MEHRA_KAI_Assignment-1

Kai Mehra

2023-01-28

Libraries

```
# Load the {tidyverse}, {ggplot2}, and {viridis} packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

Loaing Data:

```
# Load draft78 as player_picks
player_picks <- read.csv("./Data/NBA Data/draft78.csv")

# Load season78 as player_performance
player_performance <- read.csv("./Data/NBA Data/season78.csv")

# Take dimensions of the data sets
dim(player_picks)
```

```
## [1] 3642    4
```

```
dim(player_performance)
```

```
## [1] 15313     3
```

The player_picks and player_performance data sets were sourced from Kaggle. Both data sets contain data on NBA players from the 1978-79 season to the 2015-16 season. The player_picks data set has data on 3642 players encompassing their name, year they were drafted, the pick they were drafted with, and the length of their career in years. The player_performance data set has 15313 observations of players and the number of win shares they accummulated each season of their career.

Research Question

Using these data sets, my goal is to answer if top picks in NBA drafts consistently outperform lower picks? I want to understand whether top draft picks are truly worth it, or if there is a only marginal benefit between the 10th and 20th pick, for example. I hypothesize the the top 5 picks will be better on average than the other picks, but I think the remaining picks will be mostly similar in value. I also predict that the higher picks will have a higher variability in performance where the best players are incredible and the worst players are terrible.

Variables of Interest

In the player_picks data set, I am mainly focusing on the player's name, year they were drafted, pick they were drafted with, and length of their career

In the player_performance data set, I am mainly focusing on the player's name and the win shares (WS) they accumulated during a specific season.

Win Shares is a statistic that attempts to assign value to NBA players by approximating how many wins they contribute to their team in a given season. It encompasses their offensive and defensive performance and relies on comparing players to the league average for a given season. In theory, adding up the win shares of every player on a team should equal the number of wins the entire team actually earned during the season. The greatest single season win shares performances top out at about 25 win shares. However, any player with a career average around 10 is a superstar level player. A win shares value between 0-3 is an average to below average player while a negative win shares value means that a player is activley detrimental to their teams performance. This statistic is an estimate of a players value and has flaws as it attempts to boil basketball performance to one number.

Data Wrangling:

```
# Select the variables of interest, remove N/A values from the Yrs variable,
# and filter the data to only include drafts from 1980 onward.
player_picks <-
  player_picks %>%
  select(Player, Pick, Draft, Yrs) %>%
  filter(!is.na(Yrs)) %>%
  filter(Draft >= 1980)
```

I only considered drafts after the 1980 season, as this was the first season after the NBA and ABA merged which many consider the beginning of contemporary NBA history.

```
# Filter the data set to remove N/A values from the WS and Season variables,
#and filter the data to only include Win Shares data from 1980 onward.
player_performance <-
```

```
  player_performance %>%
  filter(!is.na(WS)) %>%
  filter(!is.na(Season)) %>%
  filter(Season >= 1980)
```

```
# Create a new data frame called "player_WS" that has the mean win shares each
# player accumulated over the course of their career.
player_WS <-
  player_performance %>%
  group_by(Player) %>%
  summarise(player_mean_WS = round(mean(WS), 2)) %>%
  # calculate the mean win shares each player earned over their career and add
  # it to the dataset as   player_mean_WS
  mutate(player_mean_WS) %>%
  filter(!is.na(player_mean_WS))
```

```
# Create a new data frame called "player_pick_performance" that combines the
# player_picks and player_WS data frames, and filter out all of the picks greater
# than 60
player_pick_performance <-
  inner_join(player_picks, player_WS,
             by = "Player") %>%
  filter(Pick <= 60)
```

Using the inner_join function, a data frame called "player_pick_performance" was created that contains the player's name, the pick they were drafted with, the year they were drafted, the length of their career, and the average number of win shares they accumulated per season during their career.

Additional Data Wrangling and Analysis

Calculating the mean performance by pick:

```
# create num_picks data frame
num_picks <-
  player_pick_performance %>%
  # count the frequency of each pick
  count(Pick)

# create mean_pick_performance data frame
mean_pick_performance <-
  player_pick_performance %>%
  group_by(Pick) %>%
  # calculate the mean and std deviation of mean shares for each pick
  summarise(pick_mean_WS = round(mean(player_mean_WS), 2),
            std_pick_WS = round(sd(player_mean_WS),2)) %>%
  mutate(Low_SD = pick_mean_WS - std_pick_WS,
         High_SD = pick_mean_WS + std_pick_WS)

# join the num_picks and mean_pick_performance data frames
mean_pick_performance <-
```

```
  left_join(num_picks, mean_pick_performance,
            by = "Pick")
```

Additional data wrangling was done to create a data frame with the average win shares per draft pick. Further, the standard deviation was created to be able to add error bars to the upcoming plot. Saving the frequency of each pick was important to ensure conclusions were being made on data with an adequate sample size.

Grouping picks by 5 and calcluating mean performance:

```
# creating the data frame that groups picks in by every 5 picks
mean_5_pick_performance <-
  mean_pick_performance %>%
  filter(Pick <= 60) %>%
  # assigns each pick a number 1-10 that indicates which set of 5 picks they
  # were drafted in
  mutate(group_5 = ((Pick - 1) %/% 5) + 1)
```

The mean_5_pick_performance data frame created a new variable called group_5 that indicates which set of five picks a player was drafted in. 1 corresponds to 1-5 pick, 2 to 6-10, etc. This allows the analysis to look at broader trends of top picks vs. lower picks.

```
mean_5_pick_performance_dist <-
  mean_5_pick_performance %>%
  # group by which set of 5 picks they fall into
  group_by(group_5) %>%
  # calculate mean and std deviation of these groups
  summarise(pick_5_mean_WS = round(mean(pick_mean_WS), 2),
            std_5_pick_WS = round(sd(pick_mean_WS),2)) %>%
  mutate(Low_5_SD = pick_5_mean_WS - std_5_pick_WS,
         High_5_SD = pick_5_mean_WS + std_5_pick_WS)
```

Similar to the analysis with the raw pick data, after grouping the picks into groups of 5, the means and standard deviations were calculated.

Calculating the difference between the top 10% and bottom 10% by pick

```
high_10tile_WS <-
  player_pick_performance %>%
  group_by(Pick) %>%
  filter(
    (quantile(player_mean_WS, 0.90)<=player_mean_WS) # filtering top 10%
  ) %>%
  summarise(high = mean(player_mean_WS)) # calculate the mean

low_10tile_WS <-
  player_pick_performance %>%
  group_by(Pick) %>%
  filter(
```

```
    (quantile(player_mean_WS, 0.10)>=player_mean_WS) # filter bottom 10%
  ) %>%
  summarise(low = mean(player_mean_WS)) # calculate the mean
```

These data frames contain the mean win shares for the top 10% and bottom 10% of players for each pick.

```
high_low_10tile_diff <-
  inner_join(high_10tile_WS, low_10tile_WS, by = "Pick")

high_low_10tile_diff <-
  high_low_10tile_diff %>%
  mutate(high_low_diff = high - low)
```

The high_low_10tile_diff data frame calculates the difference between the average win shares for top 10% and bottom 10% for each pick. Essentially, a larger difference implies that the pick has a higher level of variablity in performance.

Average Player Win Shares

```
avg_player_WS <- round(mean(mean_pick_performance$pick_mean_WS), 2)
# calculate the overall average wins shares of every player in the dataset
avg_player_WS
```

```
## [1] 1.66
```

In this data set, the average player earned 1.66 Win Shares per season.

Average Length of Career for Players based on draft pick

```
top_player_pick_performance <-
  player_pick_performance %>%
  filter(Pick <= 10)

mean_Yrs_top = mean(top_player_pick_performance$Yrs)
```

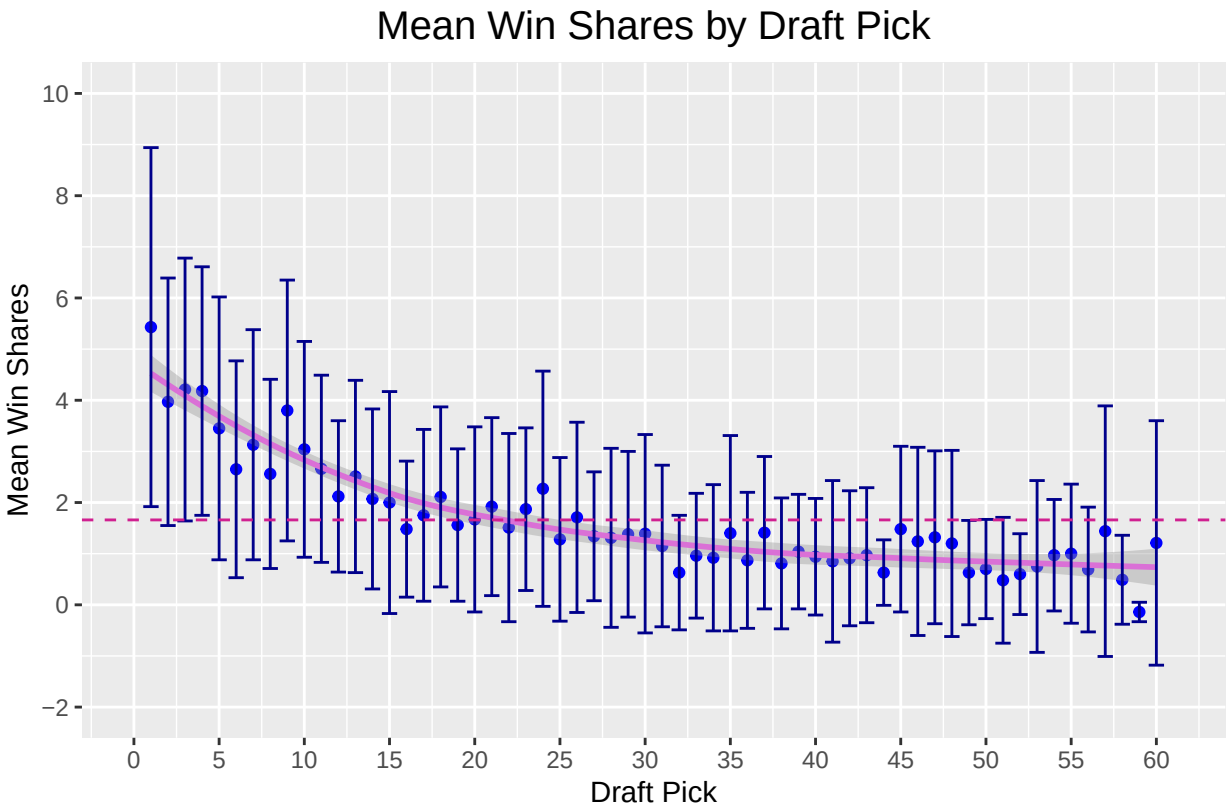
```
bottom_player_pick_performance <-
  player_pick_performance %>%
  filter(Pick >= 50)

mean_Yrs_bottom = mean(bottom_player_pick_performance$Yrs)
```

Visualisations

Plot 1: Mean WS by Pick with SE bars:

```r
mean_pick_performance %>%
  filter(n >= 5) %>% # remove draft picks with very few data points
  ggplot() +
  geom_point(
    aes(x = Pick,
        y = pick_mean_WS),
    color = "blue"
  ) + # scatterplot
  geom_smooth(
    aes(x = Pick,
        y = pick_mean_WS),
    color = "orchid"
  ) + # loess smooth line
  geom_errorbar(
    aes(x = Pick,
        y = pick_mean_WS,
        ymin = Low_SD,
        ymax = High_SD),
        color = "darkblue"
  ) + # error bars displaying plus/minus one standard deviation around the mean
  geom_hline(
    yintercept = avg_player_WS,
    linetype = 2,
    color = "violetred"
  ) + # horizontal dashed line indicating the overall average win shares value
  labs(
    x = "Draft Pick",
    y = "Mean Win Shares",
    title = "Mean Win Shares by Draft Pick",
    caption = "Note: The dashed horizontal line indicates the average player's win shares"
  ) + # better labels
  theme(
    plot.title = element_text(size = 15, hjust = 0.5),
    legend.position = "none",
    plot.caption = element_text(hjust = 0)
  ) + # center the title
  scale_x_continuous(
    limits = c(0, 61),
    breaks = seq(0, 60, 5)
  ) + # improve the x axis scale
  scale_y_continuous(
    limits = c(-2, 10),
    breaks = seq(-2, 10, 2)
  ) # improve the y axis scale
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



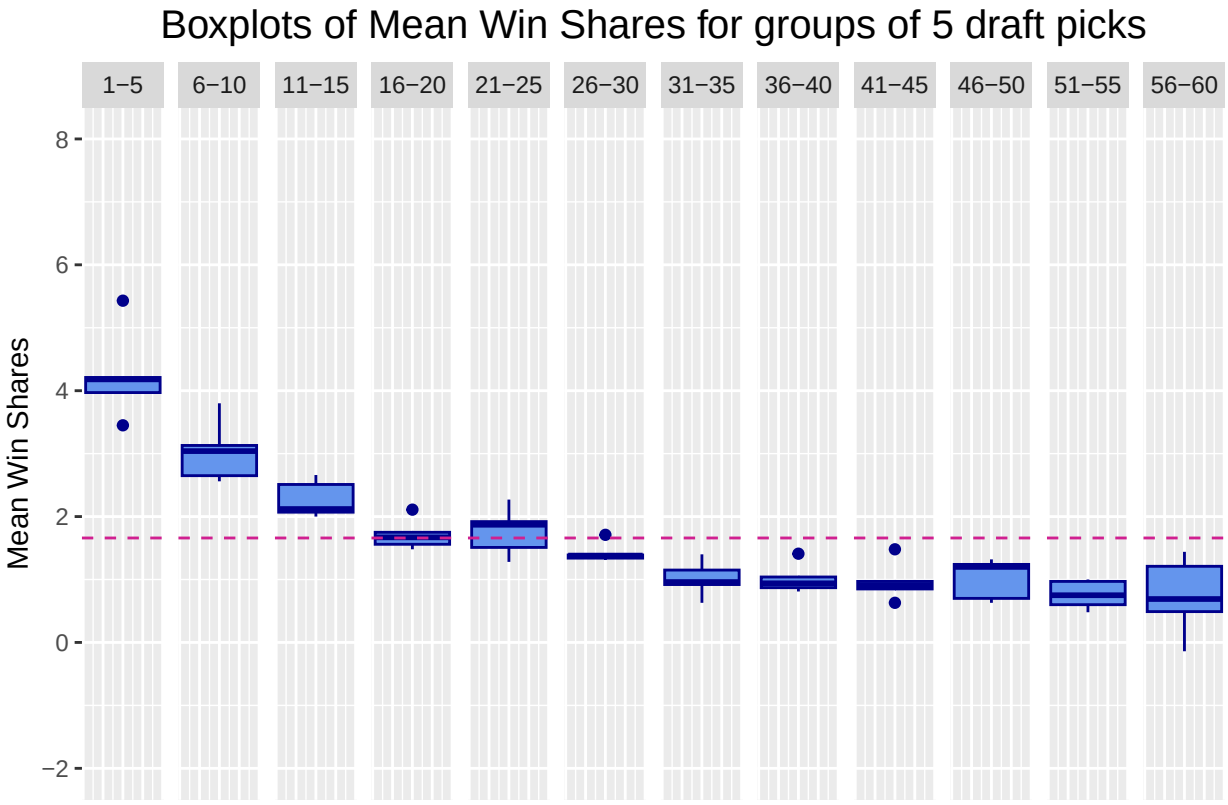
Note: The dashed horizontal line indicates the average player's win shares

Plot 2: Box plot of mean win shares by groups of 5 picks:

```
group_5.labs <- c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30",
                  "31-35", "36-40", "41-45", "46-50", "51-55", "56-60")
names(group_5.labs) <- c(1:12)
# create custom labels for each group of five picks
```

```
mean_5_pick_performance %>%
  ggplot(
    aes(
    y = pick_mean_WS
  )
  ) +
  geom_boxplot(
  color = "darkblue",
  fill = "cornflowerblue"
  ) + # boxplot
  facet_wrap(~group_5, nrow = 1, dir = "h",
             labeller = labeller(group_5 = group_5.labs)) +
  # facet wrap using our custom labels
  labs(
    y = "Mean Win Shares",
    title = "Boxplots of Mean Win Shares for groups of 5 draft picks",
```

```
    caption = "Note: The dashed horizontal line indicates the average player's win shares"
  ) + # better labels
  scale_y_continuous(
    limits = c(-2, 8),
    breaks = seq(-2, 8, 2)
  ) +
  geom_hline(
    yintercept = avg_player_WS,
    linetype = 2,
    color = "violetred"
  ) + # horizontal line indicating the average win shares overall
  theme(
    plot.title = element_text(size = 15, hjust = 0.5),
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank(),
    plot.caption = element_text(hjust = 0)
  ) # center the title and remove y-axis labels and ticks
```

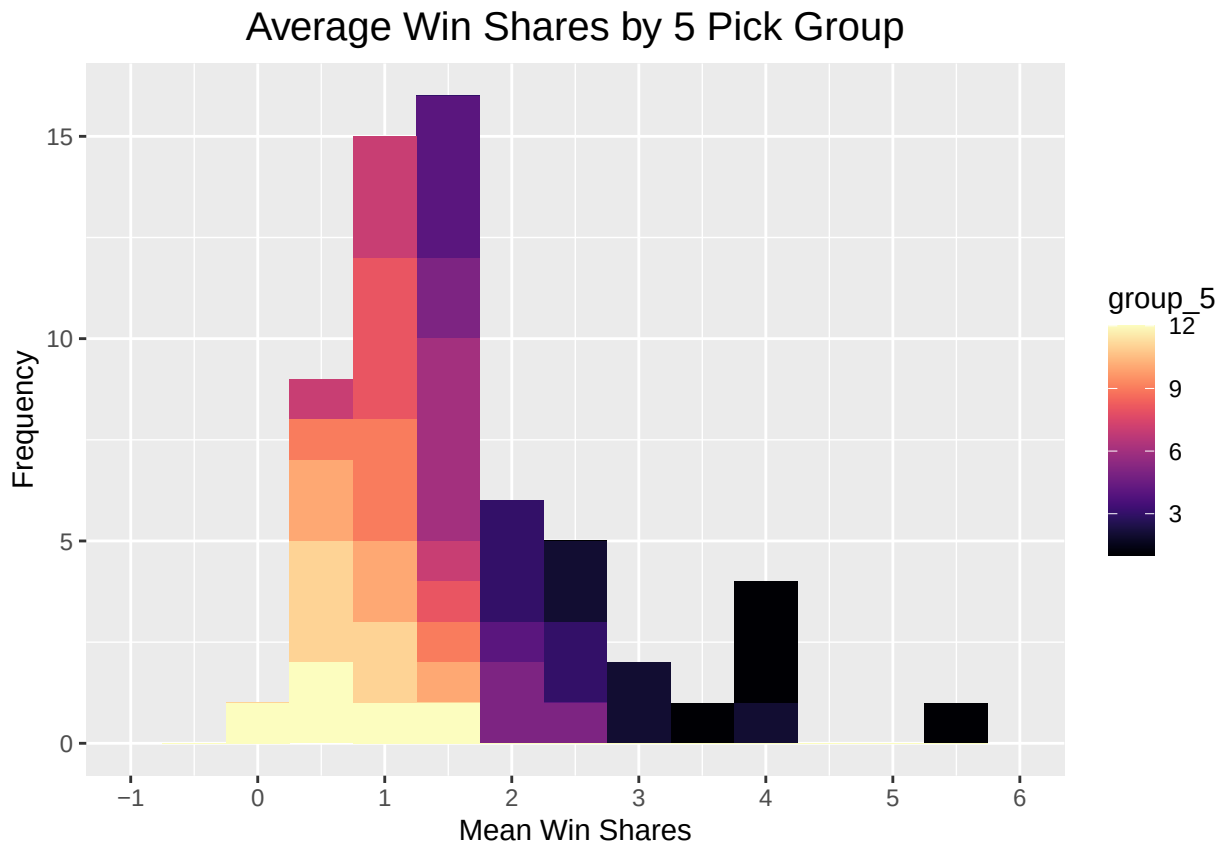


Plot 3: Histogram of average win shares by group of 5 picks

```
mean_5_pick_performance %>%
  ggplot(
    aes(
```

```
      x = pick_mean_WS,
      group = group_5,
      fill = group_5 # fill by group_5
    )
  ) +
  geom_histogram(
    bins = 15 # better bin size
      ) +
  scale_fill_viridis(option = "A") + # use viridis color pallete
  labs(
    x = "Mean Win Shares",
    y = "Frequency",
    title = "Average Win Shares by 5 Pick Group"
  ) + # better labels and title
  theme(
    plot.title = element_text(size = 15, hjust = 0.5) # centering the title
  ) +
  scale_x_continuous(
    limits = c(-1, 6),
    breaks = seq(-1, 6, 1)
  )
```

```
## Warning: Removed 24 rows containing missing values (`geom_bar()`).
```

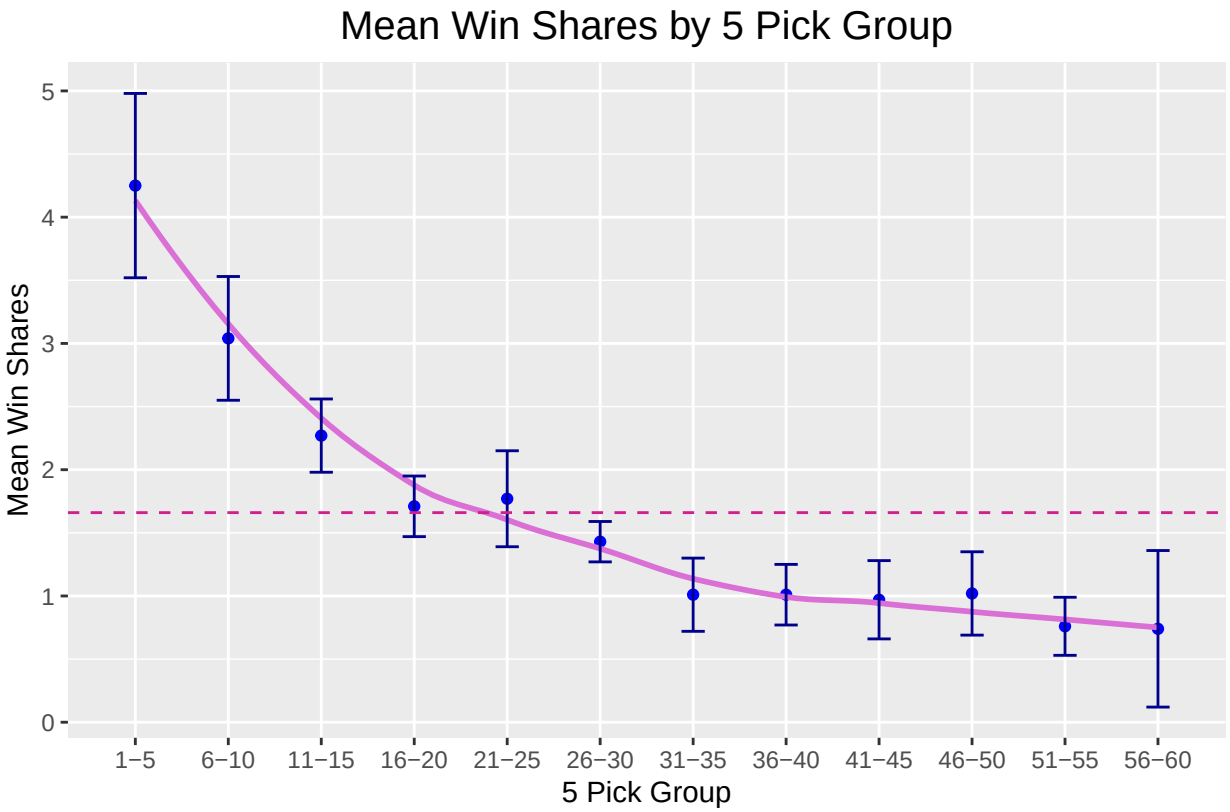


Plot 4: Scatterplot with error bars of picks in groups of 5

```
mean_5_pick_performance_dist %>%
  ggplot(
    aes(x = group_5,
         y = pick_5_mean_WS)
         ) +
  geom_point(
    aes(x = group_5,
        y = pick_5_mean_WS),
    color = "blue"
  ) + # scatterplot
  geom_smooth(
    aes(x = group_5,
        y = pick_5_mean_WS),
    color = "orchid",
    se = FALSE
  ) + # loess best fit line
  geom_errorbar(
    aes(ymin = Low_5_SD,
        ymax = High_5_SD),
    width = 0.25,
    color = "darkblue"
  ) + # error bars with plus/minus one std dev
  geom_hline(
    yintercept = avg_player_WS,
    linetype = 2,
    color = "violetred"
  ) + # horizontal line indicating the average win shares overall
  labs(
    x = "5 Pick Group",
    y = "Mean Win Shares",
    title = "Mean Win Shares by 5 Pick Group",
    caption = "Note: The dashed horizontal line indicates the average player's win shares"
  ) + # better labels and a title
  scale_x_discrete(
    limits = c(0:12),
    breaks = seq(1, 12, 1),
    labels = c("1-5", "6-10", "11-15", "16-20", "21-25", "26-30",
                  "31-35", "36-40", "41-45", "46-50", "51-55", "56-60")
  ) + # custom labels displaying the groups of 5 picks
  theme(
    plot.title = element_text(size = 15, hjust = 0.5),
    plot.caption = element_text(hjust = 0.0)
  ) # centering the title +
```

```
## Warning: Continuous limits supplied to discrete scale.
## i Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



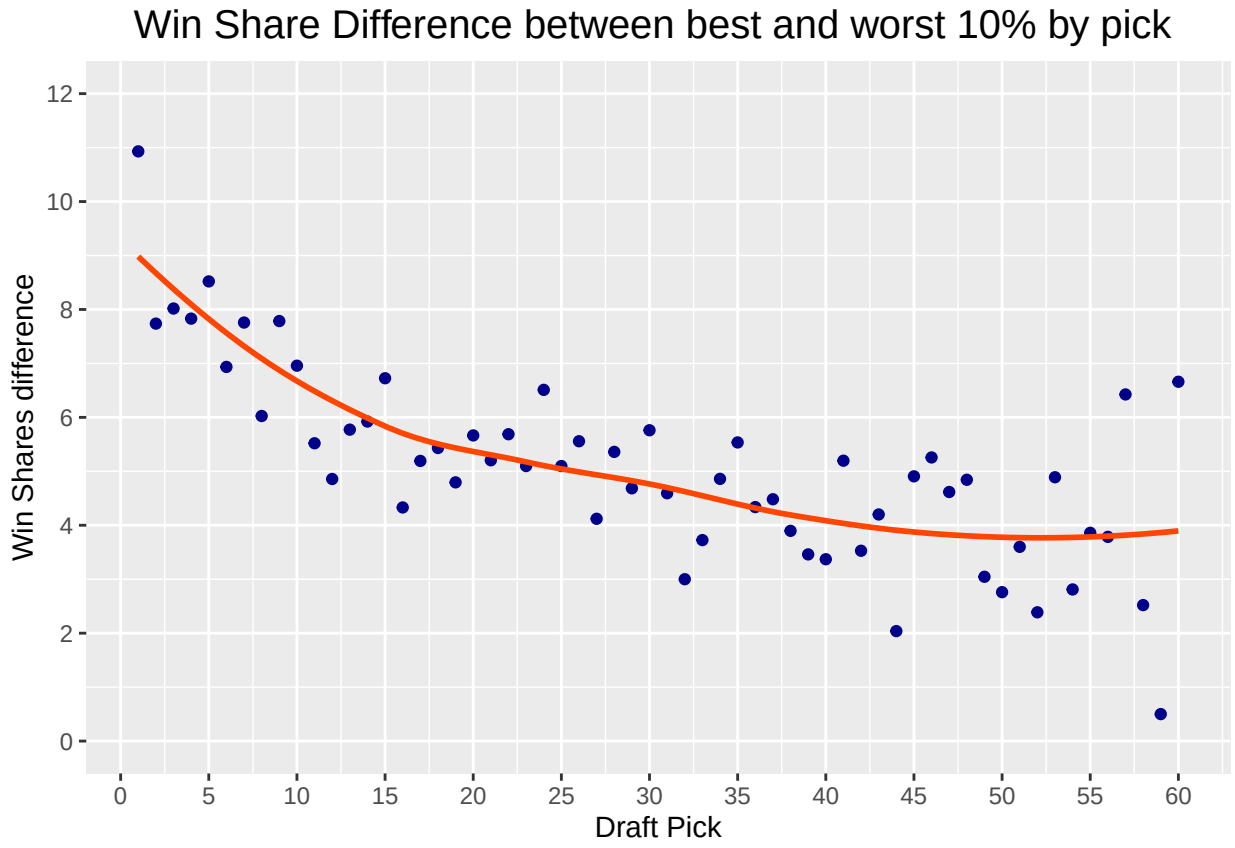
Note: The dashed horizontal line indicates the average player's win shares

Plot 5: Difference between best and worst players by pick:

```
high_low_10tile_diff %>%
  ggplot(
    aes(
      x = Pick,
      y = high_low_diff
    )
  ) +
  geom_point(
    color = "darkblue"
  ) + # scatter plot
  geom_smooth(
    color = "orangered",
    se = F # no std error
  ) + # loess best fit line
  labs(
    x = "Draft Pick",
    y = "Win Shares difference",
    title = "Win Share Difference between best and worst 10% by pick "
  ) + # better labels and a title
  scale_x_continuous(
    limits = c(1, 60),
    breaks = seq(0, 60, 5)
```

```
  ) + # better y scale
  scale_y_continuous(
    limits = c(0, 12),
    breaks = seq(0, 12, 2)
  )  + # better x scale
  theme(
    plot.title = element_text(size = 15, hjust = 0.5)
  ) # centering the title
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Discussion

These five plots were created to understand how players selected in the NBA draft provide value to their team. The analysis focused on average win shares per season for players drafted between 1980 and 2016.

The scatterplot shows the mean win shares for players drafted at each pick with error bars of one standard deviation above and below. The smooth trendline shows that early picks do perform better than later ones, but after around the 15th pick this difference is marginal. Additionally, the dahsed line, indicating the average player win shares shows that on average most players drafted in the first round (picks 1-30) perform better than average while second round picks (30-60) tend to perform worst. The error bars show that there is high variability in performance for every pick, but especially at the beginning and ends of the draft. The first few spots in the draft are characterized by superstar players or players often labelled as "busts" leading

to the large standard deviation. The end of the draft generally either has players that only last a few seasons in the NBA or turn into diamonds in the rough, outperforming their expectations. A few late pick stars are keeping the end of the draft varied and relevant.

Next, the box plots depict the average performance and distribution of player's drafted in the top 5 picks, 6th-10th pick, and so on. Similar to the scatter plot, the early picks perform the best, with performances in the later picks levelling out. The variation in the top 5 picks is minimized potentially due to the sample size increasing (better players play more seasons and more games). The late 2nd round picks (46-50, 51-55, and 56-60) still have large variability due to those overachieving sleeper players. By grouping the picks into sets of 5, it is clear that after the first 15 or so picks, performances are mostly the same.

The histogram shows how each group of five picks performs by average win shares. The darker the color, the higher the pick. This historgram is skewed left which makes sense as poor performing players are cut by their teams while very high performing players can ascend to superstar heights. This graph reinforces that top of top picks (1-15) perform much better than the rest of the draft, and it shows that very few low drafted players achieve strong results in the NBA. The data is centered around 1.5 win shares which is very close to the overall average of 1.66 win shares. This data also has a gap from 4 to 5.5 win shares showing that the elite talent in the NBA is a step above everyone else.

The next scatterplot shows the mean win shares by group of 5 picks. This plot is to the first scatterplot, but it condenses the data from 60 discrete picks to 12 groups of 5 picks. The conclusions from the plot remain the same. Top picks perform the best while after the top 15 picks performances levels off. Again the data is most varied at the top and bottom of the draft.

Finally, the last scatterplot displays the variablity in the best and worst performers at each pick. This plot shows the difference between the mean performances of the top 10% and bottom 10% of players at each pick. Essentially, it shows how much better the best players perform than the worst players at each pick. The large differences for the top 10 (1-10) and bottom 5 (55-60) picks mean that these picks have a large variability in performance. This corroborates the conclusions drawn from the previous plots, where the top and bottom have the most variability.

Overall, these plots mostly validate the predicted hypothesis. The top 10 to 15 picks in the draft perform significantly better than the rest of the draft. Then the remaining 45 to 50 picks perform relatively similar on average. However, picks at the top and bottom of the draft have much more variability, but potentially for different reasons. The top draft picks have the potential to be superstars, and if they meet that projection they can become a perennial franchise player. However, these expectations can also doom a player's career where they can turn into a "bust" and fail on their great potential. The bottom draft picks are simply trying to survive as an NBA player. On average, players selected at the bottom of the draft on last for around 4 seasons as compared to around 9.5 for top ten picks. Thus most bottom draft picks fall out of the league, but the ones who stay have to perform well. In summary, the top draft picks are worth it, but mid to late draft picks have little difference between them.