tourism demand prediction. *Complexity*. https://doi.org/10.1155/2022/6352381

Kontogianni, A., Alepis, E., Virvou, M., & Patsakis, C. (2024). *Smart tourism. the impact of artificial intelligence and blockchain*. Springer.

Kronthaler, F., & Zollner, S. (2021). *Data analysis with RStudio*. Springer.

Moreno-Bastante, M. (2023). *Comparación de modelos de machine learning para el gasto turístico tras el COVID-19* [Master's thesis, Universidad Pontificia Comillas, Facultad de Ciencias Económicas y Empresariales].

Xiaoling, R., Yanyan, L., Zhao, J., & Qiang, Y. (2021). Tourism growth prediction based on deep learning approach. *Complexity*. https://doi.org/10.1155/2021/5531754