# Predicting Learning Commons Usage: Duration and Occupancy

## A Non-Linear Approach

Naiyue Liang (Emma), Ryan Renken, Jaryt Salvo, & Jason Turk

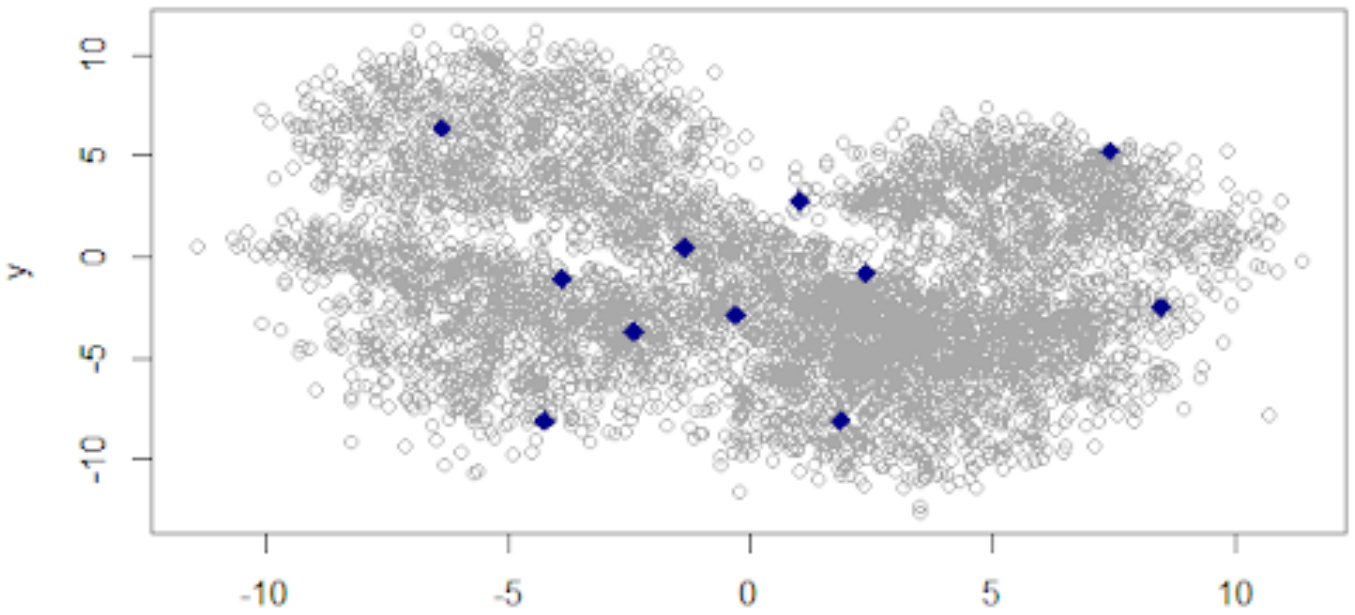
MATH 7560 Statistical Learning II || BGSU

April 24, 2025

# K-means: Data & Motivation

## Clustering Task

- **Goal**: Partition $\mathbb{R}^2$ data into $k = 11$ clusters.
- **Comparison**: Standard K-means vs K-means++ initialization.
- **K-means++ Motivation**: Better initial centroids for potentially faster/better clustering.



# K-means vs K-means++ Results

**Standard K-means**:

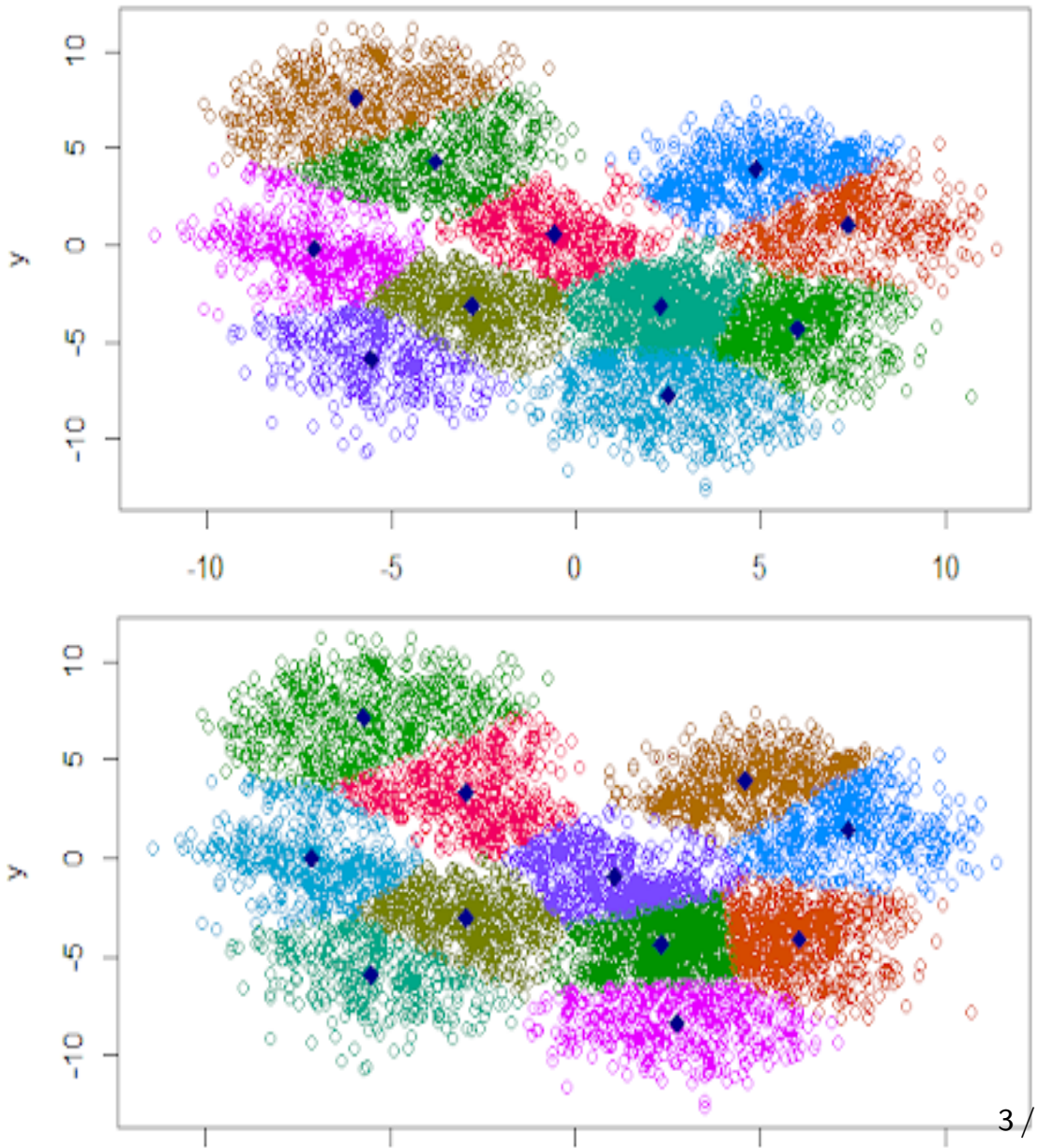
- **Stats**: 5 iterations, WCSS = 22,824.

**K-means++ Initialization**:

- **Stats**: 8 iterations, WCSS = 22,943.

**Observation**

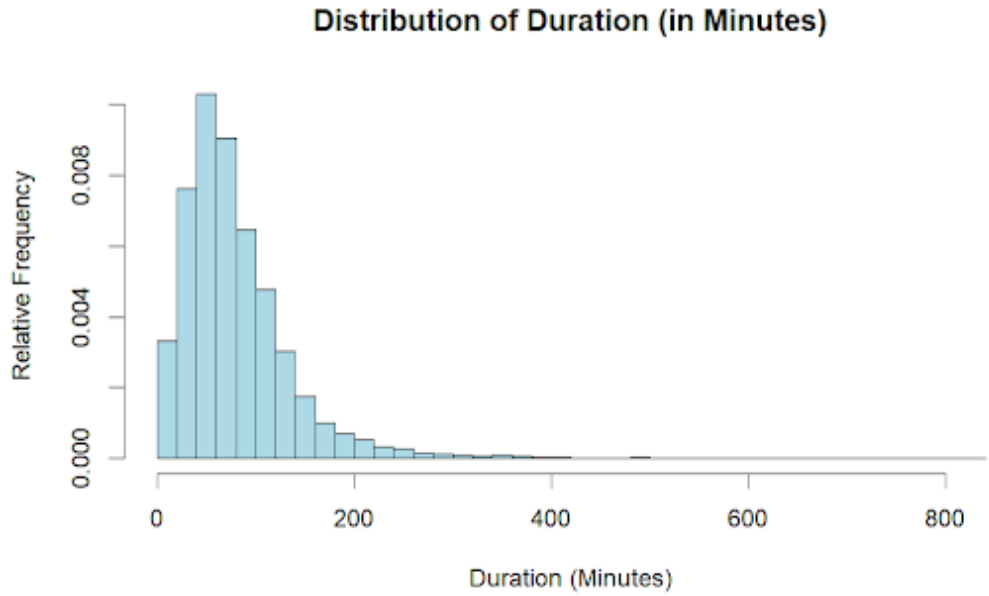
K-means++ offered no clear advantage over standard K-means for this dataset (visually or by WCSS).



# Distribution Statistics

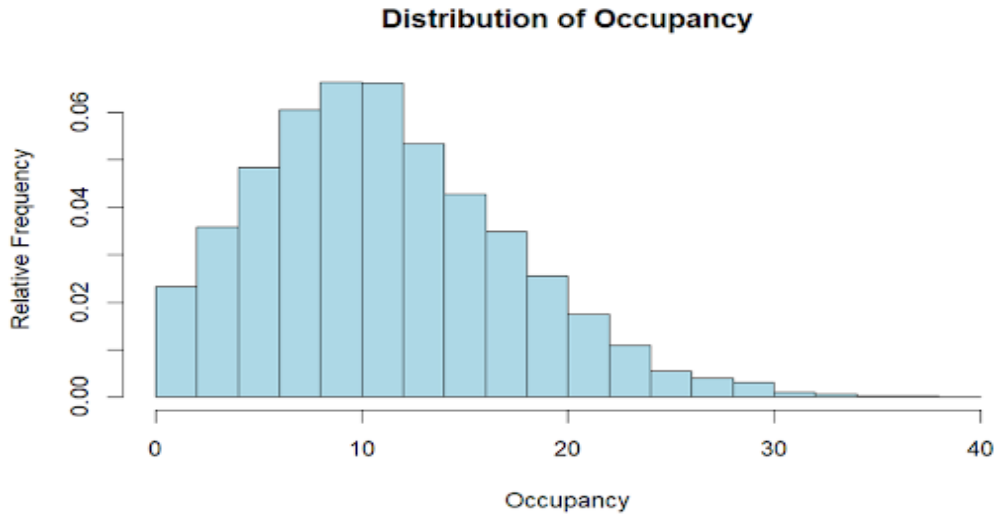
**Duration (minutes)**:

| Statistic | Value |
|---|---|
| Minimum | 6.00 |
| 1st Quartile | 44.00 |
| Median | 68.00 |
| Mean | 81.78 |
| 3rd Quartile | 103.00 |
| Maximum | 822.00 |




**Occupancy (students)**:

| Statistic | Value |
|---|---|
| Minimum | 1.00 |
| 1st Quartile | 7.00 |
| Median | 11.00 |
| Mean | 11.62 |
| 3rd Quartile | 15.00 |
| Maximum | 40.00 |



# Feature Categories Overview

| Category | Key Features |
|---|---|
| Temporal | Time of day, Day of week, Week of semester |
| Academic | Course level, GPA categories, Credit load |
| Visit | Duration patterns, Group sizes, Visit frequency |
| Course | Subject areas, Level progression, Course mix |
| Student | Major groups, Class standing, Academic progress |

| External Source | Key Features |
|---|---|
| R library 'lunar' | Moon phase data |
| R library 'openmeteo' | Hourly weather metrics (temperature, humidity, pressure, cloud cover, wind, radiation, precipitation, & soil conditions) |

# Dropped Raw Features

| Raw Feature | Engineered Into |
|---|---|
| Student_IDs | Total_Visits, Semester_Visits, Avg_Weekly_Visits |
| Class_Standing | Class_Standing_Self_Reported, Class_Standing_BGSU |
| Major | Major_Category, Has_Multiple_Majors |
| Expected_Graduation | Expected_Graduation_Date, Months_Until_Graduation |
| Course_Name | Course_Name_Category |
| Course_Number | Unique_Courses, Course_Level_Mix |
| Course_Type | Course_Type_Category |
| Course_Code_by_Thousands | Course_Level, Advanced_Course_Ratio |

**Feature Engineering Strategy**

Raw features were transformed into more informative derived features, capturing higher-level patterns and relationships in the data.

# Model Hyperparameter Tuning Ranges

## Duration Task Models

| Model | Hyperparameters |
|---|---|
| MARS | **num_terms**: [7, 15]<br>**prod_degree**: 1 |
| Random Forest | **trees**: [300, 325]<br>**min_n**: [15, 25]<br>**mtry**: [20, 25] |
| XGBoost | **trees**: [75, 100]<br>**tree_depth**: [15, 21]<br>**learn_rate**: 0.05<br>**min_n**: [10, 15]<br>**mtry**: [12, 15] |

## Occupancy Task Models

| Model | Hyperparameters |
|---|---|
| MARS | **num_terms**: [120, 130]<br>**prod_degree**: 1 |
| Random Forest | **trees**: [250, 350]<br>**min_n**: [2, 3]<br>**mtry**: [40, 45] |
| XGBoost | **trees**: [350, 450]<br>**tree_depth**: [6, 8]<br>**learn_rate**: 0.1<br>**min_n**: [2, 3]<br>**mtry**: [30, 35] |

# Model Pipeline Configurations

| CV Method | Description |
|---|---|
| kfold | Random k splits |
| rolling | Fixed-size window moving forward |
| expanding | Growing window with fixed start point |

| Pipeline | Implementation |
|---|---|
| vanilla | Scaling → Model |
| interact_select | Scaling → Interactions → SelectKBest → Model |
| pca_lda | Scaling → PCA/LDA → Interactions → SelectKBest → Model |

## Scaling Methods

- **StandardScaler**: $(x - \mu)/\sigma$ - sensitive to outliers
- **RobustScaler**: $(x - Q_2)/(Q_3 - Q_1)$ - resistant to outliers
- **MinMaxScaler**: $(x - x_{min})/(x_{max} - x_{min})$ - preserves zeros

# Best Model Configurations

**Duration Prediction**:

| Component | Value |
|---|---|
| Model | PenalizedSplines |
| Pipeline | vanilla |
| CV Method | kfold |
| RMSE | 59.47 |
| $R^2$ | 0.059 |
| Ridge $\alpha$ | 14.38 |
| Spline degree | 3 |
| Spline knots | 15 |
| Scaler | RobustScaler |

**Occupancy Prediction**:

| Component | Value |
|---|---|
| Model | PenalizedSplines |
| Pipeline | vanilla |
| CV Method | rolling |
| RMSE | 3.64 |
| $R^2$ | 0.303 |
| Ridge $\alpha$ | 29.76 |
| Spline degree | 3 |
| Spline knots | 15 |
| Scaler | RobustScaler |

**Key Insight**

Both tasks achieved best results with PenalizedSplines and vanilla features, though with different CV methods & regularization.

# Duration: Best Model Diagnostics

# Occupancy: Best Model Diagnostics

# Sanity Check

# Key Findings

**Main Results**:

- **PenalizedSplines** with *vanilla* features performed best
- **Occupancy** prediction shows promise ($R^2 = 0.303$)
- **Duration** prediction remains challenging ($R^2 = 0.059$)

**Future Directions**:

- Incorporate *weather* data
- Explore *non-linear* relationships further
- Investigate *time series* approaches

**Impact**

While duration prediction remains difficult, our occupancy model shows strong potential for a victory **#CautiousOptimism**

# *Thank You*

*For Your Attention*

---

Questions & Discussion Welcome