**Visualizing Professional Woman’s Hockey Data**

The Professional Woman’s Hockey League (PWHL) began its inaugural season in 2023-24. The league has players from 11 different countries. The league looks to expand its exposure and gain new fans for the future of the sport.

We will be investigating the player statistics from the league’s inaugural season. Our focus will be on all players including goalies which includes 147 athletes. We want to discover which age groups and positions have had the most impact on the number of points scored per game by each player.

Each position has a different role in contributing to the team. In general, the main goal of the forwards is to stay in three different lanes across the ice, moving the puck between them to make the goalie move and open scoring opportunities. The defense players compliment the forwards by positioning themselves along the boundary of the offensive zone to prevent the opposing team from moving the puck away from the zone and provide more opportunities for the forwards to score. The goalie’s focus is to guard their team’s goal by positioning themselves in front of it to prevent the opposing team from scoring.

The different age groups represent a blend of experience and athleticism at a point in the player’s career. More experience should help the player score more goals because they would have more knowledge of the game and ideas on how to score. However, more experience comes with more aging and players with more experience may be past their years of peak athleticism. That is why our goal is to find if there is a perfect blend between the two (i.e. an ideal age and position group).

**Data Description**

**PWHL\_Final** - Name of the data set

**P\_Per\_GP** -Number of points scored by the player per game played (Goals + Assists)

**Pos** – Position of the player (either Defense (D), Forward (F), or Goalie (G))

**Age** – Age of the player in years

The density plot below displays the distribution of goals per game played for each position. Use it to answer the following two questions.



1. What would you need to add to the R code below to add a title and change the x-axis label to "Points Per Game Played"?

```{r}

ggplot(data = PWHL\_Final) +

 geom\_density(aes(P\_Per\_GP, color = Pos, fill = Pos), alpha = 0.25) +

 theme\_minimal()

```

 + labs(x = "Points Per Game Played", title = "Points Scored Between Different Age Groups and Positions")

1. Describe the distribution of the Forward and Defense positions in the density plot above. Make sure to mention shape and skew. Give a possible reason why the Goalie’s curve is concentrated around zero in this visual.

Sample Answer: All three of the curves are right skewed with long tails extending towards the right (higher points per game played). The defense curve peaks around 0.075, while the forward’s curve peaks around 0.1 and is the flattest of all three. The Goalie’s curve has a prominent peak around 0 and at around 0.075 points per game and above there is approximately no area under the curve. The Goalie’s curve is concentrated around zero because given their positioning in front of their own net they have a low chance of getting points.





1. Brainstorm some ideas on which combinations of age and positions would be ideal for a player to be to maximize their number of points per game played. Explain.

Sample Answer: A combination of the Forward position and the oldest age group would be ideal because the boxplot displaying the positions vs the number of points scored per game the forward’s boxplot and median is higher meaning on average forward’s score more points per game than defense. We would choose the older age group for the same reasoning; its boxplot and median are higher on the points per game scale than the other two age groups.

1. Let’s put this all together on one graph to see the trends between the age groups and positions. First, create three different age groups “youngest” (22-25), “middle” (26-30), and “oldest” (31-36). Then, create side-by-side boxplots that display the points per game for each group. Only include Forward and Defense positions (since we already showed how few points goalies score). Make sure to add a theme and change the x and y axis labels. Refer to the data description on the second page. Once completed, get into small groups and decide whose graph displays the data the best and why. Try switching the variables between x, y, and color to see if it improves your visualization.

```{r}

PWHL\_Graph <- PWHL\_Final %>%

 mutate(Age\_Group = case\_when(

 Age >= 22 & Age <= 26 ~ "youngest",

 Age >= 27 & Age <= 30 ~ "middle",

 Age >= 31 & Age <= 36 ~ "oldest"

 )) %>%

 filter(Pos != "G")

PWHL\_Graph <- PWHL\_Graph %>%

 mutate(Age\_Group = factor(Age\_Group, levels = c("youngest", "middle", "oldest")))

```

```{r}

ggplot(data = PWHL\_Graph, aes(x = Pos, y = P\_Per\_GP, color = Age\_Group)) +

 geom\_boxplot() +

 labs(x = "Position", y = "Number of Points per Game Played", title = "Points Per Game Played Between Different Age Groups and Positions") +

 theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5))

```

Sample Graphs:







1. What trends, if any, do you see in your graph? Does it confirm your original thoughts in question 3? What seems to be the ideal position and age group to maximize goals per game played?

For forwards, as age increases, so does the number of goals per game played. For defense, the middle age has the greatest number of points per game played with youngest behind them and oldest with the least. The ideal position and age group is Forward and oldest which does confirm what we found using the boxplots in question 3.

1. Give one reason why you think older forwards have the most points per game compared to the other age groups.

Since this data represents the first season of the league, the older players likely have more experience in other professional leagues and were able to use that to their advantage and have more successful offensive seasons than the younger forwards.

1. Reread the summary on the first page and find a limitation to this dataset.

There has only been one season, so the data is limited to a small sample size. When the league has a few more seasons the data will likely be clearer and reflect what is going on better.