Amarelo- Sobre a AIAI

Azul- O que nós fazemos

Verde- Requisitos do nosso trabalho

Amazing International Airlines Inc.

Data Mining Project Guidelines

Fall Semester 2025-2026

Last Updated: 8 September 2025

---

1 Introduction

Amazing International Airlines Inc. (AIAI) is facing the challenge of designing personalized services and marketing strategies for its diverse customer base. In today's highly competitive airline industry, leveraging data-driven approaches to understand customer segments is crucial for improving satisfaction, increasing retention, and maximizing revenue potential.

Customer segmentation [4, 1] enables AIAI to identify distinct groups within their loyalty program. For instance, some customers may prioritize premium services and convenience, while others may be more cost-conscious and focused on basic travel needs. Additionally, certain groups may display seasonal travel patterns or specific route preferences. By uncovering these patterns, AIAI can tailor services, loyalty rewards, and marketing communications to meet the unique needs and behaviors of each segment.

2 Project Overview

In this project, you will act as consultants for AIAI. Your task is to analyze customer loyalty membership data and corresponding flight activity collected over a three-year period [3, 2]. Using these insights, you will develop a data-driven segmentation strategy. The dataset description is provided in Section 8.2.

The segmentation should be approached from multiple perspectives, such as:

- *Value-based segmentation,* grouping customers according to their economic contribution.
- *Behavioral segmentation,* analyzing purchasing habits and travel behaviors.
- *Demographic segmentation,* categorizing customers by age, occupation, or other attributes to reveal different interaction patterns.

The ultimate objective is to integrate these perspectives into a final segmentation framework that supports AIAI in crafting a comprehensive marketing strategy.

To structure the project, we recommend applying the **CRISP-DM** methodology [5, 6], progressing systematically from business understanding and data preparation through modeling and evaluation.

By the end of the project, students are expected to gain proficiency in:

- Applying unsupervised learning techniques for customer segmentation
- Conducting multi-perspective segmentation analysis
- Translating technical insights into actionable business strategies

CRISP-DM - (Cross-Industry Standard Process for Data Mining), que é um padrão de processo amplamente utilizado para projetos de mineração de dados e ciência de dados.
Esta metodologia é dividida em 6 fases (chat)
O CRISP-DM é cíclico, ou seja, pode-se retornar a fases anteriores conforme surgem novos insights ou problemas.

The project will be carried out in groups of up to four students. It will consist of three sequential deliverables, moving from exploratory data analysis to the development of strategic recommendations.

2.1 Deliverables Summary

Deliverable Components	Points	Description
Deliverable 1: EDA	**30**	**4 November 2025**
Jupyter Notebook	5	`GroupXX_EDA_Code.ipynb`
Report	15	`GroupXX_EDA_Report.pdf`
Infographic Poster	10	`GroupXX_EDA_Poster.pdf`
Deliverable 2: Clustering	**60**	**3 January 2026**
Jupyter Notebook	5	`GroupXX_Clustering_Code.ipynb`
Report	35	`GroupXX_Clustering_Report.pdf`
Video Presentation	20	`GroupXX_Clustering_Video` (link)
Deliverable 3: Discussion	**10**	**Date TBA** (January 2026)
Individual Assessment	10	In Person

3 References

[1] Sara Dolnicar, Bettina Grün, and Friedrich Leisch. *Market Segmentation Analysis*. Jan. 2018. DOI: 10.1007/978-981-10-8818-6. URL: https://doi.org/10.1007/978-981-10-8818-6.

[2] *Free Sample Dataset download - Airline Loyalty Program - Maven Analytics | Build Data skills, Faster*. URL: https://mavenanalytics.io/data-playground/airline-loyalty-program.

[3] *Guide to IBM Cognos Analytics Sample Data*. Dec. 2024. URL: https://community.ibm.com/community/user/browse/blogs/blogviewer?BlogKey=FF811D76-ABE0-4DF2-BCEA-917176FD72E4.

[4] Christiana Jolaoso. *Customer segmentation: the ultimate guide*. June 2024. URL: https://www.forbes.com/advisor/business/customer-segmentation/.

[5] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. "A Systematic Literature Review on Applying CRISP-DM Process Model". In: *Procedia Computer Science* 181 (2021). CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, pp. 526–534. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.01.199. URL: https://www.sciencedirect.com/science/article/pii/S1877050921002416.

[6] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester. 2000, pp. 29–39. URL: http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

4 Deliverable 1: Exploratory Data Analysis (30 points)

This deliverable lays the groundwork for segmentation by examining the airline loyalty dataset in detail. The focus is on uncovering meaningful patterns, identifying limitations, and generating initial hypotheses about customer groups.

Key Tasks:

- Conduct descriptive statistics and visualizations to highlight distributions, trends, and anomalies, while noting which variables appear most relevant for segmentation.
- Assess data quality issues and evaluate how these may affect clustering reliability.
- Identify preliminary behavioral signals that suggest distinct types of customers.
- Develop and justify engineered features. Show how these derived variables capture richer aspects of customer behavior and explain their potential contribution to clustering models.

In presenting results, consider addressing the following:

1. Which findings were most unexpected or insightful, and what do they reveal about likely customer clusters?
2. What data limitations pose the greatest risks for clustering, and how might they be mitigated?
3. Which patterns in customer activity, including those revealed by engineered features, suggest natural groupings, and what cluster characteristics do you anticipate?
4. How would you communicate these insights to non-technical stakeholders? Include a clear explanation of the expected number of clusters, the most important differentiating features, and any anticipated challenges.

4.1 Component 1: Jupyter Notebook (5 points)

This component assesses technical execution skills through documented code and analysis workflow.

Expectations:

- Jupyter notebook following file naming format `GroupXX_EDA_Code.ipynb`
- All code cells executed with visible outputs
- Systematic data exploration workflow
- Organized code, using Markdown headings to separate sections
- Clean, well-documented code with consistent variable naming
- Error-free execution in fresh environment

4.2 Component 2: Report (15 points)

This component evaluates analytical thinking and business insight as you transform technical exploration into strategic insights, following CRISP-DM methodology to prepare for segmentation analysis.

Expectations:

- Report following file naming format `GroupXX_EDA_Report.pdf`
- **Maximum** of 5 pages of content using provided framework and template
- Insights and interpretations (not just data descriptions)

Report Framework

Abstract: Provide a concise summary of your objectives, methodology, and key findings. This should give readers an immediate understanding of your study's scope and contributions.

Introduction: Establish the business context by discussing airline loyalty industry challenges and competitive pressures driving segmentation needs. Define specific analytical objectives for AIAI, explaining why customer segmentation is strategically important for this airline. This section corresponds to CRISP-DM Business Understanding and should demonstrate your grasp of the business problem.

Data Analysis: Present a systematic evaluation of data quality. Describe the dataset, variable distributions, and implications of any data anomalies discovered. Focus your exploratory analysis on understanding feature behaviors and relationships that might inform clustering decisions. This technical section aligns with CRISP-DM Data Understanding.

Results: Interpret your exploratory findings to justify specific feature engineering choices for clustering. Explain which patterns, correlations, and data characteristics lead to particular preprocessing decisions. Connect data insights, demonstrating how analysis findings inform feature selection, transformation, and scaling approaches. This section bridges CRISP-DM Data Understanding with preparation decisions.

Conclusion: This section prepares a solid foundation for the Data Preparation by establishing the analytical roadmap for clustering in the Phase 2 of the deliverables. Propose your clustering approach based on analysis findings, justify feature engineering decisions, and present specific clustering hypothesis. Outline preprocessing strategy and explain expected segmentation outcomes.

Annexes: The following are mandatory and do not count towards page limit. Examples are provided at the end of this document (Section 8.1).

- AI Usage Statement
- Contribution Statement
- Responsibility Statement

4.3 Component 3: Infographic Poster (10 points)

This component assesses your ability to communicate key insights to business stakeholders, translating technical findings into understandable visual summaries. This deliverable should be targeted at AIAI management team seeking data-driven insights for customer strategy.

Expectations:

- Single page poster following file naming format `GroupXX_EDA_Poster.pdf`
- A3 size poster in PDF
- 3-4 key findings with supporting visualizations
- Non-technical language accessible to executives
- Professional design with clear information hierarchy
- Focus on insights that inform segmentation strategy

5 Deliverable 2: Clustering Analysis (60 points)

This phase applies clustering techniques to the airline loyalty dataset in order to generate meaningful customer segments. The emphasis is on experimenting with multiple perspectives, validating results, and merging insights into a comprehensive solution that supports business objectives.

Key Tasks:

- Prepare the dataset for modeling.
- Analyze available features to determine which segmentation perspectives can be meaningfully applied.
- Perform segmentation using the identified perspectives.
- Apply at least two clustering approaches within each perspective and compare results.
- Propose a final merged segmentation solution that integrates the most important insights across perspectives into a coherent framework.

When presenting results, consider addressing the following:

1. Which clustering method(s) produced the most interpretable and stable results for each perspective?
2. How many clusters best represent the customer base overall, and what evidence supports this decision?
3. What differentiating features emerged as most important within each perspective, and how do they complement one another in the merged solution?
4. How would you describe the final set of customer segments to a business audience? Summarize defining traits, potential marketing opportunities, and any challenges encountered when integrating perspectives.

5.1 Component 1: Jupyter Notebook (5 points)

This component demonstrates technical implementation of clustering methodology through documented code workflow.

Expectations:

- Jupyter notebook following file naming format `GroupXX_Clustering_Code.ipynb`
- All code cells executed with visible outputs
- Clear code organization using Markdown structure
- Multiple clustering algorithms implemented and compared
- Clean, reproducible code with consistent documentation
- Error-free execution

5.2 Component 2: Report (35 points)

This component evaluates analytical rigor and business insight as you integrate technical clustering results with strategic recommendations for AIAI's customer segmentation.

Expectations:

- Report following file naming format `GroupXX_Clustering_Report.pdf`
- **Maximum** of 10 pages of content using provided framework and template
- Insights and interpretations (not just bulleted list of what you did)
- Evidence-based segment characterization and strategic recommendations
- Required annexes: AI usage attribution + contribution statement + responsibility statement

Report Framework

Executive Summary: Provide concise overview of key segment findings, strategic recommendations for each customer group. Target executive audience requiring actionable insights.

Introduction: Establish context by connecting Phase 1 findings to clustering objectives. Document data preparation decisions and feature engineering based on Phase 1 findings.

Methodology: Present feature selection process, algorithm selection process, parameter tuning methodology, and implementation approach comparing multiple clustering methods.

Results & Validation: Analyze clustering performance through technical metrics and statistical significance testing. Interpret business meaning of discovered segments with customer profiles and behavioral characteristics.

Strategic Recommendations: Develop actionable strategies for each customer segment, including marketing approaches, service customization, and loyalty program optimization for AIAI.

Conclusion: Integrate technical findings with business strategy, address methodology limitations, and outline future research directions for AIAI's segmentation approach.

Annexes: The following are mandatory and do not count towards page limit. Examples are provided at the end of this document (Section 8.1).

- AI Usage Statement
- Contribution Statement
- Responsibility Statement

5.3 Component 3: Video Presentation (20 points)

This deliverable should be targeted at AIAI executive leadership making strategic customer segmentation decisions. This component assesses your ability to communicate complex clustering results through compelling strategic narratives.

Expectations:

Acho que é aqui que o stor disse que podíamos fazer um ppt e gravar o ecrã com a nossa voz para fazer o vídeo

- 2-4 minute professional presentation with slide support
- Clear segment profiles with business characteristics
- Actionable strategic recommendations for each customer group
- Executive-level communication avoiding technical jargon
- Implementation timeline and expected business impact
- Online platform submission (YouTube unlisted or similar)

6 General Policies

6.1 Group composition

- Maximum of FOUR (4) students in each group. We recommend a group size of three (3).
- ALL students must be enrolled in a group on Moodle, regardless of group size.
- Students must be enrolled in a group on Moodle before the first delivery deadline.
- Changes to the group composition after the first delivery date is not recommended.

6.2 Plagiarism check

All submitted reports will undergo a plagiarism check. Ensure that your work is original and properly cite any external sources used.

6.3 Use of Generative AI tools

- The use of generative AI tools is permitted but must be fully disclosed. It is essential that the students' own contributions to the report exceed that of any AI tools used.
- Students must include a section in the annex of the report documenting the use of AI tools (Annex AI Usage Statement)
- Students must document specific AI tool usage (ChatGPT, Claude, Gemini, etc.)
- Students must specify how AI tool was used (ideation, code assistance, refining, proofreading)
- If the group did not use AI tools, this statement must explicitly state this.
- Students are fully responsible for the contents of the report they submit, including any material generated or assisted by AI tools.
- Do not just copy and paste the results if AI tools are used.
- Students may be asked to explain the meaning of any AI-generated content submitted to ensure their comprehension.

All insights, business interpretations, and strategic recommendations must represent original group analysis.

Penalties applied

A 10% penalty will be applied for each day of delay in delivering the project. Further deductions to the deliverable grade may also be applied for deviations to the guidelines.

7 Optional Bonus Components

Each deliverable can include an optional bonus component up to 20% of that deliverable's points. Bonus components are designed to reward exceptional work while maintaining core assignment focus.

7.1 Deliverable 1 Options

*Select up to two options, each contributing **up to** 10% additional credit toward the deliverable grade.*

1. External Data Integration & Validation

- Source and integrate relevant external dataset (industry benchmarks, demographic census data, etc.)
- Validate findings against external sources
- Provide comparative context for customer behavior patterns
- Include data source credibility assessment

2. Geo-Spatial Insights

- Incorporate mapping of routes, hubs, and customer origins/destinations.
- Visualize geographic concentration of loyalty members.
- Relate geographic differences to potential segmentation variables.

3. Multidimensional Outlier & Anomaly Detection

- Apply advanced anomaly detection methods.
- Identify unusual customer behaviors that may distort clustering.
- Discuss whether anomalies should be excluded, downweighted, or analyzed separately.

5. Interactive EDA Dashboard

- Build a lightweight interactive dashboard (e.g., Plotly Dash, Streamlit) for exploring customer attributes.
- Include basic filtering, drill-down, and visualization of distributions.
- Provide stakeholders with a tool for self-exploration of the dataset.

6. Business-Oriented Storytelling

- Deliver a polished business briefing document or short video targeted to executives.
- Demonstrate the ability to bridge technical and strategic communication.

7.2 Deliverable 2 Options

*Select up to two options, each contributing **up to** 10% additional credit toward the deliverable grade.*

Discussion of advanced business-focused analysis (2 pages maximum + Jupyter notebook)

1. Financial Impact Modeling (2 pages maximum + supporting code)

- Calculate ROI projections for each segment
- Include customer lifetime value estimates
- Provide cost-benefit analysis of implementation
- Create investment timeline with expected returns

Implementation and discussion of advanced clustering techniques (2 pages maximum + Jupyter notebook)

2. Fuzzy Clustering Implementation

- Apply fuzzy c-means with membership degree analysis
- Compare hard vs soft clustering interpretations
- Analyze customer overlap between segments
- Include uncertainty quantification in business recommendations

3. Deep Embedded Clustering (Autoencoders)

- Implement autoencoder-based dimensionality reduction
- Apply clustering in learned latent space
- Compare with traditional feature-based clustering
- Visualize learned representations and cluster boundaries

4. Clustering-Based Recommender System

- Develop segment-specific recommendation engines
- Implement collaborative filtering within clusters
- Cross-segment recommendation analysis
- Evaluate recommendation quality and business impact

Implementation and demonstration of interactive visualization (max 5-minute video + code hosted on GitHub)

5. Interactive Cluster Visualization Dashboard

- 3D cluster visualization with rotation and zoom capabilities
- Real-time filtering by demographic/behavioral attributes
- Customer detail pop-ups with segment characteristics
- Export functionality for stakeholder sharing

6. Dynamic Parameter Tuning Interface

- Live cluster parameter adjustment (K-values, distance metrics)
- Real-time visualization of cluster changes
- Performance metric updates during parameter modification
- Business impact calculator for different configurations

7. Interactive Dendrogram for Multi-view Clustering Acho que fizemos qq coisa parecida em AEDM

- Hierarchical clustering across multiple data perspectives
- Clickable tree navigation with cluster detail expansion
- Perspective switching (demographic vs behavioral vs temporal)
- Cross-perspective cluster comparison tools

7.3 General Bonus Guidelines

7.3.1 Quality Standards

- Bonus work should maintain the same high quality as your main deliverables.
- Completing bonus tasks should never compromise your core assignment requirements.
- Aim to add meaningful value that goes beyond the minimum expectations.

7.3.2 Documentation

- Clearly document all bonus components so others can understand your work.
- Explain why you chose this particular enhancement and what it contributes.
- Share any challenges you faced and how you solved them.

7.3.3 Grading

- Bonus points are calculated as a percentage of your base deliverable score.
- Exceptional work that meets all criteria earns full bonus points.
- Partial credit can be awarded for well-executed but incomplete bonus work.
- **Bonus points are only awarded if the core deliverable meets at least the passing standard.**

8 Annex

8.1 Annex Statements Examples

Annex AI Usage Statement

Outlines specific AI tool usage (ChatGPT, Claude, Copilot, etc.) and specify how each tool was used (ideation, code assistance, refining, proofreading, etc.)

AI tools were used for coding syntax assistance (ChatGPT for pandas operations), literature review brainstorming (Claude for industry context ideas), and report proofreading (Grammarly for grammar checks). All analytical insights, business interpretations, and strategic recommendations represent original group analysis and thinking.

Annex Contribution Statement

Individual student contributions to deliverables.

Sample Student (Student ID: 20257777)

- Fitness industry research for Business Understanding
- Introduction section, poster design
- Logistic Regression implementation

Lorem Student (Student ID: 20258888)

- Feature engineering ideation, RFM calculations
- Gym visit frequency analysis, peak hours identification
- Abstract, Conclusion sections, report coordination

Student Ipsum (Student ID: 20259999)

- Data undersampling approach
- Culinary preference insights
- Results interpretation, notebook organization

All members contributed to collaborative discussions and ideation.

Annex Responsibility Statement

We, the group members listed above, certify that this report represents our original analytical work and interpretations. While AI tools were used as specified above, all insights, conclusions, and recommendations are the result of our independent analysis and critical thinking. We take full responsibility for the accuracy and quality of this submission.

8.2 Annex: Dataset Information

`DM_AIAI_CustomerDB.csv`

Variable	Description
Loyalty#	Unique customer identifier for loyalty program members
First Name	Customer's first name
Last Name	Customer's last name
Customer Name	Customer's full name (concatenated)
Country	Customer's country of residence
Province or State	Customer's province or state
City	Customer's city of residence
Latitude	Geographic latitude coordinate of customer location
Longitude	Geographic longitude coordinate of customer location
Postal code	Customer's postal/ZIP code
Gender	Customer's gender
Education	Customer's highest education level (Bachelor, College, etc.)
Location Code	Urban/Suburban/Rural classification of customer residence
Income	Customer's annual income
Marital Status	Customer's marital status (Married, Single, Divorced)
LoyaltyStatus	Current tier status in loyalty program (Star > Nova > Aurora)
EnrollmentDateOpening	Date when customer joined the loyalty program
CancellationDate	Date when customer left the program
Customer Lifetime Value	Total calculated monetary value of customer relationship
EnrollmentType	Method of joining loyalty program

`DM_AIAI_FlightsDB.csv`

Variable	Description
Loyalty#	Unique customer identifier linking to CustomerDB
Year	Year of flight activity record
Month	Month of flight activity record (1-12)
YearMonthDate	First day of the month for the activity period
NumFlights	Total number of flights taken by customer in the month
NumFlightsWithCompanions	Number of flights where customer traveled with companions
DistanceKM	Total distance traveled in kilometers for the month
PointsAccumulated	Loyalty points earned by customer during the month
PointsRedeemed	Loyalty points spent/redeemed by customer during the month
DollarCostPointsRedeemed	Dollar value of points redeemed during the month