SPM 5708 Homework 4 (Correlations & Scatterplots)

This week, we discussed the topic of *sabermetrics* as it relates to the process of finding metrics that help us better evaluate the performances of players and teams. Metrics we would describe as "good" metrics are those that (a) accurately measure success/performance (relevancy), (b) differentiate between players of varying skill levels (discriminatory), and (c) separate skill from luck or other factors (consistency/stability). We also learned about a statistic, the *correlation coefficient,* that can be used to examine *the strength of the linear relationship between two quantitative variables,* and about the *scatterplots* that can be used to visualize them. This homework assignment further explores each of these elements. For questions that require an answer, please provide your answers in a different font color.

4.1 Exploring Advanced Batting Data Using Correlations and Scatterplots

1. Install and read the *baseballr* package into your current R session. This package provides useful functions for collecting and analyzing MLB data, including functionality for linking to Statcast, Baseball-Reference, and FanGraphs databases.

2. Let's use the following function to scrape a season's worth of batting data from the FanGraphs batting leaderboard:

   bat_data=fg_batter_leaders(startseason=2023,endseason=2023)

   If you get an error after running this, simply download the *bat_data* .csv file from Canvas and import it to R Studio before continuing with the next line of code.

3. To avoid potential outliers, filter the data to create a new data frame called *bat_use* that only contains players who made at least 50 plate appearances (*PA*) in this season.

4. Which of the following statistics displays the strongest correlation to infield hit percentage (*IFH_pct*), the percentage of batted ground balls that go for hits? Record each stat's correlation coefficient in relation to *IFH_pct* below:

   EV (exit velocity; avg. speed at which batted balls leave a player's bat)
   Spd (speed score; approximates a player's speed on a scale of 1 [slowest] to 10 [fastest])
   Contact_pct (contact percentage; the percentage of total swings that result in ball contact)
   GB_pct (groundball pct.; the percentage of a player's batted balls that are groundballs)

   Which metric displayed the *strongest* relationship with *IFH_pct*? How did you determine this?

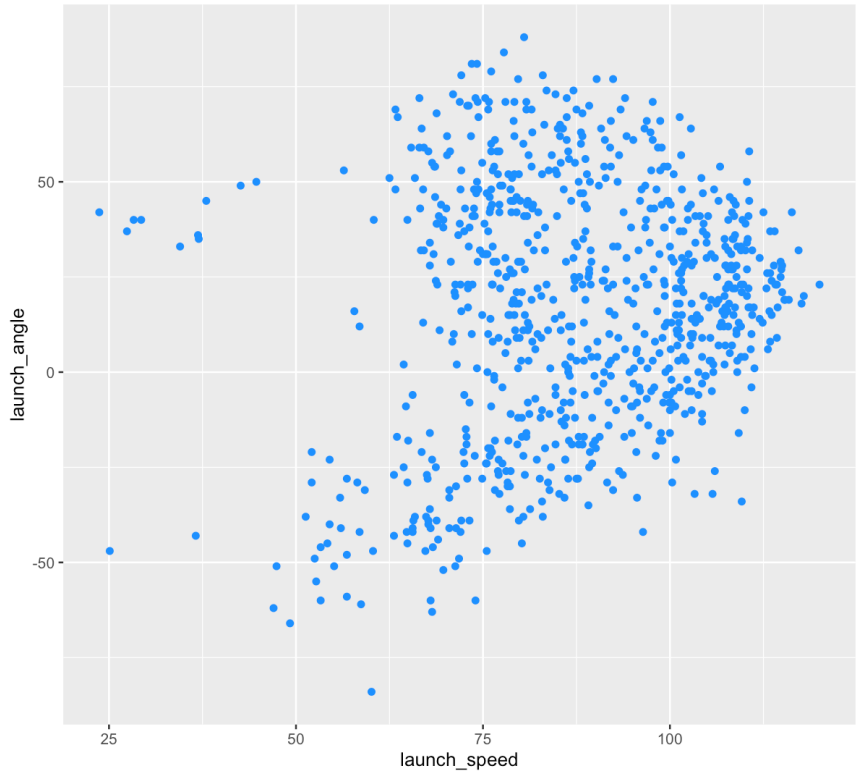
   Answer: The correlation between IFH_pct and EV is -.1049752. The correlation between Spd and IFH_pct is 0.275508. The correlation between IFH_pct and Contact_pct is -0.01508206. The correlation between IFH_pct and and GB_pct is 0.05089649. Spd displays the strongest correlation w IFH_pct with a correlation of .275508 since its pearson product (r) has the largest absolute value.

5. Import the *Ohtani_2025* dataset and find the correlation between Ohtani's exit velocity or launch speed (*launch_speed*) and his launch angle (*launch_angle*) using this code:

   cor(Ohtani_2025$launch_speed,Ohtani_2025$launch_angle,use="complete.obs")

Note that we must attach the *use="complete.obs"* function to the end of the code because there are NAs for some of the values in the data.

Launch angle is recorded with a middle value of 0 to indicate a ball hit parallel to the ground; negative values imply a downward (groundball) trajectory, and positive values imply an upward (flyball) trajectory. Thus, it probably won't display a linear relationship with launch speed (exit velocity) since balls can be hit with pace at many angles. To see this, create a scatterplot with *launch_speed* on the x-axis and *launch_angle* on the y-axis. Don't forget to read in (via *library*) the *ggplot2* package prior to creating the plot.



6. Run the following code to create a new column called *is_hr* that labels Ohtani's home runs:

   Ohtani_2025=Ohtani_2025%>%mutate(is_hr=ifelse(events=="home_run","HR","NO"))

7. Now, replace the last layer in your previous scatterplot code with this chunk of code and re-run it:

   +geom_point(aes(color=is_hr))

   What appears to be the "sweet spot" in terms of the launch angle and launch speed of Ohtani's home runs?

Answer: The sweet spot for Ohtani's home runs appears to be between launch speed = (100, 125) mph and launch angle = (20, 40)º.

8. Calculate the mean (average) launch angle and launch speed *of Ohtani's home run hits.* Report these values below.

   Answer: The mean launch angle on Ohtani's home runs is 29.61818º. The mean launch speed on Ohtani's home runs is 109.1582 mph.

4.2 Using Correlations and Scatterplots to Explore Basketball Data

1. Import the *NCAAB_Efficiency* dataset (.xlsx) and call it *NCAAB.* Start by calculating the correlation coefficients for points scored per game (*PPG*) and *WinPct*, and opponents' points per game (*OPPG*) and *WinPct*. Report them below.

   Answer: The correlation between PPG and WinPct is . 6005331. The correlation between OPPG and WinPct is -0.595546.

2. Now, find the correlations between *ORtg* and *WinPct*, and *DRtg* and *WinPct*. Comparing these results to the correlations from the prior question, what can you infer? Note that *ORtg* is a measure of points scored per 100 possessions, while *DRtg* measures points allowed (opponents' points scored) per 100 possessions.

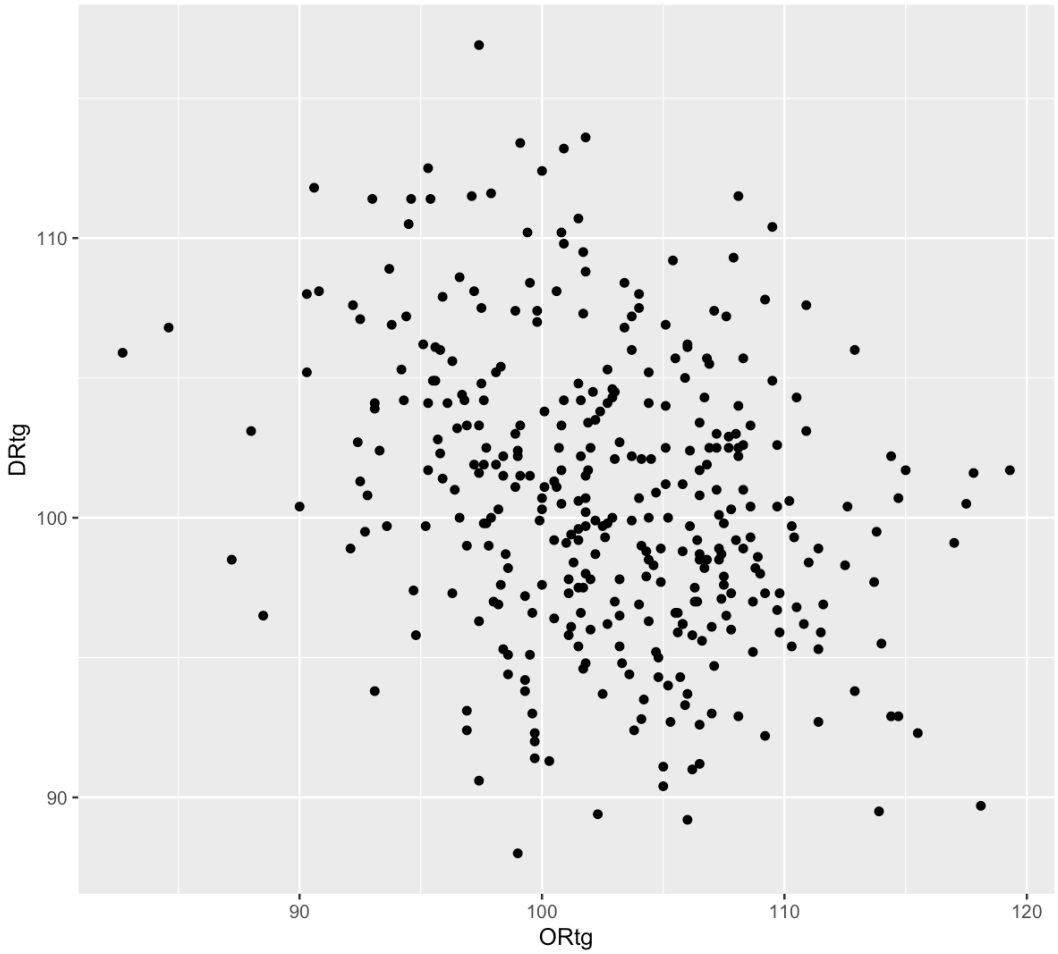
   Answer: The correlation between ORtg and WinPct is .7736732. This makes sense as we would expect our win percentage to increase as we score more points per 100 possessions. The correlation between DRtg and WinPct is -0.7261376. This also makes sense because we would expect our win percentage to decrease as we allow more points per 100

possessions. Despite both of these relationships being fairly strong, the correlation between ORtg and WinPct is slightly stronger than the correlation between DRtg and WinPct. These are both important factors in winning a game, however; it seems that having a better offense is slightly more important for increasing win percentage than having a better defense, on a linear model.

3. Is there any relationship between being an efficient offensive team (*ORtg*) and an efficient defensive team (*DRtg*)? Report evidence in support of your answer.

   Answer: There is a fairly weak, negative relationship between ORtg and DRtg, since the correlation between the two is -0.2924971.

4. Create a scatterplot that shows *ORtg* on the x-axis and *DRtg* on the y-axis.

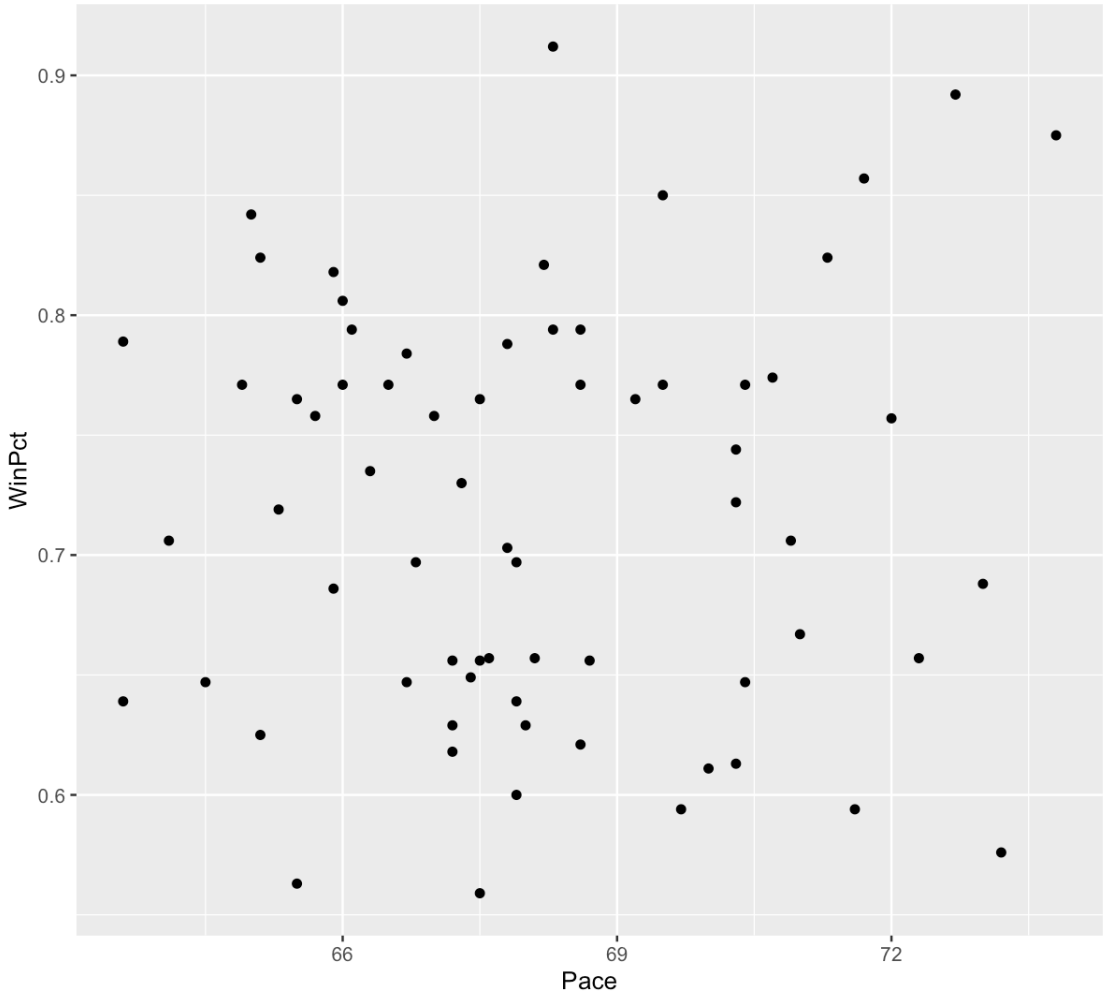


5. Next, find the correlation between the *Pace* (possessions per game) and *WinPct* (win percentage) variables. What does the resulting coefficient suggest?

Answer: The correlation between WinPct and Pace is -.05719578 suggests that there is not much of a linear relationship between WinPct and Pace (# possessions per game). I would infer that this lack of strength is due to the fact that for each game, one team's win percentage is going to increase, and the other's is going to decrease despite both playing the same number of possessions.

6. Does the correlation coefficient between *Pace* and *WinPct* change when we only examine tournament teams who have a 1 recorded in the *Tourney* column? In addition to finding the correlation coefficient, create a basic scatterplot of this relationship and paste it below.

Answer: When we only examine teams who have a 1 recorded in the tourney column, the correlation between WinPct and Pace changes to .03327786.



4.3 Using Scatterplots to Examine the Linearity of Relationships

1. Import the *PGCareers* dataset (.csv) and note that each row corresponds to the season-level performance stats of a prominent (former) NBA point guard. All stats are recorded on a per-game basis aside from minutes played (*MP*). Start by creating a new variable column called *MPG* that shows minutes-per-game for the players in each row.

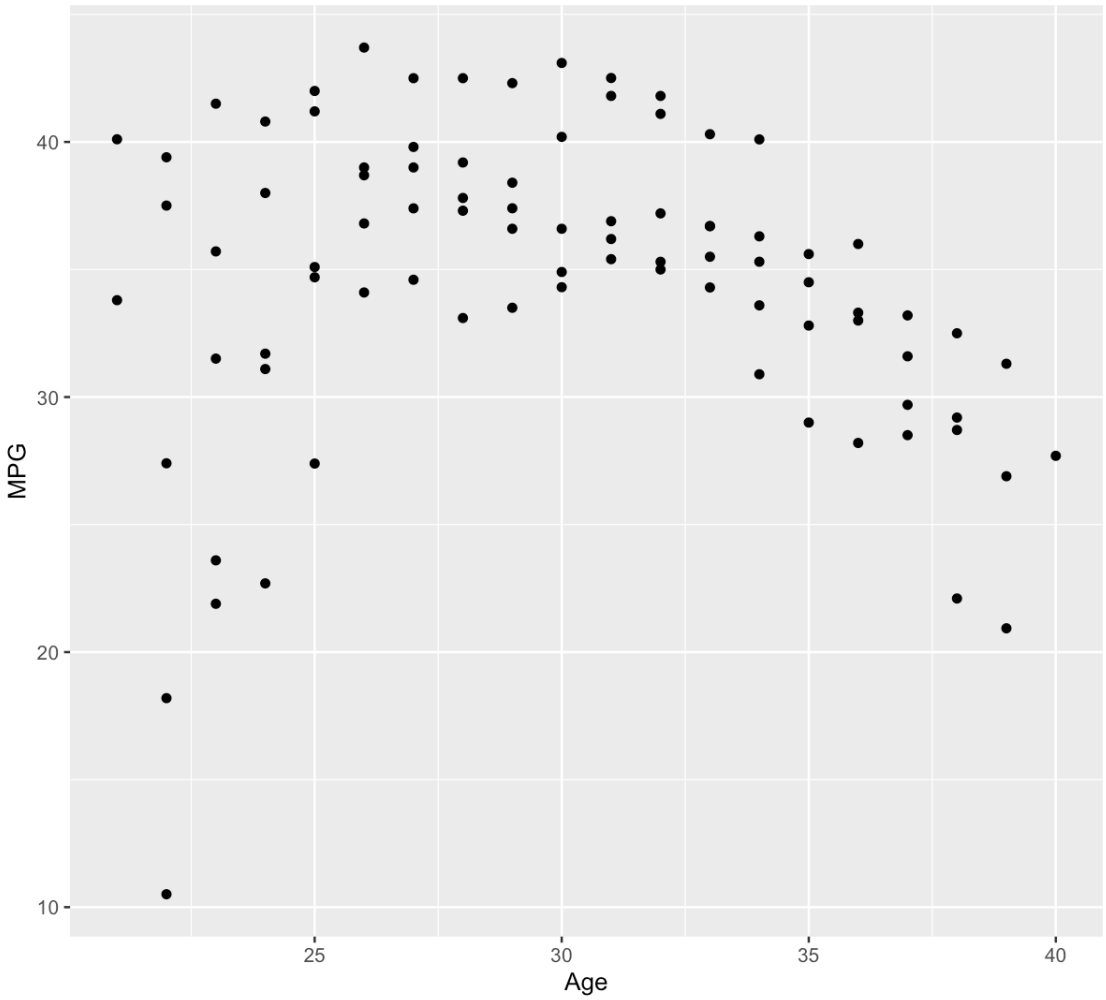
2. Would it be appropriate to run a correlation between players' *MPG* values and their *Age*? Why or why not? When you attempt this, what value do you obtain?

   Answer: The correlation between MPG and Age is -.126942. This shows that there is a very weak, linear relationship between MPG and Age. I would argue that this is not appropriate because there is not a linear relationship between MPG and Age. I would imagine that most

rookies and vets do not get as much playing time, while players in their peak play the most minutes per game.

3. Perhaps a scatterplot could help us visualize the shape of the trend between players' minutes-per-game values and their age. Create a scatterplot using the following code and paste it below:

   ggplot(PGCareers,aes(x=Age,y=MPG))+geom_point()

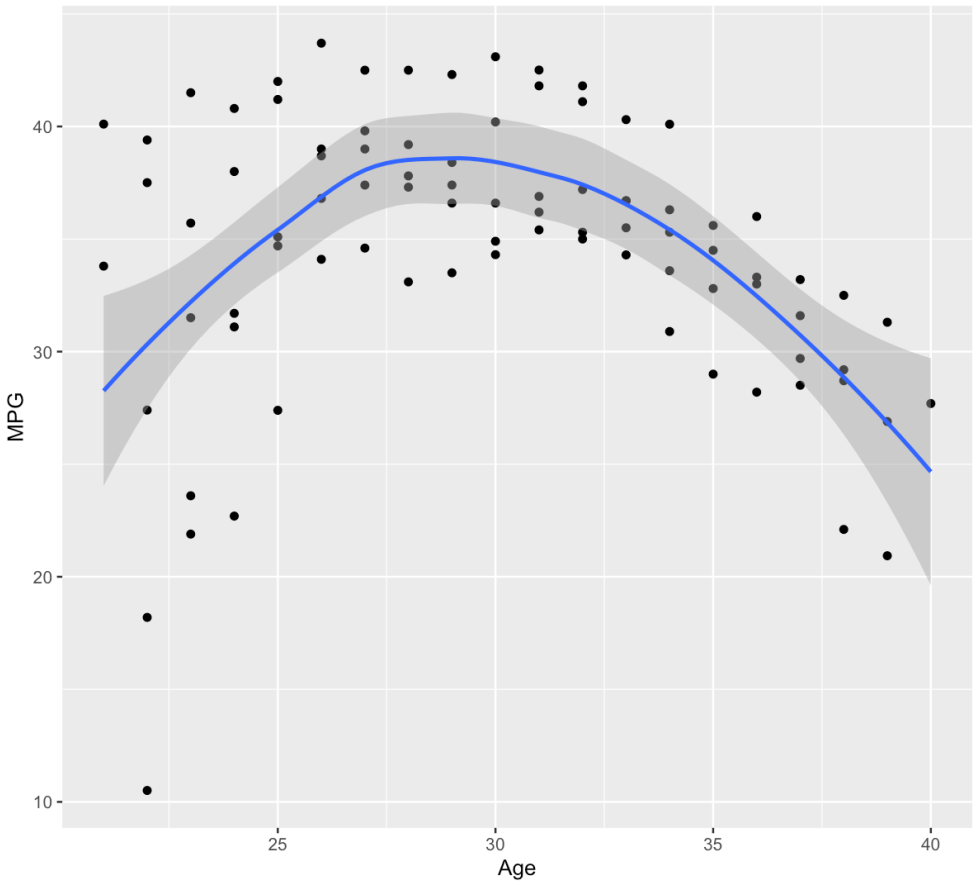


4. Note how we can add the following layer to our *ggplot* code to display a "line of best fit" that better portrays the nonlinearity of the relationship:

   ggplot(PGCareers,aes(x=Age,y=MPG))+geom_point()+geom_smooth(method="loess")

   What does the shape of this trendline suggest about the relationship between playing time and age in basketball?

   Answer: The shape of this trendline suggests that there is a negative parabolic relationship between Age and MPG. This supports my claim from part 3 that rookies and vets do not play as much as players in the peak of their careers.



Submit your completed R Script and answers to the requested questions in this Word document to the Canvas drop box by the next class.