# Data Driven Decision Making in Business Final Portfolio



Antonella Portugal Las Heras

Student Number: 213541

December 19, 2023

Table of Contents

[**Individual Assignment** 3](#_Toc153797095)

[Business Understanding 3](#_Toc153797096)

[Data Understanding 3](#_Toc153797097)

[Data Preparation 4](#_Toc153797098)

[Modelling 8](#_Toc153797099)

[Multiple Linear Regression 8](#_Toc153797100)

[K-Means Clustering 9](#_Toc153797101)

[Evaluation 11](#_Toc153797102)

[Deployment 13](#_Toc153797103)

[Reflection and Ethics 14](#_Toc153797104)

[**AI Bootcamp: LLM Micro-workshop** 15](#_Toc153797105)

[Business Understanding 15](#_Toc153797106)

[Data Understanding 15](#_Toc153797107)

[Data Preparation 15](#_Toc153797108)

[Modeling 15](#_Toc153797109)

[Evaluation 16](#_Toc153797110)

[Deployment 16](#_Toc153797111)

[Reflection and Ethics 17](#_Toc153797112)

[**Process Mining Bootcamp 18**](#_Toc153797113)

[Planning 18](#_Toc153797114)

[Extraction 19](#_Toc153797115)

[Data Processing 19](#_Toc153797116)

[Mining and Analysis 20](#_Toc153797117)

[Evaluation 26](#_Toc153797118)

[Process Improvement and support 27](#_Toc153797119)

[Reflection and Ethical Considerations 27](#_Toc153797120)

[Bibliography 28](#_Toc153797121)

# Individual Assignment

In fulfilling the requirements of the Data Driven Decision Making in Business’ final portfolio, we were tasked with an intriguing challenge: to conceive and develop a tool that not only leverages our cumulative learning but also addresses a real-world issue that is of our personal interest. For my project, I chose to focus on the world of game development. My goal was to create an innovative tool designed to provide game developers with valuable insights into their game concepts. This report details the process of this endeavor, structured around the CRISP-DM methodology. Herein, I will outline the steps taken from initial conceptualization to the final deployment, showcasing how the tool aligns with each phase of the CRISP-DM framework to effectively tackle the unique challenges faced by game developers in realizing their creative visions.

## Business Understanding

The gaming industry is one of the most relevant sectors in technology today. It is valued at over $200 billion and has evolved into both a social and cultural phenomenon. This industry encompasses various facets, including intense competition among gaming platforms like Xbox, PlayStation, and Nintendo Switch, the thriving field of competitive gaming (eSports), and the popularization of video game streaming (Daley, 2022). These aspects collectively amplify the significance of video games in modern culture.

Alongside the increasing interest in gaming, there has been a notable surge in games developed by independent (indie) developers. These developers endeavor to craft distinctive gaming experiences while contending with major game studios. Indie games, typically produced by individuals or small teams, depend on digital distribution platforms such as Steam, Itch.io, and the Epic Games Store to connect with their audience (Fungies, 2023).

The growth of indie game development can be attributed to three key factors according to Fu (2023): the abundance of online educational resources, the availability of video game assets that expedite development, and advanced tools like generative AI that enable the rapid creation of high-quality prototypes. These developments, however, also intensify competition. A recent survey by Game Developer, involving 44,000 developers, revealed that approximately 57% of them earned less than $1,000 from their Steam-published games (Game Developer, 2022). This statistic highlights a harsh reality: despite the influx of new entrants, success remains elusive for most.

In light of these challenges, the primary goal of this project is to develop a tool that aids indie game developers in gauging the potential market reception of their game concepts using the tools and topics learned from the Minor in Data Driven Decision Making in Business. Given the highly competitive nature of the industry, it is crucial for developers to have access to predictive insights on key metrics such as potential reviews, downloads, or purchases. This tool aims to bridge the gap between development and market success, offering indie developers a better understanding of their game's potential and guiding them in making informed decisions in a crowded and dynamic marketplace.

## Data Understanding

To support indie developers with a predictive tool, a comprehensive dataset was essential. The chosen dataset, sourced from Steam (Valve’s primary PC gaming platform), encompasses over 83,000 games. This dataset, obtained from Kaggle, includes diverse game-related metadata such as titles, publishers, genres, and release dates (Tomashivli, n.d.). It comprises 71,700 entries across 16 distinct features.

After an initial analysis of the dataset, the feature that appeared to have the most relevance towards the success of a game was the total amount of reviews made by users. Other notable features included game tags and specific game attributes, providing detailed insights into game types. Additionally, the game's price was considered relevant, potentially reflecting the developer's perceived value of the game as well as being a critical decision factor for a user that wishes to play a new game. A less consistently available yet interesting metric was the percentage of positive reviews, limited by Steam’s requirement for a minimum number of reviews to display this statistic.

## Data Preparation

The preparation of the dataset was a vital yet challenging step, involving cleaning and formatting:

1. **Transforming Textual Data to Numerical Values**
   * Important information in text format, like user review statistics, was converted into numerical data for analysis. For example, a phrase like “94% of the 188,617 user reviews are positive” was split into two numbers: 94% (positive review percentage) and 188,617 (total reviews).
   * Prices listed in various formats (e.g., “Free” or “$29.99”) were standardized into consistent numerical values, such as 0 for free games and 29.99 for games priced at $29.99.
   * Release dates were converted from text to a numerical year format to facilitate chronological analysis.
2. **Handling Missing Values for Recent Games**
   * Games released in the current year (2023) and unreleased games were excluded to ensure data consistency.
   * The dataset initially had a high percentage of missing values in the "Total User Reviews" column. By incorporating related data from other columns, the missing values were significantly reduced, improving data completeness (from 91% of missing values to 6% missing values).
3. **Determining Zero Reviews vs. Missing Data**

* For games with missing review data, an analysis of Steam URLs helped decide how to handle these cases. Games with "/app/" in their URL typically were smaller games with potentially no reviews. In contrast, those with "/sub/" were often part of bundles or special promotions and didn't individually track reviews. Therefore, "/app/" games with missing review data were assigned a value of 0, while "/sub/" entries were removed for clarity.

1. **Outlier Treatment**

After the features had been formatted successfully, it was now time to verify if there were any outliers that would affect the performance of the predictive model and the analysis.

* The "Total User Reviews" variable showed a highly skewed distribution, with an average of 1,590 reviews but a maximum of 7,428,921 and a standard deviation of 38,589. This indicated a sparse distribution, aligning with findings that most indie developers on Steam earn minimal revenue, with a few exceptions achieving significant success. To address this, the Interquartile Range (IQR) method was applied, which led to the removal of 16% of the dataset entries considered as outliers. IQR is a statistical measure used to identify and remove outliers by focusing on the central portion of the data distribution. This resulted in a more representative distribution of the majority of games on Steam; the new distribution now has an average of 37 reviews, a maximum of 279 and a standard deviation of 56, providing a clearer picture of typical user review patterns (Figure 1).

**Figure 1**

*Histogram and Boxplot of “Total User Reviews” before and after applying IQR method*  
A screenshot of a graph

Description automatically generated

* The game price data also exhibited unusual patterns, with some games priced exorbitantly high, reaching up to millions of dollars. After confirming all prices were in US Dollars, the IQR method was employed to remove prices significantly higher than the majority, affecting 2.7% of the dataset. This adjustment led to a more accurate representation of typical game pricing on Steam, essential for a realistic analysis of market trends (Figure 2).

**Figure 2**

*Histogram and Boxplot of “Price” before and after applying IQR method*

A screenshot of a graph

Description automatically generated

1. **Transformation of Categorical Variables**

* The categorical variables, including supported languages and game features, were transformed into dummy variables to facilitate their use in statistical models. For example, if a game is available in Spanish but not in French, two separate columns were created labeled “Spanish” and “French.” In these columns, binary indicators were used: a value of 1 under “Spanish” and 0 under “French.” This approach expanded the dataset significantly, resulting in a large number of new columns:
  + Game Features: 31 categories
  + Popular Tags: 441 categories
  + Supported Languages: 8 categories, representing the most commonly supported languages.
* The “Popular Tags” category, with its extensive range of tags, presented a challenge due to its volume and overlapping nature of some tags (e.g., “Political”, “Politics”, “Political Sim”). To efficiently condense this information without significant loss of detail, Principal Component Analysis (PCA) was applied. PCA helps in reducing dimensionality by identifying the principal components that capture the most variance in the data. According to Mukilski (2019), the optimal number of principal components can be determined by using the cumulative variance method. Setting a threshold at 95%, the PCA plot (Figure 3) indicated that 238 principal components sufficiently capture the essential information. This process effectively reduced the number of necessary columns for the “Popular Tags” by almost half, ensuring the retention of crucial data while simplifying the dataset.

**Figure 3**

*Cumulative Explained Variance Ratio as by Number of Principal Components for PCA Analysis*

A graph with a curve

Description automatically generated

1. **Creation of New Categories**

* An analysis comparing total user reviews between free-to-play (F2P) and non-F2P games showed that non-F2P games generally had fewer reviews. To capture this distinction, a new category for F2P games was introduced, considering that the price alone was not a sufficient indicator (Figure 4).

**Figure 4**

*Boxplot Comparison of Total User Reviews Between Free-to-Play (F2P) and Non-Free-to-Play (Non-F2P) Games*

A graph with a number of different colored lines

Description automatically generated with medium confidence

## Modelling

### Multiple Linear Regression

The project addresses the regression problem of predicting the number of user reviews based on various game characteristics such as the price in USD, supported languages, game features, and game tags. For this purpose, Multiple Linear Regression (MLR) is selected for its effectiveness and simplicity, as it assigns coefficients to relevant variables under the assumption of linearity and absence of multicollinearity.

The model predominantly handles categorical variables that have been converted into dummy variables, significantly reducing the potential for multicollinearity. These dummy variables represent distinct categories and are, by definition, independent of one another. Principal Component Analysis (PCA) further refines the model by consolidating correlated game tags into a smaller set of uncorrelated principal components, thus diminishing the risk of multicollinearity.

The linearity of the 'Price' variable, the only continuous variable, was tested. It was observed that games priced at zero (F2P games) accumulated a higher number of reviews than non-F2P games. While this suggested a potential non-linear relationship, the introduction of a binary variable for F2P status rectified this issue. The lowess plot, illustrated on the left in Figure 5, confirms a more consistent linear relationship when F2P games are not accounted for. However, the already introduced F2P binary variable serves as a methodological adjustment, allowing the model to differentiate the pricing dynamics unique to F2P games from those of paid games and mitigating the non-linear effects observed in the data.

This strategic inclusion of the F2P variable effectively mitigates the non-linear effects observed in the data, preserving the linear modeling framework essential for MLR.

**Figure 5**

*Lowess plots demonstrating the relationship between price and user reviews, with and without the consideration of F2P games*

A comparison of barcode data

Description automatically generated with medium confidence

### K-Means Clustering

Given the vast array of characteristics inherent to each game, clustering analysis is implemented taking into account the variable of “Total User Reviews”. This analysis is designed to offer developers further insights by comparing their game idea to similar titles. More precisely, it could provide an estimation of the percentage of positive reviews their game might receive, as well as a benchmark for the average percentage of positive reviews to expect.

Prior to clustering, any entries lacking data on the percentage of positive reviews were omitted to maintain the integrity of the analysis. K-means clustering was selected for its efficiency and ease of interpretation. This technique was applied excluding the variable of percentage of positive reviews but including "Total User Reviews" to ensure that the clusters reflect variations in review quantity rather than review sentiment.

The crucial initial step in K-means clustering is determining the optimal number of clusters, or 'k'. The elbow method was employed to ascertain this number. The resultant elbow plot, as seen in Figure 6, suggests that four clusters represent the best balance between within-cluster variation and the total number of clusters. Opting for more than four clusters would likely result in groups that are overly similar to one another.

**Figure 6**

*Elbow Plot for Determining Optimal Cluster Count in K-Means Clustering*  
A graph with a line

Description automatically generated

Once the model was trained, the distribution of the "Percentage of Positive Reviews" within each cluster was plotted, as displayed in Figure 7. The clusters exhibit distinct patterns in terms of the distribution of positive reviews. Despite a general left skew across all clusters—indicating a tendency towards higher percentages of positive reviews—cluster 2 stands out for its consistency in having the highest positive review percentages. Conversely, cluster 1 presents a more dispersed distribution, implying that games falling into this cluster may have a wider range and potentially lower percentages of positive reviews. Clusters 0 and 3 fall in between, with variations in distribution sparsity. These observations are encapsulated in Table 1, which summarizes the central tendency (median and mean) and variability (standard deviation) of the percentage of positive reviews for each cluster.

**Table 1**: *Descriptive Statistics of Percentage Positive Reviews by Cluster*

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Median** | **Mean** | **Std. Deviation** |
| 0 | 83.0 | 79.12 | 15.85 |
| 1 | 75.0 | 72.45 | 19.67 |
| 2 | 84.0 | 80.91 | 14.78 |
| 3 | 80.0 | 76.83 | 16.82 |

**Figure 7**

*Distribution of Percentage Positive Reviews Across Four Clusters*

A group of graphs showing different colored lines

Description automatically generated

## Evaluation

Evaluating the performance and statistics of the multiple linear regression (MLR) model is crucial to gauge its reliability and utility. The MLR model was trained using 70% of the dataset, and its performance was assessed using the remaining 30%. The computed metrics are as follows:

* R-squared (R2): 0.2750
* Mean Absolute Error (MAE): 30.98
* Mean Absolute Percentage Error (MAPE): 361.34%
* Root Mean Squared Error (RMSE): 47.68
* Standard Deviation of Errors: 47.67

The R-squared value of 0.28 signifies that the variables incorporated into the model can explain approximately 28% of the variance in the number of reviews. This is a noteworthy achievement, especially considering that the dataset exclusively encompasses game characteristics, devoid of additional factors like marketing, game developer recognition, and other external influences.

The MAE of 30.98 suggests that, on average, the predicted total reviews of a game can deviate by approximately 31 reviews, either more or fewer. While this level of accuracy is reasonable, it should be noted that the MAPE (361.34%) indicates significant variability in prediction errors relative to the average predicted value. In practical terms, the predicted values likely exhibit even greater dispersion than suggested by the MAE.

Furthermore, the RMSE, which places higher penalties on outliers, reveals an even wider dispersion of predicted values, with an average deviation of 47.68 reviews. The relatively high standard deviation of errors at 47.67 underscores the substantial margin of error associated with the model's predictions.

In addition to assessing overall model performance, the MLR model provides valuable insights into the influence of independent variables on the total number of user reviews. Table 2 outlines the top 10 coefficients of these variables:

**Table 2:** *Top 10 most influential variables on the MLR model based on associated coefficients.*

|  |  |
| --- | --- |
| **Feature** | **Coefficient** |
| SteamVR\_Collectibles | 51.10 |
| Steam\_Trading\_Cards | 39.94 |
| F2P | 26.72 |
| Remote\_Play\_on\_Tablet | 22.45 |
| Remote\_Play\_on\_Phone | -18.12 |
| PCA\_165 | 17.88 |
| In\_App\_Purchases | 15.18 |
| PCA\_231 | -14.61 |
| MMO | 14.60 |
| PCA\_167 | 13.93 |

Notably, features related to SteamVR Collectibles and Steam Trading Cards exhibit the highest coefficients, suggesting that games falling into these categories tend to garner more reviews. This phenomenon may be attributed to the increased community interaction facilitated by these game types, resulting in more comprehensive reviews.

As anticipated, the presence of free-to-play (F2P) games is associated with a positive coefficient, reaffirming their tendency to generate more user reviews and downloads, which aligns seamlessly with the model's findings. Interestingly, games featuring remote play functionality on tablets exhibit a positive coefficient, indicating that they tend to accumulate a higher number of reviews. Conversely, games with remote play functionality on phones receive a negative coefficient, suggesting a lower review count, possibly due to the distinct gaming experiences offered by these devices. In-app purchases also emerge as influential, with a positive coefficient, signifying their role in attracting higher review numbers. Principal components, including PCA\_165, PCA\_231, and PCA\_167, feature prominently in the top 10 coefficients, with both positive and negative values. These components, though challenging to interpret directly, underscore their significance in shaping review counts based on underlying variables. Lastly, the presence of MMO elements within games positively influences review counts, emphasizing the role of social interactions in garnering additional reviews. These observations collectively contribute to a deeper understanding of the intricate relationships between game attributes and user reviews, providing valuable guidance for game developers seeking to optimize their games' appeal in a competitive gaming landscape.

While the model provides valuable insights into factors influencing user reviews, it's important to acknowledge its limitations and the potential impact of unaccounted variables. Despite its relatively modest predictive power, the MLR model represents a valuable starting point for understanding the relationships between game attributes and user reviews. Future iterations of the model may benefit from incorporating additional external factors and employing more advanced machine learning techniques to improve predictive accuracy and better support game developers in their decision-making processes.

## Deployment

In the deployment phase of the project, the models developed during the modelling phase are integrated into a functional web application. This is done by exporting the MLR model’s coefficients and intersection, the PCA components and K-means centroids, which will then be used to perform calculations based on the user’s input.

The application’s result is intended to assist game developers by estimating the potential number of reviews their game might receive. However, using the insights from Video Game Insights (2021), the application can also employ a ratio to estimate the number of sales of a potential game from the predicted review counts. According to Carless (2020), this ratio is close to 30 in recent years, as previously in a year like 2017 it would’ve been close to 55 for example. This change across the years is mainly related to changes in Steam's UX and developers encouraging people to leave reviews.

The sales-per-review ratio chosen for this project was chosen to be 35 for optimism while staying close to the average of 30 related to recent years. This results in the estimated purchases or downloads of the game. In order to calculate the potential revenue however, the estimated amount of game owners should be multiplied by the game’s US price and by a multiplier of about 0.38. This multiplier is derived from a complex calculation considering various factors like VAT, returns, regional pricing, discounts, and platform cut, as detailed by Weinbaum (2019).

The web application, hosted at <https://antoplh.github.io/steam-predictions/> , uses index.html for user input, with script.js managing interactions, and predictor.js processing the data locally. Upon data submission, predictor.js performs PCA and feeds the transformed data into the linear regression model. After the predicted reviews, downloads and estimated revenue are calculated, cluster categorization is also applied from the centroids of the clusters. Based on the cluster the game most resembles a histogram showing the distribution of positive reviews percentage for this game is shown alongside the average as shown in Figure 8. The website itself provides detailed explanation on the results as well as the html version of the notebook. The GitHub repository can be found at <https://github.com/antoplh/steam-predictions>, it contain the files mentioned that run the web application as well as the notebook, named “[Jupyter\_Notebook\_Individual\_Project.ipynb](https://github.com/antoplh/steam-predictions/blob/master/Jupyter_Notebook_Individual_Project.ipynb)”. The dataset itself is too big to load to the GitHub repository, thus the URL is provided instead.

**Figure 8**

*Final Web Application Results.*

*A screenshot of a test results

Description automatically generated*

This system allows for an end-to-end process where indie game developers can input game details, receive predictions and insights, and understand the financial implications concerning estimated sales and revenues, thus bridging the gap between complex data models and practical, user-friendly applications.

## Reflection and Ethics

Regarding the ethical considerations of this project, one must acknowledge the unique position it occupies in the gaming industry, particularly among indie developers. The tool's intent to guide these creators through the complex market dynamics of platforms like Steam is significant. It aspires to equip them with the foresight needed to navigate a highly competitive arena, leveraging historical data to forecast potential success. Yet, this foresight must be wielded with caution; as what makes the indie gaming sector so rich and even more popular in recent years is how it aims for providing innovation and creativity. Thus over-reliance on predictive analytics risks homogenizing this diversity.

The second layer of ethical consideration arises from the impact of the project's outcomes on consumer expectations and developer practices. Predictive models can create self-fulfilling prophecies, where the focus shifts to match projections, potentially at the cost of creative risk-taking and niche experimentation. The ethical implications of the implementation of this tool implies a commitment to refine it continually, to ensure that it is a compass rather than a map, guiding without dictating the journey of indie game development. It invites a broader dialogue on incorporating varied data reflective of emerging trends, minority representation, and ethical consumption, fostering an environment where indie games thrive on both their artistic merit and market viability.

# AI Bootcamp: LLM Micro-workshop

This report outlines the implementation of a Large Language Model (LLM) workshop at the AI Bootcamp. The workshop focused on OpenAI's GPT API model, aiming to develop an automated email response system for client inquiries about component availability. The project's execution aligns with the CRISP-DM framework.

## Business Understanding

Automatic response systems have revolutionized customer support across various industries. The emergence of advanced LLMs, such as GPT-3 and GPT-4, has further enhanced this domain. These models can generate contextual responses, moving beyond traditional predefined Q&A formats with limited interactions. This capability significantly reduces response times, boosts efficiency, and improves customer satisfaction.

## Data Understanding

The workshop utilized an Excel dataset containing comprehensive details such as component names, suppliers, part numbers, availability, pricing, and delivery times. This dataset, crucial for responding to customer queries, was static for the workshop's purpose. However, in a real-world application, a dynamically updated database would ensure current and relevant information is always available.

## Data Preparation

The Excel file was processed using Python's Pandas library, converting it to a string format with .to\_string() for LLM compatibility. This conversion was essential for the model to interpret and utilize the data effectively. The dataset was already in a clean state, simplifying the preparation process.

## Modeling

 The deployment of advanced models like GPT-3.5, a cornerstone of this project, eliminates the need for additional model training. These sophisticated models are adept at handling a diverse array of tasks based on user prompts, showcasing remarkable versatility and ease of use. For our workshop, the first step was establishing an account with OpenAI, enabling access to their API through a system of paid credits. This approach is in line with OpenAI's commitment to responsible AI usage and facilitates broader access to cutting-edge technology.

Upon account setup, we integrated OpenAI's Python library into our programming environment, specifically focusing on the Chat Completions API. This API is instrumental in interpreting and responding to user inputs. The process of engaging with the API involves two-step prompting mechanism:

* **System Instruction**: The initial prompt sets the role and expectations for the model. For our project, the instruction was framed as follows:
  + *System: "The next message you get will be an email from a customer asking for quotes or information on products. Assume the role of customer support service representing BOM S.A.C Your task is to answer emails concerning quotation requests based on the provided data table:"* This directive was followed by the insertion of the relevant data table, equipping the model with the necessary context for generating responses.
  + **System function output:** The API returns 3 arguments that are required to build an email consistently through a function: subject, recipient, and body.
* **User Interaction**: The next stage involves simulating real-world user interactions. Here, a 'user' role is introduced, providing specific inquiries or commands. An example of such a user prompt might be: "What is the current availability of this component, and what is the estimated delivery time?". These user prompts are designed to reflect typical customer queries, testing the model's ability to provide accurate and relevant information.

Once the system has been prompted, the user receives a response based on their input and the data table.

Through this sophisticated, yet user-friendly interaction model, GPT-3.5 demonstrates its capability to function as an effective tool in practical applications such as automated email responses. This streamlined approach to modeling not only exemplifies the power of modern AI but also underscores the ease with which it can be adapted to real-world business needs.

## Evaluation

In order to make this into a functioning email response system for customers interested in quotes or any other information for components requires the access to an email that should receive.

## Deployment

To transform our concept into a practical email response system, we integrated it with an email service that manages both incoming and outgoing correspondence. For our project, we chose Gmail due to its widespread usage and robust API capabilities.

The integration process involves leveraging a special file, known as token.js. This file plays a crucial role as it contains authentication tokens and credentials necessary for our Python application to interact with Gmail. These tokens serve as digital keys, allowing our application to access the Gmail account with specific permissions. These permissions include reading incoming emails and sending responses, crucial for the functionality of our automated system.

The token.js file typically includes the client ID, client secret, and access tokens obtained from the Google API Console.

At the culmination of our workshop, we presented a functioning demo, encapsulated within our Python workbook. This demo illustrates the end-to-end process of our email response system:

1. **Message Composition:** Users can compose a message within the Python environment, querying about specific components or requesting quotes. This message simulates a typical customer email.
2. **Processing Through Chat Completion API:** The composed message is then relayed to the Chat Completion API of the GPT-3.5 model. Here, the model processes the text, understanding the query in the context of the provided data.
3. **Generation of Response:** Based on the query, the model generates an appropriate response. This response leverages the comprehensive understanding of the model, ensuring accurate and relevant information is provided.
4. **Email Dispatch:** Once the response is generated, it is automatically sent through the Gmail account to the intended recipient. This step completes the cycle of receiving a query and sending an informative response, mimicking a real-world customer service interaction.

The implementation and deployment of this automated email response system showcases a practical application of LLMs in business communication. It underscores the potential of aI enhancing customer service and provides a model that can be adapted to various industries and business needs. An example output from the user input and the model’s response in shown in Figure 9. It shows that the system is able to effectively understand what component in the database the user is asking about and provide relevant information.

**Figure 9**

*Python output of a example of a user message and GPT model’s response based on data.*

|  |
| --- |
| DEMO: REQUEST COMPONENTS TO SUPPLIER  **\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***  **User input**  **\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***  Subject: Quote on hyperdriver ignitor  Recipient: [bom@sac.com](mailto:bom@sac.com)  Body:  Hello,  I'd like a quote on buying 4 hyperdriver ignitors, are these still available? how long would they take to arrive?  Best Regards,  Antonella  **\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***  **GPT Model response based on data**  **\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***  Subject: Quote Request: Hyperdrive Ignitor  Recipient: [antonellaportugal00@gmail.com](mailto:antonellaportugal00@gmail.com)  Body:  Hello Antonella,  Thank you for your inquiry. We have the Hyperdrive Ignitor in stock and available for purchase. The expected delivery time for 4 units is 10 days.  Please find the details below:  Component Name: Hyperdrive Ignitor  Supplier: Quantum Nexus  Supplier Part Number: QN-101  Part Number: Part-101  Quantity: 4  Price (EUR): 8.20 per unit  Expected Delivery Time (Days): 10  If you have any further questions or would like to proceed with the purchase, please let us know.  Best Regards,  BOM S.A.C. |

## Reflection and Ethics

One of the foremost ethical considerations in this project involved data privacy. Handling customer emails requires strict adherence to data protection laws and regulations. Ensuring that customer data is used solely for the intended purpose of responding to inquiries, and not for any unsanctioned analysis or distribution, is crucial. Additionally, obtaining explicit consent from users for their data to be processed by an AI system is a key ethical requirement.

Transparency is another critical ethical aspect. Users interacting with the system should be made aware that they are communicating with an AI. This transparency ensures that users understand the nature of the interaction and can adjust their expectations accordingly. It also fosters trust in the system and the organization deploying it.

In any AI system, there is a possibility of errors or misinterpretations by the model. Establishing clear accountability for such instances is a significant ethical concern. It's important to have mechanisms in place for users to report inaccuracies or inappropriate responses as well as relatively frequent moderation so these issues can be addressed promptly and effectively.

Finally, this project also served as a reminder that AI should be viewed as a tool to augment human capabilities, not replace them. Ethical deployment of AI in business should focus on enhancing and supporting human workers, providing them with more time for tasks that require human empathy and understanding, rather than completely automating all aspects of customer interaction.

# Process Mining Bootcamp

This section of the report documents the methodology and outcomes of the Process Mining Bootcamp, utilizing the Celonis software tool. Our approach is structured around the PM² methodology, offering a comprehensive view of the project's lifecycle from planning to analysis.

## Planning

**Accounts Payable P2P Tutorial**

Our initial engagement with Celonis involved exploring the "Accounts Payable DEMO" dashboard, a critical tool for managing B2B sales and risk through invoice process analysis. This exploration was instrumental in acquainting us with key functionalities of Celonis, such as process flow visualization in the variant explorer, analysis techniques, conformance checking, and examining invoice automation rates. This phase was crucial for building foundational knowledge in Celonis, facilitating a more effective engagement in subsequent stages.

**Pizzeria Business Case: Problem Identification and Research Questions**

The Pizzeria case study was a focal point of our project, encapsulating challenges in customer satisfaction and profitability. A thorough understanding of the business case revealed issues such as inconsistent customer experiences and inefficiencies in product offerings impacting profit margins. In response, our team formulated specific research questions:

* *Profit Margin Analysis*: "How do different products contribute to the pizzeria's profit margins, and which items might be underperforming or superfluous?"
* *Customer Satisfaction Study*: "Who are our primary customer segments, and what are their preferences and dissatisfactions with our service and product offerings?"

Our analysis was structured into two focused subgroups, each addressing distinct aspects of the business case:

|  |  |
| --- | --- |
| **Subgroup Members** | **Assigned Tasks** |
| Karolis, Job and Krishna | * Create variant analysis * Create business analysis * Create conformance analysis * Revise the As-Is process * Understand deviations * Understand most common variants and their characteristics * Impact of deviations on performance * Analyze cost factors * Analyze profit margin * Revenue breakdown |
| Antonella, Sophie and Lana | * Create the BPMN model of the process * Analysis of customer’s preferences * Analysis of customer satisfaction * Make customer analysis dashboard * Customer behavior analysis * Analysis of deviations of the process |

My role was pivotal in analyzing customer preferences and satisfaction. This involved creating detailed dashboards that not only highlighted customer behaviors but also provided actionable insights and recommendations. My focus was primarily on the analysis phase of the PM² process, collaborating closely with my team to ensure a comprehensive understanding of customer dynamics and their influence on the pizzeria's performance.

## Extraction

The Extraction phase, typically involving data collection and preparation, was not a primary focus in our workshop, as the data was pre-prepared and made available to us. This step, typically applied in real life scenarios, was bypassed to allow for a more immediate dive into the practical applications of process mining with Celonis.

## Data Processing

Our engagement with data processing began with the foundational task of creating a data pool and a corresponding data model in Celonis.

1. **Creating the Data Pool**

The first step involved constructing a data pool from the ground up. This process required us to meticulously upload and integrate various files containing comprehensive data about cases, customer interactions, and event logs. Attention was paid to resolving any formatting issues, particularly concerning the use of headers, ensuring the data's integrity and usability.

**2. Developing the Data Model**

With the data pool in place, the focus shifted to the development of a robust data model. This model was pivotal in interlinking the information from the uploaded files to facilitate process analysis. The key steps in this phase included:

* **Configuring the Model**: We initiated by selecting the primary activity table (pizza\_event.xlsx), crucial for mapping the process flow.
* **Identifying Essential Columns**: Columns representing "Case ID", "Activity", and "Timestamp" were marked as they form the backbone of Celonis' process mapping capabilities.
* **Linking Tables**: The next crucial step involved establishing relational links between the activity table and the other datasets. This was achieved by using "Case ID" and "Customer ID" as foreign keys, creating a network of interconnected data points.
* **Finalizing the Data Model**: The culmination of this phase was specifying the case table, enabling Celonis to process and analyze the data effectively, thus setting the stage for a comprehensive exploration of existing business processes.

This approach to data processing laid the groundwork for the analysis of the pizzeria's operational processes, which will be outlined in the next stage.

## Mining and Analysis

In this stage, we leveraged the powerful analytics capabilities of Celonis to dissect the pizzeria's operational data. This analysis was structured around four main activities: discovery, conformance, enhancement, and analytics. Below, we detail the tools, the rationale behind their use, our visualization strategies, and the PQL scripts that supported our analysis.

**1. Variant Analysis**

Using the Variant Explorer, we identified the most common process path initiated by phone orders, with an average throughput time of 43 minutes. This provided a benchmark for the 'happy path' in the process.

**Figure 10**

A screenshot of a chat

Description automatically generated*Happy path shown on the Variance Analysis*

**2. Conformance Analysis**

The Conformance Analysis tool provided critical insights into the process integrity. The statistics shown in the first image below shows that 42% of all cases where actually conforming (including additional processes that were manually defined as allowed deviations from the happy path). The tool allowed to understand the nature of some of the most common non-allowed deviations. Key findings included:

* Instances where the 'Baking pizza ready' stage was followed by 'Start preparing pizza,' indicating potential rework or order corrections.
* Cases where 'Start preparing pizza' was followed by a 'Call Customer' step, suggesting possible order clarification or late modifications to the order.
* Occurrences where 'Baking pizza ready' was followed by 'Departure pizza,' reflecting a possible delay in order dispatching.

These violations (shown in the second image below) highlighted areas where the process could be streamlined to reduce unnecessary steps.

**Figure 11**

A screenshot of a phone

Description automatically generated*Statistics on Conformance Analysis*

**Figure 12**

A screenshot of a computer

Description automatically generated*Most common violations in the Conformance Analysis*

**3. Cost Variance Dashboard Analysis**

The Cost Variance Dashboard was a pivotal tool in our analysis, revealing significant insights into product profitability and guiding our recommendations for menu optimization.

* **Small Pizzas**: The data showed that small pizzas, despite their popularity, contributed the least to the profit margin. Given their disproportionate costs compared to revenue generated, we concluded that they might not be a cost-effective offering for the pizzeria.
* **Paprika Pizza**: Further scrutiny revealed that the Paprika pizza, across all sizes, had the least popularity and profitability. This insight suggests that the Paprika pizza could be removed from the menu, which could streamline the kitchen processes and reduce ingredient costs.
* **Profit Margins and Throughput Time**: The overall average profit margin of 34% provided a benchmark for the pizzeria's financial health. We noticed a correlation between the average throughput time for pizza preparation and the revenue generated, indicating that efficiency in the kitchen directly affects profitability.
* **Strategic Pricing**: The insights led us to propose a strategic pricing adjustment. Specifically, we suggested the elimination of small pizzas from the menu and aligning the price of medium pizzas to the former small size to potentially increase profit margins through operational simplification and cost-effectiveness.

By incorporating these data-driven insights into the operational strategy, the pizzeria can streamline its menu offerings, improve the efficiency of its processes, and enhance profitability.

**Figure 13**

*Cost Variance Dashboard.*

A screenshot of a computer screen

Description automatically generated

**4. Business Analysis Dashboard**

The Business Analysis Dashboard provided invaluable insights into customer demographics and purchasing patterns, which are critical for tailoring the pizzeria's business strategy.

* **Customer Demographics**: Students emerged as the dominant customer group, indicating that the pizzeria's location or menu might be particularly appealing to this segment. This insight could be leveraged for targeted marketing campaigns, student discounts, or loyalty programs to further capitalize on this customer base.
* **Spending Patterns of Seniors**: Despite the prevalence of student customers, the data revealed that seniors are the highest spenders on average. This suggests an opportunity for creating tailored promotions or menu items that appeal to seniors, encouraging increased spending from this demographic.
* **Revenue Analysis**: The dashboard also provided a detailed breakdown of revenue by customer type and pizza category. By analyzing the total revenue and average revenue per pizza type, we can discern which items are the most and least profitable. For instance, Calzone and Fungi pizzas show a higher revenue, suggesting they are well-received and could be featured more prominently in marketing efforts.
* **Strategic Recommendations**: Based on these findings, we recommend the implementation of specific discount days or promotional deals aimed at seniors and students. Additionally, considering the popularity and profitability of certain pizzas, a menu reconfiguration might be beneficial to prioritize high-revenue items and potentially phase out less profitable ones.
* **Menu Adaptation**: It is essential to monitor the sales trends of different pizza types continuously. If certain pizzas consistently yield lower revenue despite high sales volumes, it may indicate a need for price adjustment or cost reduction in preparation.

**Figure 14**

*Business Analysis Dashboard*

A screenshot of a computer

Description automatically generated

**Customer Satisfaction Dashboard Analysis**

The Customer Satisfaction Dashboard served as a critical tool for assessing the pizzeria's service quality from the customers' viewpoint, revealing an average satisfaction score of 2.31 out of 5. This metric, while reflective of the overall experience, requires a deeper dive to understand its underlying factors.

* **Wait Times and Satisfaction**: A significant insight from the dashboard was the negative correlation between wait times and customer satisfaction. Customers who waited over 30 minutes rated their experience between 0-3, highlighting the urgency of improving service speed to enhance customer satisfaction.
* **Consistency in Product Quality**: The analysis indicated that the type or size of pizza did not significantly influence satisfaction ratings, suggesting that the pizzeria maintains a consistent quality across its offerings. This is a positive indicator of the kitchen's ability to deliver a uniform customer experience.
* **Demographic-Specific Preferences**: Diving further into demographic data, we uncovered distinct preferences that could inform targeted marketing strategies:
  + Adults preferred Paprika and Salami pizzas but were less satisfied with Calzones, indicating a potential to optimize the menu to adult tastes.
  + Seniors favored Speciale and Paprika pizzas but rated Veggie pizzas lower. This preference presents an opportunity to cater more closely to senior customers' preferences.
  + Students displayed relatively uniform satisfaction across all pizza types, suggesting a varied palate that could be engaged with a broad range of offerings.
  + Teenagers showed a marked preference for Veggie pizzas over Margherita, which could guide product development and marketing campaigns aimed at this younger demographic.

The senior demographic's favorable rating for Paprika pizza poses a strategic question given its lower profitability. This feedback necessitates a cost-benefit analysis to decide whether to retain this menu item, perhaps by reducing production costs, or to discontinue it altogether.

* **Impact of Non-Conformance on Satisfaction**: When focusing on non-conforming cases, the average satisfaction plummeted to 1.95, underlining the impact of operational inconsistencies on customer perceptions. Addressing these deviations could lead to significant improvements in customer satisfaction levels.

To address the identified issues, we suggest implementing operational optimizations to reduce wait times, such as revising the order-to-delivery process or increasing staffing during peak hours. Considering the unique preferences identified within customer demographics could also inform special promotions or menu customizations to enhance satisfaction.

**Figure 15**

*Customer Satisfaction Dashboard*A screenshot of a graph

Description automatically generated

**Figure 16**

*Customer Satisfaction Dashboard after filtering only non-conforming cases*

A screenshot of a computer

Description automatically generated

**PQL Scripts Utilized**

To quantify our findings, we wrote and executed several PQL scripts that were used for different plots:

* **Profit Margin Calculation:**

(AVG("Pizza\_Case\_Pizza\_Case"."REVENUE")-AVG("Pizza\_Case\_Pizza\_Case"."COSTS"))/AVG("Pizza\_Case\_Pizza\_Case"."REVENUE")

* **Revenue Calculation:**

(SUM("Pizza\_Case\_Pizza\_Case"."REVENUE")-SUM("Pizza\_Case\_Pizza\_Case"."COSTS"))

* **Average Revenue per Variant Calculation:**

(SUM("Pizza\_Case\_Pizza\_Case"."REVENUE")-SUM("Pizza\_Case\_Pizza\_Case"."COSTS"))/COUNT("Pizza\_Case\_Pizza\_Case"."VARIANT")

* **Day of the Week:**

CASE WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 0 THEN 'Sunday' WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 1 THEN 'Monday' WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 2 THEN 'Tuesday' WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 3 THEN 'Wednesday' WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 4 THEN 'Thursday' WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 5 THEN 'Friday' WHEN DAY\_OF\_WEEK("Pizza\_Event\_Pizzeria\_Event"."EVENTTIME") = 6 THEN 'Saturday' ELSE NULL END

* **Order Method Segmentation:**

CASE WHEN "Pizza\_Event\_Pizzeria\_Event"."ACTIVITY\_EN" = 'Order by phone' THEN 'Phone' WHEN "Pizza\_Event\_Pizzeria\_Event"."ACTIVITY\_EN" = 'Start order website' THEN 'Website' WHEN "Pizza\_Event\_Pizzeria\_Event"."ACTIVITY\_EN" = 'Order at the counter' THEN 'Shop' ELSE NULL END

These scripts facilitated the extraction of key performance indicators and provided a granular view of the operational data, driving our analysis and recommendations.

## Evaluation

In this evaluation section, we summarize the potential improvements derived from our analysis using Celonis, linking these suggestions back to the initial business case objectives established in the Planning phase. Each recommendation is supported by relevant Key Performance Indicators (KPIs) to demonstrate the anticipated impact on the pizzeria's performance.

**1. Streamlining Product Offerings:** Our analysis suggested the elimination of small-sized pizzas and the Paprika pizza, which were identified as low-margin items that complicate the menu and operational process. By simplifying the menu offerings, we anticipate an increase in the average profit margin and a reduction in the average throughput time for order fulfillment. These changes should positively influence financial KPIs such as the cost of goods sold and operational efficiency.

**2. Targeted Marketing Strategies:** The data revealed key demographic segments – students and seniors – that offer opportunities for targeted marketing. By implementing student discounts and promotions tailored to seniors, we expect to see an uplift in customer visit frequency and an increase in the average revenue per customer. These strategies are aligned with the goal of enhancing customer acquisition and retention rates.

**3. Improving Customer Satisfaction:** Wait times emerged as a critical factor affecting customer satisfaction. By optimizing kitchen operations and staff scheduling to reduce the average wait time, the pizzeria can expect to see an improvement in customer satisfaction scores. This operational enhancement is particularly crucial given the direct correlation between satisfaction rates and repeat business.

**4. Cost Reduction Without Compromising Quality:** The preference for Paprika pizzas among seniors suggests that if this item is to be retained, cost-reduction measures should be explored. Initiatives could include renegotiating supplier contracts or revising the recipe to maintain flavor while lowering costs. Success in this area would be measured by a maintained or increased gross margin for the Paprika pizza, without a decline in customer satisfaction.

During our presentation, we utilized several Celonis sheets to support our findings and recommendations:

* **Variant Explorer Sheet**: Illustrated the most common process path and areas where simplification could occur.
* **Conformance Analysis Sheet**: Identified where deviations from the expected process were occurring and the impact on customer satisfaction.
* **Cost Variance Dashboard**: Highlighted the less profitable products and informed our recommendation to remove them from the menu.
* **Business Analysis Dashboard**: Showed customer demographics and spending patterns, which supported our targeted marketing recommendations.
* **Customer Satisfaction Dashboard**: Provided insights into customer satisfaction levels and the factors influencing them, underpinning our recommendations for operational improvements.

Each sheet played a critical role in formulating our improvement ideas and provided a clear visual link between the data analysis and the actionable recommendations made. By following through on these recommendations, the pizzeria can expect to see measurable improvements in both customer satisfaction and financial performance, ensuring that the business case objectives are met, and that the pizzeria is positioned for continued success.

## Process Improvement and support

As this was a simulated business case study, the ideas are not able to be implemented, thus, this stage is not applicable for this project.

## Reflection and Ethical Considerations

As we conclude this report on the Process Mining Bootcamp, it is imperative to integrate a reflection on the learnings and ethical considerations that have emerged from this experience.

The bootcamp provided a comprehensive exposure to the PM² methodology, which was instrumental in shaping our understanding of structured process analysis and improvement. The hands-on application of this methodology using the Celonis tool enriched our skills in data analysis, translating complex datasets into actionable insights. The project, though a simulated business case, underscored the importance of each phase in the PM² process, from planning to analysis, in driving informed business decisions.

Collaboration within the team was a key learning aspect, highlighting the value of clear communication, teamwork, and individual contributions. My role in analyzing customer preferences and satisfaction offered a deeper understanding of how data-driven insights can significantly influence business strategies. This experience has equipped me with valuable skills and knowledge that are transferable to real-world scenarios.

Looking ahead, the ethical considerations encountered during the project serve as a foundation for future engagements in process mining. While this project was a case study with no real-life data privacy concerns, it highlighted the importance of ethical considerations in real-world scenarios, especially those involving customer data. Key learnings include:

* **Data Privacy and Security**: In any future project involving real data, ensuring data privacy and security will be paramount. The handling of customer data must comply with relevant data protection laws and ethical standards, maintaining customer trust and legal compliance.
* **Responsible Data Usage**: The responsible use of data is crucial. Future projects will require a careful balance between leveraging data for business insights and respecting the privacy and rights of individuals. This includes being mindful of how data is interpreted and the potential impacts of data-driven decisions on customers and stakeholders.
* **Transparency and Accountability**: Upholding transparency in how data is used and being accountable for the outcomes of data analysis are essential ethical practices. These principles will guide future engagements, ensuring credibility and integrity in process mining projects.

This bootcamp has not only enhanced my technical prowess in process mining but also instilled a keen awareness of the ethical dimensions of working with data. These insights and skills are invaluable for my future professional journey, where I anticipate applying these learnings to real-life business challenges, always with a keen eye on ethical implications and the responsible use of data.

# Bibliography

Carless, S. (2020). *How that game sold on Steam, using the “NB number”.* https://newsletter.gamediscover.co/p/how-that-game-sold-on-steam-using

Daley, S. (2022). *A Look Into the Gaming Industry: Types of Video Games and Gaming Jobs | Built In*. https://builtin.com/gaming

Fu, K. (2023). *The Indie Game Renaissance*. https://www.superjumpmagazine.com/why-its-easier-than-ever-for-indie-games-to-launch/

Fungies. (2023). *Amazing opportunities of indie gaming market - Fungies.io*. https://fungies.io/2023/03/16/amazing-opportunity-of-indie-gaming-market-size/

Game Developer. (2022). *There are 44,000 game developers on Steam. Who are they?* https://www.gamedeveloper.com/blogs/there-are-44-000-game-developers-on-steam-who-are-they-#close-modal

Mikulski, B. (2019). *PCA — how to choose the number of components?* https://mikulskibartosz.name/pca-how-to-choose-the-number-of-components

Tomashivli, N. (n.d.). *Steam Games Dataset*. Retrieved December 11, 2023, from https://www.kaggle.com/datasets/nikatomashvili/steam-games-dataset

Video Game Insights. (2021). *How to Estimate Steam Video Game Sales? | Video Game Insights*. https://vginsights.com/insights/article/how-to-estimate-steam-video-game-sales

Weinbaum, D. (2019). *Genre Viability on Steam and Other Trends - An Analysis Using Review Count 2*.