Understanding and Benefitting from Yellowdig Partnerships

Drexel University LeBow College of Business

Ai Vi Truong, Kate Alden, Sean Li, Luqing Qi

August 23, 2021

Agenda

- Goals & Objectives
- Data
- Analysis
  - Predicting Community Health
    - Multiple Regression
    - Decision Tree
  - Community Engagement
    - Cluster Analysis
  - Segmenting Customers
    - Cluster Analysis
- Summary & Recommendations

Goals & Objectives

Objectives and Goals

- Develop classification models to:
  - Identify key decision points that lead to optimal community engagement
  - Segment partners based on different important aspects including how client relationships are organized and projects are managed
  - Measure current clients community engagement
- Develop predictive models to:
  - Determine whether a class is “healthy” or not in terms of engagement with subject matter
  - Identify key decision points that lead to optimal community engagement

DATA

Data Sets

Yellowdig Datasets

- Closed Won Since 01-01-201
- Point Settings
- Current Clients Report
- Total Pipeline Database
- Community Health All Time

Data: Clients

Data Sets: Closed Won, Current Clients Report, and Total Pipeline Data Set

- Can be used to understand which aspects of each account contribute to length of each contract
- Overall revenue provided to Yellowdig
- Relationship between client and salesperson
- Relationship between Yellowdig and its partners

Closed Won	• 174 observations • 15 variables • Quality: Missing Values, Duplicates Variables, and Outliers
Current Clients Report	• 100 observations • 9 variables • Quality: Missing Values, Redundant Varialbes
Total Pipeline	• 111 observations • 17 variables • Quality: Missing Values

Closed Won Data

- Pivot Table Based on Closed Won Dataset
  - Tyler handles majority of Account Types, and is responsible for most Client Expansion-type contracts
  - *New Logo – New* has the largest share of Account Types with 78 observations, followed by Client Expansion, with 57

	Type Count
Bob Ertischek	**6**
Client Expansion	1
New Logo – New	5
Gerry Meyle	**2**
New Logo – New	2
Jim Gandolfo	**26**
Client Expansion	7
New Logo – Existing	1
New Logo – New	18
Kailie Starr	**18**
Client Expansion	5
Client Renewal	1
New Logo – New	12
Randy Sealy	**20**
Client Expansion	4
Client Renewal	4
New Logo – New	12
Ryan Nemetz	**4**
New Logo – New	4
Steve Davis	**1**
New Logo – New	1
Tyler Rohrbaugh	**97**
Client Expansion	40
Client Renewal	27
Existing Business	1
New Logo – Existing	2
New Logo – New	24
One Time Fee (OTF)	3

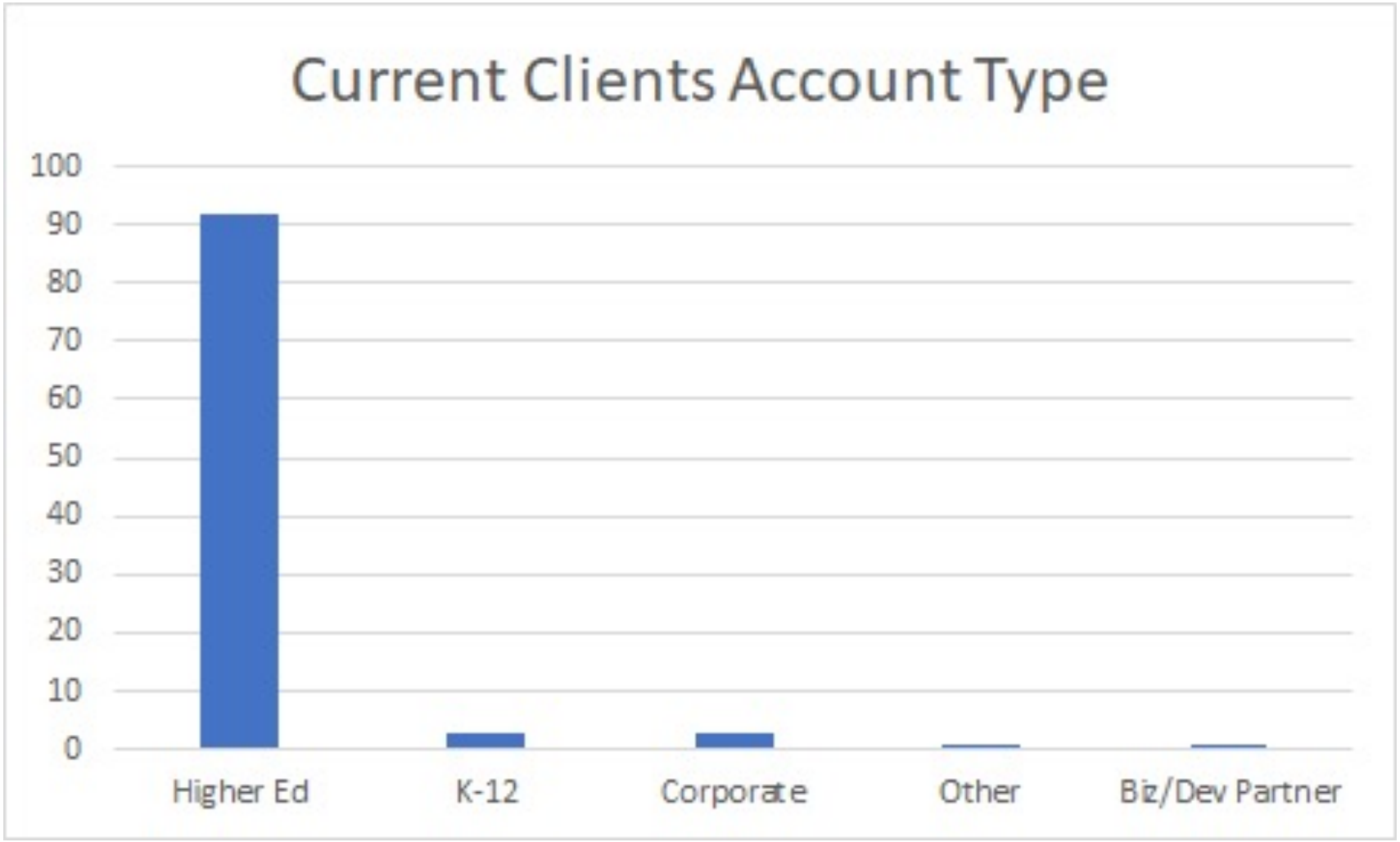
Current Clients Data

- Current Clients are mostly in Northeast and Mid-West of the U.S

DREXEL UNIVERSITY
LeBow
College of Business

Current Clients Data

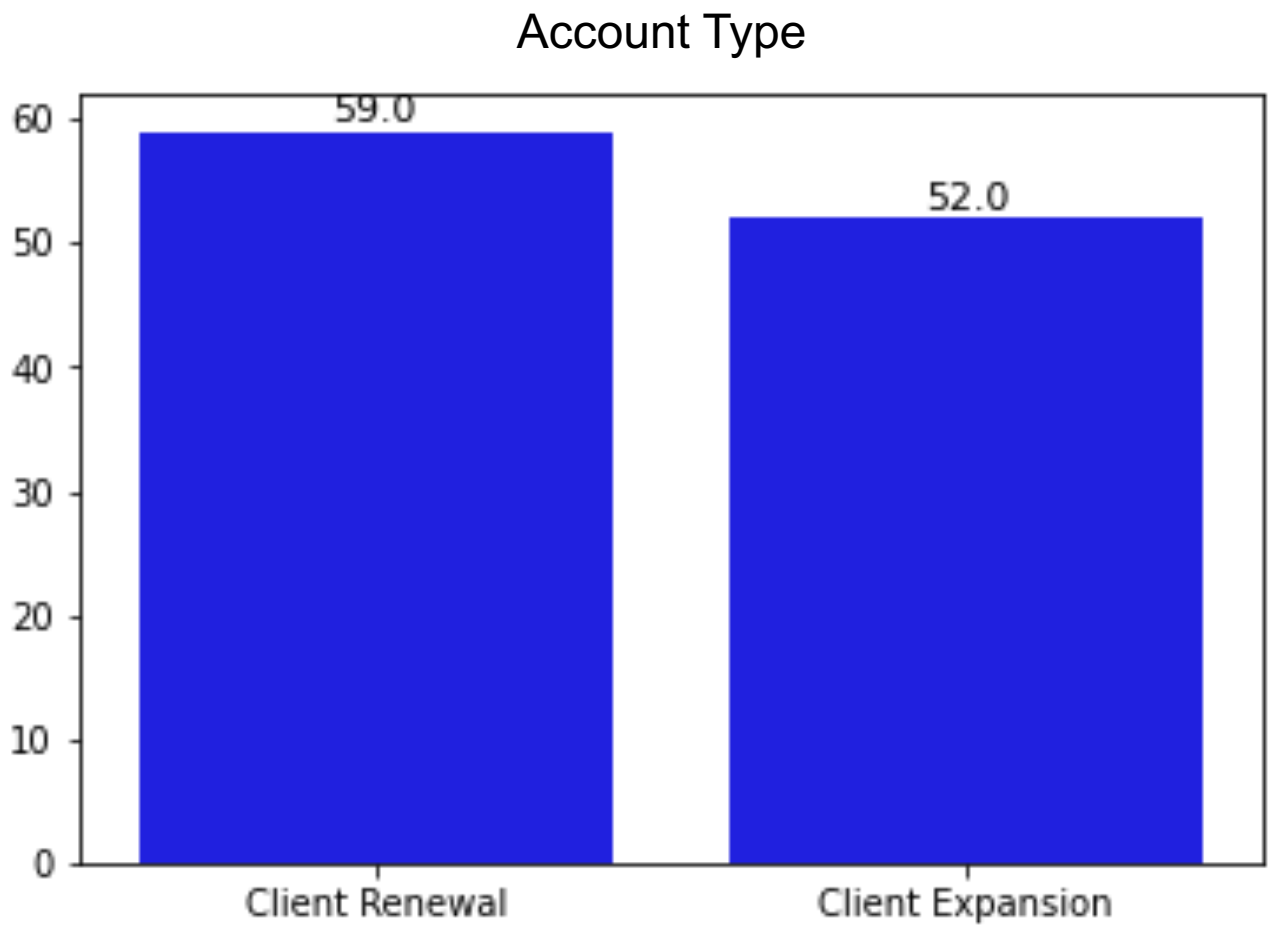
- 92% of the current clients account type is High Education
- 8% is the other account types including K-12, Corporate



DREXEL UNIVERSITY
LeBow
College of Business

Total Pipeline Report

- Based on Total Pipeline Report (Sales Prediction)
  - 47% of current clients most likely to expand their contracts
  - 53% of current clients are mostly likely to renew their contracts



DREXEL UNIVERSITY
LeBow
College of Business

Community Data

- **Data Sets: Point Settings and Community Health**
- Can be used to understand the community engagement based on points/score
- Study how to leverage data and study community engagement

Point Settings

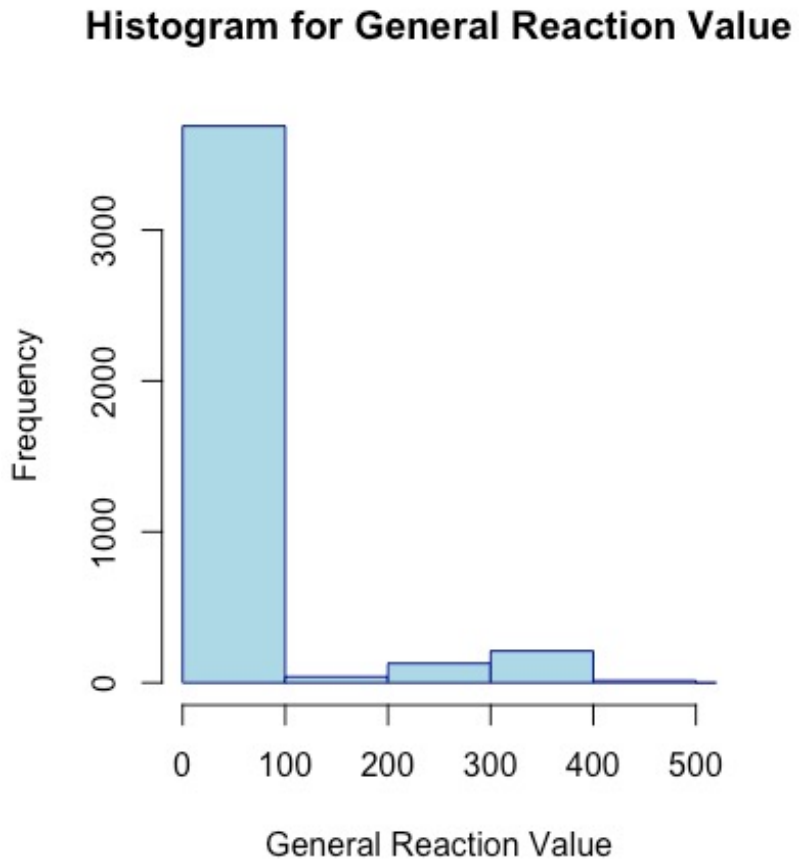
- 4759 observations
- 17 variables
- Quality: Missing Values, Duplicates Variables, and Outliers

Community Health

- 4435 observations
- 45 variables
- Quality: Missing Values, Redundant Variables, Date Format

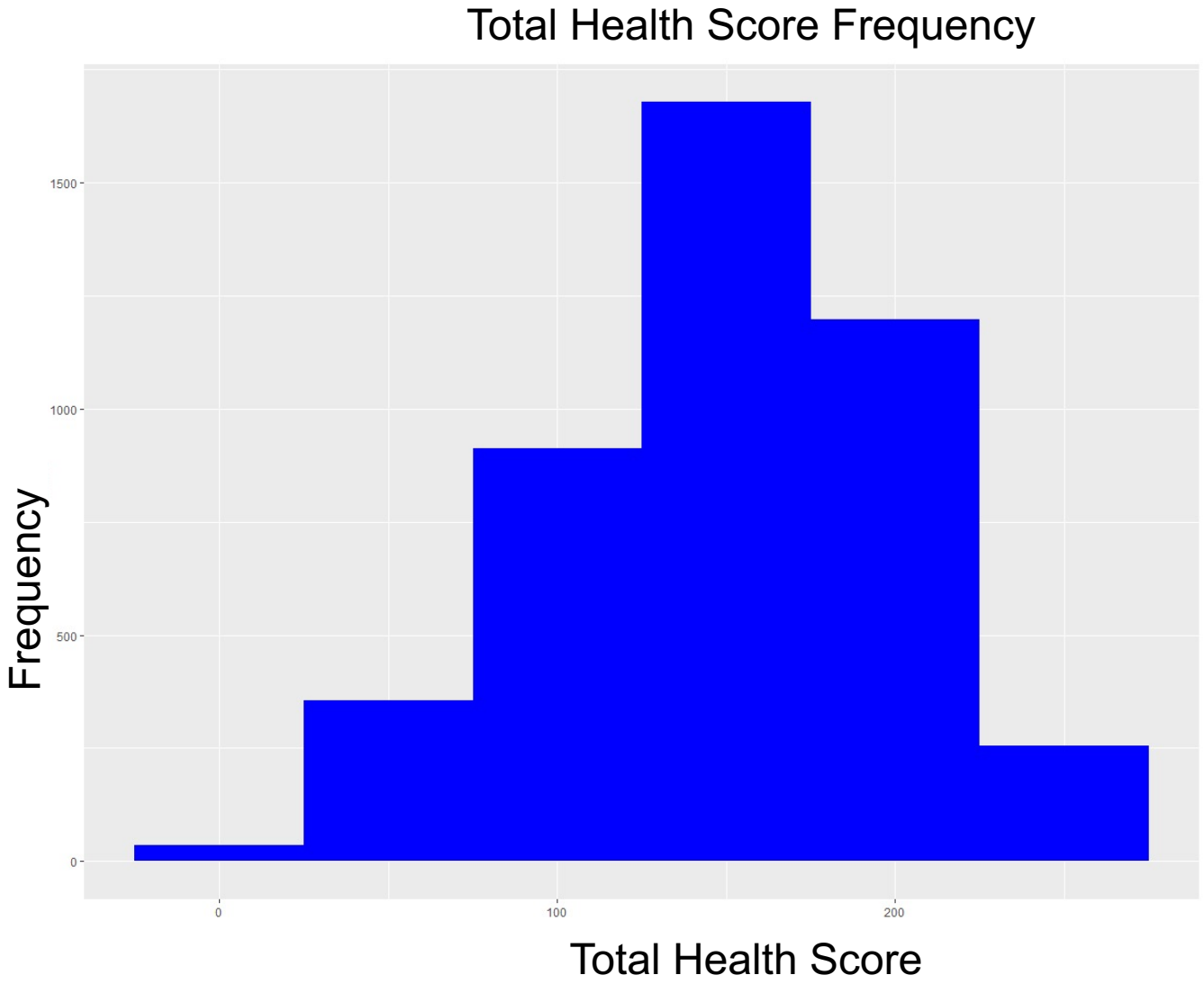
Point Settings Data

- Most General Reaction Values are in the 0-100 range



Community Health Data

- Most of the board achieved the total health score range from 140-160

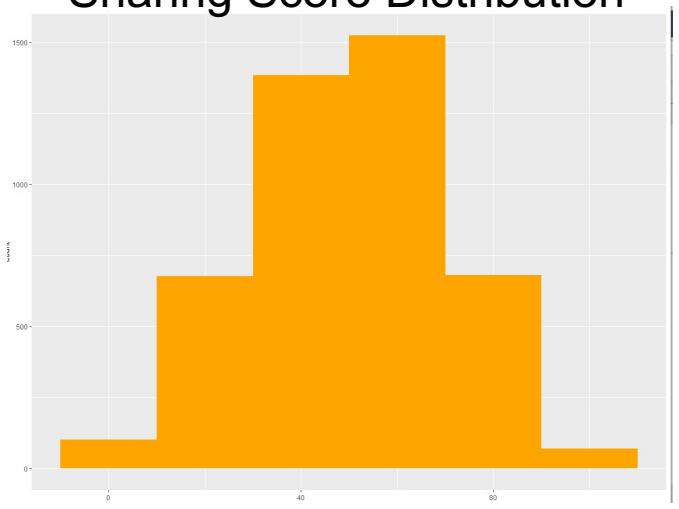


DREXEL UNIVERSITY
LeBow
College of Business

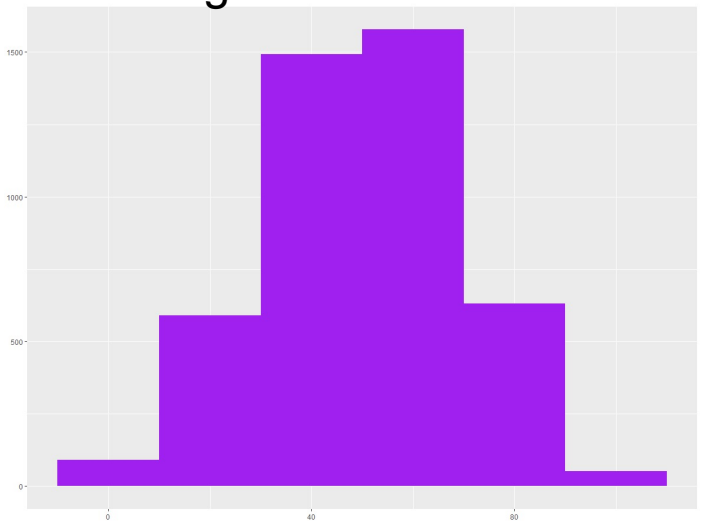
Community Health Data

- Total Health Score is a sum of sharing, listening, and interacting scores
  - All three variables have similar distribution

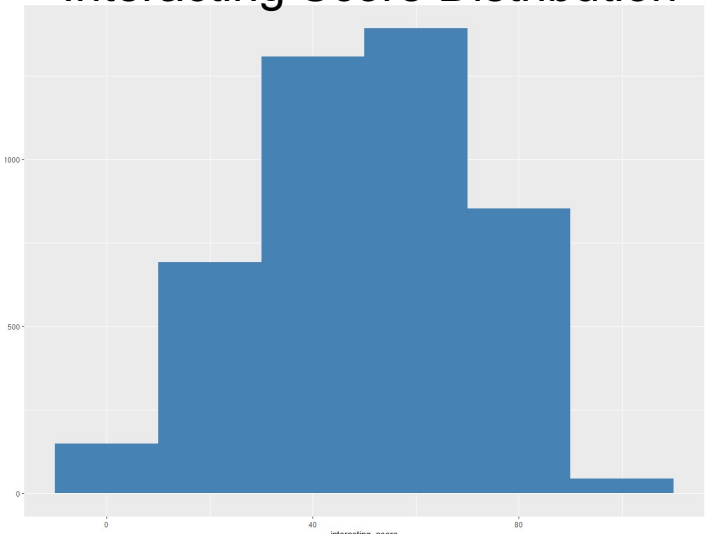
Sharing Score Distribution



Listening Score Distribution



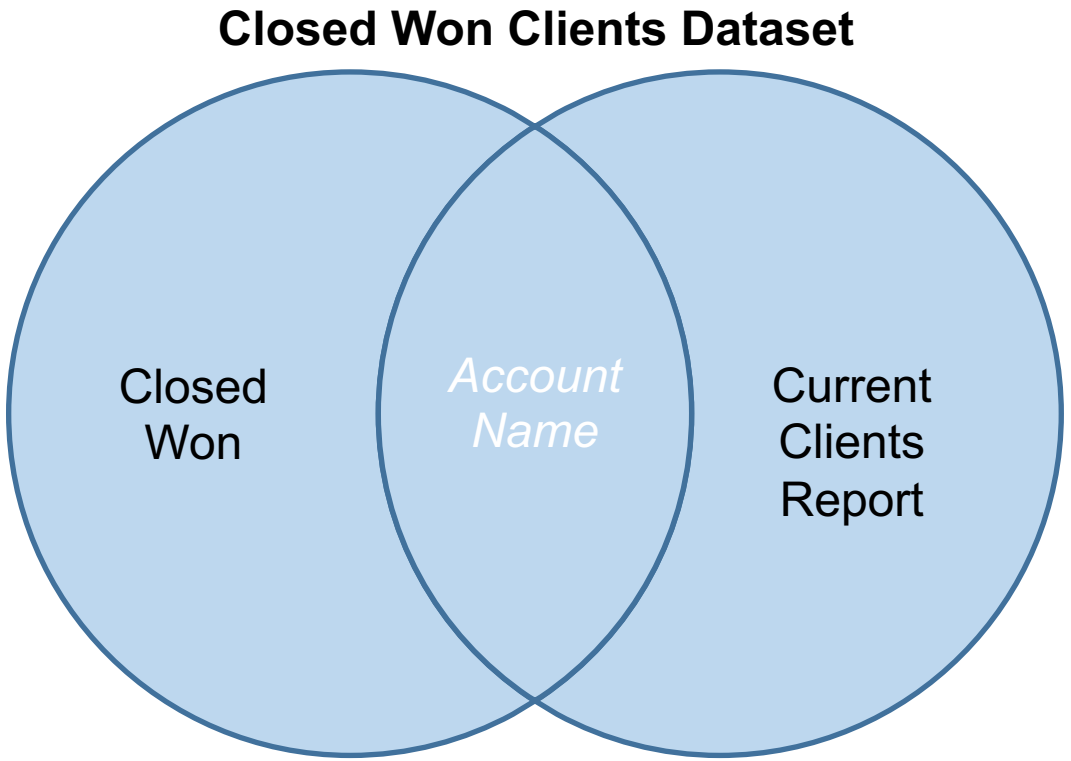
Interacting Score Distribution



DREXEL UNIVERSITY
LeBow
College of Business

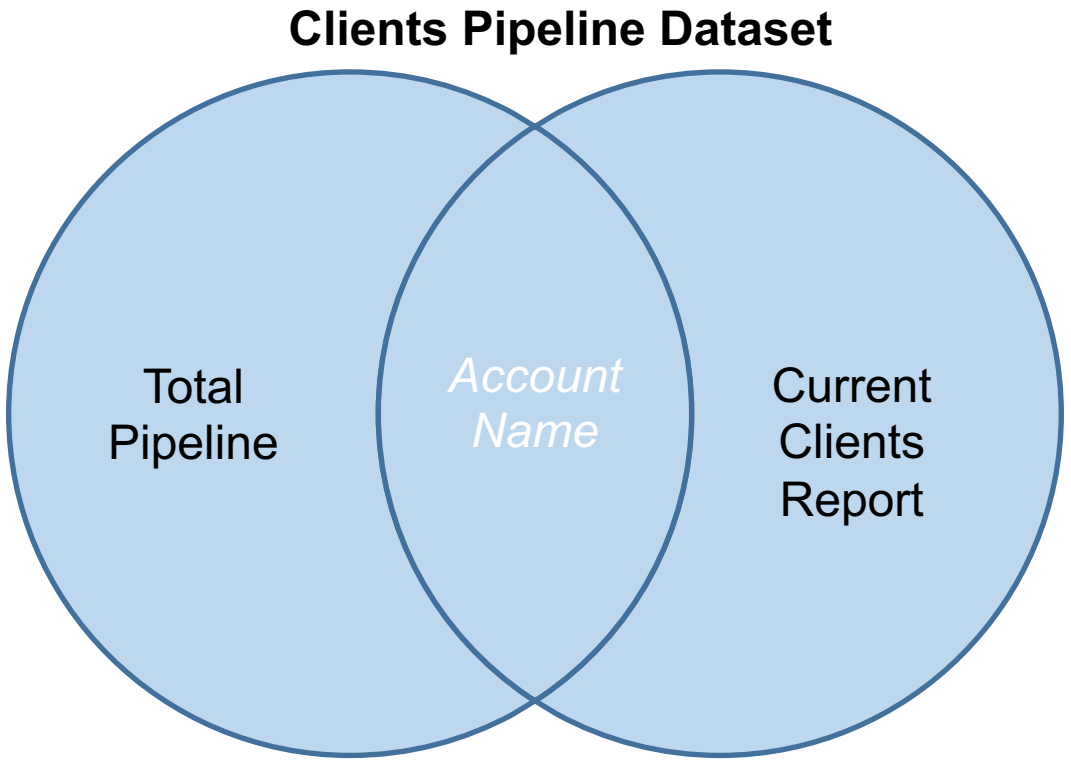
Closed Won Clients Dataset

- **Combined two datasets: Closed Won and Current Clients Report**
- Full join on **Closed Won and Current Clients Report** based on Account Name



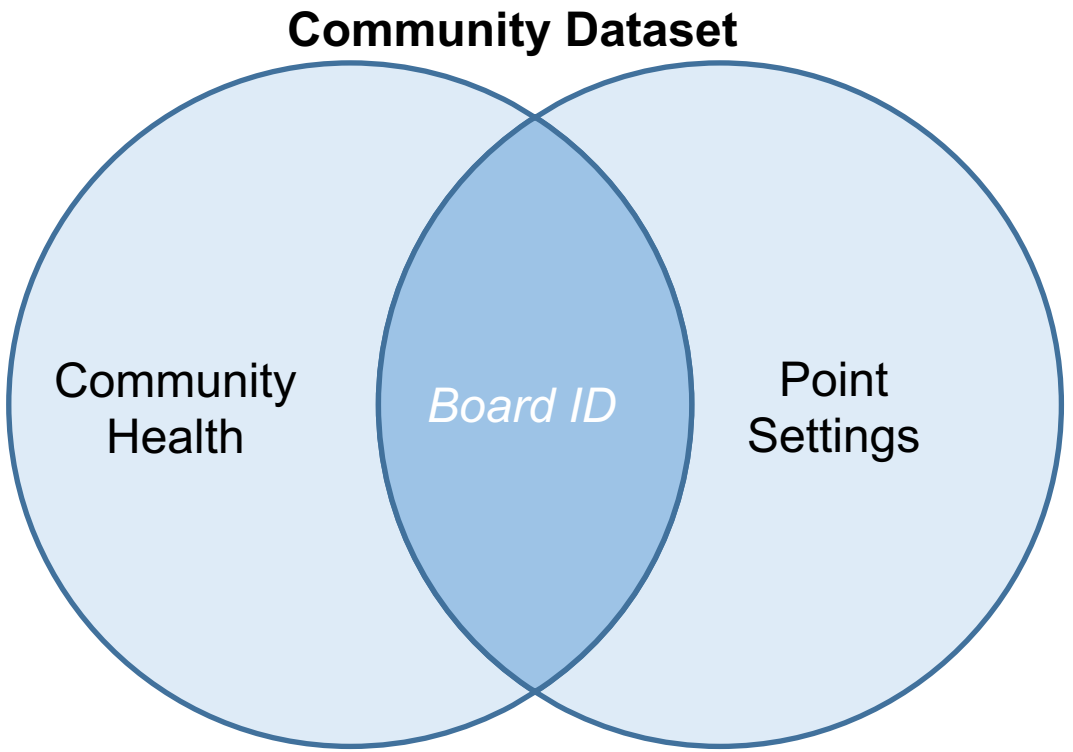
Clients Pipeline Dataset

- **Combined two datasets: Total Pipeline and Current Clients Report**
- Full join on Total Pipeline and Current Clients Report based on Account Name



Community Dataset

- **Combined two datasets: Point Settings and Community Health**
- Inner join on Community Health and Point Setting based on Board ID
- Understanding intersections between points assigned to actions and each community's health



Analysis

Objectives and Plans for Analysis

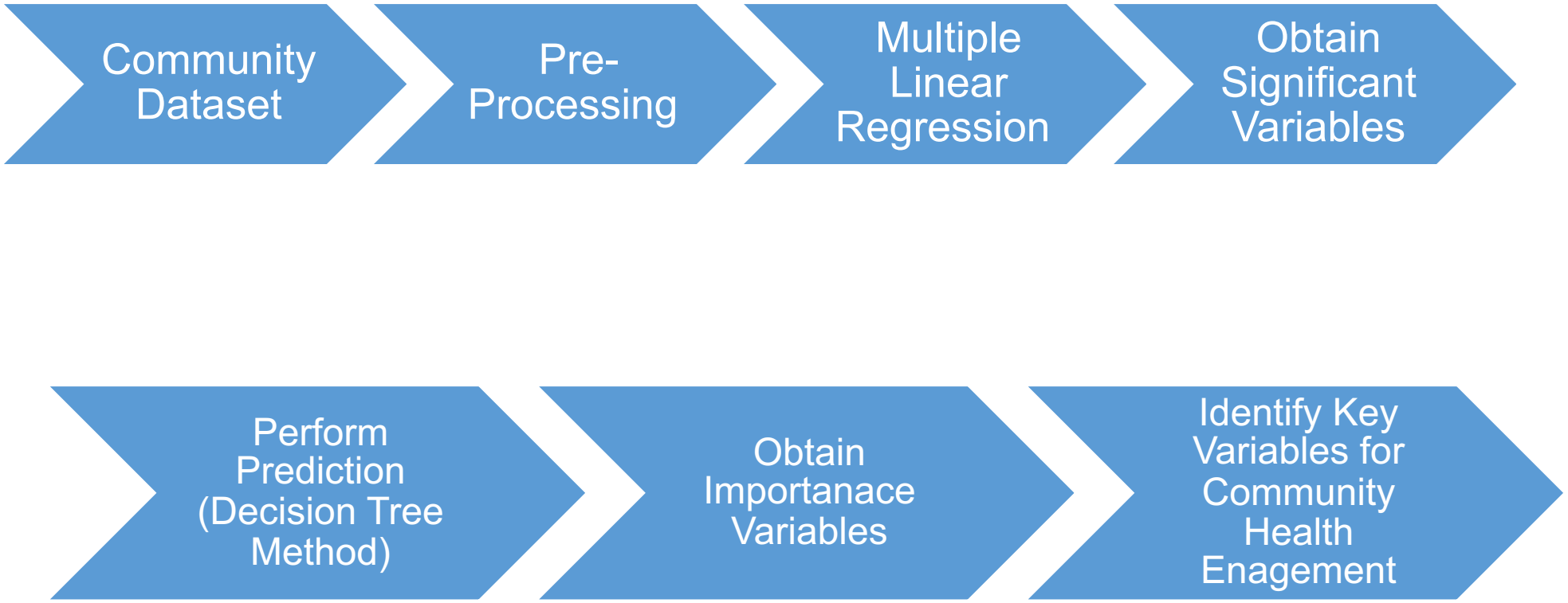
- Predicting Community Health
  - Use Multiple Linear Regression to find significant variables for predicting success in total health score
  - Construct a Decision Tree to understand thresholds for a “healthy” engagement level in a classroom
- Segmenting Clients
  - Use Cluster Analysis to determine how is the current community engagement for Yellowdig’s current clients

Predicting Community Health

Multiple Linear Regression & Decision Trees

DREXEL UNIVERSITY
LeBow
College of Business

Predicting Community Health Process

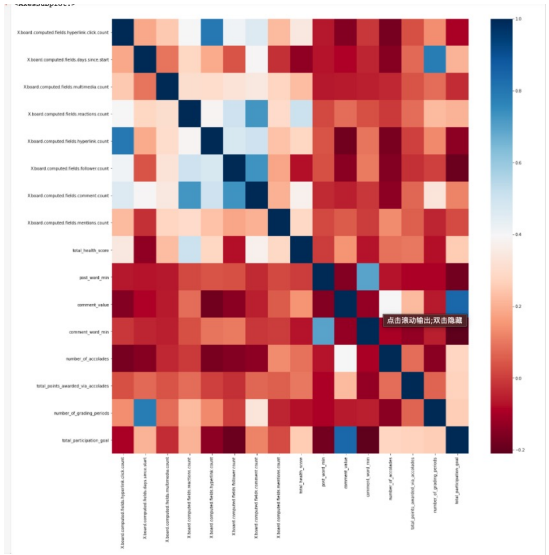


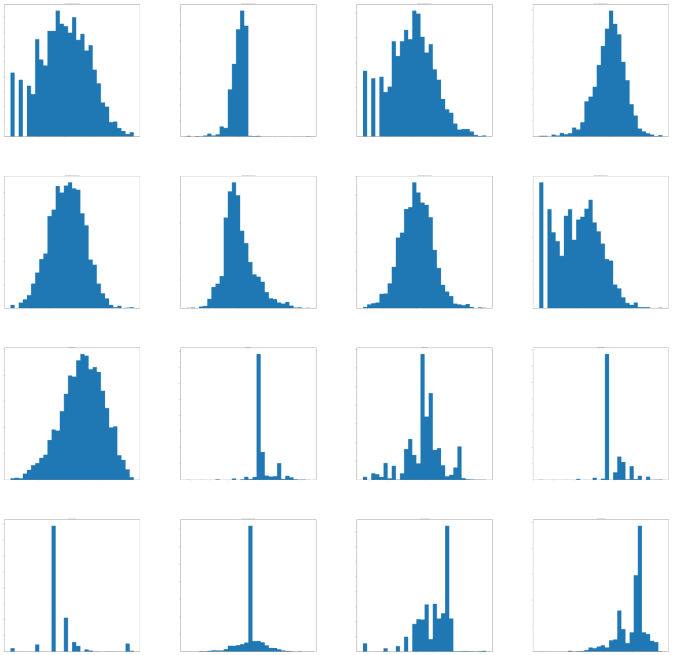
Multiple Linear Regression

- **Data**: Community Dataset
- **Objective**: Determine significant variables in contributing to Total Health Score values
- **Target Variable**: Total Health Score

Multiple Linear Regression

- **Data Preprocessing Steps:**
  - VIF to remove highly correlated variables
  - Imputed for NA values using median imputation
  - Data standardization & log transformation
  - Performed feature selection using stepwise selection
  - Remove variables that have 3 standard deviations from mean
  - 20% data in testing, 80% in training





Significance Results

- Significant predictors of Total Health Score have p-values less than 0.05
- Log Transformation

Variable	Coefficient	P-Value
Days Since Start	-0.7904	0.000
Multimedia Count	0.1185	0.000
Hyperlink Count	0.2526	0.000
Follower Count	-1.2813	0.000
Comment Count	1.1905	0.000
Mentions Count	0.0497	0.000
Reactions Count	0.3508	0.000
Comment Value	-0.0787	0.000
Comment Word Minimum	0.0479	0.000
Number of Grading Periods	-0.0484	0.000
Total Participation Goal	0.0988	0.000

R-square	Adjusted R-Square
0.922	0.922

MSE	0.0743
MAE	0.211
RMSE	0.273

DREXEL UNIVERSITY
LeBow
College of Business

Multiple Linear Regression

- **Recommendations:**
  - The Professors who use Yellowdig should put more hyperlinks and multimedia resources in the community.
  - The Professors should pay more attention to students' reactions.

Predicting Community Health

Decision Tree

DREXEL UNIVERSITY
LeBow
College of Business

Decision Tree

- **Data**: Community Dataset
- **Objective:** To classify users as highly engaged or not so, determining significant variables that contribute to high engagement
- **Target Variable**: Engagement Level
- 11 Predictor variables included in analysis
  - Significant variables result from multiple regression model
  - Board computed values
- Train/Test Split: 80/20

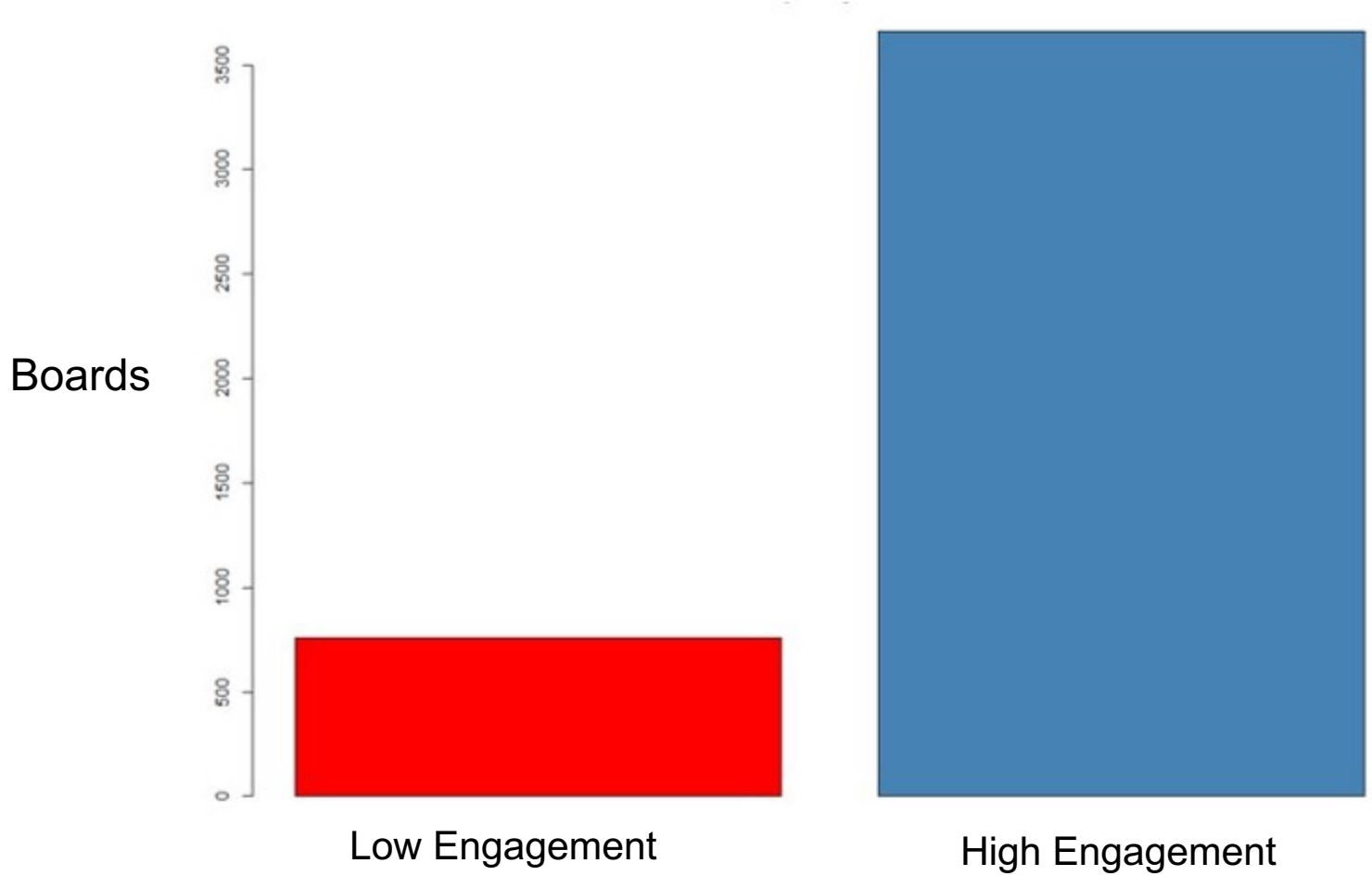
Engagement Level

- Focus Variable: Total Health Score
- Binarized Total Health Score variable, based on statistics
  - Assign 0 for any Total Health Score < 100
  - Assign 1 for any Total Health Score > 100
- 0 – Low in Engagement
- 1 – High in Engagement

	Total Health Score
Mean	149.95
Standard Deviation	50.06
Minimum	0
25%	117.53
50%	153.26
75%	186.14
Max	271.75

Engagement Level

Boards By Engagement Level



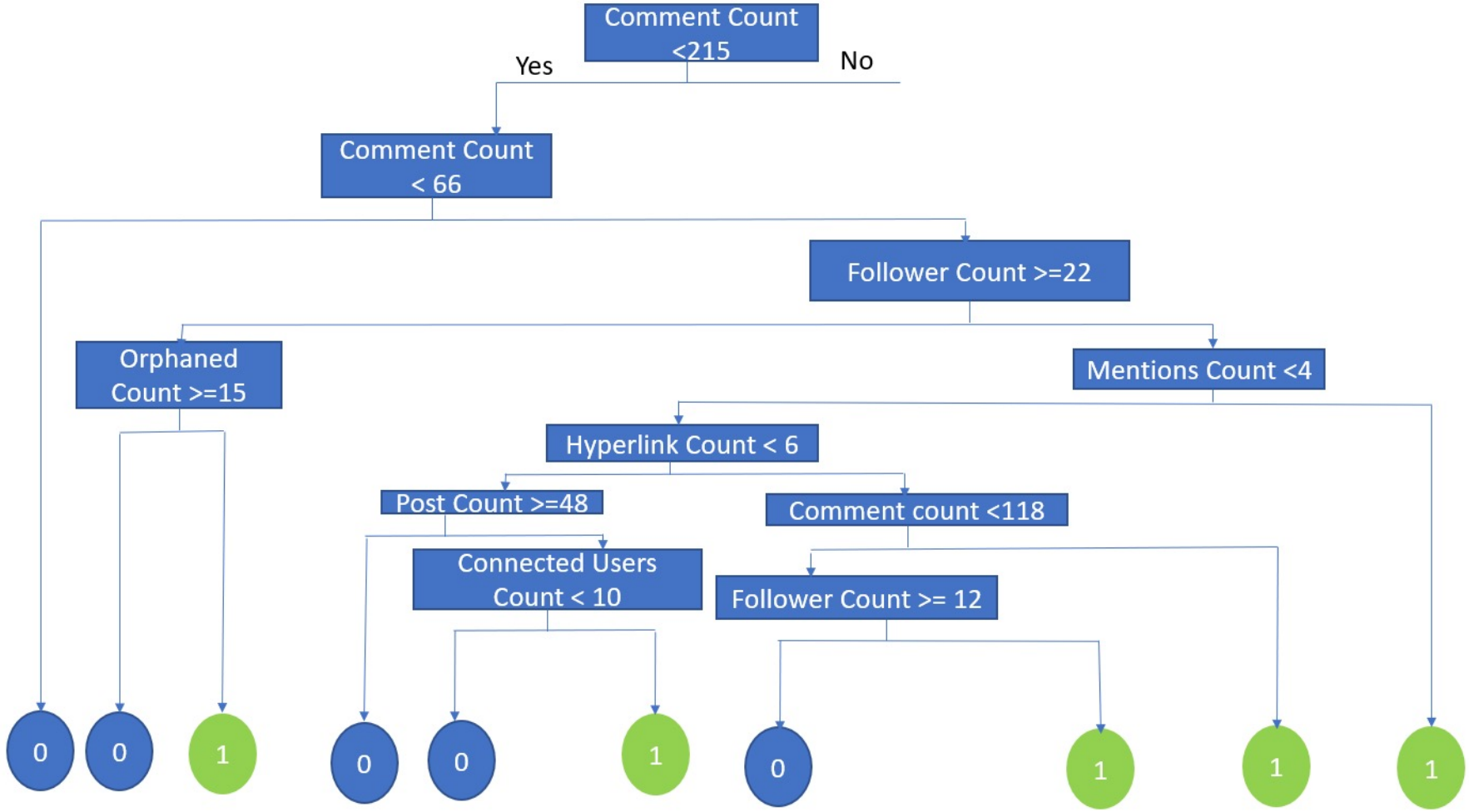
- Most of Boards have total health score larger than 100

DREXEL UNIVERSITY
LeBow
College of Business

Decision Tree – Left Branch

0 Low In Engagement

1 High In Engagement

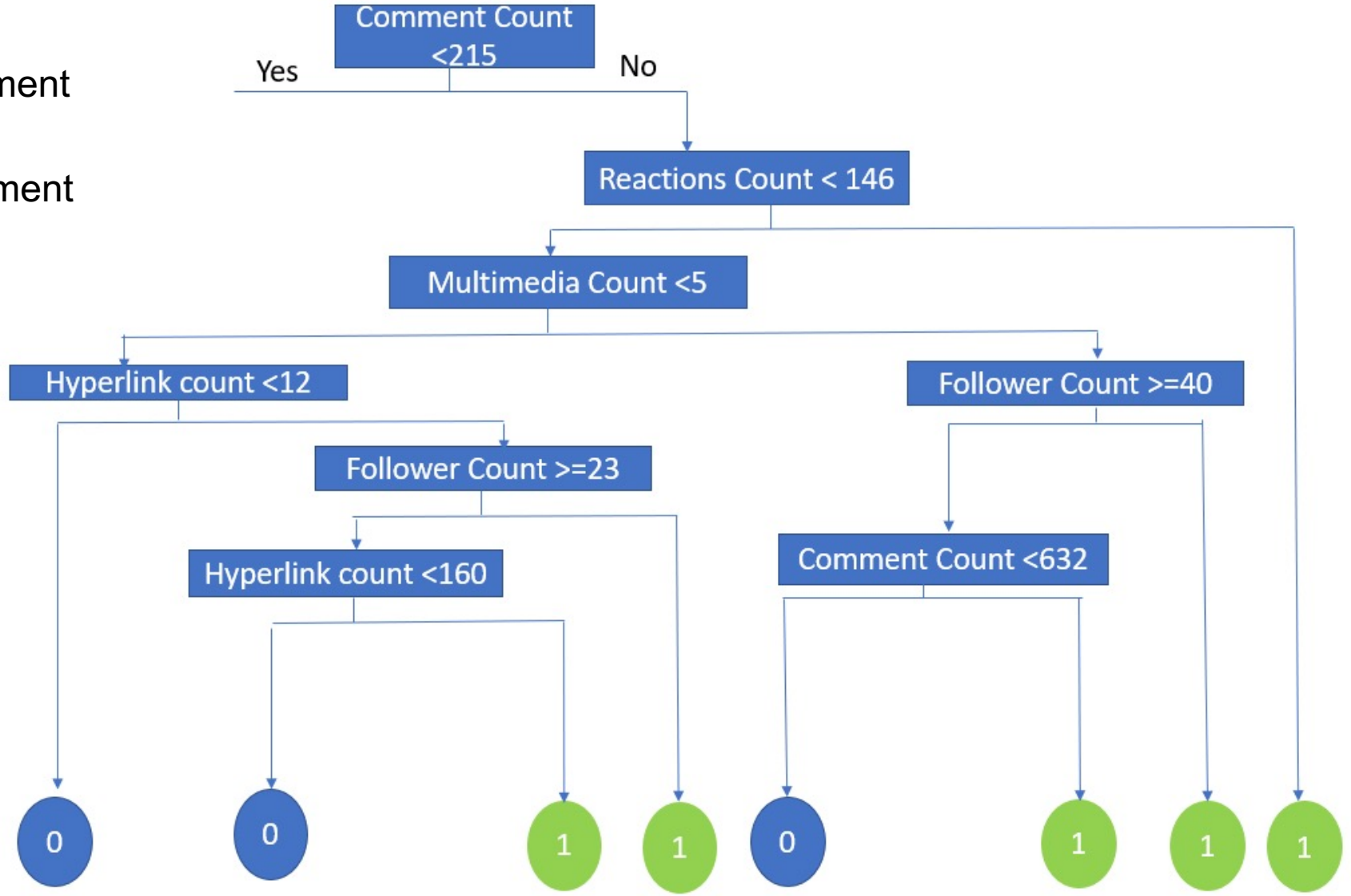


DREXEL UNIVERSITY
LeBow College of Business

Decision Tree – Right Branch

0 Low In Engagement

1 High In Engagement



Decision Tree – Important Variables

Variable	Importance Level (1 – Highest Importance Level, 13– Lowest Importance Level)
Comment Count	1
Word Count	2
Post Views Count	3
Reactions Count	4
Post Count	5
Connected Users Count	6
Follower Count	7
Hyperlink Click Count	8
Multimedia Count	9
Mentions	10
General Reaction Value	11
Total Points Awarded via Accolades	12
Average Accolade Value	13

Decision Tree – Overview

- Accuracy:
  - Training dataset: 87%
  - Testing dataset: 86%
- Board computed variables could be used as key decision points that lead to optimal community engagement
  - By setting higher score in comment count, word count, reaction count, and other important variables from decision tree result

Community Engagement

Cluster Analysis

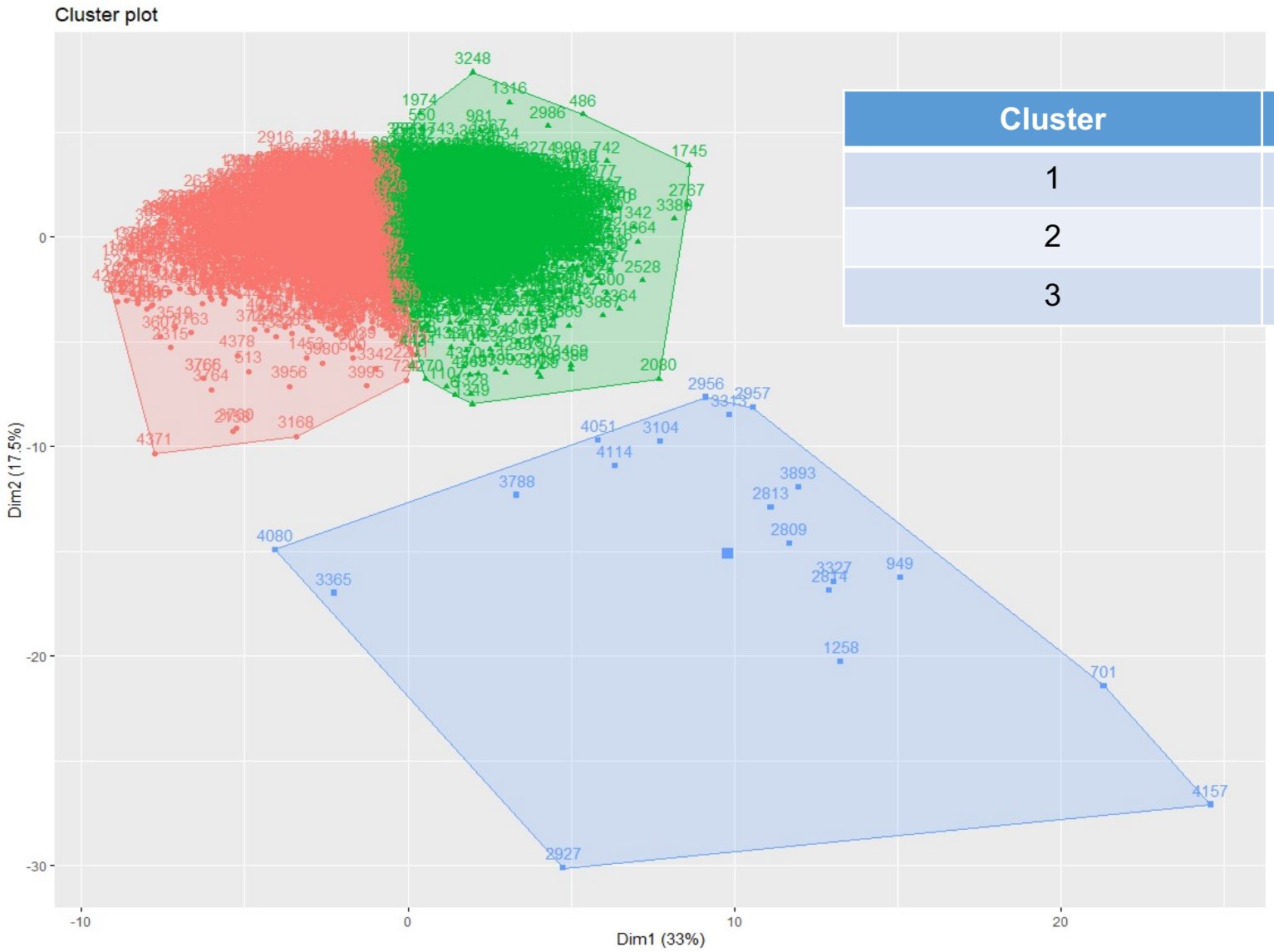
DREXEL UNIVERSITY
LeBow
College of Business

Clustering

- **Data**: Community Dataset
- **Objective:** Segment boards by engagement levels to understand community engagement
- **Analysis**: K-means Cluster Analysis
- 21 variables included in analysis (Derived variables)
- 3 Clusters chosen using Within-Cluster Sum of Squares (WSS)

- Data Preprocessing and Transformation:
  - Removed Outliers using standardization (z-score method)
  - Standardized numeric variables

Clustering Result



Cluster	Total Number of Boards
1	1877
2	2539
3	19

DREXEL UNIVERSITY
LeBow
College of Business

Clustering Result - Describe

- Cluster 1: Lowest score across all clusters
- Cluster 2: Average group
  - Highest average score in conversation ratio and interacting
- Cluster 3: High performance group in terms of engagement

Cluster	Average Word Count	Average Comment Count	Average Sharing Score	Average Listening Score	Average Interacting	Average Conversation Ratio	Average Total Health Score
1	28.900	0.241	36.552	35.223	31.683	3.013	103.372
2	52.716	0.608	59.577	60.778	63.580	5.672	183.935
3	644.597	4.116	92.355	68.825	47.535	2.273	208.714

SEGMENTING CUSTOMERS

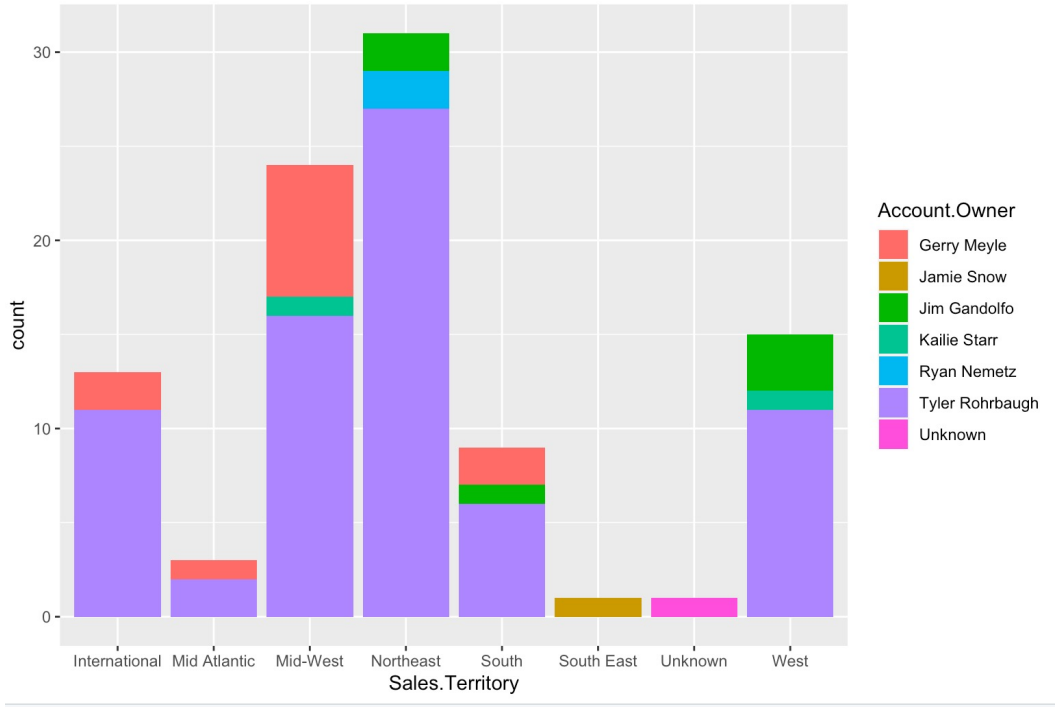
Cluster Analysis

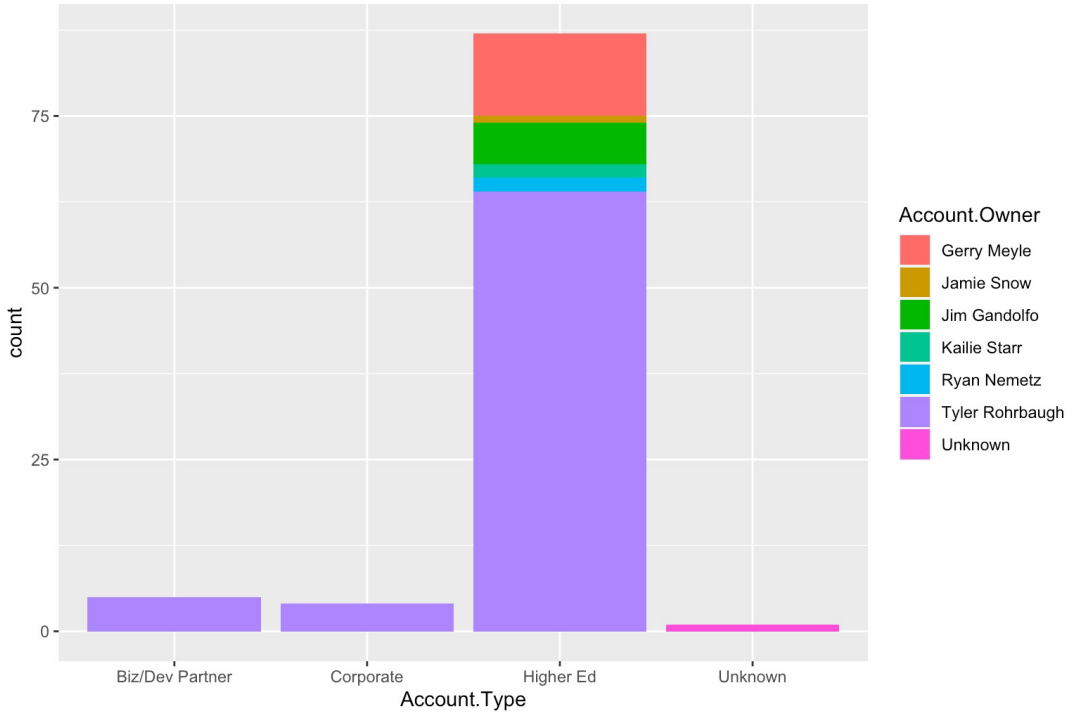
DREXEL UNIVERSITY
LeBow
College of Business

Cluster Analysis

- **Data:** Closed Won Clients
- **Objective:** Understand industries and education types and how staffing decisions can be made to capture and maintain business
- **Analysis:** k-Medoids Cluster Analysis
- 19 Variables included in analysis
- 3 Clusters chosen using Average Silhouette
- **Preprocessing**:
  - Mean Imputation
  - YeoJohnson

Closed Won Clients - Cluster 1



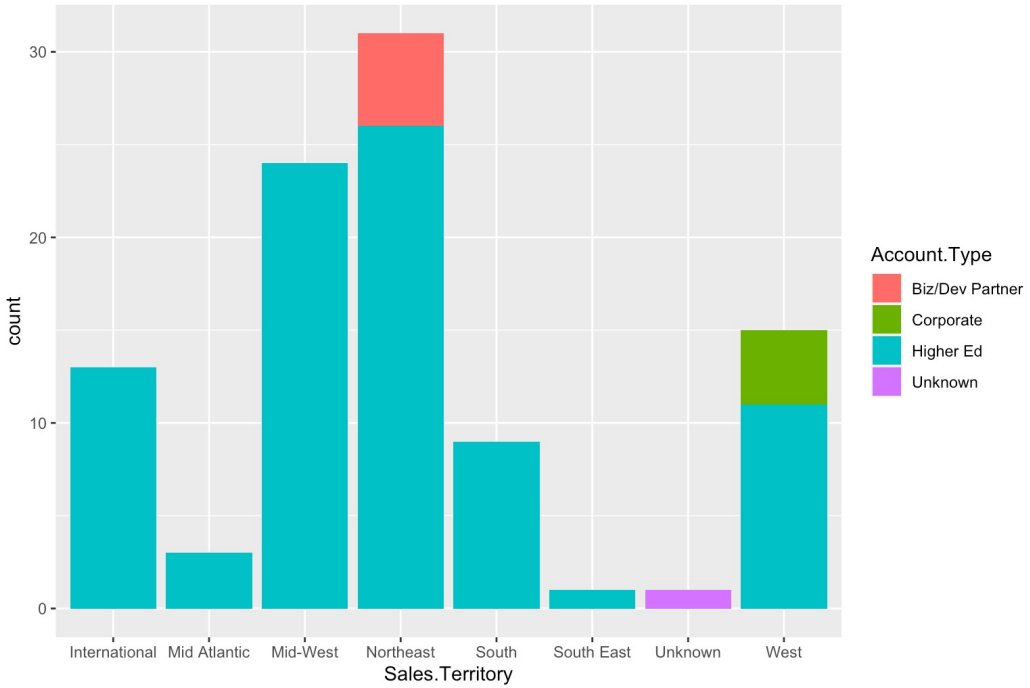


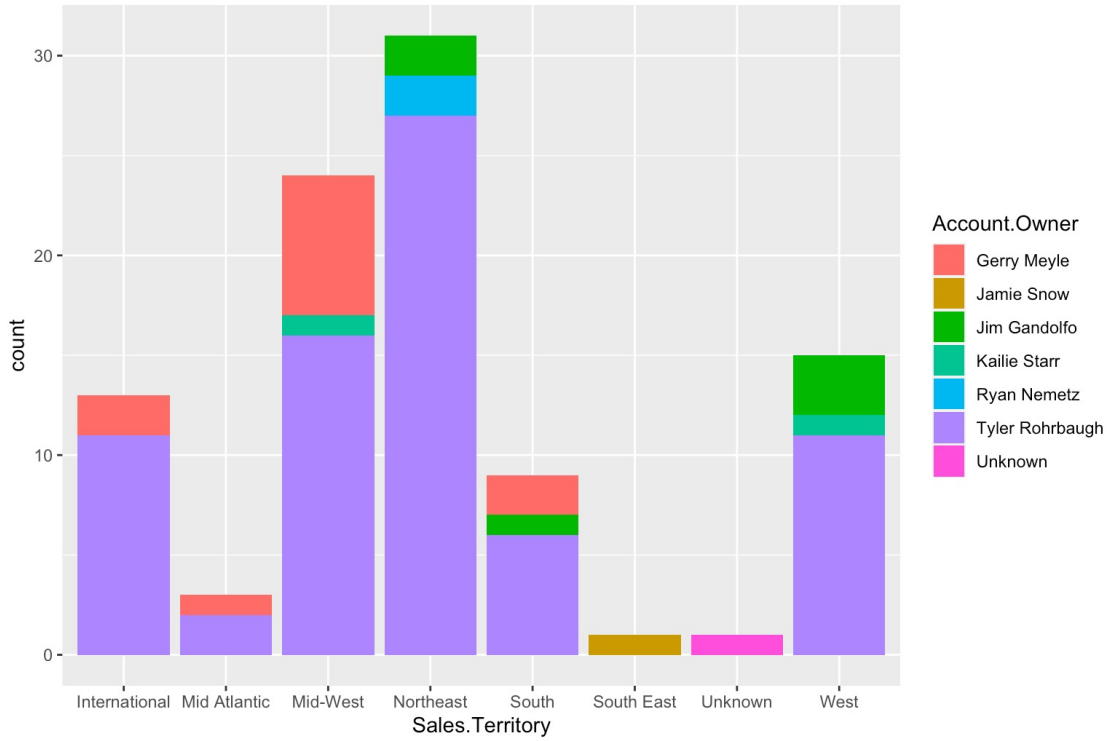
Observations:

- Tyler contributes to most contracts in Cluster 1.
- Tyler R. has a very strong network in International, Mid-West, Northeast, South and West.

DREXEL UNIVERSITY
LeBow
College of Business

Closed Won Clients - Cluster 1





Observations:

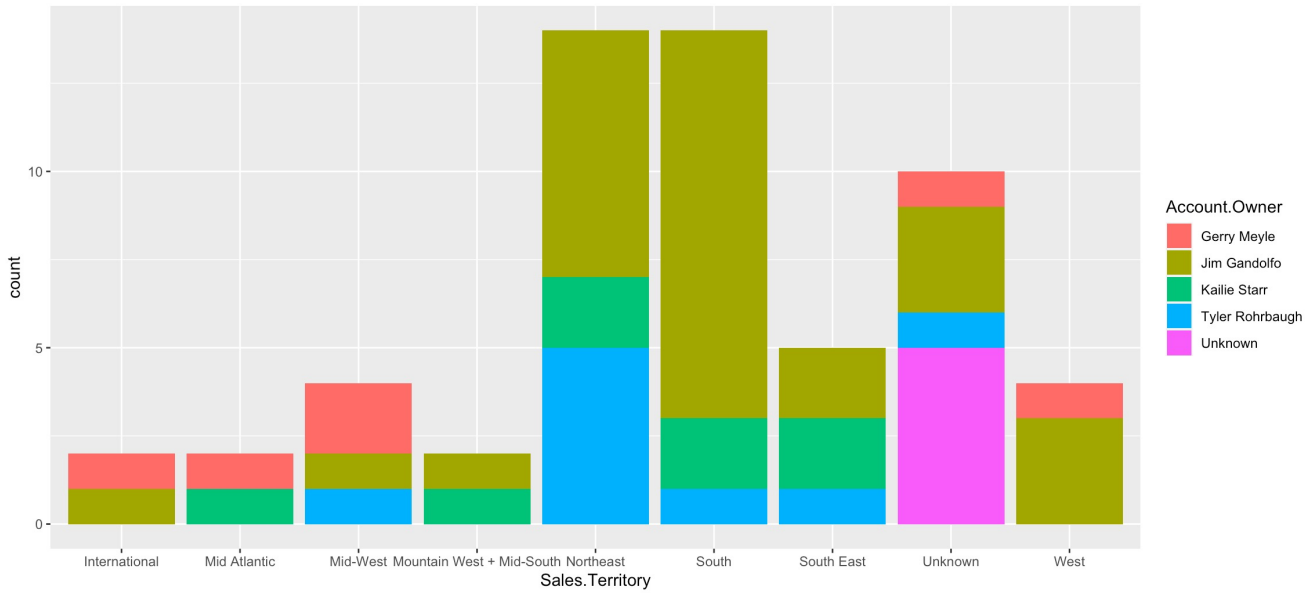
- Industries like Biz/Dev Partner and Corporate are located in Northeast and West.
- Education Industry is the highest represented account type

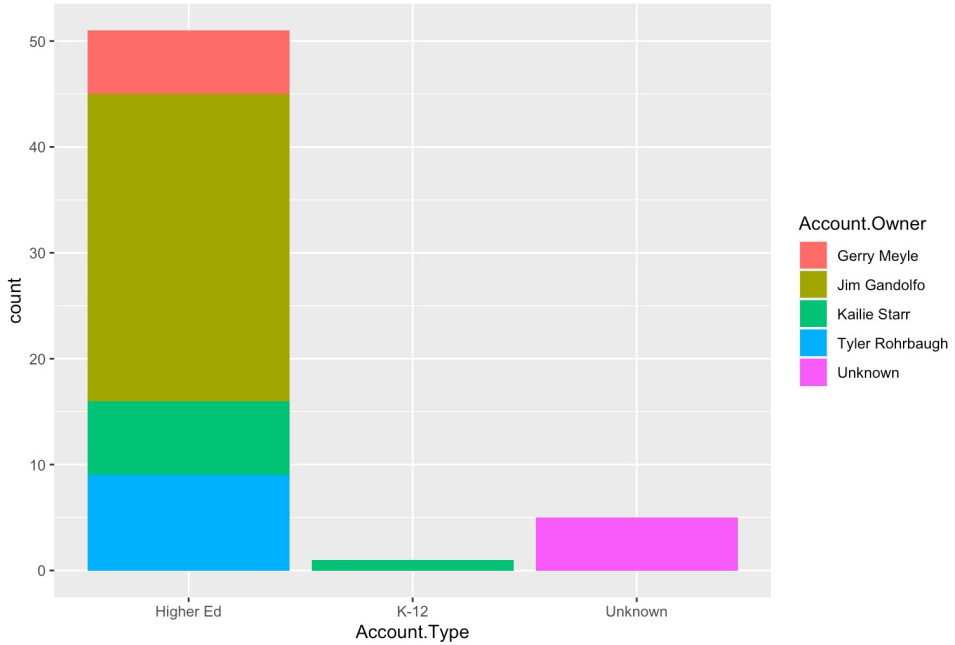
Recommendations:

- Tyler R. as Regional Manager
- Start relationships with Corporate and Biz/Dev Partners

DREXEL UNIVERSITY
LeBow
College of Business

Closed Won Clients - Cluster 2





Observations:

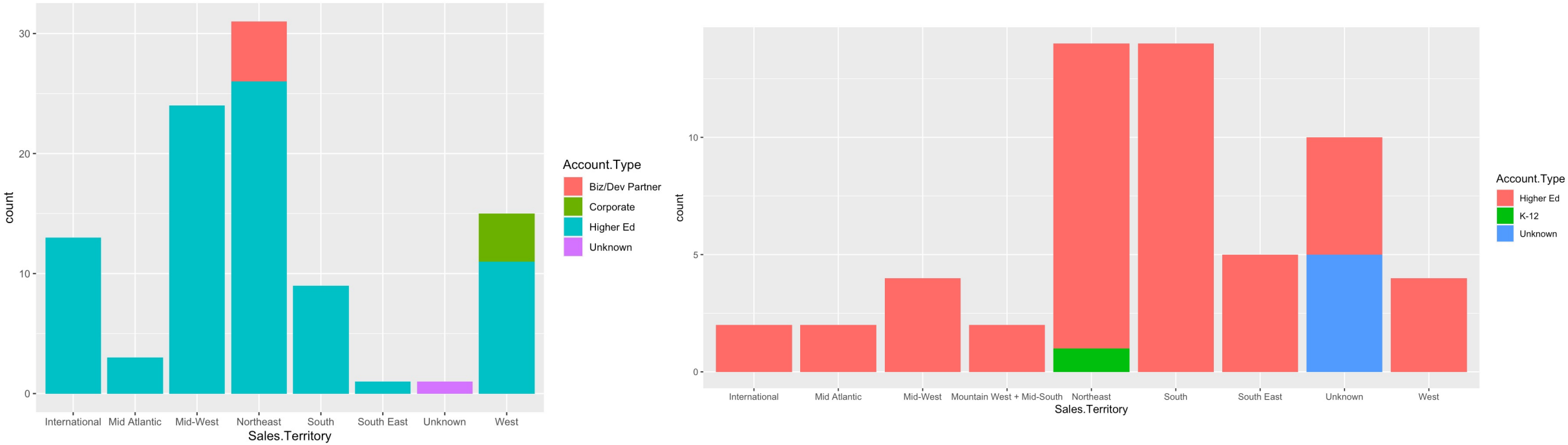
- Higher Education also has the greatest share in Cluster 2
- Jim G. has a strong network in Northeast, South, South-East, Mid-West and West territories

Recommendation:

- Establish Jim G. as Regional Manager of South-East and West territories

DREXEL UNIVERSITY
LeBow
College of Business

Closed Won Clients - Clusters 1 & 2



Observation:

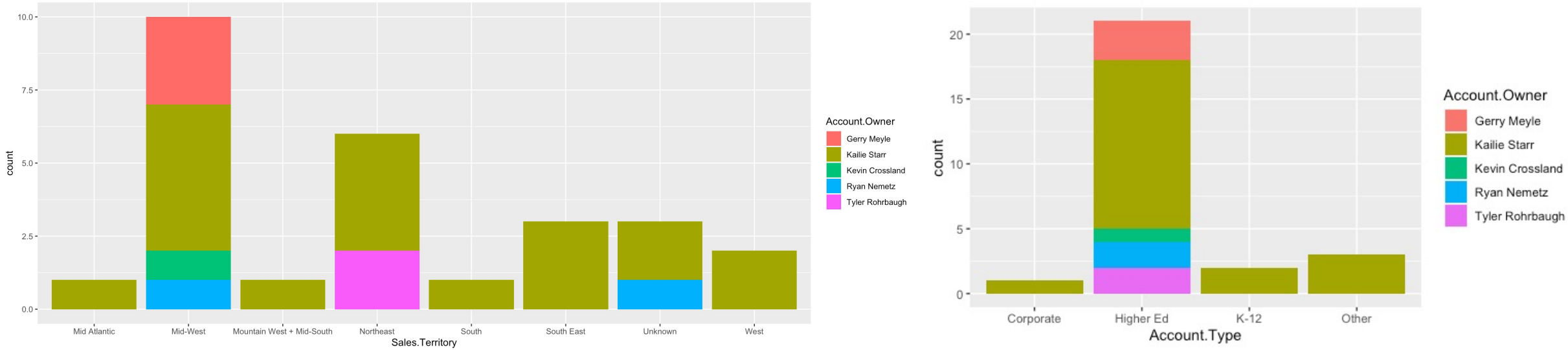
- Northeast is the most prevalent territory.

Recommendation:

- Have Tyler R. focus on Biz/Dev and Corporate partners in this region
- Have Kailie S. focus on K-12 business in this region

DREXEL UNIVERSITY
LeBow
College of Business

Closed Won Clients - Cluster 3



Observations:

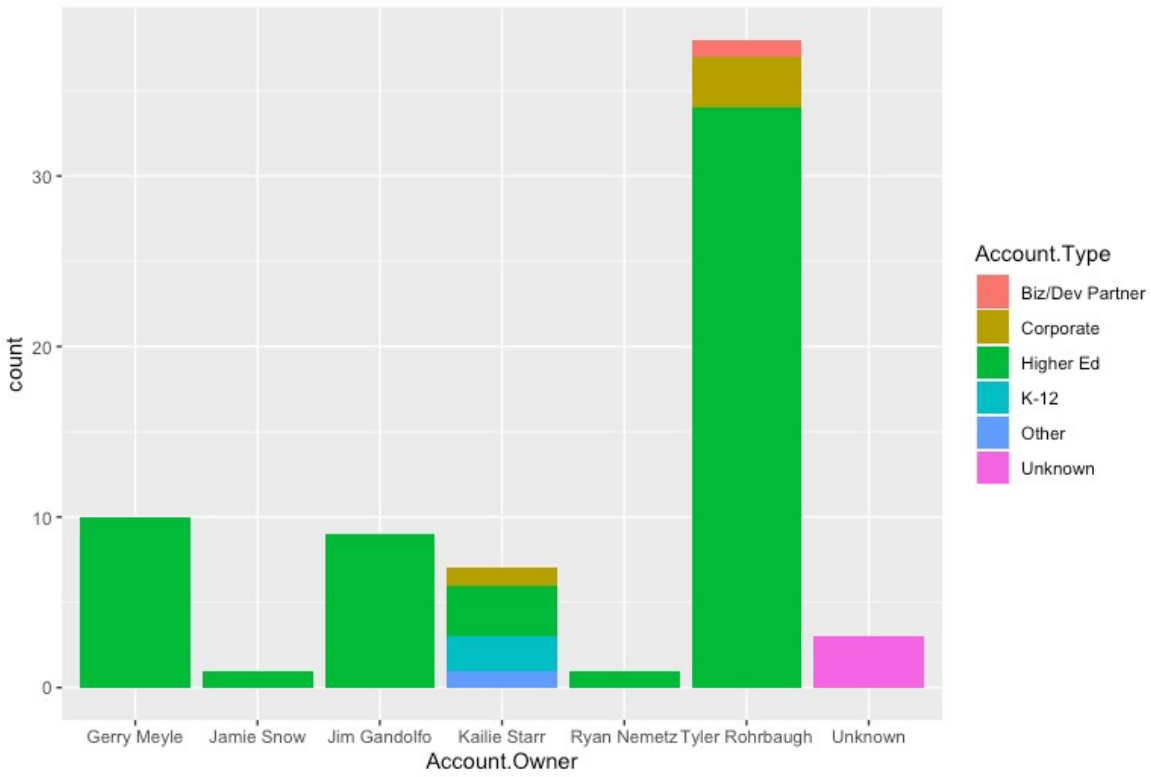
- Higher Education Industry is the main customer of the company.
- Kailie S. has a strong connection in Mountain West and Mid-South.

DREXEL UNIVERSITY
LeBow
College of Business

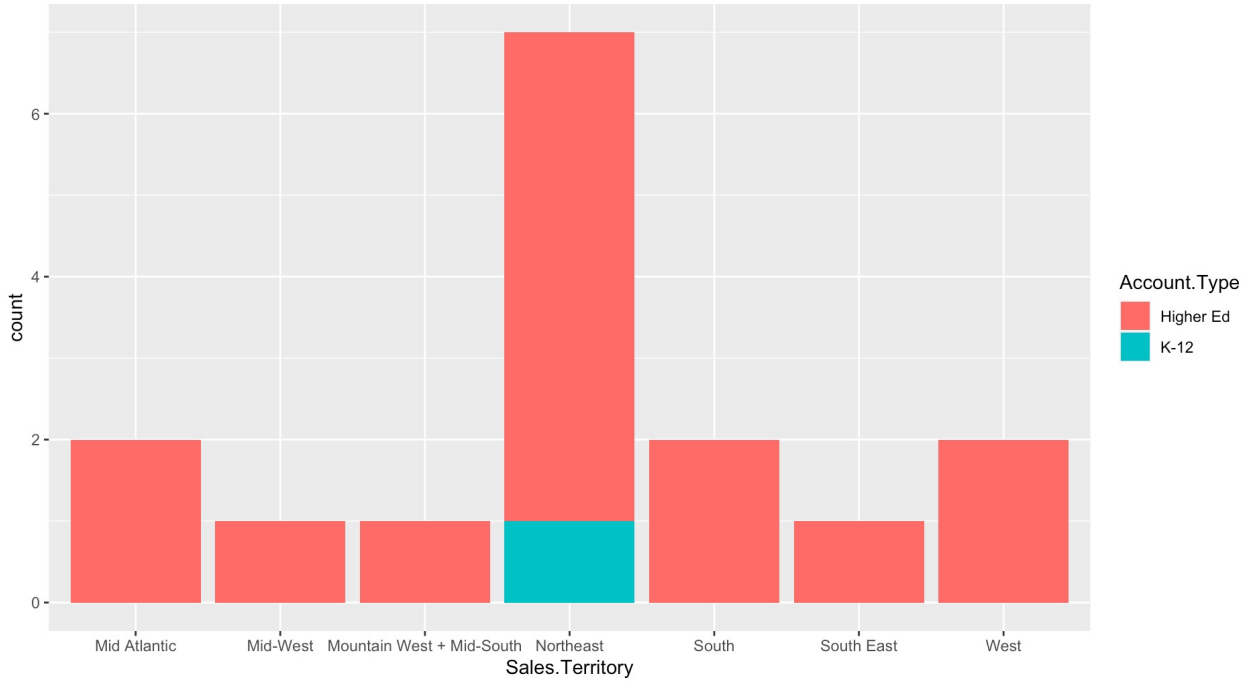
Clients Pipeline

- **Data:** Clients Pipeline
- **Objective:** Understand industries and education types and how staffing decisions can be made to capture and maintain business
- **Analysis:** k-Medoids Cluster Analysis
- 22 Variables included in analysis
- 3 Clusters chosen using Average Silhouette
- **Preprocessing**:
  - Mean Imputation
  - YeoJohnson

Clients Pipeline



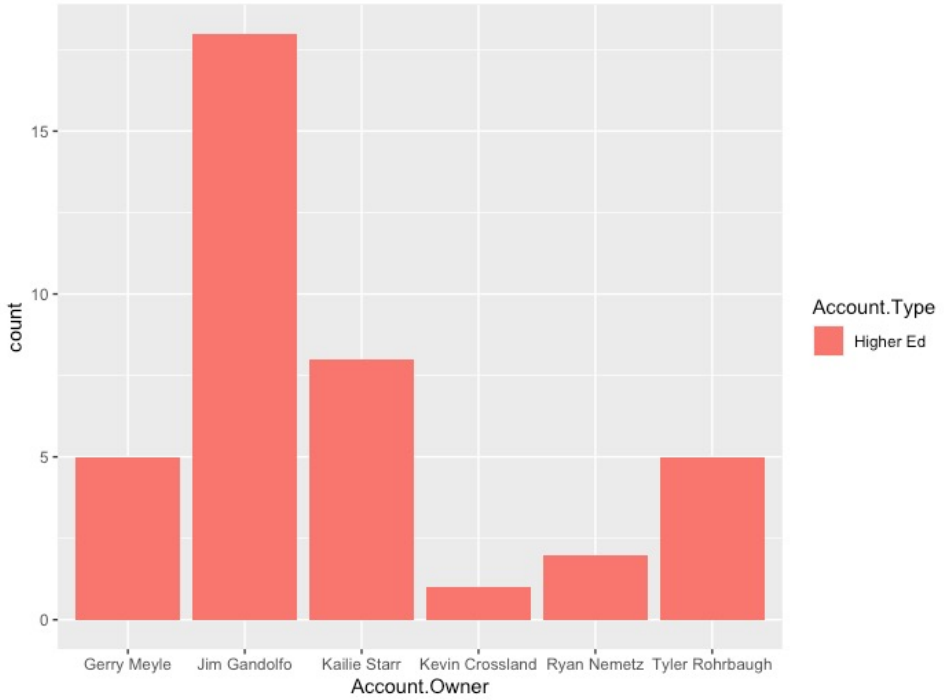
Cluster 1 focuses on High Ed, but we can see different industries were distributed across Account Owner.

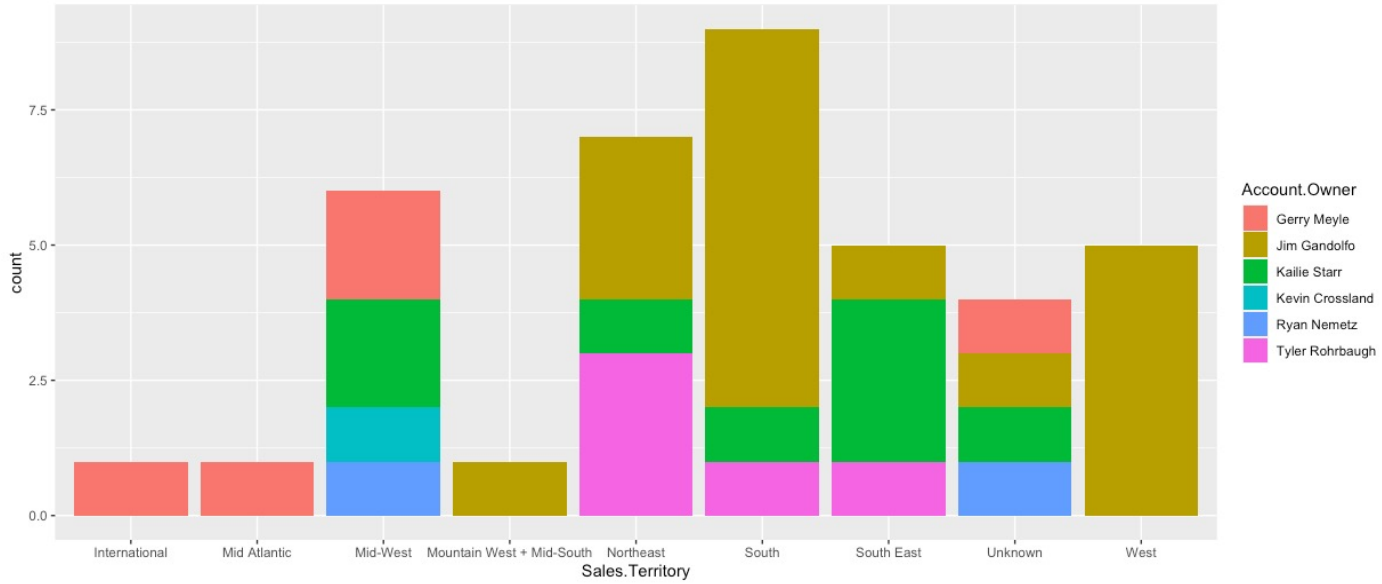


In Cluster 3, K-12 businesses, managed by Kailie, are clustered in NE. There is opportunity to grow business in this area.

DREXEL UNIVERSITY
LeBow
College of Business

Clients Pipeline





Observation:

- Cluster 2 focuses on Higher Education, especially in Midwest, NE, S, SE and W territories.

Recommendation:

- Staff Jim G. as regional manager for Higher Education in these regions

DREXEL UNIVERSITY
LeBow
College of Business

Cluster Recommendations

Staffing recommendations based on Territory and Industry.

Person	Territory	Industry
Tyler Rohrbaugh	All	Higher Education
Tyler Rohrbaugh	Northeast, West	Corporate, Biz/Dev
Jim Gandolfo	Northeast, South, Southeast, Mid-West, West	Higher Education
Kailie Starr	Northeast	K-12
Kailie Starr	Mountain West, Mid-South	Higher Education

SUMMARY & RECOMMENDATIONS

Summary & Recommendations

- Current clients are highly engaged according to our analysis
- Based on the MLR model, use more hyperlinks and multimedia resources in class boards to increase engagement.
- Based on the Cluster model, Implement a regional manager system to organize how business is managed throughout the company
- Based on the decision tree prediction model, recommended the boards to set higher score in comments count, word count, and other important variables from decision tree result

Questions, Comments?