Predicting Skaters' Plus Minus Using Key Statistics

MATH2319 - Machine Learning
Course Project

Ben Cole - s3412349

Print Date: 20/05/2019

Contents

1 Phase 1 - Introduction, Cleaning, and Exploration

1.1 Outline

Sports statistics is an area where many data points are collected that contribute to a wide range of statistics. Primarily, two main objectives arise for the use of these statistics; to either predict a result of a given match or to quantify the performance of a given participant. One such sport where a large amount of data is collected is the National Hockey League (here on **NHL**) in North America. Game statistics are collected for both skaters and goaltenders for a range of performance indicators ranging from hits/body checks to face-off wins to short-handed goals for and against. However, not all of these statistics are easily compared across players from differing positions. Comparing goals blocked between a defensemen and a left wing is unfair, whilst comparing penalty minutes may be a more accurate comparison to see which skaters are costing the team. Points are simply a tally of the goals and assists generated by a skater, and so may be weighted towards the attacking lines. As such, it would seem that Plus/Minus (or **+/-**) is one of the simplest ways to compare all skaters, regardless of their playing position.

Plus/Minus is an overall measure of the number of goals scored by a skater's team they are on the ice minus the number of goals scored against the skater's team while they are on the ice. The formula used to calculate Plus/Minus for a given skater is given by:

$$PM_s = \sum_{s=1}^{t} (GF_{t,s} - GA_{t,s})$$

Where $s =$ given skater, $t =$ team in which skater s played, $GF =$ goals for team t while skater s was on the ice, and $GA =$ goals against team t while skater s was on the ice.

1.1.1 Nature of the Data

Data was source from the website **Inalitic**. Two files are available for download from Inalitic; an MS Excel file containing several tables across several spreadsheets, and one comma separated values file that contains statistics on a large number of skaters. The latter file was the only file downloaded and used from Inalitic.

The .csv file contained a large number of statistics on NHL players between the years 1940 - 2018 inclusive. The file contained only *skaters* and did not include *goaltenders*. As such, the data table was an exhaustive list of every NHL player to whom a Plus/Minus statistic was appropriate to apply between the above years.

Below is a brief description of each of the variables included in the data table, as provided by the Inalitic website:

```
library(pacman)

suppressWarnings(p_load(c("knitr",
                          "kableExtra",
                          "rvest",
                          "xml2"),
                        character.only = T))

Inalitic <- read_html("http://inalitic.com/datasets/nhl%20player%20data.html")

Variables <- as.data.frame(html_table(html_nodes(Inalitic, "table")[4],
                                      fill = T, trim = T))

colnames(Variables) <- Variables[1, ]
Variables <- Variables[2:nrow(Variables), ]

column_spec(kable_styling(kable(Variables, row.names = F,
                    caption = "Variable descriptions as provided by Inalitic"),
              latex_options = "striped", full_width = F, font_size = 9),
            4, width = "12.5cm")
```

Table 1: Variable descriptions as provided by Inalitic

NAME	TYPE	FORMAT	DEFINITION
SEASON	Date	yyyy	Year corresponding to the season stats based on the when season ended, (i.e. 2015 - 2016 equates to 2016).
PLAYER	Character	First Name/Last Name	Player Name.
AGE	Number	00	Age of player during that corresponding season of play.
TM	Character	-	Three character code corresponding to the team that player played for in that season.
POS	Character	-	Postion played by player.
GP	Number	#0	Games played by that player in corresponding season.
G	Number	#	Goals scored by that player in corresponding season.
GPG	Number	#.0000	Goals per game by that player in corresponding season.
A	Number	#	Assists tallied by that player in corresponding season.
PTS	Number	#	Points tallied by that player in corresponding season.
+/-	Number	+/-#0	Plus/minus by player based on team Goals For (+) when player is on the ice vs. team Goals Against (-) when player is on the ice.
PIM	Number	#	Penalities in minutes given to that player in the season.
EVG	Number	#	Goals scored by player while team is even strength. First recorded in the 1968 season.
PPG	Number	#	Goals scored by player while team is on the powerplay. First recorded in the 1968 season.
SHG	Number	#	Goals scored by player while team is on the penalty kill (shorthanded). First recorded in the 1968 season.
GWG	Number	#	Game winning goals scored by player in that season. First recorded in the 1968 season.
EVA	Number	#	Assists tallied by player while team is even strength. First recorded in the 1968 season.
PPA	Number	#	Assists tallied by player while team is on the powerplay. First recorded in the 1968 season.
SHA	Number		Assists tallied by player while team is on the penalty kill (shorthanded). First recorded in the 1968 season.
S	Number	#	Total number of shots taken by player in that season. First recorded in the 1968 season.
S%	Number	#0.0%	Shooting percentage of player calculated from goals scored (G) divided by shots taken (S). First recorded in the 1968 season.
TOI	Number	#0	Total time on the ice by that player in corresponding season. First recorded in the 1999 season.
ATOI	Number	#.00	Average time on ice per game by player. First recorded in the 1999 season.
BLK	Number	#	Number of blocked shots. First recorded in the 2008 season.
HIT	Number	#	Number of hits. First recorded in the 2008 season.
FOwin	Number	#	Number of faceoffs won. First recorded in the 2008 season.
FOloss	Number	#	Number of faceoffs lost. First recorded in the 2008 season.
FO%	Number	#0.0%	Faceoff percentage calculated from total faceoffs won divided by total faceoffs taken. First recorded in the 2008 season.
ADJ FACTOR	Number	#.00	Average goals per game in corresponding season calculated from all players seasonal GPG.
ADJ GPG	Number	#.00	Adjusted Goals per Game of each player calculated by the Adjustment Factor multiplied by their GPG for that season.

1.1.1.1 Target Feature

Plus Minus was selected as the target feature in the data set.

1.1.1.2 Descriptive Features

All other variables in the data set were chosen as descriptive features, with the exemption of `Season` and `Player`.

1.2 Data Processing

1.2.1 Libraries

```
library(pacman)                          ## for loading multiple packages

suppressMessages(p_load(character.only = T,
                        install = F,
                        c("tidyverse",  ## thanks Hadley
                          "lubridate",  ## for handling dates
                          "forcats",    ## for categorial variables, not for felines
                          "zoo",        ## some data cleaning capabilities
                          "lemon",      ## add ons for ggplot
                          "rvest",      ## scraping web pages
                          "knitr",      ## knitting to RMarkdown
                          "kableExtra", ## add ons for knitr tables
                          "scales",     ## quick and easy formatting prettynums
                          "grid",       ## for stacking ggplots
                          "gridExtra",  ## also for stacking ggplots
                          "e1071",      ## for skew and kurtosis
                          "janitor",    ## cleaning colnames
                          "beepr")))    ## plays a beep tone
```

1.2.2 Loading Data

```
Skater_Stats <- read_csv(file.path(getwd(),
                                   "skater_stats.csv"),
                         na = c("NA","na", "N/A", "n/a", "-", ""))
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   X1 = col_character(),
##   Player = col_character(),
##   Tm = col_character(),
##   Pos = col_character(),
##   TOI = col_number(),
##   ATOI = col_time(format = "")
## )
```

```
## See spec(...) for full column specifications.
```

```
Skater_Stats <- rename(Skater_Stats,
                       "PM" = `+/-`)

Skater_Stats <- clean_names(Skater_Stats,
                            "parsed")

Skater_Stats <- select(Skater_Stats,
                       -X1)

Skater_Stats[, c(7:length(colnames(Skater_Stats)))] <- mutate_all(Skater_Stats[, c(7:length(colnames
                                                                  as.double)
```

The data set provided by Inalitic was, so little data preprocessing was needed. The variables `S_percent` and `FO_percent` were expressed as numbers between 0 and 100, so these columns were modified to be a value between 0 and 1.

```
Skater_Stats[ , c("S_percent", "FO_percent")] <- Skater_Stats[ , c("S_percent", "FO_percent")]/100

Skater_Sample <- sample_n(Skater_Stats, 20)

kable_styling(kable(Skater_Sample[ , c(1:15)],
                    caption = "Sample of `Skater\\_Stats` data frame"),
              latex_options = c("striped"), font_size = 9)
```

```
kable_styling(kable(Skater_Sample[ , c(1:2, 16:28)],
                    caption = "Sample of `Skater\\_Stats` data frame continued"),
              latex_options = c("striped"), font_size = 9)
```

Table 2: Sample of 'Skater_Stats' data frame

Season	Player	Age	Tm	Pos	GP	G	GPG	A	PTS	PM	PIM	EVG	PPG	SHG
2017	Ben Lovejoy	32	NJD	D	82	1	0.0122	6	7	-7	39	1	NA	NA
2004	Andrej Nedorost	23	CBJ	LW	9	2	0.2222	NA	2	NA	6	2	NA	NA
2001	Jaroslav Spacek	26	TOT	D	62	7	0.1129	19	26	3	28	4	3	NA
1992	Dave Snuggerud	25	TOT	LW	66	3	0.0455	16	19	-15	40	3	NA	NA
1983	Doug Shedden	21	PIT	C	80	24	0.3000	43	67	-20	54	19	4	1
1973	Larry Sacharuk	20	NYR	D	8	1	0.1250	NA	1	-1	NA	1	NA	NA
2006	Jim Campbell	32	TBL	RW	1	NA	0.0000	NA	NA	NA	2	NA	NA	NA
1977	Tom Lysiak	23	ATF	C	79	30	0.3797	51	81	3	52	24	5	1
2003	Jason Allison	27	LAK	C	26	6	0.2308	22	28	9	22	4	2	NA
1997	Todd Simpson	23	CGY	D	82	1	0.0122	13	14	-14	208	1	NA	NA
1998	Philippe Boucher	24	LAK	D	45	6	0.1333	10	16	6	49	5	1	NA
2000	Mark Mowers	25	NSH	RW	41	4	0.0976	5	9	NA	10	4	NA	NA
2000	David Van Drunen	24	OTT	D	1	NA	0.0000	NA	NA	NA	NA	NA	NA	NA
2013	Andrej Meszaros	27	PHI	D	11	NA	0.0000	2	2	-9	2	NA	NA	NA
2009	Tim Wallace	24	PIT	RW	16	NA	0.0000	2	2	2	7	NA	NA	NA
1998	Bryan McCabe	22	TOT	D	82	4	0.0488	20	24	19	209	2	1	1
1983	Derek Smith	28	DET	C	42	7	0.1667	4	11	-7	12	6	1	NA
2017	Patrick Wiercioch	26	COL	D	57	4	0.0702	8	12	-18	23	4	NA	NA
2015	Mike Cammalleri	32	NJD	LW	68	27	0.3971	15	42	2	28	16	9	2
1998	Dave Manson	31	MTL	D	81	4	0.0494	30	34	22	122	2	2	NA

1.2.3 Descriptive Statistics

1.2.3.1 Target Feature

```
kable_styling(kable(t(summarise(Skater_Stats,
                                "Mean PM" = mean(PM, na.rm = T),
                                "Std Dev PM" = sd(PM, na.rm = T),
                                "Min PM" = min(PM, na.rm = T),
                                "Q1 PM" = quantile(PM, 0.25, na.rm = T),
                                "Median PM" = median(PM, na.rm = T),
                                "Q3 PM" = quantile(PM, 0.75, na.rm = T),
                                "Max PM" = max(PM, na.rm = T))),
                    caption = "Descriptive Statistics for Target Feature PM
                                (Plus/Minus)",
                    digits = 3,
                    format.args = list(nsmall = 2, scientific = F)),
              latex_options = "striped", font_size = 10, position = "center")
```

```
Skater_Stats_Long <- gather(Skater_Stats,
                            c(Age, GP:PTS, PIM:FO_percent),
                            key = "Variable",
                            value = "Statistic")

kable_styling(kable(summarise(group_by(Skater_Stats_Long,
                                       Variable),
                              "Mean" = mean(Statistic, na.rm = T),
                              "Std Dev" = sd(Statistic, na.rm = T),
                              "Min" = min(Statistic, na.rm = T),
                              "Q1" = quantile(Statistic, 0.25, na.rm = T),
                              "Median" = median(Statistic, na.rm = T),
                              "Q3" = quantile(Statistic, 0.75, na.rm = T),
                              "Max" = max(Statistic, na.rm = T)),
                    digits = 2,
```

Table 3: Sample of 'Skater_Stats' data frame continued

Season	Player	GWG	EVA	PPA	SHA	S	S_percent	TOI	ATOI	BLK	HIT	F_Owin	F_Oloss	FO_percent
2017	Ben Lovejoy	1	6	NA	NA	84	0.01	1703	74760	149	82	0	0	NA
2004	Andrej Nedorost	NA	NA	NA	NA	16	0.13	117	46920	NA	NA	NA	NA	NA
2001	Jaroslav Spacek	1	9	7	3	106	0.07	1306	75840	NA	NA	NA	NA	NA
1992	Dave Snuggerud	NA	15	1	NA	94	0.03	NA	NA	NA	NA	NA	NA	NA
1983	Doug Shedden	2	26	17	NA	175	0.14	NA	NA	NA	NA	NA	NA	NA
1973	Larry Sacharuk	NA	NA	NA	NA	9	0.11	NA	NA	NA	NA	NA	NA	NA
2006	Jim Campbell	NA	NA	NA	NA	NA	NA	9	31620	NA	NA	NA	NA	NA
1977	Tom Lysiak	3	37	13	1	277	0.11	NA	NA	NA	NA	NA	NA	NA
2003	Jason Allison	3	12	10	NA	46	0.13	562	77760	NA	NA	NA	NA	NA
1997	Todd Simpson	NA	13	NA	NA	85	0.01	NA	NA	NA	NA	NA	NA	NA
1998	Philippe Boucher	NA	8	2	NA	80	0.08	NA	NA	NA	NA	NA	NA	NA
2000	Mark Mowers	NA	5	NA	NA	50	0.08	449	39480	NA	NA	NA	NA	NA
2000	David Van Drunen	NA	NA	NA	NA	NA	NA	5	16680	NA	NA	NA	NA	NA
2013	Andrej Meszaros	NA	1	1	NA	18	NA	203	66480	24	18	0	0	NA
2009	Tim Wallace	NA	2	NA	NA	17	NA	130	29220	1	37	2	1	0.667
1998	Bryan McCabe	NA	11	8	1	123	0.03	NA	NA	NA	NA	NA	NA	NA
1983	Derek Smith	NA	4	NA	NA	50	0.14	NA	NA	NA	NA	NA	NA	NA
2017	Patrick Wiercioch	NA	8	NA	NA	63	0.06	950	60000	57	52	1	0	1.000
2015	Mike Cammalleri	8	13	2	NA	156	0.17	1247	66000	19	29	114	157	0.421
1998	Dave Manson	NA	22	8	NA	148	0.03	NA	NA	NA	NA	NA	NA	NA

Table 4: Descriptive Statistics for Target Feature PM (Plus/Minus)

Mean PM	-0.416
Std Dev PM	12.876
Min PM	-82.000
Q1 PM	-7.000
Median PM	-1.000
Q3 PM	5.000
Max PM	124.000

```
            format.args = list(scientific = F, big.mark = ",", nsmall = 2),
            caption = "Descriptive Statistics for all numeric descriptive
                       features"),
        latex_options = "striped", font_size = 9, full_width = F)
```

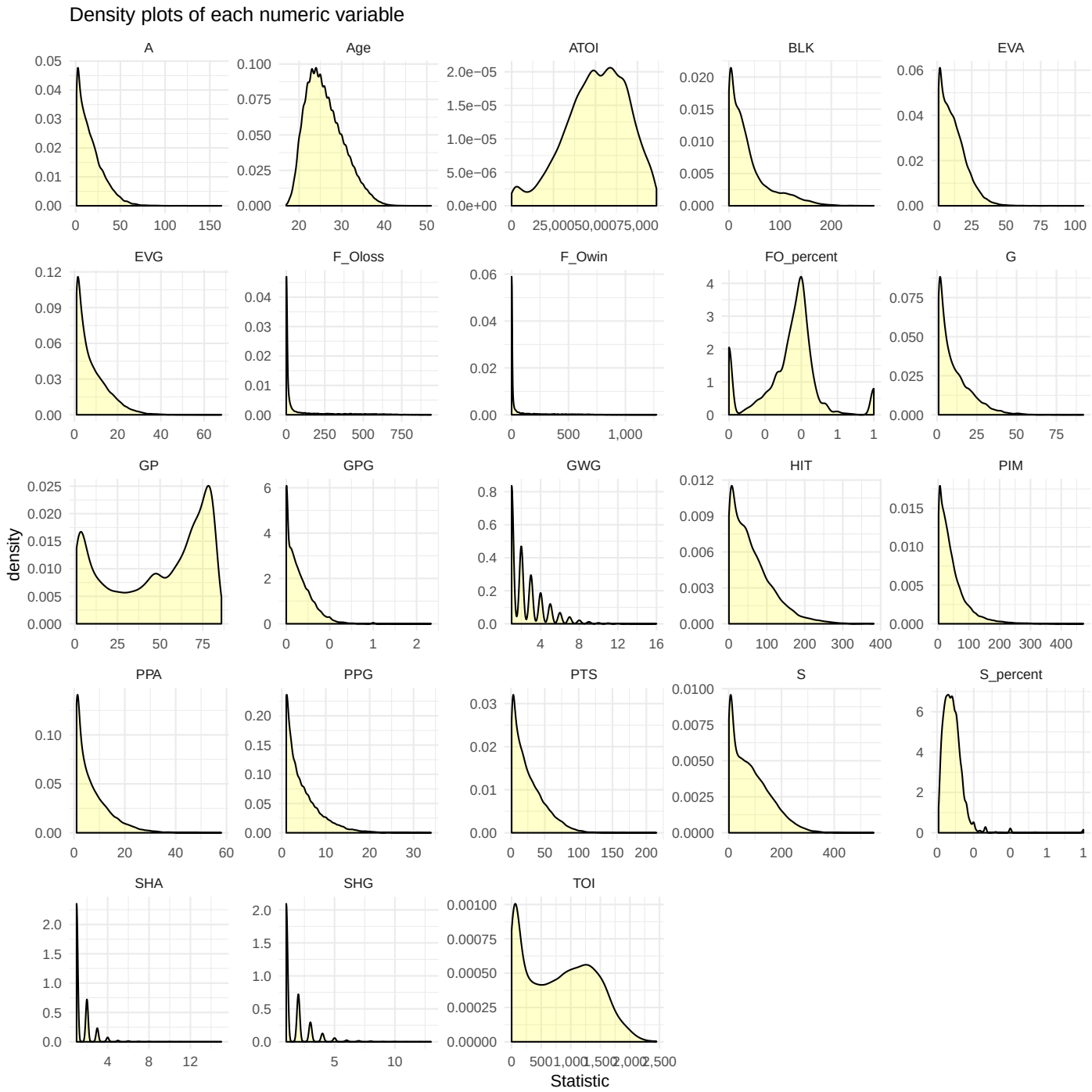
Table 5: Descriptive Statistics for all numeric descriptive features

Variable	Mean	Std Dev	Min	Q1	Median	Q3	Max
A	16.40	14.22	1.00	5.00	13.00	24.00	163.00
Age	26.06	4.34	17.00	23.00	25.00	29.00	51.00
ATOI	50,807.98	18,254.02	0.00	38,880.00	52,080.00	64,740.00	86,340.00
BLK	37.51	40.42	0.00	8.00	24.00	51.00	283.00
EVA	11.86	9.20	1.00	4.00	10.00	17.00	106.00
EVG	7.96	7.20	1.00	2.00	6.00	12.00	68.00
F_Oloss	79.38	167.92	0.00	0.00	3.00	42.00	941.00
F_Owin	79.38	177.17	0.00	0.00	2.00	35.00	1,273.00
FO_percent	0.42	0.21	0.00	0.33	0.46	0.52	1.00
G	10.74	10.29	1.00	3.00	7.00	16.00	92.00
GP	48.06	28.07	1.00	20.00	56.00	74.00	86.00
GPG	0.14	0.15	0.00	0.03	0.10	0.22	1.67
GWG	2.48	1.84	1.00	1.00	2.00	3.00	16.00
HIT	60.17	54.59	0.00	17.00	47.00	88.00	382.00
PIM	45.79	46.92	2.00	14.00	32.00	61.50	472.00
PPA	7.19	6.87	1.00	2.00	5.00	10.00	58.00
PPG	4.50	4.03	1.00	1.00	3.00	6.00	34.00
PTS	25.51	23.18	1.00	7.00	19.00	38.00	215.00
S	85.46	72.50	1.00	23.00	71.00	131.00	550.00
S_percent	0.11	0.08	0.01	0.06	0.10	0.14	1.00
SHA	1.48	0.90	1.00	1.00	1.00	2.00	15.00
SHG	1.67	1.15	1.00	1.00	1.00	2.00	13.00
TOI	807.07	579.26	1.00	235.00	816.00	1,286.00	2,449.00

1.2.4 Density Plots

1.2.4.1 Numeric Variables

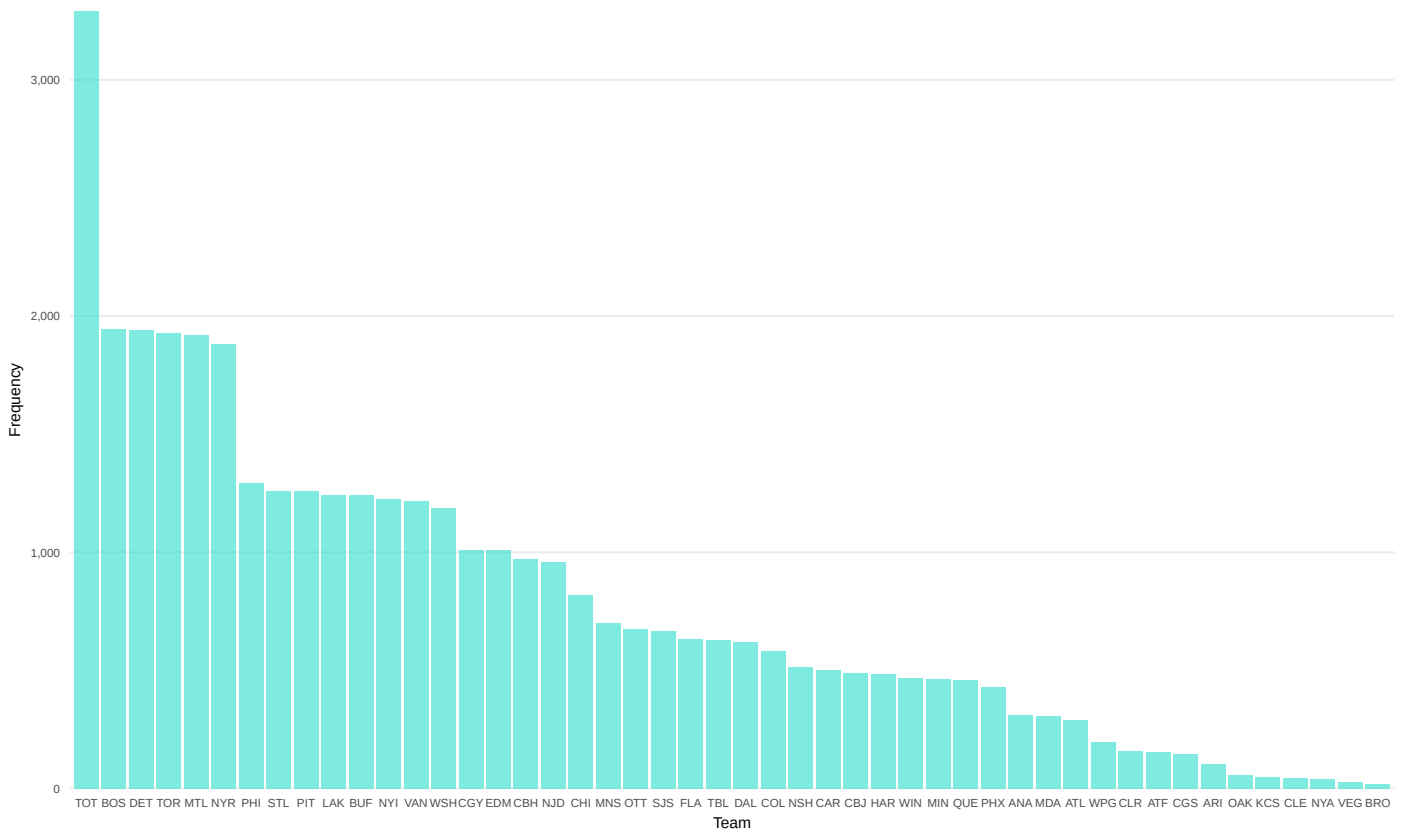
```
ggplot(Skater_Stats_Long) +
  geom_density(aes(x = Statistic), na.rm = T,
               fill = "yellow", alpha = 0.2) +
  facet_rep_wrap(~Variable, scales = "free") +
  scale_x_continuous(labels = comma) +
  labs(title = "Density plots of each numeric variable") +
  theme_minimal()
```



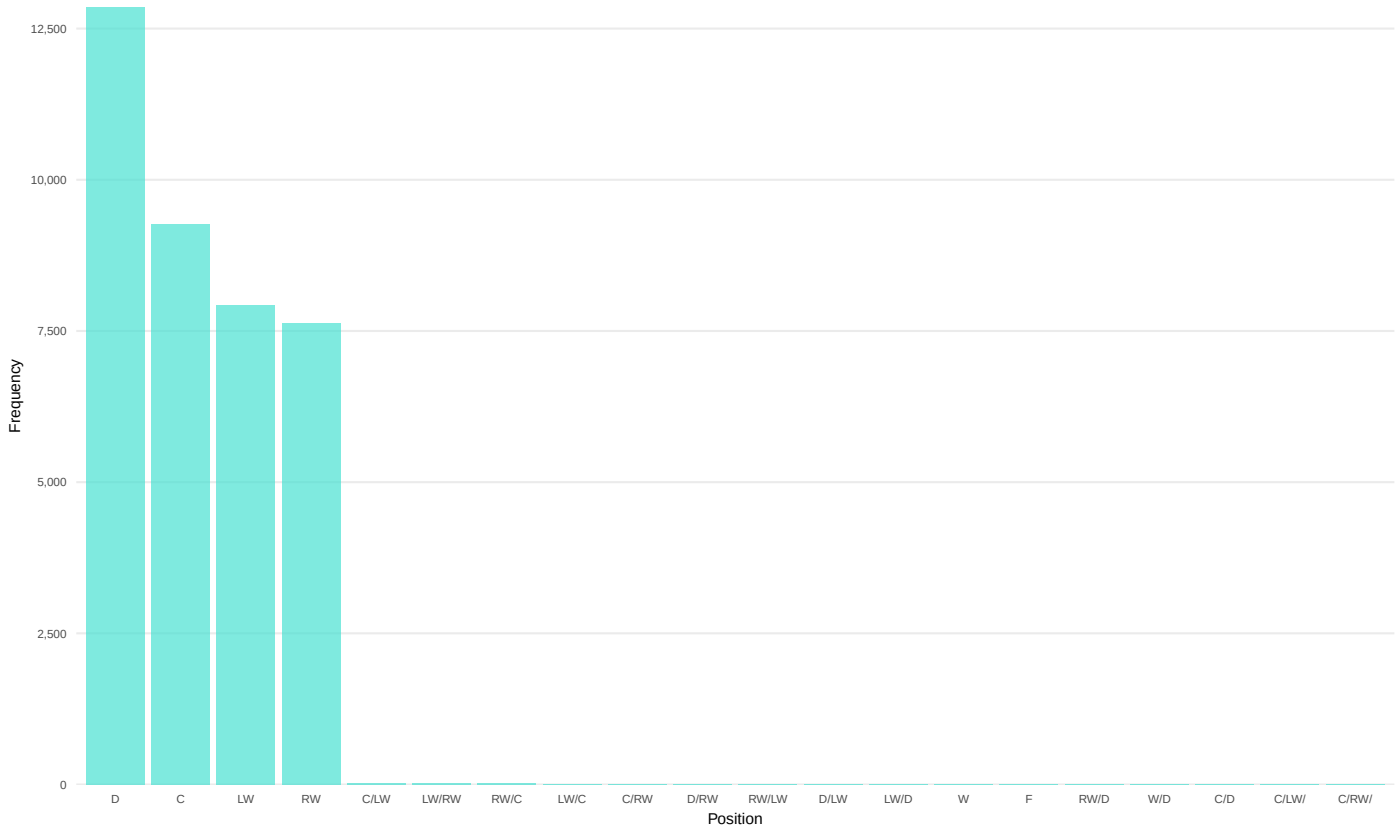
1.2.4.2 Qualitative Variables

```
ggplot(Skater_Stats) +
  geom_bar(aes(x = fct_infreq(Tm)),
           fill = "turquoise", alpha = 2/3) +
  scale_y_continuous(labels = comma,
                     "Frequency",
                   expand = expand_scale(c(0,0))) +
  scale_x_discrete("Team") +
  labs("Frequencies of each Team") +
  theme_minimal() +
  theme(panel.grid.minor = element_blank(),
```

```
        panel.grid.major.x = element_blank(),
        panel.border = element_blank(),
        text = element_text(size = 8))
```




```
ggplot(Skater_Stats) +
  geom_bar(aes(x = fct_infreq(Pos)),
           fill = "turquoise", alpha = 2/3) +
  scale_y_continuous(labels = comma,
                     "Frequency",
                   expand = expand_scale(c(0,0))) +
  scale_x_discrete("Position") +
  labs("Frequencies of each Skating Position") +
  theme_minimal() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.border = element_blank(),
        text = element_text(size = 8))
```



1.2.5 Autocorrelation of Doubles

```
Skater_double <- Skater_Stats[, c(3, 6:length(colnames(Skater_Stats)))]

for (i in colnames(Skater_double)) {

  print(acf(Skater_double,
            na.action = na.pass,
            lag.max = nrow(Skater_double),
            main = colnames(Skater_double[, i]),
            ylab = "Autocorrelation",
            xlab = "Lag Position"))

}
```

1.3 References

- http://inalitic.com/datasets/nhl%20player%20data.html