Written Report: Assisting Scouts in Identifying Top Talent (CKM Sports Mangement)

INSY 442: Data Analysis & Visualization

Dr. Geneviève Bassellier

Alexander Sukhanov (260988130)

Emily Hagelauer (260892609)

Jonathan Steinberg (260924571)

Justin Ma (260898526)

Thomas Atchison (260986081)

DESAUTELS Faculty of Management
Faculté de gestion

Introduction

CKM Sports Management is a Vancouver-based full-service hockey agency founded in 2010. The organization supports amateur and professional hockey players throughout North America and internationally through a trusted network of scouts, trainers, agents, and player support staff to help players achieve success on and off the ice. CKM provides a holistic range of developmental services to its clients including nutrition and strength training, post-secondary education and financial planning, assistance with finding professional and junior hockey jobs as well as hockey analytics. CKM's rapid growth can be attributed to their innovative approach and superior client services, which has enabled them to represent clients in over seven professional hockey leagues globally, including the WHL, OHL, QMJHL, AHL, ECHL, SPHL, and various European hockey leagues. CKM's partners include Project Sport Agency, KeySport Agency, and Sport Agon.

CKM leverages technology and data to determine a players' strength, weaknesses, tendencies, and what works and doesn't work for them. The use of hockey analytics has practical applications that can be exploited by individuals for both development and negotiation utilization. Player analytic performance reports can be a key component to accelerated development, with individual statistical tendencies shared with coaching staff for better integration within team coaching strategy. For negotiation, analytics are highly valuable to agents, scouts, and professional hockey programs. Tracked data can serve as a negotiating tool by comparing their client's inputs and outputs players with statistical cohorts.

CKM's data can also be utilized for scouting purposes, where data-driven insights are used to identify and select top talent. Assisting scouts in identifying top talent is important because it helps upper management make informed decisions about who to draft or sign. By using data and advanced analytics to identify top performers and future stars, teams can gain a competitive edge and improve their chances of success. In addition, selecting the right talent can impact a team's revenue through increased ticket sales, merchandise, and sponsorship deals.

For this project, we constructed dashboards to assist scouts in identifying the best hockey talent at the minor league level. Our dashboards equip scouts and team executives with the information they need to make informed decisions about player selection and provides them with the flexibility to use the tool as they see fit. We assembled two dashboards, one for forwards and one for defense. The graphs within each dashboard highlight specific aspects of a player's performance relevant to their position, such as their playmaking vs goal scoring tendencies, shooting ability, as well as their offensive and defensive impact. Some of the visualizations represent different archetypes of players for each position, such as puck luck, discipline, heavy hitters, and power play specialists, enabling scouts to assess players on a holistic set of key performance indicators and identify players who align with their team's objectives. Our dashboards also allow scouts to effectively compare players by filtering across age groups, leagues, teams, and player names across all visualizations. Certain graphs are equipped with their own filters, allowing scouts to customize the relationship they want to focus on. In addition, the dashboards provide a seperate section for player analysis, making it easier for scouts to focus on key information and gain a better understanding of players' strengths and weaknesses.

Dataset and Methodology

The dataset used for our dashboards was extracted from InStat, an online database for players and teams in a variety of sports that is popular amongst the hockey analytics community. Accessing the InStat database requires a paid-for subscription, therefore the dataset was provided to us by CKM Sports. The dataset was composed of 4,538 players across 177 teams in 16 different minor leagues across Canada. Each of these minor leagues either consists of players under 16, 17, or 18 years of age.

Player were evaluated on 31 statistics, most of which are related to their ice time, points, shots, and faceoffs. These statistics were given to us in average per game form, so we calculated each of their totals by multiplying their averages by the number of games played by each player. This included calculating their total time on ice in seconds, which the totals were then divided by and subsequently multiplied by 3,600 to obtain statistics per 60 minutes. Hockey statistics are often displayed in "Per 60" form as it is a normalized metric that facilities a fairer comparison between players with differences in ice time.

In addition to these player-centric statistics, we were able to extrapolate totals for each team by summing the statistics of all players on the same team. We then divided each player's total by that of their team for each statistic to create a "relative to team" metric. Statistics displayed as a proportion of their team's total allowed us to gain insight on the most dominant players on each team, however they should be analyzed with caution: Players on teams with a lower overall talent level may have a greater share of their team's statistics, which can make them seem better than they actually are. On the other hand, players on teams with a high concentration of talent may have a lower share of their team's statistics, thus appearing worse than they are in reality.

Challenges: Duplicate Players

There were 382 player names that appeared more than once in our dataset. These duplicate players pose several issues, one of which is distinguishing between instances of the same player that appear in multiple rows and different players who coincidentally share the same name. In the former case, the same player would appear more than once in our visualizations, which can confuse the end-user and limit their ability to analyze these players holistically.

To remedy this issue, we analyzed the values for Team, League, Number, and Position attributes between the rows of each duplicate player. Given our limited domain knowledge of players in the dataset, we used these attributes to find patterns to help uncover the reason for these duplicate players. Our results were as follows:

- 57 players had same values for Team and same values for League
- 188 players had same values for Team and different values for League
- 62 players had different values for Team and same values for League
- 75 players had different values for Team and different values for League

All the players in the subset of 57 were evidently the same player appearing in more than one row in the dataset. Some players were duplicated because they had played both forward and defense positions, while others had several jersey numbers. For all 57 players, we merged their statistics by taking the sum between their duplicate rows and created an additional column for those that had multiple jersey numbers. We did not have to create an additional column for Position since we opted to create sperate dashboards for forwards and defensemen anyway.

The group of 188 players predominately consisted of those who play in several age groups with similar league abbreviations. For example, some talented 16-year-old players play in both the CSSHL U16 and CSSHL U17 leagues. However, the difference between league abbreviations were sometimes starker; some players played in the CSSHL, a country-wide league, and the SMAAAHL, a Saskatchewan-based league. Regardless, CKM instructed us to keep these players as separate observations in the dataset since to account for the skill disparity between leagues; often a 16-year-old will perform at a much higher level in their own league than in higher age groups.

When handling the players in the group of 62 and 75 players that played in different leagues, we took a hybrid approach. First, we extracted players who are separate individuals that coincidentally share the same name. Most of these players had different values for Team, League, and Number between their duplicate rows. Then, we merged similar team abbreviations under a common name, specifically their team abbreviation belonging to their oldest age group. For example, the Mississauga U16 team in the GTHL league is called the Reps, while their U18 team is referred to as the Rebels, so we assigned the team name Rebels to all rows of this type. Lastly, the difference between some team abbreviations between duplicate were quite distinct, and were clearly not part of the same organization. In this case, we were instructed to retain the observation that players logged the most games with and remove the row containing the team they played for the least. We discovered that, on average, players played 17.9 games with their primary team and only 2.3 games with their secondary team. Therefore, the removal of their secondary-team data would have a negligible affect on their player evaluation.

Now, the only type of duplicate players remaining in the dataset are those who share the same team abbreviation and different league abbreviation between their rows. Typically, our dashboard would be able to dynamically display their statistics depending on what filters are activated. For a player in a U16 and U17 league, our dashboard would only display their U17 stats when the U17 filter is activated, and the summation of their stats when no filters are activated. However, since most of our data is in Per 60 or Relative to Team form, summing statistics between league would be erroneous. Therefore, we opted to keep these players as separate entities in our visualizations.

Challenges: Data-Related

Data quality posed several challenges, such as inappropriate formatting and missing data points. To ensure readability, some columns needed to be appropriately formatted. For instance, the "Time on Ice" column followed the DD-MM-YYYY HH:MM:SS format, which meant that players with ice time over 24 minutes needed to be manually adjusted to a more readable format. Additionally, some columns had missing data points, which hindered the ability to draw reliable conclusions. For example, the "Passes to the Slot" column only had 1,042 out of 4,549 observations, and was thus dropped to reduce the impact of missing data bias. Furthermore, many teams played a small number of games (41 teams out of 184 had less than five games played), which may lead to a small sample bias. Therefore, when developing charts, we were cautious when including players with a small number of games played.

Another challenge we faced was the limited availability of certain attributes in the dataset provided by CKM. We noticed that some features that were widely available on InStats were missing in our dataset which would have enables us to develop more comprehensive graphs, and help scouts identify future talents more efficiently. Analyzing additional features related power plays, penalty kills, and possession time would have provided our analysis with more diversity and allowed us to evaluate players more comprehensively, enabling the identification of hidden talents that are often missed through traditional metric analysis. We understand that most of these additional features were excluded to ensure comparability across leagues as they are not measured uniformly in all leagues. Overall, we prioritized the comparability of data across leagues over delving into more high-depth statistics. Our main objective was to ensure that potential talents from all leagues were not excluded simply because their league did not provide advanced statistics.

Lastly, we had small issues with communication. While our group was able to communicate well together, the lack of interactivity in Tableau made it hard to work all together, especially when creating dashboards. Yet, our group was able to subdivide the work efficiently and benefit from synergy while working individually as some members attempted to develop features and then met to build on each finding.

APPENDIX

Graph	Objective	Challenges
Puck Luck Graph	This graph is to identify the players more likely to regress in the future. The % of Secondary Assists [Secondary Assist / Total Assists] (Y-axis) emphasizes players that had few primary ones (more impactful and repeatable). The Shooting Percentage [Goals / Shots on Goal] (X-axis) single out high shooting percentage players (less likely to be sustained).	Had to create calculated fields for both axis.
Defensive Forwards	This graph aims to identify players with the toughest assignments. The DZone Start % [Faceoff in DZone / Total Faceoff] emphasize players with defense-first assignments, while the number of shots on the Penalty Kill emphasizes players with more minutes played at 4 VS 5 (also a tougher assignment). The size and color of the bubble represents how well they perform in these situations.	This chart only graphs centers since they are the only one taking Faceoffs and had to make sure the chart was not too clutter.
Relative to Team	This chart is to add context to the performance of players through the integration of the team statistics. In fact, by displaying the % of points from players compared to their team, it is easy to identify players on bad team which may have alter their statistics.	Doing the toggle to display the changing fields required the use of parameters + calculated fields.
+/- Relative to Team	This chart is to add context to the performance of players through the integration of the team statistic. In fact, by displaying the +/- from players compared to their team, it is easy to identify players which did much better than their teammates as they have a better goal differential than their teammates.	This visualization required much manipulation beforehand to find the team statistics and then the player relative.
Primary Points Correlation	This chart shows the performance of players on the most important dimensions for forwards (Points per 60, Assists Per 60, Primary Assists Per 60, and Secondary Assists per 60) on a scatterplot. This chart will enable to see the best player (and style) rapidly as they will appear in the top right of the scatterplot.	(see Primary Points Relative to Team Bar Chart)
Goal scorers VS playmakers	This chart enables in the glimpse of an eye to see if a player is a playmaker or goal scorer. Through a stacked bar chart showing their number of goals per 60 (in orange) and assists per 60 (in blue) one can identify the playing style of players (e.g., high proportion of goals represents a goal scorer).	Figure out the multi-axing for the stacked chart.
Offensive Impact	Our “catch-all statistic” enables users to see the overall offensive contribution of a player. This is the addition of the	A lot of data pre-processing was

	percentile rank of each player on the 3 important offensive metrics (Goals, First Assists, and Inner Slot Shots). The boxplot with percentile enables users to see the difference between individual data points (mostly through outliers).	necessary before using the data in Tableau to create our field.
Faceoff Effectiveness	The goal of this chart is to identify a new archetype: the faceoffs specialists. An underrated statistics when it comes to forward – it defines whether you start with possession or not.	Making sure the points have consistent size.
Sharp Shooters	This graph is to show who are the best pure goal scorers in the dataset by ranking them by shooting percentage [Goals / Shots on Goal] (left horizontal bar chart) and which player gets the most shots from high danger areas – more likely to score goals (right horizontal bar chart).	Ensure the readability of the side-by-side chart which can give data overload easily.
Heavy Hitters	This chart shows a new archetype of players: the players who are hard to play against. They take many penalties, they have many hits per game, you do not want to be there at the same time as they are!	Ensure the proper readability of the plot despite having different axis scale.

Table 1: Forwards Graph

Graph	**Objective**	**Challenge**
Assists Correlation	This chart is to show the performance of players on the most important dimensions for an offensive defenseman (Assists per 60, TOI Per 60, +/- Per 60) on a scatterplot. This chart will enable to see the best player rapidly as they will appear in the top right of the scatterplot.	(see Primary Points Relative to Team Bar Chart)
Scoring Chance Creator	This graph is to show who are the best defenseman at getting puck to the net – an undervalued statistic for defenseman. First, by looking at the % of shots that hit the net [Shots on Goal / Shots (left horizontal bar chart) and which player gets the most proportion of his team shots on the PP (if the PP is centered around him).	(see Forward Shots Breakdown Chart)
Defensive Usage	Time one ice – although slightly biased – is a logical statistic to use when evaluating a Defenseman because of its universality. Scouts can easily understand this statistic as it is commonly used. The data displayed as a histogram also enables us to see the stats relative to others quickly.	Adjusting the highlighter to efficiently visualize filtered player.
Defenseman Impact	Our "catch-all statistic" that enables users to see the overall contribution of a player. This is the addition of the percentile rank of each player on the 3 important metrics (Time on Ice, Primary Points, and +/-). The boxplot with percentile enables users to see the difference between individual data points (mostly comparing outliers).	A lot of data pre-processing was necessary before using the data in Tableau to create our field.

Plus/Minus (Per Game)	This chart is to show the performance of players on one of the most important dimensions for defenseman: +/-. The objective is to see who the best all-around defenseman are.	Select the most appropriate way to display the value.

Table 2: Defensemen Graph

Graph	**Description**	**Challenges**
Primary Points Breakdown Stacked Chart, Offensive Impact, Primary Points Correlation, Faceoff Effectiveness, Sharp Shooters	This dashboard is the first layer of analysis for forwards: it enables scouts to identify the best talent. However, it does so by presenting different archetype of players: The playmakers, the goals scorers, the overall offensive gem, and the Faceoff specialists.	The interconnectivity of every graph (i.e., with the filters and highlighters)

Table 3: Forwards – Scouting Dashboard

Graph	**Description**	**Challenges**
Puck Luck Graph, Defensive Forwards, Relative to Team, ¬/- Relative to Team Heavy Hitters	This dashboard is the second layer of analysis: once the best players were identified, this dashboard adds context to the data and enables to understand if the player performance is likely going to stay the same or regress.	Finding enough metrics to create a 2nd dashboard that remains insightful.

Table 4: Forwards – Player Situational Data

Graph	**Description**	**Challenges**
Defenseman Impact, Defensive Usage Assists Correlation, Scoring Chance Creator Plus/Minus (Per Game)	This dashboard enables scouts to identify the best talent. However, it does so by presenting different archetypes of players: The overall gems, the offensive defenseman, and the best defensive defenseman.	There was a lack of metrics for defenseman.

Table 5: Defensemen – Scouting Dashboard