# IST 782 Applied Portfolio - DRAFT

Francisco Franco Arenas

## 1. introduction

Data science stands at the intersection of statistical analysis, computational techniques, and domain-specific knowledge, offering powerful tools for extracting meaningful insights from vast datasets. As an economist at the Ministry of Economic Affairs in Spain, I recognized the transformative potential of data science to enhance decision-making and policy formulation. Pursuing a Master of Science in Applied Data Science was a strategic decision to augment my expertise with advanced data analysis skills. This program has equipped me with the ability to harness complex data, apply sophisticated machine learning algorithms, and develop predictive models, thereby improving my capacity to analyze economic trends, evaluate policies, and make data-driven recommendations. The integration of these skills is crucial in today's data-driven world, enabling me to contribute more effectively to the mission of fostering economic stability and growth.

As I reflect on my journey through the Master of Science in Applied Data Science program at Syracuse University, this paper serves as a comprehensive demonstration of how I have met the program's learning goals through various assignments and project deliverables. Each section of this document links specific learning objectives to the practical work I have completed, showcasing not only my academic achievements but also my preparedness for a career in data science. By examining the projects I have undertaken, the methodologies I have employed, and the insights I have generated, this paper provides a holistic view of my growth as a data scientist and my readiness to tackle real-world challenges.

## 2. Learning goals

The main learning goals of the program are as follows:

1. Collect, store, and access data by identifying and leveraging applicable technologies
2. Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
3. Apply visualization and predictive models to help generate actionable insight
4. Use programming languages such as R and Python to support the generation of actionable insight
5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

## 3. Projects

### 3.1 Project 1: IST 719 Information Visualization: Poster on transitions from University to employment in Spain

Associated learning goals: 1,2,3,4,5

In the IST 719 Information Visualization course, we explored advanced visualization techniques, primarily utilizing R's ggplot2 library, while grounding our practice in the theoretical principles of effective data visualization. The course also introduced us to Adobe Illustrator, enabling us to design visually compelling posters that effectively communicate complex data insights.

For my project, I developed an in-depth poster using data from the University Placement Survey of Spain. This project aimed to evaluate the effectiveness of various fields of study in facilitating the transition from university to employment. It provided a comprehensive analysis of how different academic disciplines prepare students for the job market. Through this project, I gained significant experience in handling and transforming public survey data into meaningful and impactful visual representations.

The poster featured several advanced visualization techniques that extended beyond the standard curriculum. I created an alluvial plot to illustrate the flow of graduates into various employment sectors, capturing the pathways from education to career. Ridgeline plots were employed to compare wages across different fields of study, offering a clear and concise overview of which disciplines have the highest and lowest wages. Additionally, Cleveland dot plots were used to highlight key statistics, providing a straightforward comparison of employment outcomes for graduates in different fields.

This project was a crucial learning experience that honed my technical skills in data visualization and my ability to convey complex information effectively. It underscored the importance of clear and impactful visual communication in the field of data science, particularly in making data-driven insights accessible and understandable to a broader audience. Through this project, I not only applied the skills learned in class but also expanded my proficiency in creating advanced and informative visualizations, which are essential for data-driven decision-making in my role as an economist at the Ministry of Economic Affairs in Spain.

### 3.2 Project 2: IST 718 Big Data: ML classifier of financial complaints

Associated learning goals: 1,2,3,4,6

In the IST 718 Big Data course, we focused on handling large datasets using Apache Spark. We began by delving into the theory behind the MapReduce architecture to understand Spark's underlying mechanisms. Subsequently, we learned data wrangling and machine learning techniques within the Spark framework. A crucial aspect of this learning process was grasping Spark's lazy evaluation, which enabled us to program more efficiently by understanding how Spark executes code.

For our project, we utilized an open dataset of financial complaints with the objective of classifying them into categories based on the textual content of the complaints. This dataset comprised over 10 million instances, necessitating the application of efficient big data techniques. We employed Spark's MLlib library to preprocess the text data, extract features, and train a multi-class classifier to categorize the complaints accurately.

To achieve accurate classification, we experimented with various types of text embeddings, ranging from the simpler TF-IDF (Term Frequency-Inverse Document Frequency) to more sophisticated BERT embeddings. Handling the class imbalance within the data was a significant challenge, as the main category encompassed over 50% of the instances. To address this, we implemented a two-stage classifier. The first stage predicted the main category, and for the instances that did not belong to this category, the second stage fine-tuned the classification for the remaining categories.

This approach not only improved the accuracy of our classification but also ensured that the model was robust across all categories. The final model achieved an AUC (Area Under the Curve) score of XXX, reflecting its ability to distinguish between different categories effectively. This project was instrumental in enhancing our skills in managing big data, implementing machine learning algorithms on large datasets, and addressing challenges such as class imbalance. These skills are invaluable in my role as an economist, enabling me to analyze large-scale financial data and derive meaningful insights for informed decision-making at the Ministry of Economic Affairs in Spain.

### 3.3 Project 3: IST 737 Visual Analytic Dashboards: Europe's quality of life visualization dashboard

Associated learning goals: 1,2,3,4,5

In the IST 737 Visual Analytic Dashboards course, I focused on creating interactive visual analytic dashboards using Tableau. We started with the basics and advanced to complex topics such as window functions. The course emphasized not only creating visualizations but also communicating stories effectively through interactive data.

For my project, I developed an interactive dashboard that visualizes the quality of life in the European Union. The goal was to provide a comprehensive and interactive tool for exploring various aspects of quality of life across EU countries.

The project began with the creation of a data pipeline to automatically ingest data from Eurostat using the Eurostat API. This required a deep understanding of API integration and data extraction techniques. Once the data was ingested, I prepared it for Tableau by ensuring it was efficiently modeled for optimal performance. This involved cleaning and transforming the data to ensure it was in the correct format and structure for analysis.

The dashboard focused on three main topics: health, job quality, and life satisfaction. Each topic was explored through multiple interconnected dashboards:

- Health: I created visualizations to display metrics such as healthcare resources, good habits and prevalence of various health conditions. Interactive elements allowed users to compare health outcomes across different countries and demographic groups.
- Job Quality: This section highlighted data on job satisfaction, job security, and working conditions. Visualizations included comparative bar charts, heatmaps, and trend lines to show how job quality varied across the EU.
- Life Satisfaction: I used surveys and subjective well-being data to create visualizations that depicted overall life satisfaction. This included mapping satisfaction scores across regions and demographic segments, using color-coded maps and dynamic charts.

Through these dashboards, I aimed to tell a compelling story that went beyond the raw data, providing context and insights into the quality of life in the EU. The interactive nature of the dashboard allowed users to explore the data in a meaningful way, revealing patterns and trends that are not immediately apparent in static reports.

This project significantly enhanced my expertise in Tableau, allowing me to leverage its powerful features to create engaging and informative visualizations. However, I have encountered some difficulties in transferring these skills to PowerBI, the software used in my current job. Despite this challenge, the project has been invaluable in developing my ability to create meaningful and interactive data presentations, a crucial skill for my role as an economist at the Ministry of Economic Affairs in Spain.