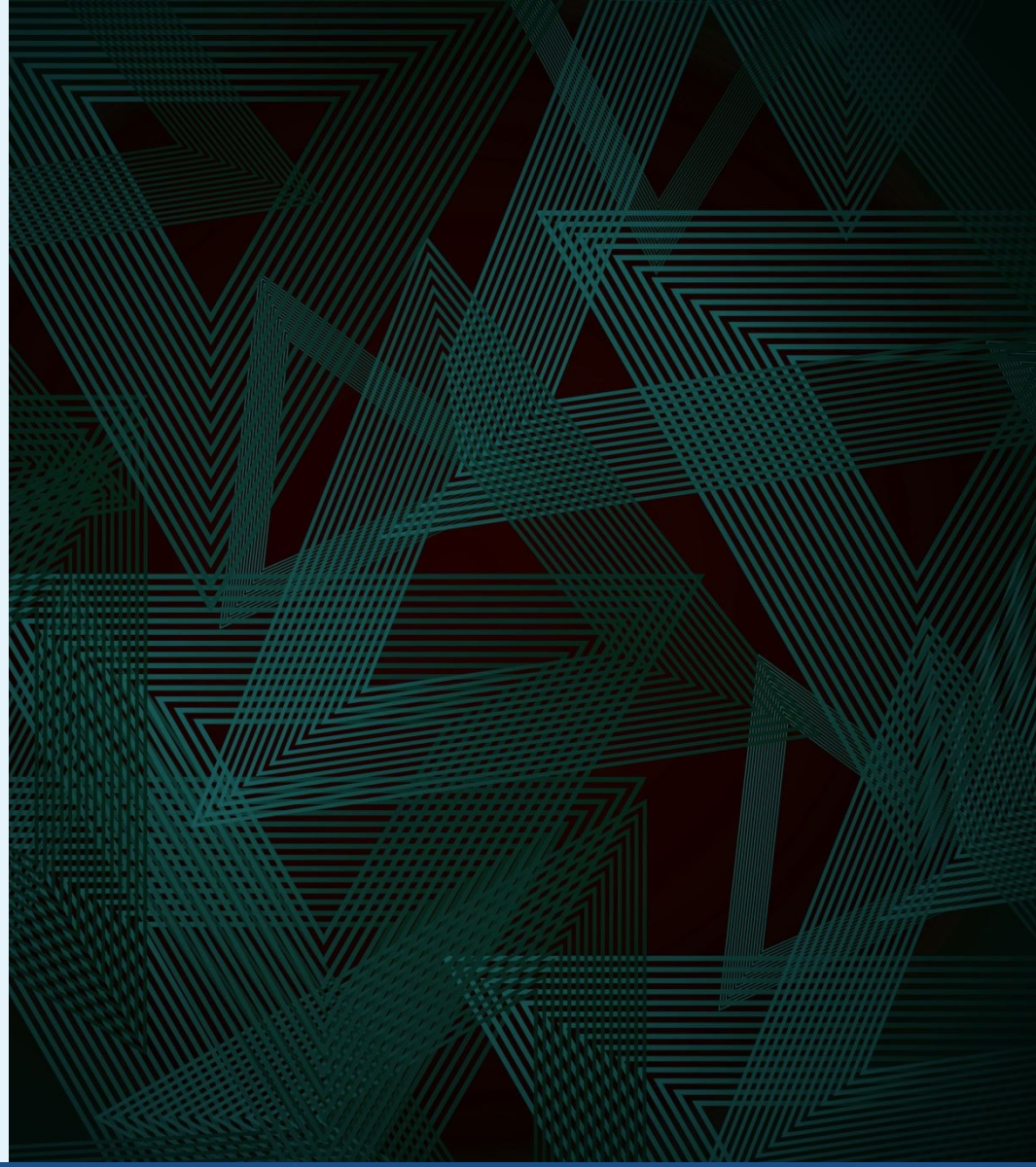
MS - Applied Data Science Portfolio

By Gabe Herz



Introduction

- The Master's of Science in Applied Data Science at Syracuse University's School of Information Studies is an interdisciplinary degree providing students the opportunity to learn in a broad range of areas pertaining to data science.
- The program outlines six main goals students will come away with at the conclusion of their degree. These goals are:
  - Collect, store, and access data
  - Create actionable insights across a range of contexts
  - Apply visualizations and predictive models to help generate actionable insight
  - Use programming languages such as R, SQL, and Python
  - Communicate insights to a broad range of audiences
  - Apply ethics in the development, use, and evaluation of predictive models and data.

Applied Data Science Program Overview

- The ADS program consists of 34 credits
  - 18 Credits of Core Classes
  - 6 Secondary Track (Concentration) Credits
  - 9 Free Elective Credits
  - 1 Portfolio Credits

Course Overview

- Five Courses were highlighted in this Portfolio
  - IST 659 – Intro to Database Management (Core)
  - IST 687 – Intro to Data Science (Core)
  - IST 722 – Data Warehousing (Secondary Track)
  - IST 769 – Advanced Big Data Management (Secondary Track)
  - IST 707 – Applied Machine Learning (Core)

- The chosen secondary track was Data Pipelines & Platforms

IST 659 – Intro to Database Management

Term: Fall 2022

Programming Languages: SQL

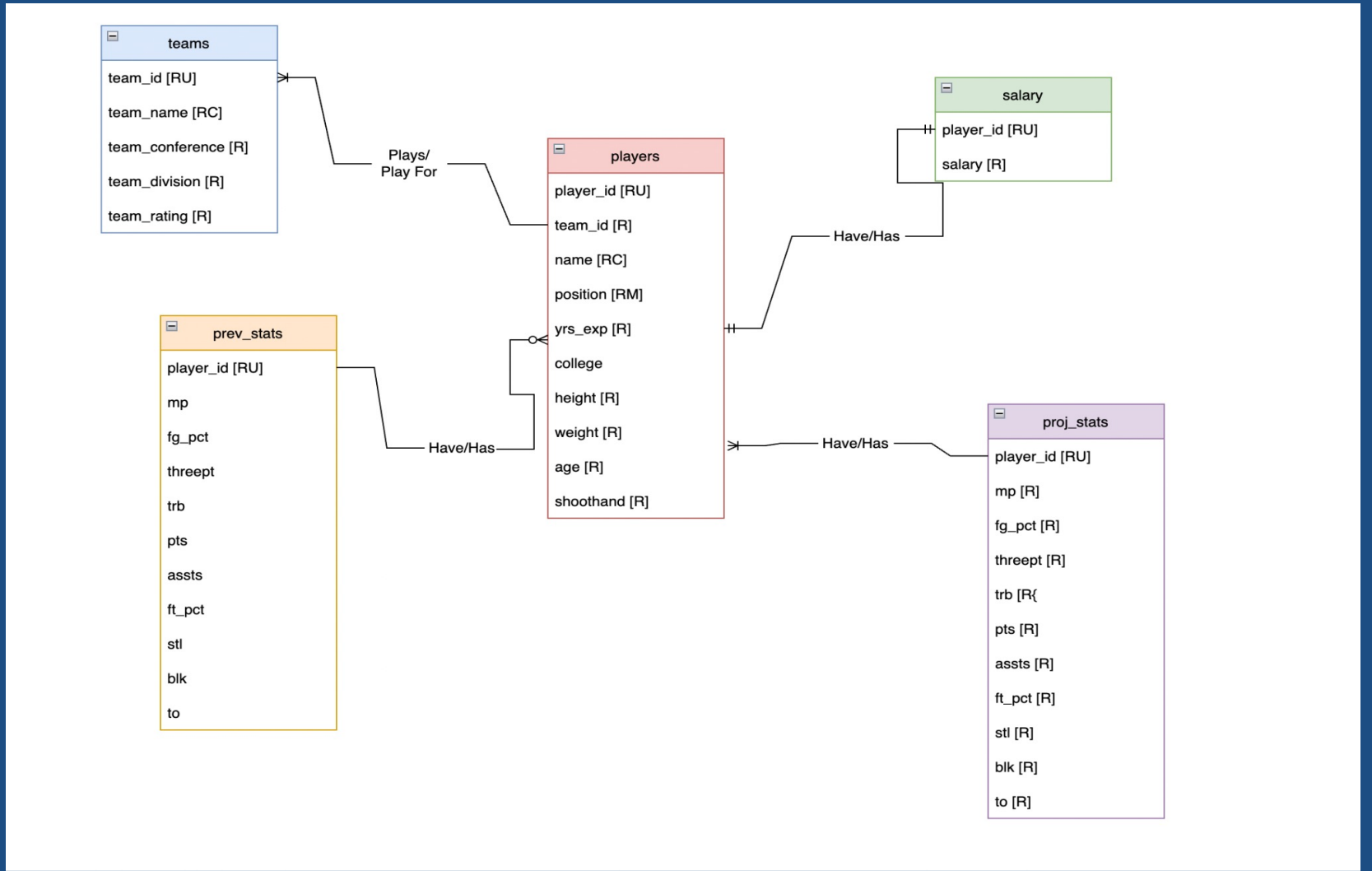
Tools/Software: Microsoft SQL Server

Course Structure: Asynchronous Lectures, In-person review sessions, lab assignments, and final project

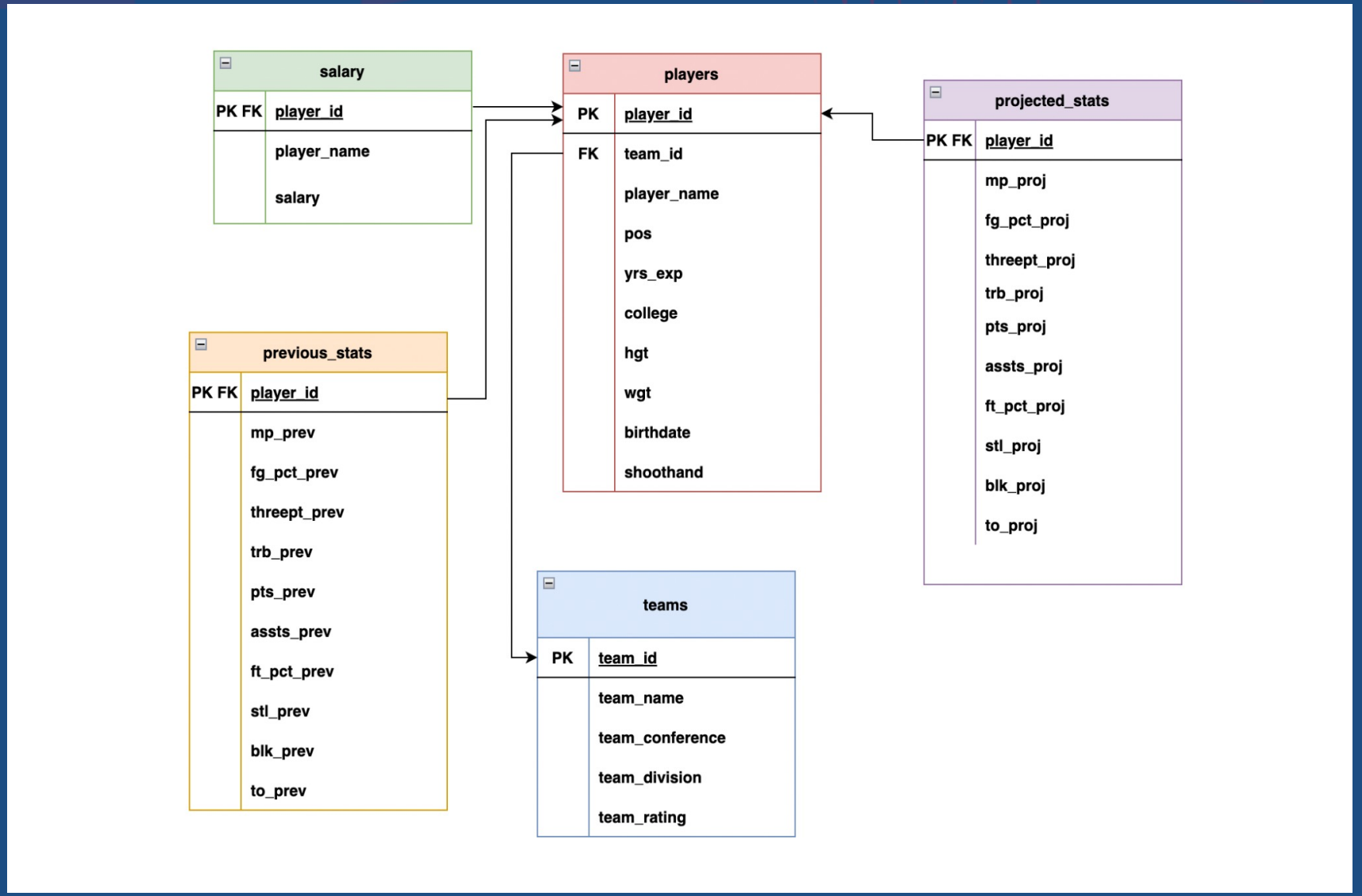
IST 659 – Final Project

- **Final Project Idea**: Database for Fantasy Basketball Website
- **Use-Case**: Serve as a repository for information on players, their statistics, their cost (known as “salary” in Fantasy Basketball), and their team’s information.
- **Process**:
  - Entity-Relationship table created in Excel and then visually illustrated through Conceptual and Logical Models
  - Up/down scripts created in SQL to initialize database tables, relationships, and populate with data.
  - Used Canva to create animated demo of app
  - Created PowerPoint presentation to communicate results and showcase product

Conceptual Diagram



Logical Diagram



IST 659 – Final Project Takeaways

- Clear to see how main course goals are highlighted
- Data was collected, accessed, and stored using SQL commands within Microsoft SQL Server.
- Implementation of the power app allowed for the creation of visualizations, aiding the ability to create and generate actionable insights.
- The presentation helped explain some nuances of both database management and fantasy basketball in layman's terms, allowing the project to resonate with broad audiences.
- Additionally, the presentation provides snippets of SQL commands used to power the application on the back-end.

IST 687 – Intro to Data Science

Term: Fall 2022

Programming Languages: R

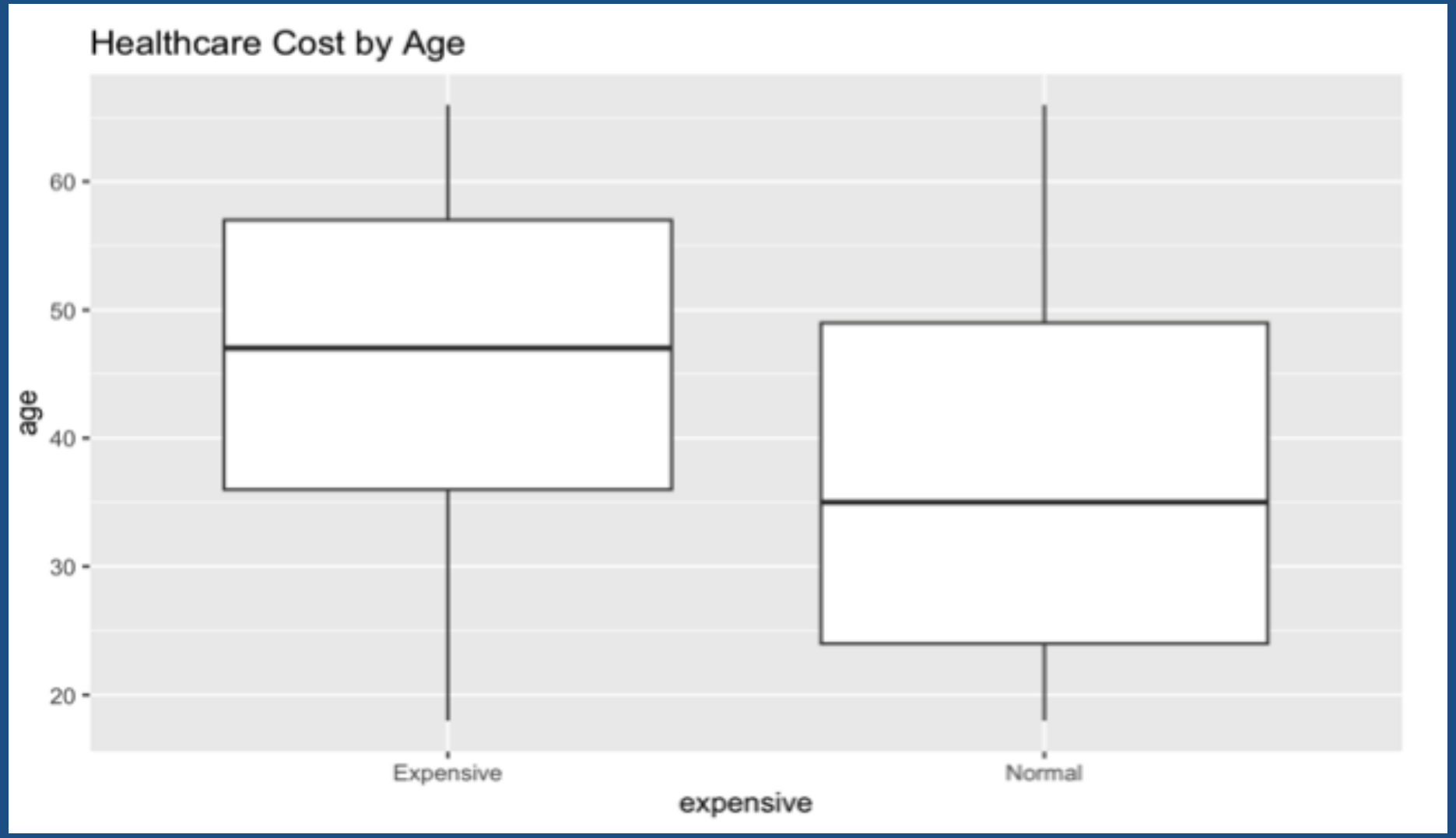
Tools/Software: R-Studio, R Shiny

Course Structure: Lecture class, Lab class, homework assignments, lab assignments, one exam, and final project

IST 687 – Final Project

- **Project Topic:** Analysis for a Health Maintenance Organization (HMO) on factors that lead to higher healthcare costs in the United States, and how HMO can adjust for these factors
- **Process**:
- Using R code, the data was loaded into R-Studio and data cleaning was performed
- Categorical variables indicating “Normal” and “Expensive” costs of healthcare were added
- Partitioned data into test and training set
- Created SVM and Rpart models
- Created Visualizations and Interactive Shiny App to help share insights
- Created PowerPoint Presentation to summarize all findings

Example Visualization



IST 687 - Final Project Takeaways

Recommendations to HMO included the following:

- Charge higher premiums to smokers
- Charge higher premiums to those who live in New York
- Charge higher premiums to older people

This project tapped into all the learning goals of the ADS program

- To do the analysis, it was imperative to properly collect, store, and access data as well as use R
- The ethics of data science was very important
- When dealing with different demographics of people, it's imperative to make sure the source of data collection is accurate, and all groups are represented proportionally.
- Otherwise, the data can be skewed and the insights from the machine learning analysis will be biased.

IST 722 – Data Warehousing

Term: Winter 2022

Programming Languages: SQL

Software/Tools: Microsoft SQL Server, Microsoft SQL Server Integration Services, Microsoft Analysis Services, Power BI, Excel

Course Structure: Four days of 8-hour long lectures, assignments, and Final Project

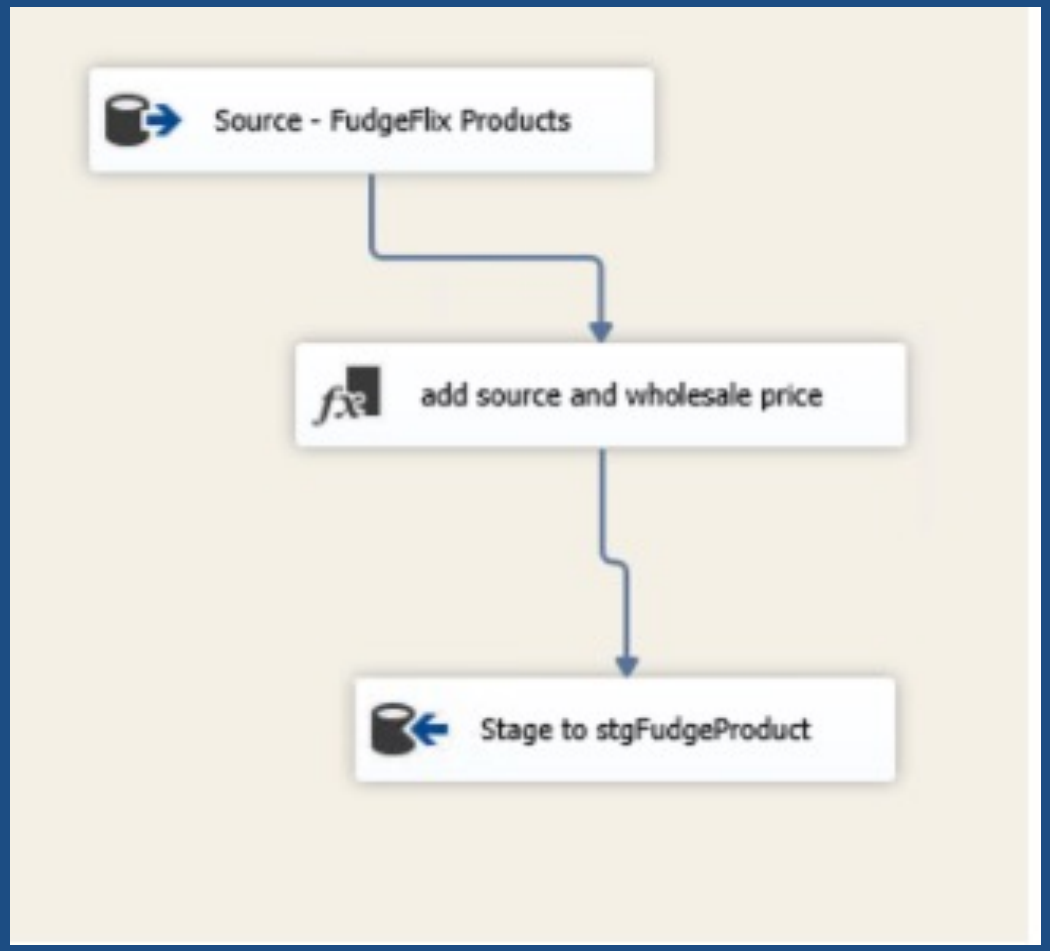
IST 722 – Final Project

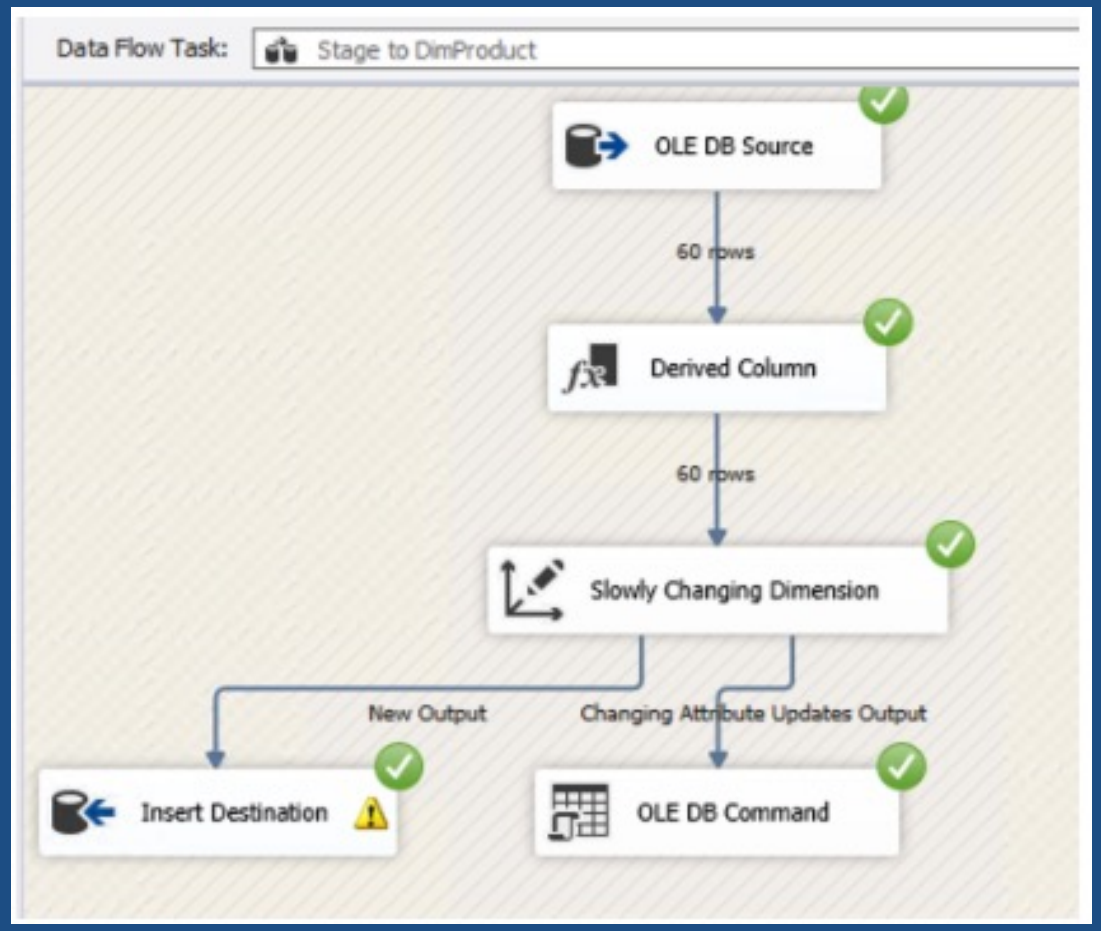
Task: Create Data Warehouse and Business Intelligence Analysis for the merger of two companies: FudgeMart and FudgeFlix

Process:

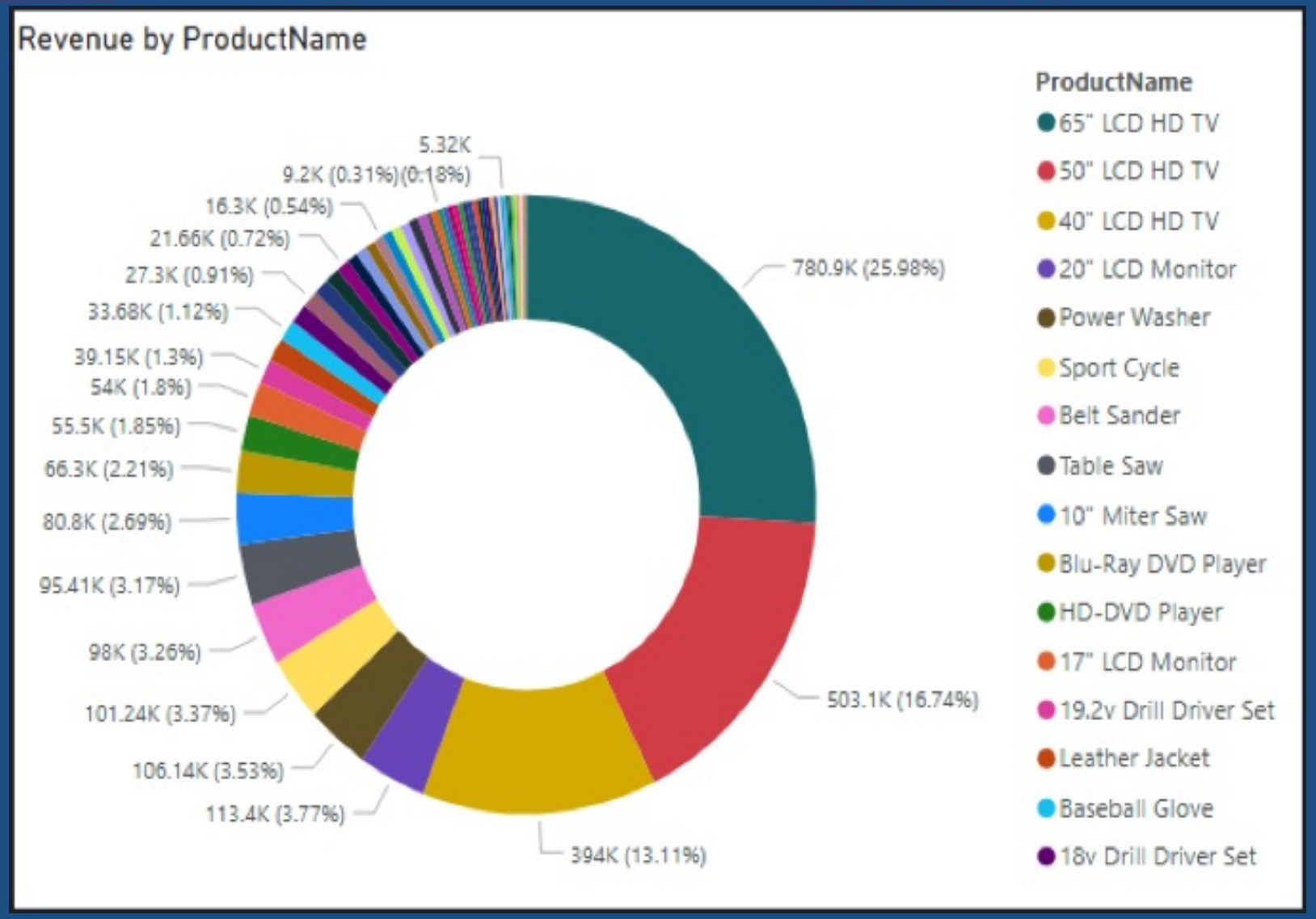
Define business process (Sales Analysis)	High-level dimensional modeling worksheet	Detailed-Dimensional modeling worksheet	Populate SQL Script using Macros to initialize tables and relationships	Retrieve data from source to stage tables, and then from stage tables to dimensional tables and fact table	Create ROLAP and MOLAP Cubes	Power BI analysis	Presentation

ETL Processes





Business Intelligence



IST 722 - Final Project Takeaways

Intensive course and final project emphasized all the key learning goals of the ADS program.

Excel, Microsoft SQL Server, Microsoft Analysis Server, and SQL programming language were used to collect, store, and access data.

Power BI aided in applying visualizations and creating actionable insights into the sales analysis of the merged company.

Presentation combined the use of these visualizations with examples from the ETL pipeline to communicate these insights to the audience in an easy-to-understand manner.

IST 769 - Advanced Big Data Management

Term: Spring 2023

Programming Language: Python, SQL, Spark

Tools/Software: Juptyer Notebook, Docker, Minio, Hadoop, Neo4j, MongoDB, Cassandra, Kafka, Bash

Course Structure: Lectures, Labs, Two Exams, and Final Project

IST 769 – Final Project

Task: Create Docker Environment for database of choosing (Apache Hbase)

Process:

Create Docker-Compose Yaml file	Include most-recent image of Apache Hbase	Include images of Hbase's dependencies (Hadoop name nodes, data nodes, history server)	Add image of Apache Drill to connect to database with SQL commands	Add image of Jupyter Notebook with PySpark to connect via code

IST 769 – Final Project Takeaways

- This project was different from most other projects within the program
  - Doesn't explicitly touch on all the key goals
  - Still provided a challenging avenue to highlight some of the key goals with a higher intensity than others
- Project required a lot of research, trial, and error to collect, store, and access data
- Creating a docker file and virtual environment from scratch was thought provoking, and having no prior knowledge of HBase made the feat of storing and accessing the data complex
- SQL and Python were used heavily within this project to connect, create, modify, and query to and from data tables
- The presentation served as an avenue to give insight into what a software, foreign to the audience, was and how it could be properly used

IST 707 – Applied Machine Learning

Term: Spring 2023

Programming Languages: R and Python

Tools/Software: Juptyer Notebook & R-Studio

Course Structure: Lectures, Assignments, and Final Project

IST 707 – Final Project

Task: Conduct a Comprehensive Analysis on Multi-Million Row Dataset on MLB Pitches

Questions to Answer:

Which pitches are thrown most frequently in various game scenarios?

What physical properties of pitches lead to the most optimal outcome?

What characteristics of fastballs and sliders make them more, or less, successful?

How can teams better prepare for matchups with unfamiliar opponents?

Use-Case: Provide actionable insights for MLB teams and players

Process:

Download and Clean Dataset

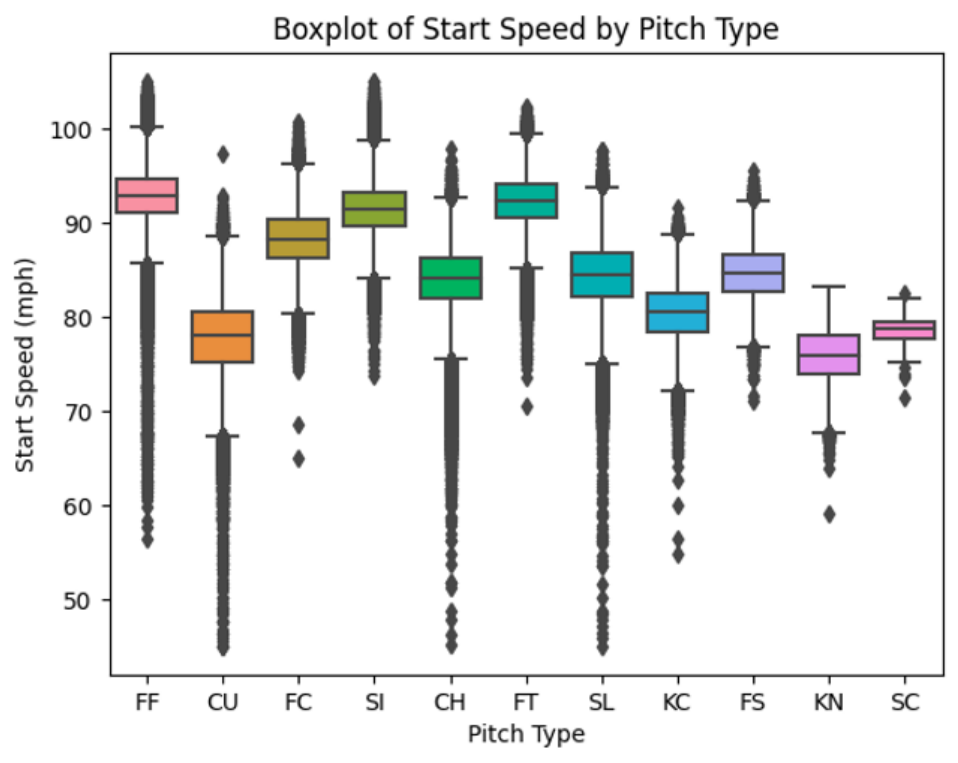
Exploratory Data Analysis

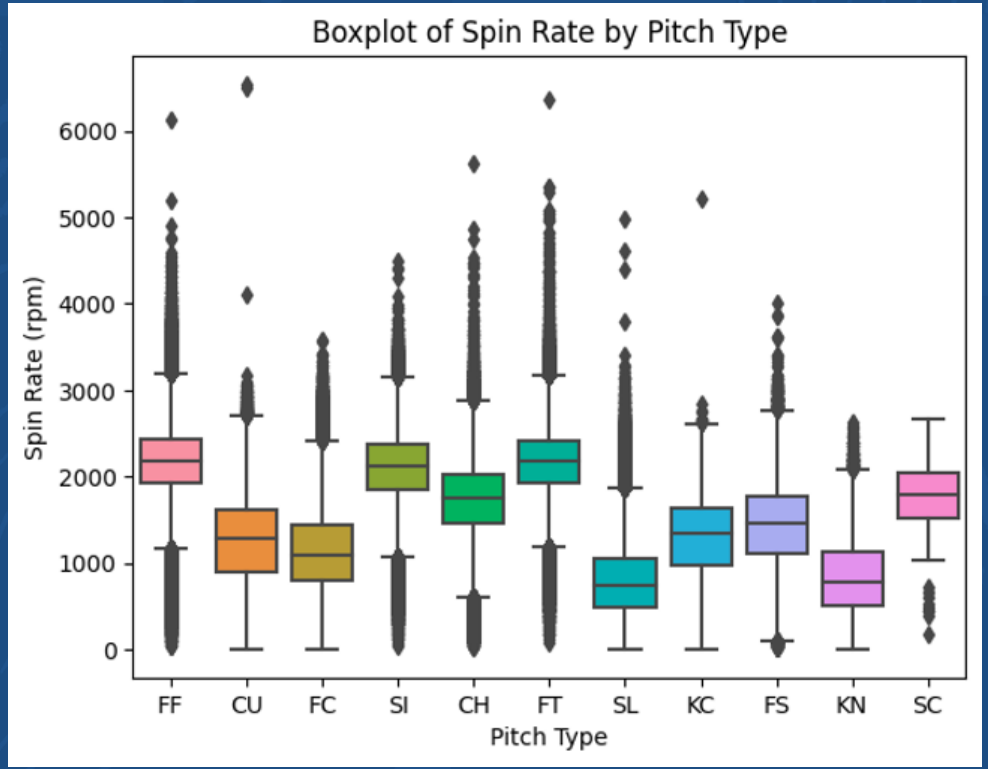
Modeling

K-Means Clustering and Association Rule Mining

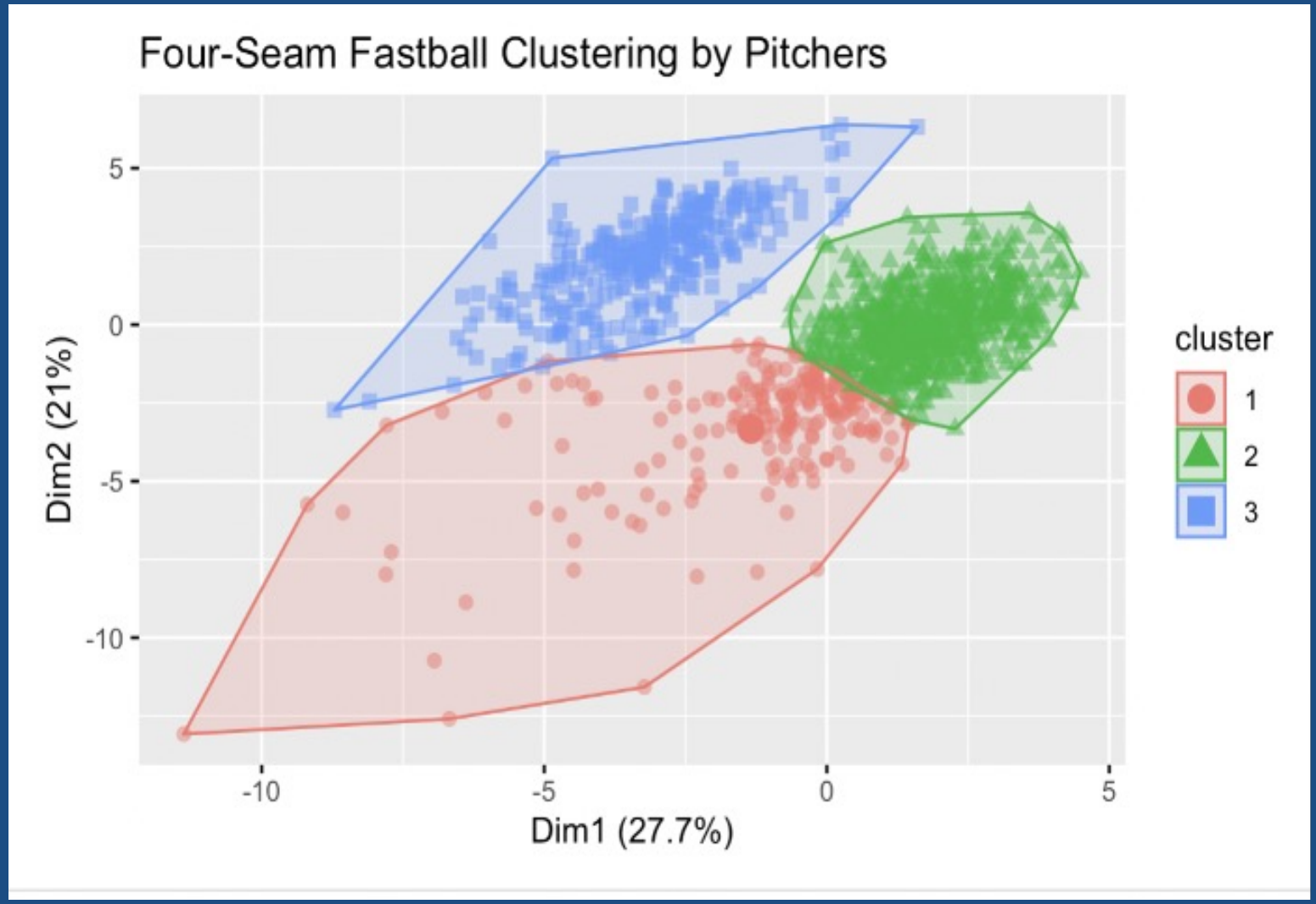
Presentation and Written Paper

Exploratory Data Analysis





Clustering Algorithm Output



IST 707 – Final Project Takeaways

- From the Association Rules Mining
  - Were able to identify specific game scenarios where pitchers were most likely to throw a specific pitch type
  - Uncovered statistical relationships between pitch metrics and success rates
- From the k-Means Clustering
  - Were able to help teams better prepare for unfamiliar opponents by clustering pitchers into skill buckets
- Collecting, storing, and accessing data was of the utmost importance throughout this project
  - Largest dataset used throughout the ADS program
- Managing, editing, and joining multiple data files with millions of rows within two distinct code environments (R Studio and Jupyter Notebook) demonstrated mastery of the skill
- Histograms, box plots, and cluster plots were generated in both R and Python to create actionable insights and supplement the insights shown in raw form by the algorithm outputs
- Written report and PowerPoint presentation were necessary in translating a complex project and subject matter to an audience largely unfamiliar with the sport of baseball
- Paying deep attention to the datasets to ensure there was no bias was very important, given the vast differences in observations per pitch

Conclusion

- Portfolio demonstrated thorough understanding and completion of six learning goals of ADS program.
- Proficiency in data collection, storing, and access across Python, R, and SQL
- Creation of actionable insights across a range of contexts was evidenced
- Mastery of applying visualizations and machine learning models was shown
- Fluency in R, SQL, and Python was demonstrated through three SQL-based projects (including a SQL based concentration), two Python-based projects, and two R-based projects.
- Insights were effectively communicated to a broad range of audiences, with one or more written reports, demos, and presentations completed for each project.
- The application of ethics in the development, use, and evaluation of predictive models and data was applied to each project

Thank You