# Team Tech Deliverable #2

## Detailed Summary

Our problem is an open-ended one: how can we help the New York Islanders understand their Twitter followers, and provide insights to help the organization better engage fans on this platform? As a large consumer-facing brand, maintaining an effective social media presence is a big priority for the Islanders. Driving fan engagement, while an intangible metric, helps to drive ticket revenue, merchandise revenue, and creates a larger fan audience that can entice valuable sponsorships for the Islanders as well.

Most major sports teams lack the resources to approach social media from an advanced analytical perspective, which means there is a lot of untapped potential for our team to add value. Being among the first to explore this space is a double-edged sword: on one hand, almost anything we can provide the Islanders will be new and valuable, but on the other hand, we will have to carefully manage the project scope. Furthermore, there is no publicly available precedent for us to build off of, at least specifically related to sports.

Our major resource is a trove of data obtained through Twitter's API using a Twint github package. We are collecting information on tweets, interactions, and profiles of accounts that tweet certain keywords or follow the official @NYIslanders account. We are also partnering with Sam's team to get their insights into things like: comparable Islanders-focused Twitter accounts, reasonable ways to categorize tweet content, and expanded list of keywords and hashtags. As far as tools go, we are storing our data on BigQuery, and using Google cloud tools such as CoLab, Sheets, and Docs, as well as Slack to collaborate on our various workflows.

**Project Breakdown**
We have broken our project into two main teams, with two sub-teams:
1. **Prong 1:** Understand Islanders Twitter activity
    a. **1a:** Compile relevant summary statistics about the fans, and calculate more complex variables that we can feed into the clustering algorithm.
    b. **1b:** Analyze fan engagement based on tweet timing, content, and other variables.
2. **Prong 2:** Cluster most active Islanders Twitter fans
    a. **2a:** Apply clustering techniques to identify fan clusters.
    b. **2b:** Measure fan interactions across the different clusters.
    c. **2c:** Extract insights and make recommendations.

**Data Used**
With the help of Ferran, we have been able to store Twitter data stored in BigQuery. The data we have access to can be broken down into the following tables:
1. All tweets from the Islanders account (2018-2019 season and 2019-2020 season)

2. All tweets containing Islanders-related keywords and/or hashtags (2018-2019 season and 2019-2020 season)
3. All tweets from Islanders' "avid" fans (defined as Twitter users that have Tweeted with "#Isles" at least twice) (August 2019 - Present)
4. Notable Twitter accounts (e.g. players, prospects, fan pages, media members, etc.) and who their followers are

**Preprocessing Strategies**

Using the datasets mentioned above, we are able to create a structured dataset that contains profile and Twitter activity information for every Islanders follower. We will use these features to build a clustering algorithm, which we can use to identify specific social media strategies for certain types of fans.

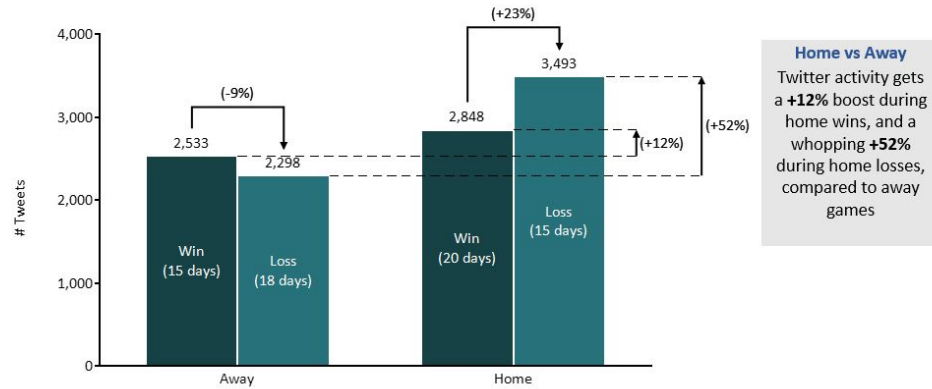**Initial Feature List**

Our initial features are below:

| Followers | Following | Total tweets | Total Likes | Islanders related tweets | @NYIslanders tweets RT |
|---|---|---|---|---|---|
| @NYIslanders tweets replied | Total tweets in dataset | Media accounts following | Alumni accounts following | Prospects accounts following | Team officials following |
| Fan pages following | Meetup pages following | Twitter fans accounts following | Other accounts following | Players accounts following | Has Bio? |
| Bio related to Islanders? | Account verified? | Private account? | Has location? | Profile Picture? | Bio characters |
| Language | Account age | | | | |

Features were selected by observing key and readily available information from the scraped twitter user data, extraction algorithms have not yet been employed to reduce feature space

With this feature list, we are attempting to capture a variety of summary account features (language, verified, number of followers, etc.) as well as many different metrics that can factor into a fan's avidity (tweet volume, tweet content, engagement with Islanders content, etc.). For the preliminary clustering, we wanted to get an initial feature set that was reasonably easy to obtain, but also that represents enough possible measures of avidity to inform meaningful clusters. From there, we used feature reduction and extraction to improve upon our results and identify meaningful clusters.

# Preliminary Results
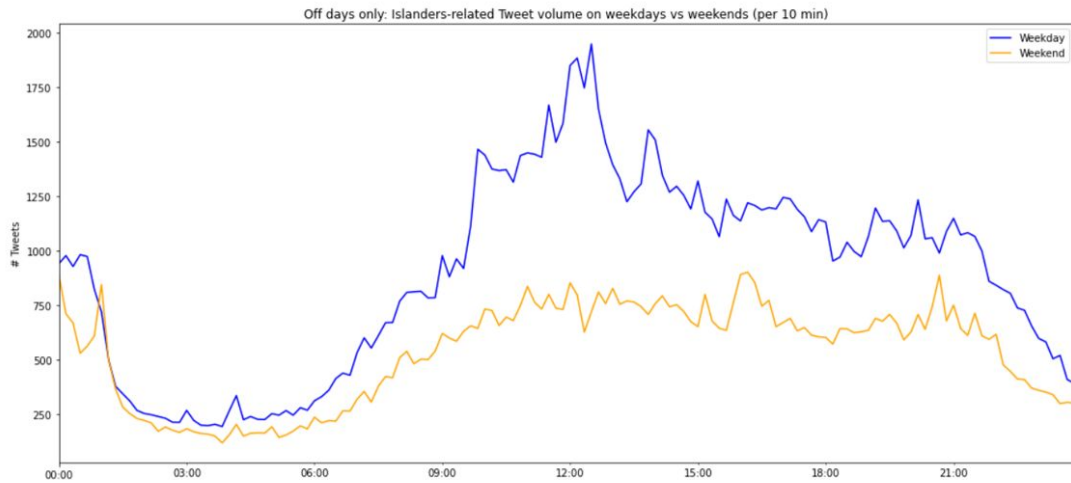
### The Home Game Effect: Wins vs Losses



*Home games generate **29%** more Twitter activity than road games, mostly driven by losses.*
*For away games, a loss **reduces** Twitter activity, whereas a loss at home **increases** Twitter activity.*

*Data excludes retweets, Sep 1 2019 – April 1 2020

### Off-day Twitter Activity: 24H map



*On off-days, weekday Twitter activity peaks over lunch (11a-1p), while weekend activity is more steady*

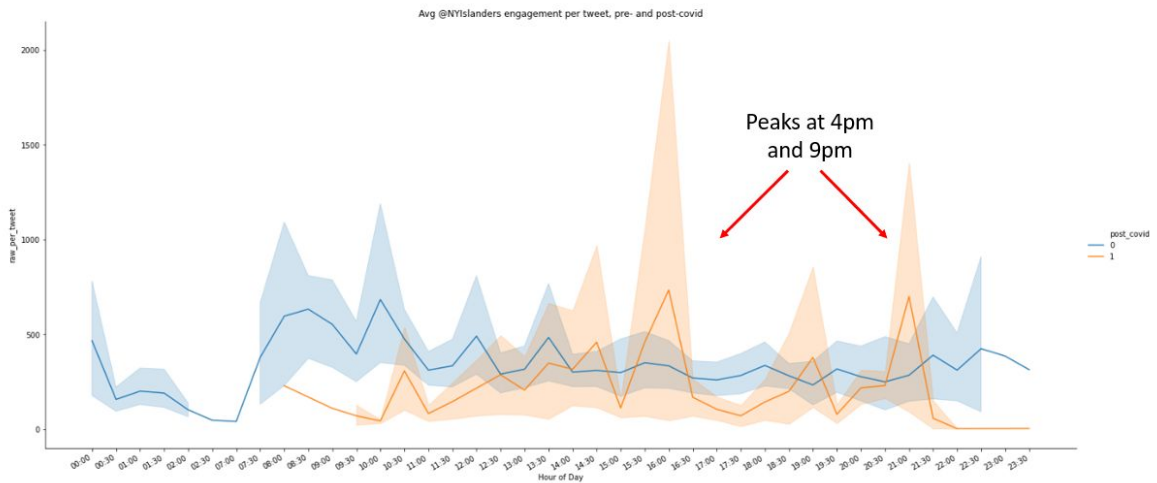*Data excludes retweets, Sep 1 2019 – April 1 2020

## Most-used hashtags by @NYIslanders

| hashtag | relative_frequency | hashtag | relative_frequency |
|---|---|---|---|
| isles | 0.540741 | goring91 | 0.014141 |
| lgi | 0.166330 | hockeywithaheart | 0.013468 |
| nhlallstar | 0.023569 | isitoctoberyet | 0.012121 |
| nhl | 0.020875 | hockeyfightscancer | 0.010774 |
| islesdailychallenge | 0.019529 | islesfamily | 0.008754 |
| islesrookiegame | 0.017508 | nhlmovienight | 0.008754 |
| tonelli27 | 0.014815 | isleswallpaperwednesday | 0.008081 |
| | | islesrewind | 0.006734 |

*#isles is by far the most-used hashtag, representing 54% of all hashtag usage, followed by #lgi*

*Data excludes retweets, Sep 1 2019 – April 1 2020

## @NYIslanders tweet engagement on off-days, before and after NHL shutdown



Avg @NYIslanders engagement per tweet, pre- and post-covid

Peaks at 4pm and 9pm

post_covid
— 0
— 1

*Overall engagement is down since NHL shutdown, but 4pm and 9pm show higher spikes of engagement*

*Data excludes retweets, Sep 1 2019 – April 1 2020, y-axis measures total RTs + Replies + Favorites
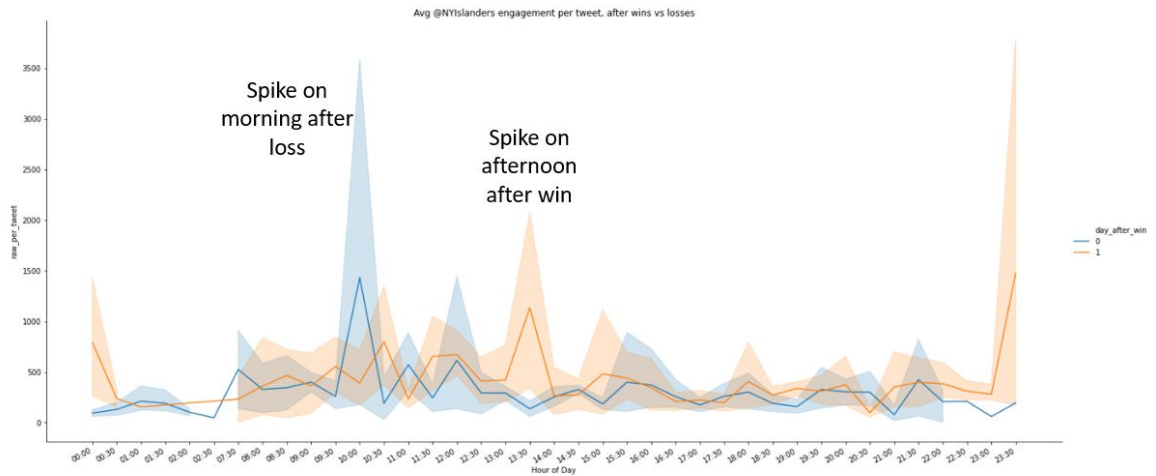
## @NYIslanders depth score on off-days, before and after NHL shutdown



Avg @NYIslanders depth score per tweet, pre- and post-covid

New peak over lunch

post_covid
— 0
— 1

depth_per_tweet

Hour of Day
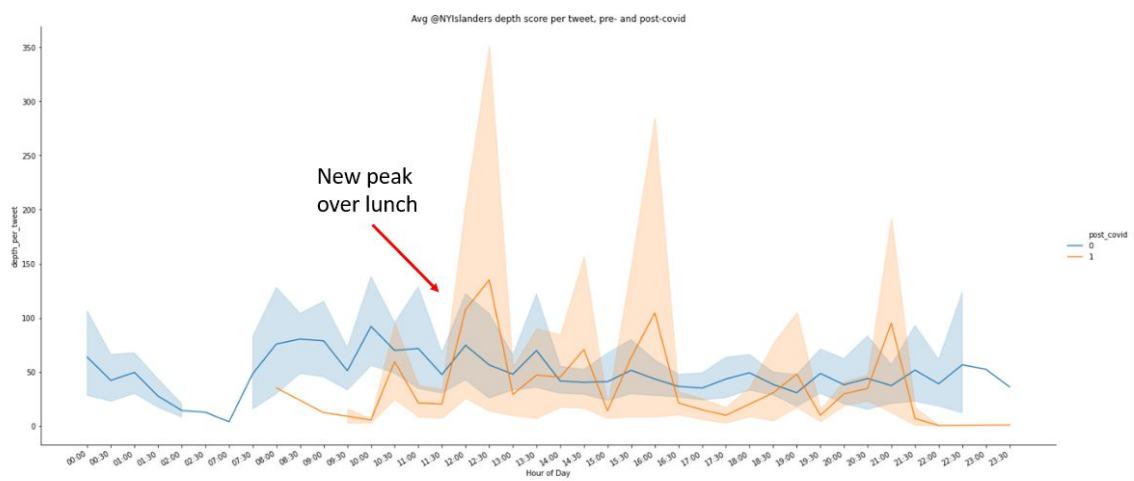
*Post shutdown, there is a peak in follower responsiveness over lunch and around 4pm, that doesn't show in overall engagement*

*Data excludes retweets, Sep 1 2019 – April 1 2020, y-axis measures (RTs*0.25) + Replies + (Favorites*0.1)

## @NYIslanders tweet engagement on off-days, after wins vs losses



Avg @NYIslanders engagement per tweet, after wins vs losses

Spike on morning after loss

Spike on afternoon after win

day_after_win
— 0
— 1

raw_per_tweet

Hour of Day

*Engagement spikes the morning after a loss, and the early afternoon after a win.*

*Data excludes retweets, Sep 1 2019 – April 1 2020, y-axis measures total RTs + Replies + Favorites

**Data Visualization using PCA**



6 clusters with Silhouette Coefficient of 0.3785698837390404
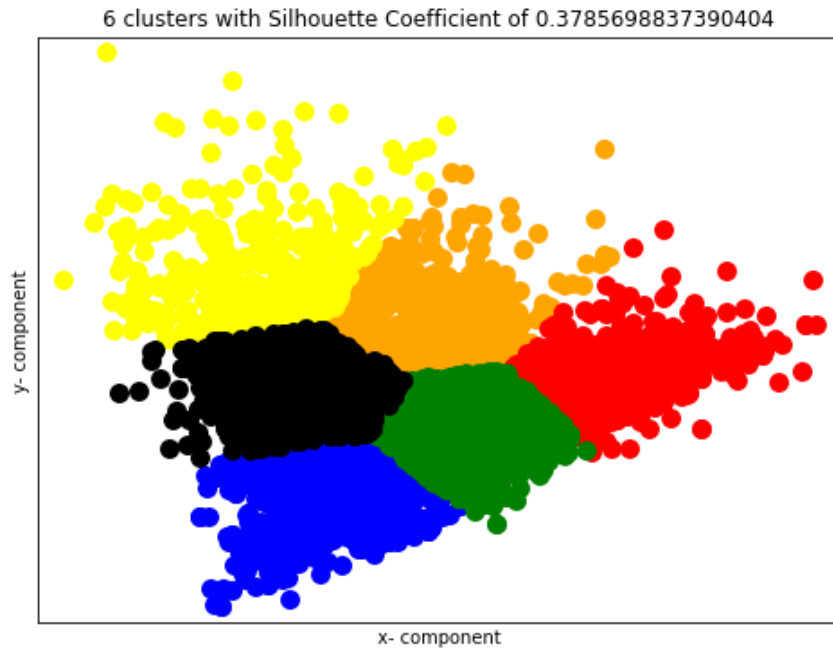
## There are clear clusters of accounts with high Twitter 'influence'

| Cluster Label | % w/ Islanders Mentioned in Bio | % w/ Profile Pic | % w/ Location Enabled | Avg Followers | Avg Following | Avg # of Notable Accounts Followed | Count |
|---|---|---|---|---|---|---|---|
| 0 | 52.0% | 96.0% | 56.0% | 353 | 1,331 | 54.0 | 25 |
| 1 | 31.3% | 91.5% | 36.3% | 147 | 202 | 18.6 | 281 |
| 2 | 53.3% | 100.0% | 66.7% | 312 | 407 | 37.5 | 15 |
| 3 | 27.3% | 99.8% | 57.3% | 324 | 429 | 22.6 | 524 |
| 4 | 51.1% | 100.0% | 76.5% | 26,744 | 1,315 | 54.9 | 268 |
| 5 | 55.2% | 100.0% | 65.0% | 435 | 831 | 66.6 | 317 |
| 6 | 68.0% | 99.3% | 74.0% | 1,331 | 1,566 | 112.9 | 150 |
| 7 | 15.1% | 99.7% | 72.2% | 13,825 | 2,048 | 14.7 | 331 |
| **Grand Total** | **37.5%** | **98.5%** | **62.1%** | **6,439** | **968** | **40.1** | **1,911** |

NHL ✓
@NHL

TSN Hockey ✓
@TSNHockey

Shannon Hogan ✓
@Shannon_Hogan
Sportscaster covering the @NYIslanders on @MSGNetworks – Imported from Detroit -Mizzou Alum.....insta: Instagram.com/ShannonHoganSp...

MSG Networks ✓
@MSGNetworks

## Cluster results are also identifying super fans based on Twitter engagement

| Cluster Label | Avg Total Tweets | Avg Tweets Mentioning Islanders | Avg Replies to Islanders Account | Avg Retweets of Islanders Account | Count |
|---|---|---|---|---|---|
| 0 | 590.6 | 220.8 | 4.6 | 80.0 | 25 |
| 1 | 79.0 | 15.6 | 2.4 | 0.6 | 281 |
| 2 | 438.1 | 129.5 | 58.6 | 5.9 | 15 |
| 3 | 89.1 | 10.2 | 1.1 | 0.8 | 524 |
| 4 | 311.8 | 64.8 | 1.7 | 2.8 | 268 |
| 5 | 58.5 | 12.2 | 1.0 | 0.8 | 317 |
| 6 | 236.4 | 70.3 | 1.5 | 3.7 | 150 |
| 7 | 323.5 | 10.7 | 0.6 | 1.0 | 331 |
| **Grand Total** | **175.2** | **27.5** | **1.8** | **2.4** | **1,911** |

# Next Steps

For the first prong on understanding the Islanders Twitter activity, we have several avenues we still want to explore. The first is building out a regression or linear model in R to try and get deeper understanding on what impacts the interaction to the Islanders tweets. Hopefully this gives us more insight into how strong each factor is and what the Islanders can focus on going forward. Since there are so many confounding variables, a linear regression could help us explain to the Islanders how each decision they make, from tweet timing to content, individually impacts the expected engagement.

Another next step is looking into the Islanders Twitter activity post COVID and the cancellation of the season. We will investigate how their activity changes via time of day and see how they can change their activity schedule to adjust to the fan interaction. This is an especially timely analysis that the Islanders are eager to receive, as it has the potential to make an immediate impact on how they run their Twitter account.

We are also looking into comparing the Islanders content schedule to how people interact with their tweets in general to see if there are any major gaps in how they interact on twitter with their fans.

Lastly, we hope to use the clusters once they are finished and overlap them with our existing data frames in place to gain more insight into how those avid fans interact with different content and understand their activity patterns. Hopefully this will help us get actionable insights for the Islanders on how to help and interact with their most loyal fans.