ETL PROJECT

GROUP 4: MARK BURTON, JAMMY LO, SCOTT FRAZIER, HOA ROACH

Topic: Scraping the data of hockey players in skater position from National Hockey League (NHL) website. The relevant data of players, player stats and teams will be pulled from website, transformed and then loaded to cloud database for public access

Data Source:

Hockey players data: http://www.espn.com/nhl/statistics/player/

NHL Player Points Statistics - 2019-20

Statistics: Points | Shooting | Goaltending | Defensive | Time On Ice | Faceoffs | Major Penalties | Minor Penalties

Season: 2019-20 Regular Season League: NHL

Splits: Total Positions: All

Points Leaders - All Players

RK	PLAYER	TEAM	GP	G	A	PTS	+/-	PIM	PTS/G	SOG	PCT	GWG	PP G	PP A	SH G	SH A
1	Leon Draisaitl, C	EDM	71	43	67	110	-7	18	1.55	218	19.7	10	16	28	0	0
2	Connor McDavid, C	EDM	64	34	63	97	-6	28	1.52	212	16.0	6	11	32	0	0
3	David Pastrnak, RW	BOS	70	48	47	95	21	40	1.36	279	17.2	10	20	18	0	0
	Artemi Panarin, LW	NYR	69	32	63	95	36	20	1.38	209	15.3	4	7	17	0	0
5	Nathan MacKinnon, C	COL	69	35	58	93	13	12	1.35	318	11.0	4	12	19	0	0
6	Brad Marchand, LW	BOS	70	28	59	87	25	82	1.24	185	15.1	5	5	23	1	2
7	Nikita Kucherov, RW	TB	68	33	52	85	26	38	1.25	210	15.7	6	4	21	0	0
8	Patrick Kane, RW	CHI	70	33	51	84	8	40	1.20	275	12.0	2	8	15	0	0
9	Auston Matthews, C	TOR	70	47	33	80	19	8	1.14	290	16.2	5	12	13	0	0
10	Jack Eichel, C	BUF	68	36	42	78	5	34	1.15	227	15.9	9	11	16	1	0

RK	PLAYER	TEAM	GP	G	A	PTS	+/-	PIM	PTS/G	SOG	PCT	GWG	PP G	PP A	SH G	SH A
	Jonathan Huberdeau, C	FLA	69	23	55	78	5	30	1.13	152	15.1	3	5	24	0	0
12	Mika Zibanejad, C	NYR	57	41	34	75	9	14	1.32	208	19.7	6	15	12	3	2
	John Carlson, D	WSH	69	15	60	75	12	26	1.09	189	7.9	6	2	24	0	0
14	Evgeni Malkin, C	PIT	55	25	49	74	7	58	1.35	171	14.6	5	7	17	0	0
15	Kyle Connor, LW	WPG	71	38	35	73	4	34	1.03	239	15.9	7	9	8	1	1
	Mark Scheifele, C	WPG	71	29	44	73	2	45	1.03	170	17.1	6	10	10	0	0
17	J.T. Miller, C	VAN	69	27	45	72	11	47	1.04	165	16.4	3	9	16	0	2
18	Alex Ovechkin, LW	WSH	68	48	19	67	-12	30	0.99	311	15.4	3	13	5	0	0
	Mitch Marner, C	TOR	59	16	51	67	6	16	1.14	154	10.4	2	6	18	0	1
20	Sebastian Aho, LW	CAR	68	38	28	66	10	26	0.97	206	18.4	5	8	9	4	1

RK	PLAYER	TEAM	GP	G	A	PTS	+/-	PIM	PTS/G	SOG	PCT	GWG	PP G	PP A	SH G	SH A
	Max Pacioretty, LW	VGS	71	32	34	66	18	44	0.93	307	10.4	5	8	11	0	0
	Steven Stamkos, C	TB	57	29	37	66	14	22	1.16	176	16.5	6	10	9	0	0
	Elias Pettersson, C	VAN	68	27	39	66	16	18	0.97	162	16.7	5	8	16	0	0
24	Blake Wheeler, RW	WPG	71	22	43	65	1	37	0.92	180	12.2	2	4	18	1	1
	Roman Josi, D	NSH	69	16	49	65	22	41	0.94	260	6.2	1	4	19	0	0
26	Brayden Point, C	TB	66	25	39	64	28	11	0.97	141	17.7	4	8	5	0	0
27	Patrik Laine, RW	WPG	68	28	35	63	8	22	0.93	226	12.4	1	8	8	0	0
	Mark Stone, RW	VGS	65	21	42	63	15	27	0.97	168	12.5	3	6	11	0	1
	Teuvo Teravainen, LW	CAR	68	15	48	63	20	8	0.93	182	8.2	1	4	17	1	3
30	Anze Kopitar, C	LA	70	21	41	62	6	16	0.89	135	15.6	5	7	14	1	0

RK	PLAYER	TEAM	GP	G	A	PTS	+/-	PIM	PTS/G	SOG	PCT	GWG	PP G	PP A	SH G	SH A
	Aleksander Barkov, C	FLA	66	20	42	62	2	18	0.94	172	11.6	2	7	13	0	1
32	Travis Konecny, C	PHI	66	24	37	61	-1	28	0.92	141	17.0	3	5	18	0	0
	Andrei Svechnikov, RW	CAR	68	24	37	61	9	54	0.90	183	13.1	5	6	14	0	0
	Matthew Tkachuk, LW	CGY	69	23	38	61	-5	74	0.88	188	12.2	4	5	14	0	0
	Tomas Tatar, LW	MTL	68	22	39	61	5	36	0.90	162	13.6	1	8	6	0	0
	Ryan Nugent-Hopkins, C	EDM	65	22	39	61	1	33	0.94	172	12.8	4	7	17	0	0
	Ryan O'Reilly, C	STL	71	12	49	61	11	10	0.86	118	10.2	3	2	16	1	0
38	John Tavares, C	TOR	63	26	34	60	-7	24	0.95	197	13.2	4	7	14	0	0
	David Perron, LW	STL	71	25	35	60	2	52	0.85	166	15.1	9	9	18	0	0
	Mathew Barzal, C	NYI	68	19	41	60	5	44	0.88	171	11.1	2	4	8	0	0

798 Results 1 of 20

Hockey teams and abbreviation: https://www.kaggle.com/martinellis/nhl-game-data

Hockey teams: https://en.wikipedia.org/wiki/National_Hockey_League

List of teams [edit]

Division	Team	City	Arena	Capacity	Founded	Joined	General manager	Head coach	Captain
Eastern Conference									
Atlantic	Boston Bruins	Boston, Massachusetts	TD Garden	17,850	1924		Don Sweeney	Bruce Cassidy	Zdeno Chara
	Buffalo Sabres	Buffalo, New York	KeyBank Center	19,070	1970		Kevyn Adams	Ralph Krueger	Jack Eichel
	Detroit Red Wings	Detroit, Michigan	Little Caesars Arena	19,515	1926		Steve Yzerman	Jeff Blashill	Vacant
	Florida Panthers	Sunrise, Florida	BB&T Center	19,250	1993		Bill Zito	Joel Quenneville	Aleksander Barkov
	Montreal Canadiens	Montreal, Quebec	Bell Centre	21,302	1909	1917	Marc Bergevin	Claude Julien	Shea Weber
	Ottawa Senators	Ottawa, Ontario	Canadian Tire Centre	18,652	1992		Pierre Dorion	D. J. Smith	Vacant
	Tampa Bay Lightning	Tampa, Florida	Amalie Arena	19,092	1992		Julien BriseBois	Jon Cooper	Steven Stamkos
	Toronto Maple Leafs	Toronto, Ontario	Scotiabank Arena	18,819	1917		Kyle Dubas	Sheldon Keefe	John Tavares
Metropolitan	Carolina Hurricanes	Raleigh, North Carolina	PNC Arena	18,680	1972	1979*	Don Waddell	Rod Brind'Amour	Jordan Staal
	Columbus Blue Jackets	Columbus, Ohio	Nationwide Arena	18,144	2000		Jarmo Kekalainen	John Tortorella	Nick Foligno
	New Jersey Devils	Newark, New Jersey	Prudential Center	16,514	1974*		Tom Fitzgerald	Lindy Ruff	Vacant
	New York Islanders	Uniondale, New York	Nassau Coliseum	13,917	1972		Lou Lamoriello	Barry Trotz	Anders Lee
	New York Rangers	New York City, New York	Madison Square Garden	18,006	1926		Jeff Gorton	David Quinn	Vacant
	Philadelphia Flyers	Philadelphia, Pennsylvania	Wells Fargo Center	19,500	1967		Chuck Fletcher	Alain Vigneault	Claude Giroux
	Pittsburgh Penguins	Pittsburgh, Pennsylvania	PPG Paints Arena	18,387	1967		Jim Rutherford	Mike Sullivan	Sidney Crosby
	Washington Capitals	Washington, D.C.	Capital One Arena	18,506	1974		Brian MacLellan	Peter Laviolette	Alexander Ovechkin
Western Conference									
Central	Chicago Blackhawks	Chicago, Illinois	United Center	19,717	1926		Stan Bowman	Jeremy Colliton	Jonathan Toews
	Colorado Avalanche	Denver, Colorado	Ball Arena	18,007	1972	1979*	Joe Sakic	Jared Bednar	Gabriel Landeskog
	Dallas Stars	Dallas, Texas	American Airlines Center	18,532	1967*		Jim Nill	Rick Bowness	Jamie Benn
	Minnesota Wild	Saint Paul, Minnesota	Xcel Energy Center	17,954	2000		Bill Guerin	Dean Evason	Vacant
	Nashville Predators	Nashville, Tennessee	Bridgestone Arena	17,113	1998		David Poile	John Hynes	Roman Josi
	St. Louis Blues	St. Louis, Missouri	Enterprise Center	18,724	1967		Doug Armstrong	Craig Berube	Vacant
	Winnipeg Jets	Winnipeg, Manitoba	Bell MTS Place	15,321	1999*		Kevin Cheveldayoff	Paul Maurice	Blake Wheeler
Pacific	Anaheim Ducks	Anaheim, California	Honda Center	17,174	1993		Bob Murray	Dallas Eakins	Ryan Getzlaf
	Arizona Coyotes[fn 3]	Glendale, Arizona	Gila River Arena	17,125	1972	1979*	Bill Armstrong	Rick Tocchet	Oliver Ekman-Larsson
	Calgary Flames	Calgary, Alberta	Scotiabank Saddledome	19,289	1972*		Brad Treliving	Geoff Ward	Mark Giordano
	Edmonton Oilers	Edmonton, Alberta	Rogers Place	18,347	1972	1979	Ken Holland	Dave Tippett	Connor McDavid
	Los Angeles Kings	Los Angeles, California	Staples Center	18,230	1967		Rob Blake	Todd McLellan	Anze Kopitar
	San Jose Sharks	San Jose, California	SAP Center	17,562	1991		Doug Wilson	Bob Boughner	Logan Couture
	Vancouver Canucks	Vancouver, British Columbia	Rogers Arena	18,910	1945	1970	Jim Benning	Travis Green	Bo Horvat
	Vegas Golden Knights	Paradise, Nevada	T-Mobile Arena	17,356	2017		Kelly McCrimmon	Peter DeBoer	Vacant

Extract:

Our original data sources included gathering data the NHL section on the official ESPN website. From this website we found a data table consisting of a plethora of data pertaining to the many available statistics for the top players that are currently in the NHL. This data was then scraped to our jupyter notebook by utilizing html, and xml after finding the source code for the table on the ESPN website in the html inspect tool.

The hockey team's information is easily found in Wikipedia website. There are total of 31 teams. We then convert the table to csv file for importing and transforming the data.

Transform:

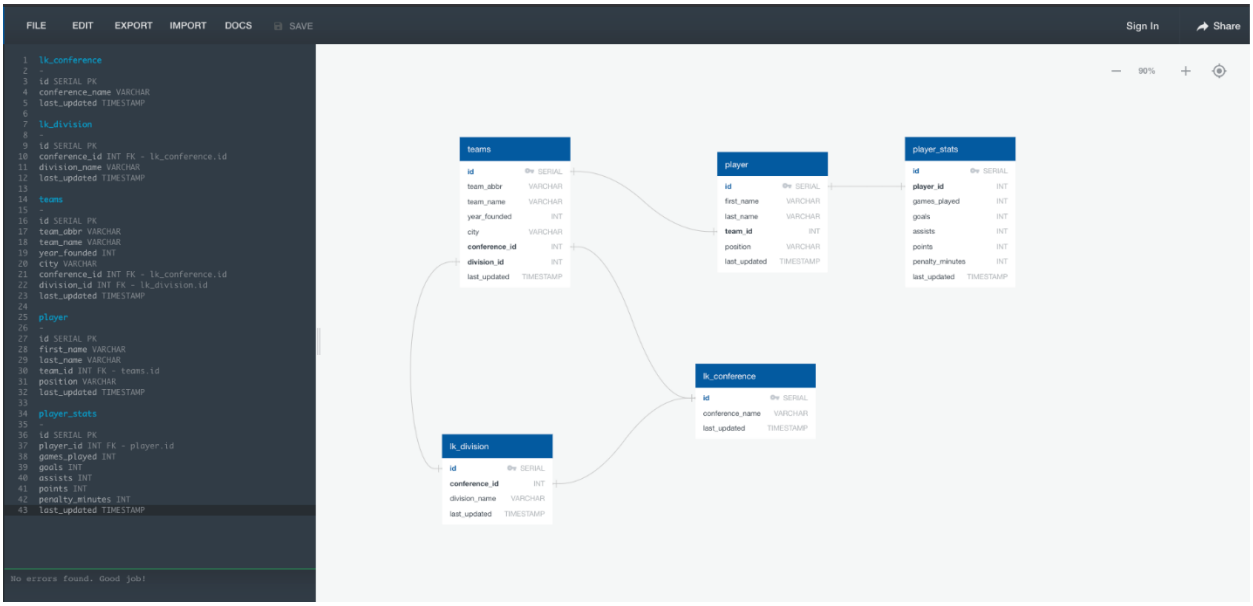
For the web scraping portion of this project, we cleaned our data by pulling in all rows and columns of data from a specific data table on the website we were scraping from. We then decided to create our final table by utilizing seven columns of data and dropping the remaining ten columns, because we would not be utilizing those in our final table. Once we had all of the desirable data, we continued to parse through the data by creating new columns such as a first name, last name, and position column from the original "Player" column which contained all of this information originally. To finish off the web scraping data table, we finally rearranged the data in the desired order that would make the most sense to anyone who viewed the data, such as putting the first and last name columns at the beginning of the table. The final data

scraped from website then saved to 2 tables of player information and player stats. These tables then saved to csv files for transformation and upload.

Most of the transformation work requires us to join different tables and rename the columns for appropriate data structures. Some of columns require extracting the certain characters from columns.

Load:

Final database that we created has the following structure represented in ERD diagram:



The tables are all uploaded and viewed in pgAdmin as follow:

postgres/postgres@smu-etl-project

Query Editor | Query History

```sql
select
    p.first_name,
    p.last_name,
    p.position,
    t.team_abbr
from
    player as p
join teams as t
    on p.team_id = t.id
where
    t.team_abbr = 'DAL'
```

Data Output | Explain | Messages | Notifications

	first_name character varying	last_name character varying	position character varying	team_abbr character varying
1	Tyler	Seguin	C	DAL
2	Jamie	Benn	LW	DAL
3	Miro	Heiskanen	D	DAL
4	Alexander	Radulov	RW	DAL
5	Roope	Hintz	LW	DAL
6	John	Klingberg	D	DAL
7	Joe	Pavelski	C	DAL
8	Denis	Gurianov	RW	DAL
9	Esa	Lindell	D	DAL
10	Jason	Dickinson	C	DAL
11	Mattias	Janmark	C	DAL
12	Corey	Perry	RW	DAL
13	Radek	Faksa	C	DAL
14	Blake	Comeau	LW	DAL
15	Andrew	Cogliano	C	DAL
16	Jamie	Oleksiak	D	DAL
17	Taylor	Fedun	D	DAL
18	Andrej	Sekera	D	DAL

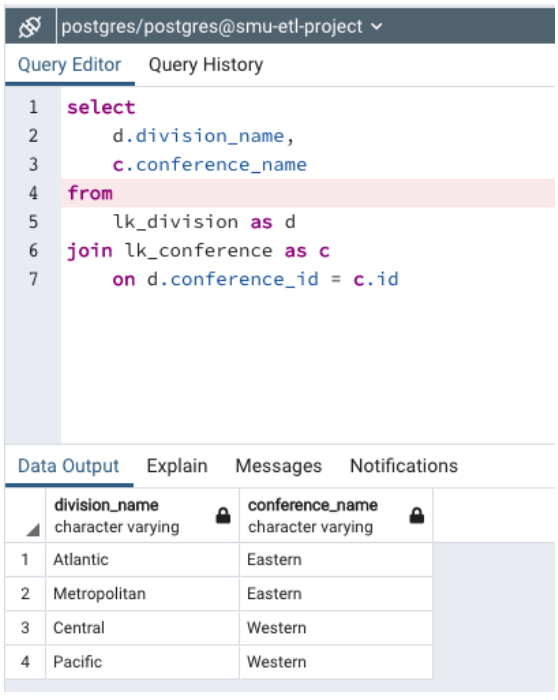
postgres/postgres@smu-etl-project

Query Editor | Query History

```sql
select
    first_name,
    last_name,
    position,
    team_id,
    games_played,
    goals,
    assists,
    points
    penalty_minutes
from
    player
join player_stats as ps
    on player.id = ps.player_id
```

Data Output | Explain | Messages | Notifications

	first_name character varying	last_name character varying	position character varying	team_id integer	games_played integer	goals integer	assists integer	penalty_minutes integer
1	Leon	Draisaitl	C	27	71	43	67	110
2	Connor	McDavid	C	27	64	34	63	97
3	David	Pastrnak	RW	1	70	48	47	95
4	Artemi	Panarin	LW	13	69	32	63	95
5	Nathan	MacKinnon	C	18	69	35	58	93
6	Brad	Marchand	LW	1	70	28	59	87
7	Nikita	Kucherov	RW	7	68	33	52	85
8	Patrick	Kane	RW	17	70	33	51	84
9	Auston	Matthews	C	8	70	47	33	80
10	Jack	Eichel	C	2	68	36	42	78
11	Jonathan	Huberdeau	C	4	69	23	55	78
12	Mika	Zibanejad	C	13	57	41	34	75
13	John	Carlson	D	16	69	15	60	75
14	Evgeni	Malkin	C	15	55	25	49	74
15	Kyle	Connor	LW	23	71	38	35	73
16	Mark	Scheifele	C	23	71	29	44	73



Bonus:

We also worked on building the Flask API for all our database tables. From the home page, we add the links for each table as easy access. Moreover, if you want to query the teams under specific division, you can dynamically operate it by adding the division id on the URL.