Manipulation of Professional Hockey Database

Jiayao Wu, Section 415, jwu86@jhu.edu
Helen Collins, Section 315, hcollin8@jhu.edu

December 2017

Manipulate and query a Professional Hockey Database with specializations in data mining and complex data extraction

Database Final Project Phase II

Introduction

This project queries and manipulates data imported from Kaggle. The database contains 20 tables of statistics on Hockey leagues, games, players, and coaches from 1917 to 2011. The data is robust, and is therefore a rich source on which to query.

Development Environment

Rather than use MySQL on dbase on the ugrad machines, we downloaded MySQL to our local machines for more efficient manipulation of data files. In addition, we used Sublime Text to write .sql and .python files and Terminal to execute them.

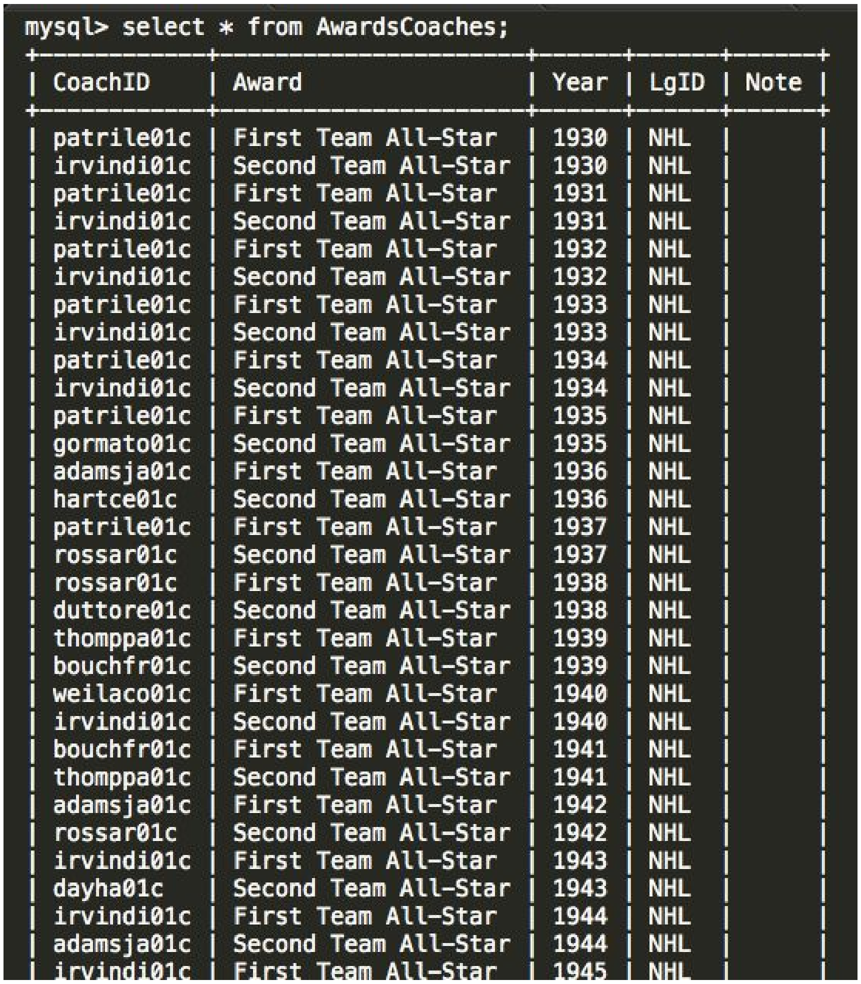
Load Dataset

The data was loaded from CSV files into tables in MySQL through the LOAD DATA LOCAL INFILE command.
An example: LOAD DATA LOCAL INFILE '/Users/HelenCollins/601.415/Abbrev.csv' INTO TABLE Abbrev FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 LINES;

Areas of specialization

Major Area of Specialization: Data Mining
We implemented multiple classification models to predict whether a player would win the First or Second Team All-Star award based on a players' seasonal statistics. We used the following machine learning algorithms: Logistic regression, SVM, Decision Tree, Naive Bayes, K-nearest neighbors and Random Forest.
Note: We modelled how we handled players with multiple positions off of an approach used by Kevin Kazmierczak in his own data mining project for predicting NHL Hall of Famers on the same dataset on Kaggle. All code was our own.

Minor Area of Specialization: Complex extraction of real data from online sources
All data was contained in CSV files for this project. We extracted only the relevant information from these CSV files into tables for this project.
Complications included: handling null values in the data, formatting large csv files properly into sql tables, handling line termination and delimiting errors appropriately, handling the large variation of variable types and sizes, researching and handling multiple warnings resulting from loading such large and complex datasets.
An example of a partial table created in MySQL from information contained in a CSV file is shown below:



```
mysql> select * from AwardsCoaches;
+------------+----------------------+------+------+------+
| CoachID    | Award                | Year | LgID | Note |
+------------+----------------------+------+------+------+
| patrile01c | First Team All-Star  | 1930 | NHL  |      |
| irvindi01c | Second Team All-Star | 1930 | NHL  |      |
| patrile01c | First Team All-Star  | 1931 | NHL  |      |
| irvindi01c | Second Team All-Star | 1931 | NHL  |      |
| patrile01c | First Team All-Star  | 1932 | NHL  |      |
| irvindi01c | Second Team All-Star | 1932 | NHL  |      |
| patrile01c | First Team All-Star  | 1933 | NHL  |      |
| irvindi01c | Second Team All-Star | 1933 | NHL  |      |
| patrile01c | First Team All-Star  | 1934 | NHL  |      |
| irvindi01c | Second Team All-Star | 1934 | NHL  |      |
| patrile01c | First Team All-Star  | 1935 | NHL  |      |
| gormato01c | Second Team All-Star | 1935 | NHL  |      |
| adamsja01c | First Team All-Star  | 1936 | NHL  |      |
| hartce01c  | Second Team All-Star | 1936 | NHL  |      |
| patrile01c | First Team All-Star  | 1937 | NHL  |      |
| rossar01c  | Second Team All-Star | 1937 | NHL  |      |
| rossar01c  | First Team All-Star  | 1938 | NHL  |      |
| duttore01c | Second Team All-Star | 1938 | NHL  |      |
| thomppa01c | First Team All-Star  | 1939 | NHL  |      |
| bouchfr01c | Second Team All-Star | 1939 | NHL  |      |
| weilaco01c | First Team All-Star  | 1940 | NHL  |      |
| irvindi01c | Second Team All-Star | 1940 | NHL  |      |
| bouchfr01c | First Team All-Star  | 1941 | NHL  |      |
| thomppa01c | Second Team All-Star | 1941 | NHL  |      |
| adamsja01c | First Team All-Star  | 1942 | NHL  |      |
| rossar01c  | Second Team All-Star | 1942 | NHL  |      |
| irvindi01c | First Team All-Star  | 1943 | NHL  |      |
| dayha01c   | Second Team All-Star | 1943 | NHL  |      |
| irvindi01c | First Team All-Star  | 1944 | NHL  |      |
| adamsja01c | Second Team All-Star | 1944 | NHL  |      |
| irvindi01c | First Team All-Star  | 1945 | NHL  |      |
```

Users Guide

1) Create a new database in MySQL.
2) In tables.sql, specify the correct path to your local CSV files (all 20 CSV files were included in the submission on Gradescope).
3) Login to MySQL, specify the database you are to use that you created in Step 1, and run table.sql to load the csv files into their appropriate tables. This step will create and populate tables in the database which you have created.
4) Run FinalHockeyQueries.sql to see the outputs for all the 15 queries.
5) Run python dataMining.py to see the outputs for each classification from our Data mining portion.
6) sqlResults.txt is an example of all of the above steps, as well as the output generated from running the queries. The Python data mining results can be found below in the "Data Mining" section of this write-up.

Project Strengths

Data Mining:

- For players with similar statistics, the models were able to accurately determine which player received the all-star award for a specific season
- Used multiple classification algorithms in order to determine which method works best for the given dataset
- Used cross validation accuracy to avoid overfitting of the training data

Efficient and accurate conversion of CSV files to sql files

Multiple SQL queries required advanced arithmetic, subqueries, and query manipulation without the use of the Join statement.

Project Weaknesses

Data mining:

- We read in the data from the CSV files rather than using a SQL package. In the future, we could incorporate more SQL into the python data mining portion.
- If a player played multiple positions in any given year, we chose the first listen position as their position for that season. We did not take into consideration any of the other positions they played as a possibility for entry into an All-Star team
    - If player was position W or F, then we chose the second option for the position because we weren't sure what those positions would necessarily translate into

Relational Table Specification

(See Phase I for complete relational table specification. We modified some parts of it)

Queries

The following 15 English queries were implemented in SQL found in the FinalHockeyQueries.sql file attached.

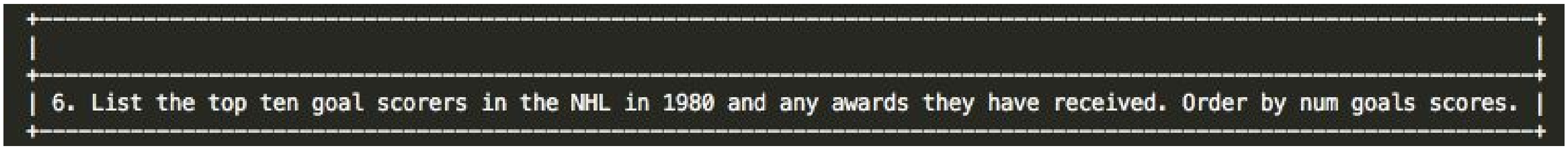
1) List the name of all Players who have a nickname and were given at least two different awards in the same year.
2) List the team name with the highest number of wins in each league from 1909 to 2011
3) List the first name, last name, and playerID of each goalie whose coach won the First Team All-Star.
4) List the coach's first name, last name and year of each team ranked #1 in the NHA(keep the years ordered).
5) List IDs of all goalies elected into the hall of fame whose coach is also in the hall of fame.
6) List the top ten goal scorers in the NHL in 1980 and any awards they have received. Order by num goals scores.
7) List the name, height and weight of the top 5 goalies with the most shutouts in their entire career. Include the number of shutouts.
8) List the players who scored the most points at the Stanley Cup whose coach is dead.

9) List the names of both players who tied for an award, as well as the year the award was given.
10) List the names of players with the most game deciding goals for every year between 2000 and 2010 and their position.
11) List the average number of shots made over all 1962 Hall of Famer players hockey careers.
12) List all teams in the NHL in 2007 who ranked higher in the second half of the season than the first half.
13) List all coaches who coach the same team that they previously played on.
14) List the average number of "goals against" per game for each goalie in 1999.
15) List the coach of the team with the player with the most awards in 2010.

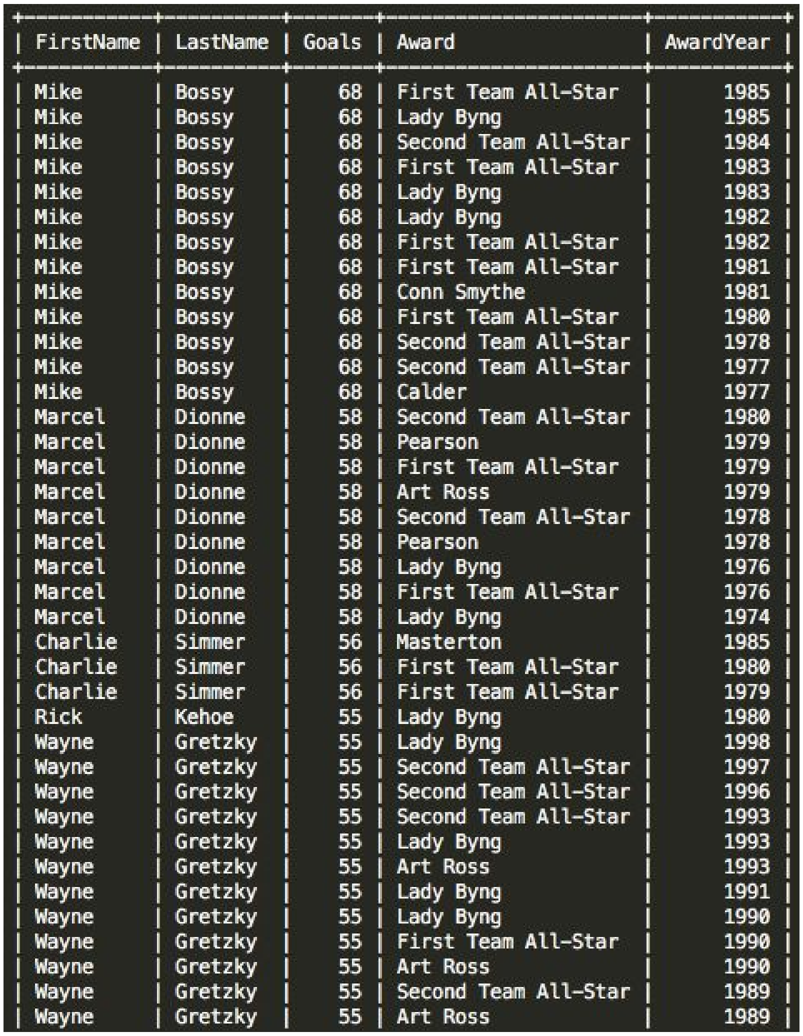
Example output for more complex queries:

Below is the output for Query 6: List the top ten goal scorers in the NHL in 1980 and any awards they have received. Order by num goals scores.

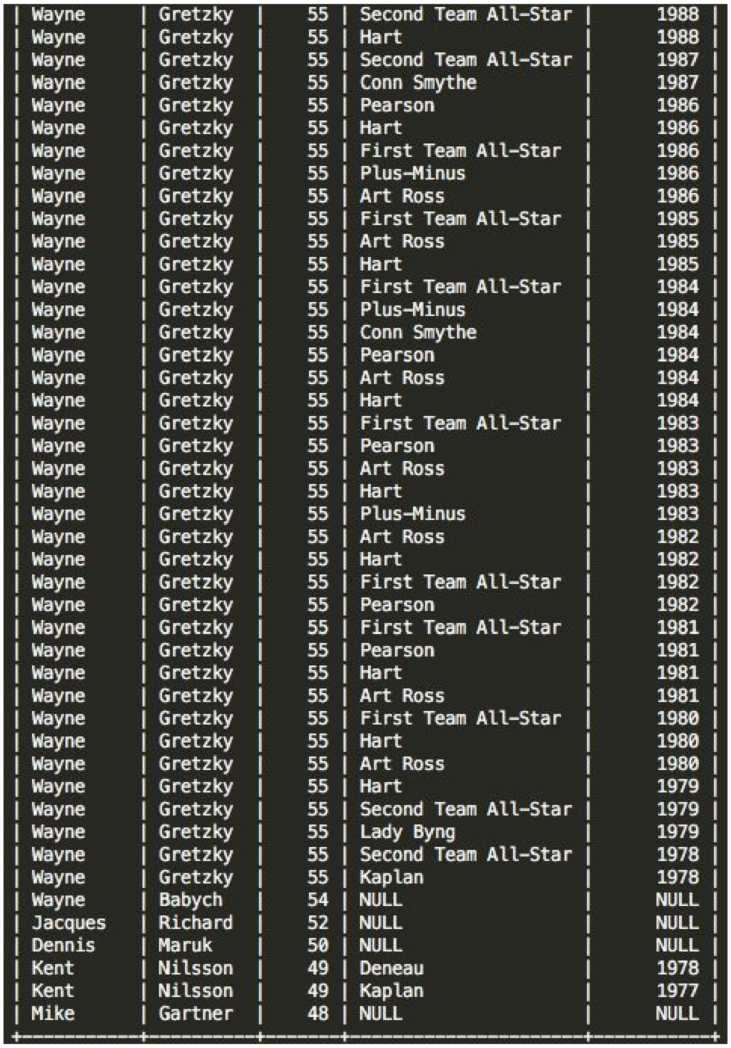
Implications of the query results: The hockey players with the highest goals scored in 1980 are usually high performers in other seasons as well (as evidenced by the number of awards received over the span of many seasons). Wayne Gretsky, though he wasn't the highest goal scorer for the 1980 season (as this was one of his first years in the NHL), wound up being one of the highest decorated hockey players in the NHL. Furthermore, some of the highest goal scorers in 1980 received no awards in their hockey career, indicating that points scored is not the only important factor in receiving an award.



```
+-------------------------------------------------------------------------------------------------------------------+
|                                                                                                                   |
+-------------------------------------------------------------------------------------------------------------------+
| 6. List the top ten goal scorers in the NHL in 1980 and any awards they have received. Order by num goals scores. |
+-------------------------------------------------------------------------------------------------------------------+
```

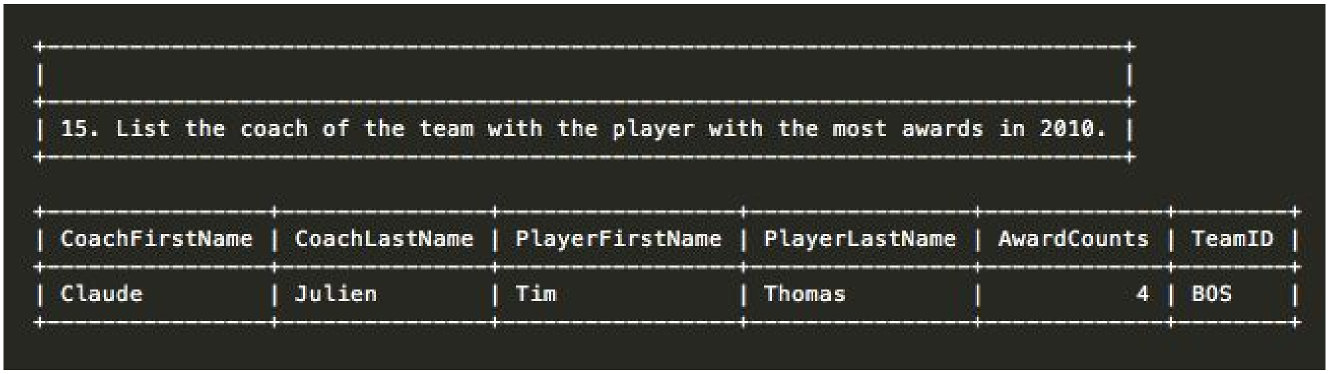


```
+-----------+----------+-------+----------------------+-----------+
| FirstName | LastName | Goals | Award                | AwardYear |
+-----------+----------+-------+----------------------+-----------+
| Mike      | Bossy    |    68 | First Team All-Star  |      1985 |
| Mike      | Bossy    |    68 | Lady Byng            |      1985 |
| Mike      | Bossy    |    68 | Second Team All-Star |      1984 |
| Mike      | Bossy    |    68 | First Team All-Star  |      1983 |
| Mike      | Bossy    |    68 | Lady Byng            |      1983 |
| Mike      | Bossy    |    68 | Lady Byng            |      1982 |
| Mike      | Bossy    |    68 | First Team All-Star  |      1982 |
| Mike      | Bossy    |    68 | First Team All-Star  |      1981 |
| Mike      | Bossy    |    68 | Conn Smythe          |      1981 |
| Mike      | Bossy    |    68 | First Team All-Star  |      1980 |
| Mike      | Bossy    |    68 | Second Team All-Star |      1978 |
| Mike      | Bossy    |    68 | Second Team All-Star |      1977 |
| Mike      | Bossy    |    68 | Calder               |      1977 |
| Marcel    | Dionne   |    58 | Second Team All-Star |      1980 |
| Marcel    | Dionne   |    58 | Pearson              |      1979 |
| Marcel    | Dionne   |    58 | First Team All-Star  |      1979 |
| Marcel    | Dionne   |    58 | Art Ross             |      1979 |
| Marcel    | Dionne   |    58 | Second Team All-Star |      1978 |
| Marcel    | Dionne   |    58 | Pearson              |      1978 |
| Marcel    | Dionne   |    58 | Lady Byng            |      1976 |
| Marcel    | Dionne   |    58 | First Team All-Star  |      1976 |
| Marcel    | Dionne   |    58 | Lady Byng            |      1974 |
| Charlie   | Simmer   |    56 | Masterton            |      1985 |
| Charlie   | Simmer   |    56 | First Team All-Star  |      1980 |
| Charlie   | Simmer   |    56 | First Team All-Star  |      1979 |
| Rick      | Kehoe    |    55 | Lady Byng            |      1980 |
| Wayne     | Gretzky  |    55 | Lady Byng            |      1998 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1997 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1996 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1993 |
| Wayne     | Gretzky  |    55 | Lady Byng            |      1993 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1993 |
| Wayne     | Gretzky  |    55 | Lady Byng            |      1991 |
| Wayne     | Gretzky  |    55 | Lady Byng            |      1990 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1990 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1990 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1989 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1989 |
```



```
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1988 |
| Wayne     | Gretzky  |    55 | Hart                 |      1988 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1987 |
| Wayne     | Gretzky  |    55 | Conn Smythe          |      1987 |
| Wayne     | Gretzky  |    55 | Pearson              |      1986 |
| Wayne     | Gretzky  |    55 | Hart                 |      1986 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1986 |
| Wayne     | Gretzky  |    55 | Plus-Minus           |      1986 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1986 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1985 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1985 |
| Wayne     | Gretzky  |    55 | Hart                 |      1985 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1984 |
| Wayne     | Gretzky  |    55 | Plus-Minus           |      1984 |
| Wayne     | Gretzky  |    55 | Conn Smythe          |      1984 |
| Wayne     | Gretzky  |    55 | Pearson              |      1984 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1984 |
| Wayne     | Gretzky  |    55 | Hart                 |      1984 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1983 |
| Wayne     | Gretzky  |    55 | Pearson              |      1983 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1983 |
| Wayne     | Gretzky  |    55 | Hart                 |      1983 |
| Wayne     | Gretzky  |    55 | Plus-Minus           |      1983 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1982 |
| Wayne     | Gretzky  |    55 | Hart                 |      1982 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1982 |
| Wayne     | Gretzky  |    55 | Pearson              |      1982 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1981 |
| Wayne     | Gretzky  |    55 | Pearson              |      1981 |
| Wayne     | Gretzky  |    55 | Hart                 |      1981 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1981 |
| Wayne     | Gretzky  |    55 | First Team All-Star  |      1980 |
| Wayne     | Gretzky  |    55 | Hart                 |      1980 |
| Wayne     | Gretzky  |    55 | Art Ross             |      1980 |
| Wayne     | Gretzky  |    55 | Hart                 |      1979 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1979 |
| Wayne     | Gretzky  |    55 | Lady Byng            |      1979 |
| Wayne     | Gretzky  |    55 | Second Team All-Star |      1978 |
| Wayne     | Gretzky  |    55 | Kaplan               |      1978 |
| Wayne     | Babych   |    54 | NULL                 |      NULL |
| Jacques   | Richard  |    52 | NULL                 |      NULL |
| Dennis    | Maruk    |    50 | NULL                 |      NULL |
| Kent      | Nilsson  |    49 | Deneau               |      1978 |
| Kent      | Nilsson  |    49 | Kaplan               |      1977 |
| Mike      | Gartner  |    48 | NULL                 |      NULL |
+-----------+----------+-------+----------------------+-----------+
```

Below is the output for Query 15: List the coach of the team with the player with the most awards in 2010. Implications of the query results: This was a complicated query to execute. It required nested queries and a more complicated utilization of the tables in our database in order to extract the exact information requested. The final result is found below, with the name of the player as well as the number of awards they won in 2010, along with the coaches name and the team they both belong to.



```
+------------------------------------------------------------------------+
|                                                                        |
+------------------------------------------------------------------------+
| 15. List the coach of the team with the player with the most awards in 2010. |
+------------------------------------------------------------------------+

+----------------+---------------+-----------------+----------------+-------------+--------+
| CoachFirstName | CoachLastName | PlayerFirstName | PlayerLastName | AwardCounts | TeamID |
+----------------+---------------+-----------------+----------------+-------------+--------+
| Claude         | Julien        | Tim             | Thomas         |           4 | BOS    |
+----------------+---------------+-----------------+----------------+-------------+--------+
```

Data Mining

Our goal for this portion of the project was to attempt to predict which NHL players would make First or Second Team All-Star given seasonal statistics on Games Played (GP), Goals (G), Assists (A), Points (Pts) and Penalty Minutes (PIM) from the Scoring and AwardsPlayers schemas. We selected relevant columns and dropped NULL values for cleaner datasets. If a player was documented as playing more than one positions for a given season, then we chose the first position as their position for that season. Then for each player PID in the year Y at the position P, in the AwardsPlayers table, if we found a row of this PID in year Y at position P winning First Team All-Star, then we labelled this player as 1; else if we found a row of this PID in year Y at position P winning Second Team All-Star, then we labelled this player as 2; else we labelled this player as 0 (meaning neither First or Second Team All-Star).

Here is a sample row of the matrix which we trained and tested:

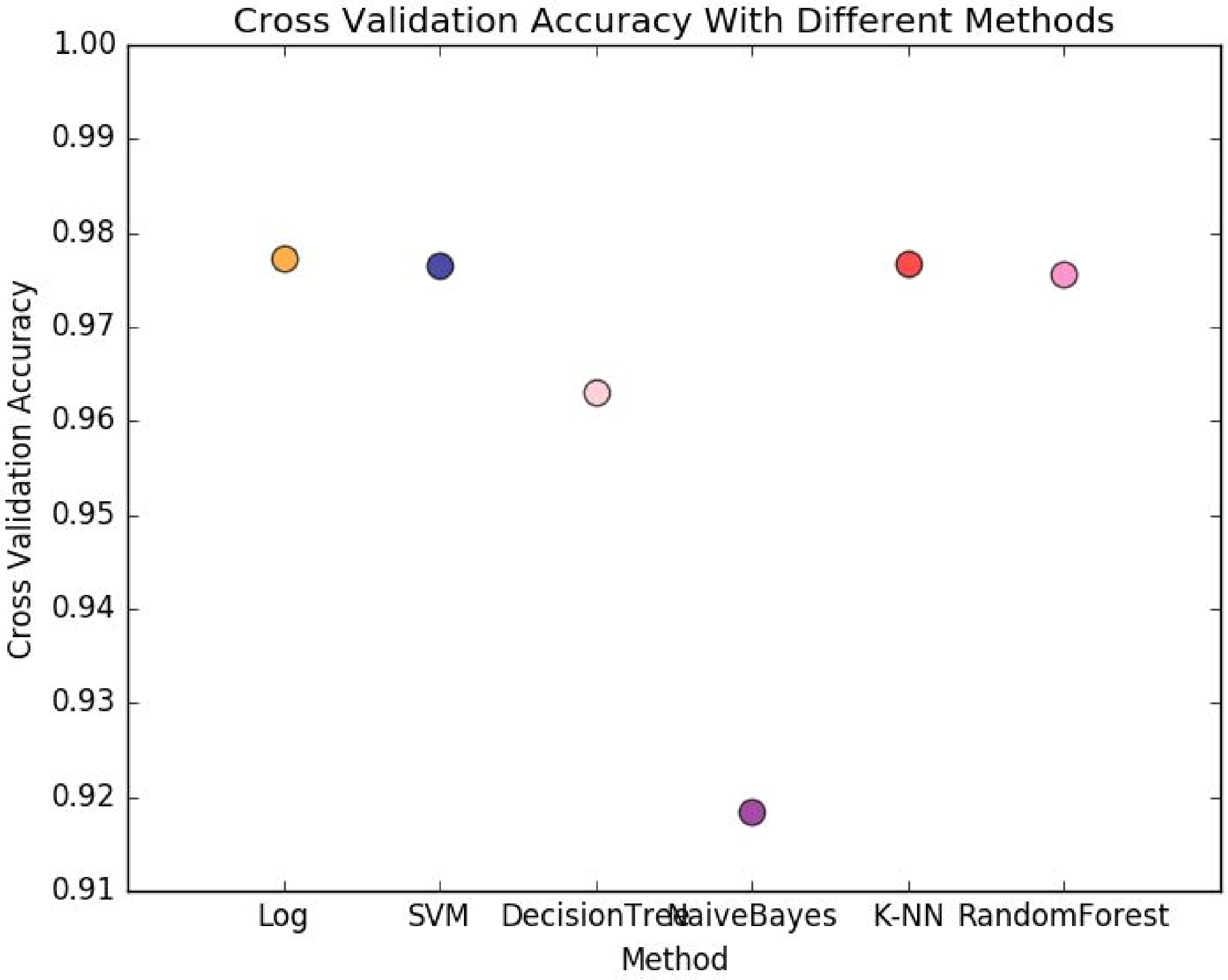
PlayerID	GP	G	A	Pts	PIM	Label
aaltoan01	73	3	5	8	24	0

This matrix was then split into 70% training and 30% testing. We then applied different classification algorithms: Logistic regression, SVM, Decision Tree Classifier, Naive Bayes, K-Nearest Neighbors and Random Forest.

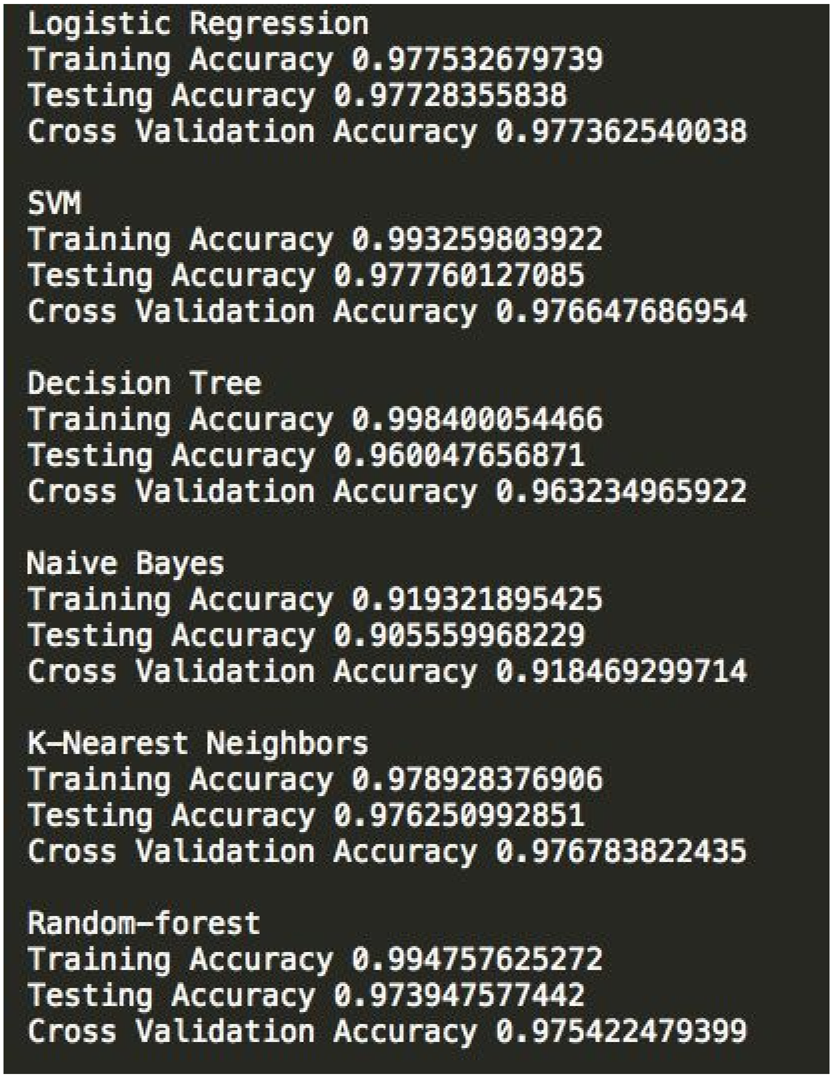
The following results are the accuracies from the different classification algorithms:

Accuracy Summary for Different Classification Algorithms						
	Logistic Regression	SVM	Decision Tree	Naive Bayes	K-Nearest Neighbors	Random Forest
Training Accuracy	0.9775	0.9933	0.9984	0.9193	0.9789	0.9948
Testing Accuracy	0.9773	0.9778	0.9600	0.9056	0.9763	0.9740
Cross Validation Accuracy	0.9774	0.9766	0.9632	0.9185	0.9768	0.9754

The following plots the cross-validation accuracy of the classification algorithms to visually demonstrate which algorithms were more accurate at predicting which hockey players would be awarded First-Team All-Star or Second-Team All-Star. It is clear from this plot that the classifications with the highest accuracy were: Logistic Regression, SVM, K-nearest Neighbors and Random Forest, which suggests there is no overfitting. Naive Bayes did not perform well as expected. The reason might be that Naive Bayes assumes that any feature is independent to the value of any other features. This might not be the case with our dataset because features such as GP, G and PIM could be correlated with each other. Perhaps using Feature Extraction technique such as PCA that transforms data into linearly uncorrelated principal components can help increase accuracy for Naive Bayes.



The raw output from the dataMining.py file:



```
Logistic Regression
Training Accuracy 0.977532679739
Testing Accuracy 0.97728355838
Cross Validation Accuracy 0.977362540038

SVM
Training Accuracy 0.993259803922
Testing Accuracy 0.977760127085
Cross Validation Accuracy 0.976647686954

Decision Tree
Training Accuracy 0.998400054466
Testing Accuracy 0.960047656871
Cross Validation Accuracy 0.963234965922

Naive Bayes
Training Accuracy 0.919321895425
Testing Accuracy 0.905559968229
Cross Validation Accuracy 0.918469299714

K-Nearest Neighbors
Training Accuracy 0.978928376906
Testing Accuracy 0.976250992851
Cross Validation Accuracy 0.976783822435

Random-forest
Training Accuracy 0.994757625272
Testing Accuracy 0.973947577442
Cross Validation Accuracy 0.975422479399
```

Extensions to this Project

- Parameter tuning for the machine learning algorithms. For example, for SVM, different kernel functions can be implemented which may give us even higher accuracy.
- Further predictions using data mining such as: predicting which team will win the Stanley Cup based off of a season's team/player statistics, Predicting the performance of a team in the second half of a season based off of the first half of a season's statistics
- Natural Language query processing. Handling sentence queries of our database, including parsing the sentence and returning the requested information back to the user.

Additional Notes:

Database Final Project Phase I

1. Team members:

Jiayao Wu (jwu86)
Helen Collins (hcollin8)

2. Target Domain

Professional Hockey Database from Kaggle
Link: https://www.kaggle.com/open-source-sports/professional-hockey-database/data

3. English Queries (15)

1) List the name of all Players who have a nickname and were given at least two different awards in the same year.
2) List the team name with the highest number of wins in each league from 1909 to 2011
3) List the first name, last name, and playerID of each goalie whose coach won the First Team All-Star.
4) List the coach's first name, last name and year of each team ranked #1 in the NHA (keep the years ordered).
5) List IDs of all goalies elected into the hall of fame whose coach is also in the hall of fame.
6) List the top ten goal scorers in the NHL in 1980 and any awards they have received. Order by num goals scores.
7) List the name, height and weight of the top 5 goalies with the most shutouts in their entire career. Include the number of shutouts.
8) List the players who scored the most points at the Stanley Cup whose coach is dead.
9) List the names of both players who tied for an award, as well as the year the award was given.
10) List the names of players with the most game deciding goals for every year between 2000 and 2010 and their position.
11) List the average number of shots made over all 1962 Hall of Famer players hockey careers.
12) List all teams in the NHL in 2007 who ranked higher in the second half of the season than the first half.
13) List all coaches who coach the same team that they previously played on.
14) List the average number of "goals against" per game for each goalie in 1999.
15) List the coach of the team with the player with the most awards in 2010.

4. Relational Data Model with primary keys

Underlined attribute is the primary key. Since all the referential attributes' names are the same (e.g. CoachID, PlayerID), we don't have any referential constraints.

Abbrev: Abbreviations used in Teams and SeriesPost tables

Type	Code	Fullname
Playoffs	NHAF	Lost NHA finals

AwardsCoaches: Coaches awards, trophies, postseason all-star teams

CoachID	Award	Year	LgID	Note
patrile01c	First Team All-Star	1930	NHL	Null

AwardsPlayers: Player awards, trophies, postseason all-star teams

PlayerID	Award	Year	LgID	Note	Pos
malonjo01	Art Ross	1917	NHL	Null	Null

Coaches: Coaching statistics

CoachID	Year	TmID	LgID	Stint	Notes	G	W	L	T	Postg	Postw	Postl	Postt
abelsi01c	1952	CHI	NHL	1		70	27	28	15	7	3	4	0

CombinedShutouts: List of combined shutouts.

Year	Month	Date	TmID	OppID	RorP	IDgoalie1	IDgoalie2
1929	3	14	TOR	NYA	R	chabolo01	grantbe01

Goalies: Goaltending statistics

PlayerID	Year	TmID	LgID	GP	Min	W	L	TorOL	SHO	ENG	SHO	GA	SA
benedcl01	1914	OT1	NHA	3	180	0	3	0	0	26	0	0	0

GoaliesSC: Goaltending for Stanley Cup finals, 1917-18 through 1925-26

PlayerID	Year	TmID	LgID	GP	Min	W	L	T	SHO	GA
benedcl01	1914	OT1	NHA	3	180	0	3	0	0	26

GoaliesShootout: Goaltending statistics for shootouts

PlayerID	Year	Stint	TmID	W	L	SA	GA
aebisda01	2005	1	COL	2	1	10	2

HOF:Hall of Fame information

Year	HofID	Name	Category
1945	bakerho01h	Hobey Baker	Player

Master: Names and biographical information (removed some columns in the original csv)

PlayerID	CoachID	HofID	FirstName	LastName	NameNote	NameGiven	NameNick	Height	Weight	ShootCatch	LegendsID	FirstNHL	LastNHL	FirstWHA	LastWHA	Pos	BirthYear	BirthMon	BirthDay	BirthCountry	BirthState	BirthCity	DeathYear	DeathMon	DeathDay	DeathCountry	DeathState	DeathCity
abelsi01	abelsi01c	abelsi01h	Sid	Abel	Null	Antti	BootNose	71	170	L	P196901	1938	1953	null	null	C	1918	2	22	Canada	SK	Melville	null	null	null	null	null	null

Scoring: Scoring statistics (removed some columns in the original csv)

PlayerID	Year	Stint	TmID	LgID	Pos	GP	G	A	Pts	PIM	PlusOrMinus	PPG	PPA	SHG	SHA	GWG	GTG	SOG
aaltoan01	1997	1	ANA	NHL	C	3	0	0	0	0	0	0	0	0	0	0	0	0

ScoringSC: Scoring for Stanley Cup finals, 1917-18 through 1925-26

PlayerID	Year	TmID	LgID	Pos	GP	G	A	Pts	PIM
adamsbi01	1920	VML	PCHA	R	4	0	0	0	0

ScoringShootout: Scoring statistics for shootouts

PlayerID	Year	Stint	TmID	S	G	GDG
adamske01	2006	1	PHO	1	0	0

Teams: Team regular season statistics (removed some columns in the original csv)

Year	LgID	TmID	FranchID	ConfID	DivID	Rank	Playoff	G	W	L	T	OTL	Pts	SoW	SoL	GF	GA	Name	PIM	BenchMinor	PPG	PPC	SHA	PKG	PKC	SHF
2006	NHL	MIN	MIN	WC	NW	2	CQF	82	48	26	0	8	104	10	7	235	191	Minnesota Wild	862	12	72	380	8	48	342	9

TeamsHalf: First half / second half standings, 1917-18 through 1920-21

Year	LgID	TmID	Half	Rank	G	W	L	T	GF	GA
1916	NHA	MOC	1	1	10	7	3	0	58	38

TeamsSC: Team Stanley Cup finals statistics, 1917-18 through 1925-26

Year	LgID	TmID	G	W	L	T	GF	GA	PIM
1918	NHL	MTL	5	2	2	1	10	19	24

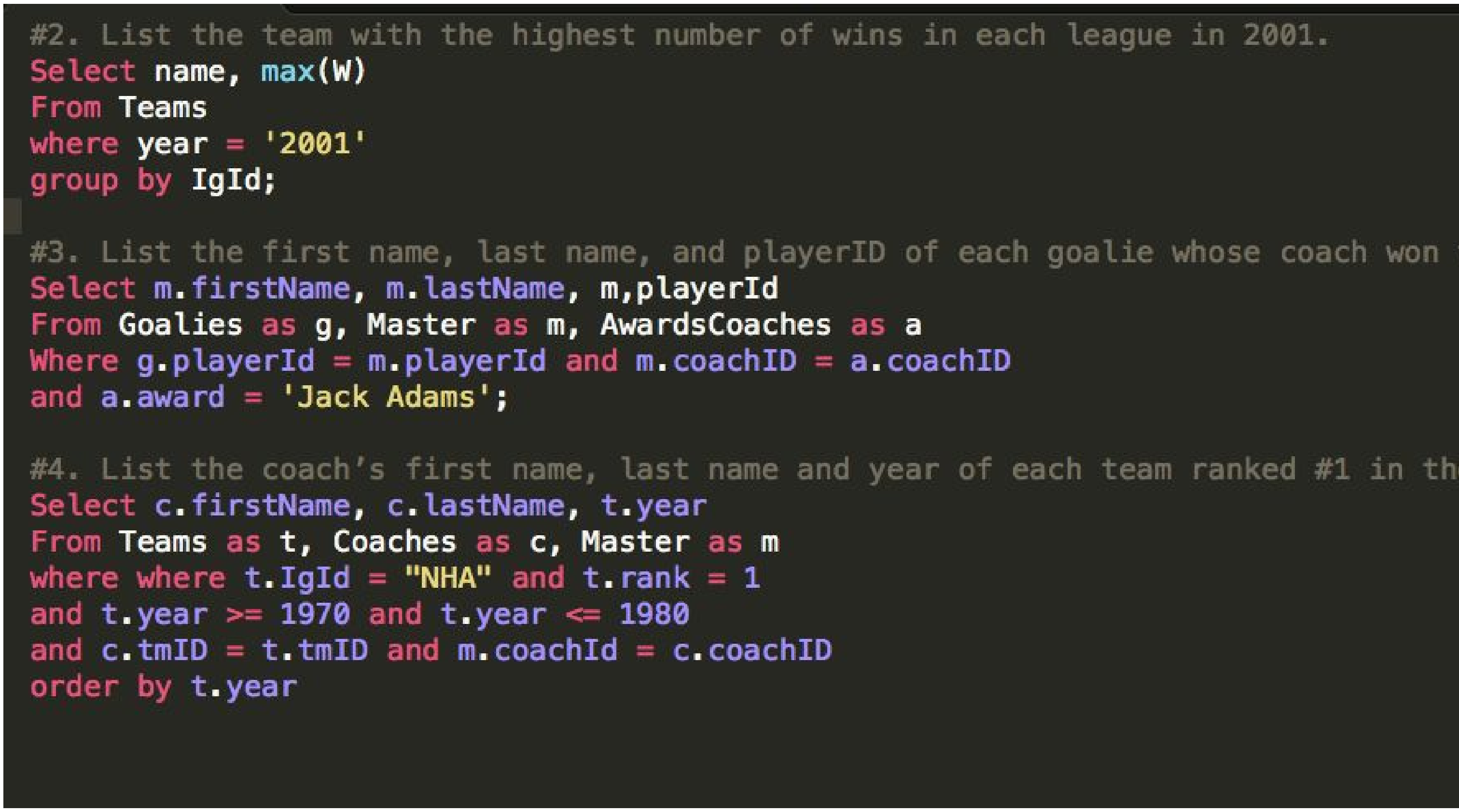
TeamVsTeam: Team vs. team results

Year	LgID	TmID	OppID	W	L	T	OTL
1909	NHA	COB	HAI	1	1	0	

Tables ignored for now (but we may use them in Phase II):

- SeriesPost: Postseason series
- TeamsPost: Team postseason statistics
- TeamSplits: Team home/road and monthly splits

5. SQL statements



```
#2. List the team with the highest number of wins in each league in 2001.
Select name, max(W)
From Teams
where year = '2001'
group by IgId;

#3. List the first name, last name, and playerID of each goalie whose coach won
Select m.firstName, m.lastName, m,playerId
From Goalies as g, Master as m, AwardsCoaches as a
Where g.playerId = m.playerId and m.coachID = a.coachID
and a.award = 'Jack Adams';

#4. List the coach’s first name, last name and year of each team ranked #1 in th
Select c.firstName, c.lastName, t.year
From Teams as t, Coaches as c, Master as m
where where t.IgId = "NHA" and t.rank = 1
and t.year >= 1970 and t.year <= 1980
and c.tmID = t.tmID and m.coachId = c.coachID
order by t.year
```

6. Plan on how to load values

Again, here is the link to our database:
https://www.kaggle.com/open-source-sports/professional-hockey-database/data

Since the data sources are all in csv, we need to convert from csv into mysql. ‘mysqlimport’ function will do this for us (reference: https://dev.mysql.com/doc/refman/5.7/en/mysqlimport.html) One of the challenges we might face is that some values are null in the database, so we need to see if this will cause any issues. Also, we need to consider how to deal with these null values when generating output using SQL code.

7. Special plans with output/views

If we find some queries that need the same join of same tables, then we might create a few views for these joins.

As for the output, we will format it nicely. Maybe we will try to implement some HTML and PHP (just like assignment 3) so that the outputs are formatted nicely.

8. Specialized/advanced topics to focus

- Major advanced Topics to focus on: Data mining, i.e. prediction using regression, clustering etc.
- Minor advanced topics to focus on: Complex data extraction (since our raw files are csv files)