# Semester 4 Reflection: AI, ML & Data Learning Journey

**Weeks 1-4 Review | Anton Horvat | Fontys University**

## Overview

The first four weeks of this semester focused on building foundational competencies in data science methodology, visualization techniques, and machine learning algorithms through systematic exercises that directly supported my movie success prediction project. Each instructor’s guidance proved invaluable in developing both technical skills and project direction.

Contents

[Semester 4 Reflection: AI, ML & Data Learning Journey 1](#_Toc209526566)

[Overview 2](#_Toc209526567)

[Key Learning Experiences and Project Alignment 4](#_Toc209526568)

[Data Visualization Mastery (Week 2) 4](#_Toc209526569)

[Wine Dataset Analysis - Data Provisioning Foundation 4](#_Toc209526570)

[k-Nearest Neighbors Implementation - Algorithm Understanding 4](#_Toc209526571)

[Decision Trees - Explainable AI Focus 5](#_Toc209526572)

[Alpha Corporation Group Project - Collaborative Data Analysis 5](#_Toc209526573)

[Data Integration Exercise - Multi-Source Analysis 5](#_Toc209526574)

[Research Methodology Development 6](#_Toc209526575)

[Teacher Feedback Integration and Weekly Guidance 6](#_Toc209526576)

[Hans - Weekly Motivation and Direction 6](#_Toc209526577)

[Technical Instructors - Concept Mastery 6](#_Toc209526578)

[Strategic Decisions and Rationale 7](#_Toc209526579)

[Algorithm Selection 7](#_Toc209526580)

[Feature Engineering Approach 7](#_Toc209526581)

[Visualization Strategy 7](#_Toc209526582)

[Current Project Status - Movie Success Prediction 8](#_Toc209526583)

[Data Collection Phase - COMPLETED ✅ 8](#_Toc209526584)

[Integration Techniques Applied 8](#_Toc209526585)

[Success Classification System 8](#_Toc209526586)

[Next Phase - Data Provisioning 8](#_Toc209526587)

[Competency Development Progress 9](#_Toc209526588)

[Learning Reflections 10](#_Toc209526589)

[What Worked Well 10](#_Toc209526590)

[Key Insights 10](#_Toc209526591)

[Areas for Continued Focus 10](#_Toc209526592)

[Next Steps 10](#_Toc209526593)

[Conclusion 10](#_Toc209526594)

# Key Learning Experiences and Project Alignment

### Data Visualization Mastery (Week 2)

The matplotlib and pandas visualization exercises proved instrumental for my project’s data provisioning phase. Working with city population data taught me essential visualization principles: - Multiple chart types for different data patterns (scatter plots, histograms, boxplots) - Proper labeling and professional presentation standards - Comparing distributions across different categories

**Project Impact**: These skills directly transferred to analyzing movie datasets, particularly for understanding budget distributions across genres and success rate patterns over time.

**Teacher Support**: The instructor’s emphasis on proper chart labeling and professional presentation standards elevated my visualization quality from basic plots to presentation-ready graphics.

### Wine Dataset Analysis - Data Provisioning Foundation

The wine classification exercise established my systematic approach to data understanding: - Correlation matrix analysis for feature relationships - Distribution analysis through multiple visualization techniques  
- Missing value assessment and data quality evaluation - Feature scaling importance for algorithms

**Critical Insight**: The wine exercise taught me that “more features” doesn’t always mean better performance - a lesson that influenced my movie project’s feature selection strategy.

**Teacher Guidance**: The instructor’s detailed walkthrough of the systematic data provisioning methodology helped me understand why each step matters, preventing me from rushing into modeling without proper data understanding.

### k-Nearest Neighbors Implementation - Algorithm Understanding

The iris flower classification provided hands-on experience with: - Train/test splitting and reproducibility (random\_state importance) - Feature selection through systematic experimentation - Hyperparameter tuning with k-values - Distance-based algorithm behavior and preprocessing requirements

**Project Application**: Understanding k-NN’s interpretability (“similar movies performed like this”) made it an ideal baseline algorithm for my movie prediction system, where stakeholders need explanations for investment decisions.

**Learning Enhancement**: The instructor’s explanation of why k-NN requires feature scaling clicked when I saw the distance calculation examples. This understanding proved crucial for implementing proper preprocessing in my movie project.

### Decision Trees - Explainable AI Focus

The Titanic survival prediction exercise demonstrated transparent machine learning: - Feature importance visualization and interpretation - Decision path explanation through tree diagrams - Overfitting recognition and prevention techniques - Business-relevant evaluation metrics

**Strategic Value**: This exercise validated my project’s emphasis on explainable AI - crucial for movie industry stakeholders who need to understand why certain predictions are made.

**Instructor Impact**: The teacher’s focus on interpreting the decision tree visualizations helped me realize that model explainability isn’t just about accuracy - it’s about building trust with business users who need to understand the “why” behind predictions.

### Alpha Corporation Group Project - Collaborative Data Analysis

Working with team members on variance analysis developed: - Professional visualization standards for business reporting - Collaborative data interpretation skills - Clear communication of analytical findings - Group coordination on technical tasks

**Team Contribution**: I handled the data loading and statistical analysis components while coordinating with teammates on visualization design and interpretation. The exercise taught me the importance of consistent formatting and clear documentation when working in teams.

**Collaborative Learning**: The instructor’s feedback on our group presentation emphasized the importance of telling a coherent story with data, not just showing individual charts - a principle I’ve applied to structuring my movie project analysis.

### Data Integration Exercise - Multi-Source Analysis

The Croatia GDP integration exercise taught critical data combination techniques: - Union operations for combining similar datasets - Join operations (inner, left, outer) for enriching primary data - Handling different data granularities (quarterly to yearly aggregation) - Managing missing data patterns across integrated sources

**Technical Growth**: This exercise prepared me for combining movie data from multiple APIs (TMDB, OMDb) with different coverage periods and data quality levels. Understanding when to use different join types prevented data loss in my movie dataset integration.

**Practical Application**: The web scraping component showed me how to work with real-world data sources that aren’t perfectly clean - directly applicable to my movie project’s API data collection challenges.

# Research Methodology Development

The “On Research” exercise enhanced my systematic approach by:

- Distinguishing experimental vs. observational research strategies

- Formulating precise research questions

- Identifying research sub-questions and assumptions

- Understanding content analysis methodology

**Project Relevance**: These skills improved my movie project’s domain research approach and helped structure my analysis of entertainment industry patterns.

# Teacher Feedback Integration and Weekly Guidance

### Hans - Weekly Motivation and Direction

Hans’s weekly check-in conversations proved invaluable for maintaining momentum and clarity:

- **Motivation Maintenance**: When I didn’t know what too really do the first week, he explained to me exactly what to do

- **Priority Clarification**: His guidance helped me focus on what to work on next, preventing me from getting lost in interesting but tangential research

- **Confidence Building**: Regular affirmation that I was on the right track kept me motivated during challenging technical implementation phases

**Key Impact**: Without Hans’s weekly guidance, I would have struggled with project scope management and maintaining consistent progress. His approach of asking “what specific thing will you work on this week?” kept me accountable and focused.

### Technical Instructors - Concept Mastery

Each technical instructor contributed unique insights:

- **Roopali** Helped me understand that charts must tell a story, not just display data

- **Priyanka**: Clarified why certain preprocessing steps are essential and when to use different algorithms

- **John**: Emphasized the importance of systematic approaches and proper documentation

**Collective Impact**: The combination of technical depth from subject matter experts and strategic guidance from Hans created a comprehensive learning environment that addressed both skills development and project management.

# Strategic Decisions and Rationale

### Algorithm Selection

Chose k-NN as baseline because:

- Natural interpretability aligns with explainable AI requirements

- Distance-based logic matches business intuition (“similar movies”)

- Provides clear performance benchmark for future algorithm comparisons

### Feature Engineering Approach

Applied wine dataset insights:

- Systematic correlation analysis before feature selection

- Testing feature combinations rather than assuming “more is better”

- Balancing technical performance with business interpretability

### Visualization Strategy

Implemented lessons from city data exercise:

- Multiple visualization types for comprehensive data understanding

- Professional presentation standards for stakeholder communication

- Clear axis labeling and legend design for accessibility

# Current Project Status - Movie Success Prediction

### Data Collection Phase - COMPLETED ✅

Successfully implemented multi-API data collection system:

- **2,813 movies collected** from TMDB and OMDb APIs

- **Quality filtering applied** - retained movies with complete budget and revenue data

- **2,696 clean movies** after removing pandemic-affected years (2020-2021)

- **55 comprehensive features** including financial, content, people, and external validation data

### Integration Techniques Applied

* **Union Integration**: Combined collection batches, removed duplicates
* **Inner Join**: Movies with both TMDB and OMDb data for high-quality subset
* **Left Join**: All TMDB movies with OMDb data where available
* **Exclusion**: Systematic removal of anomalous data points

### Success Classification System

Implemented industry-standard profitability categories:

- **Flop**: Revenue < 1x Production Budget (immediate loss)

- **Break-even**: Revenue = 2-2.5x Production Budget (covers all costs)

- **Hit**: Revenue > 2.5x Production Budget (actual profit)

### Next Phase - Data Provisioning

Ready to begin comprehensive analysis with 12-16 visualizations following systematic methodology learned from wine exercise.

# Competency Development Progress

**Professional Standard**: Systematic methodology application and clear documentation demonstrate professional-level work quality. Weekly feedback sessions with Hans reinforced the importance of consistent professional practices.

**Personal Leadership**: Independent decision-making in algorithm selection, feature engineering, and project direction shows emerging leadership capabilities.

**Explainable AI**: Focus on interpretable models and clear explanation of predictions addresses industry need for transparent AI systems. Instructor guidance on decision trees reinforced this priority.

**Data Preparation & Analysis**: Comprehensive data provisioning methodology ensures robust foundation for modeling work. Wine exercise insights continue to guide approach.

**Model Engineering**: Progressive learning from simple to complex algorithms builds solid technical foundation while maintaining focus on business value.

# Learning Reflections

### What Worked Well

* **Systematic approach** learned from exercises translated directly to project work
* **Teacher guidance** provided both technical depth and strategic direction
* **Hands-on practice** with real datasets built confidence for independent work
* **Weekly check-ins with Hans** maintained momentum and prevented scope creep

### Key Insights

* **Integration is crucial**: Each exercise built capabilities needed for the larger project
* **Methodology matters**: Systematic approaches prevent costly mistakes and rework
* **Teacher support amplifies learning**: Having both technical expertise and motivational guidance accelerated progress significantly
* **Real-world application**: Working on a personally meaningful project (movies) increased engagement and retention

### Areas for Continued Focus

* **Advanced modeling techniques** while maintaining explainability focus
* **Stakeholder communication** of technical findings
* **Systematic experimentation** and results documentation
* **Business value demonstration** alongside technical metrics

## Next Steps

The foundation built in these four weeks, supported by excellent instructor guidance, positions me well for advancing to more sophisticated modeling techniques while maintaining focus on explainable, business-relevant predictions. Hans’s continued weekly guidance will be essential for maintaining progress momentum as the project complexity increases.

The systematic approach developed through these exercises, reinforced by consistent teacher support, ensures continued methodical progress toward project completion.

# Conclusion

These first four weeks established a solid foundation for the remainder of the semester through a carefully structured progression of technical skills and project application. The integration of systematic exercises with real-world project work created a learning environment where theoretical concepts immediately found practical application.