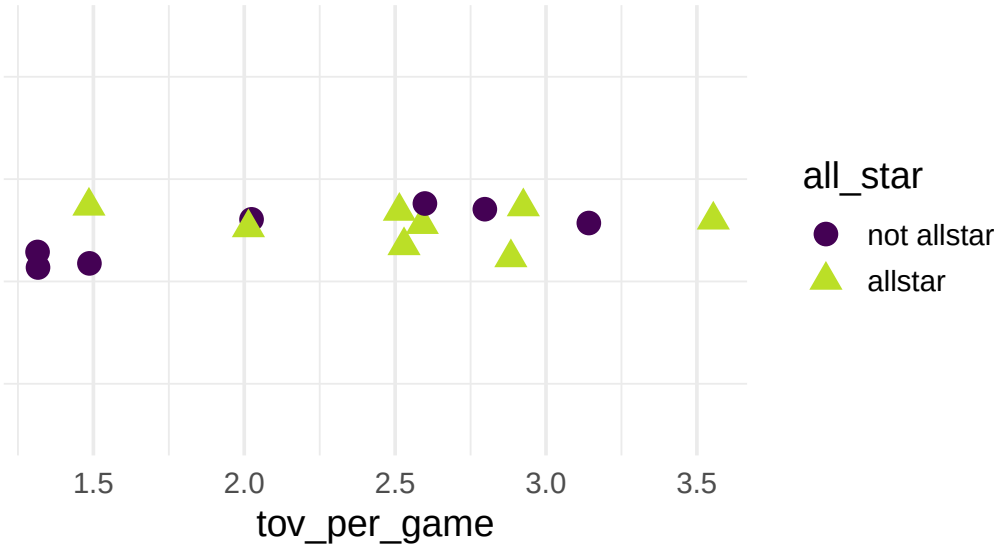
Worksheet on k-Nearest-Neighbors (SOLUTIONS)

Part 1 (After First Video on Introducing knn)

In the video, you saw how the k-nearest-neighbors algorithm works with a `pts_per_game` predictor to predict whether or not NBA players made an all-star team.

Here, you will apply the algorithm using a different predictor: turnovers per game (called `tov_per_game`).



Exercise 1. With $k = 1$, predict whether or not the following 10 players make the all-star team. Write in either `allstar` or `not allstar` as a new column called `pred_1` in the printout below (at the turnover value of 1.5, assume that the blue circle has a higher turnover value than the green triangle. Additionally, you might have to make an educated guess about which observation is "closer" for some of the players below, which is just fine :)).

```
# A tibble: 10 x 4
   player             tov_per_game pred_1                   pred_3
   <chr>                     <dbl> <chr>                    <chr>
 1 Keita Bates-Diop            0.8 not allstar              not allstar
 2 Kelly Olynyk                2.5 allstar                  allstar
 3 Kenyon Martin Jr.           1.1 not allstar              not allstar
 4 Kevin Durant                3.3 not allstar (or allstar) allstar
 5 Kevon Looney                0.5 not allstar              not allstar
 6 Lauri Markkanen             1.9 allstar                  not allstar
 7 LeBron James                3.2 not allstar              allstar
 8 Markelle Fultz              2.3 allstar (also close)     allstar
 9 Mason Plumlee               1.5 not allstar              not allstar
10 Michael Porter Jr.          1.1 not allstar              not allstar
```

Exercise 2. With $k = 3$, predict whether or not the 10 players above make the all-star team. Write in either `allstar` or `not allstar` as a new column called `pred_3` in the printout above.

See above.

Exercise 3. From the graph, it looks like a lot of players with a large amount of turnovers still made the all-star team, even though giving up a turnover is "bad." Come up with a hypothesis for why this might be.

One logical hypothesis is that players who are good just have the ball a lot so are more likely to turn the ball over for that reason!

Exercise 4. Recall that there are 15 total NBA players shown on the graph. selected to be on the all-star team? What would happen if you set $k = 15$, the total number of players shown on the graph? How would the algorithm classify each of the 10 players in the data printout?

If k was 15, then the the algorithm would classify each of the 10 players as whatever the most popular category is of the 15 players. In this case, there are 8 all stars and 7 "non" allstars, so knn would classify all 10 players as all stars.

Exercise 5. Consider the $k = 1$ setting again. How might you break ties if two players were exactly the same distance away from the player you are trying to classify?

A couple of reasonable answers might be: (1) use the next nearest neighbor to break the tie, (2) break the tie randomly.

Part 2 (After Second Video on Classification Rates)

Exercise 1. Using the turnovers variable as the predictor with $k = 1$ and your predictions from Part 1, construct a confusion matrix by filling in the table below.

Recall from the slides that, of the 10 players in the printout, only Durant, Markkanen, and James were actually selected as all-stars (the other 7 players were not selected as allstars that year). You might want to make a new column called `all_star` in the printout on the previous page that gives the value `allstar` for those 3 players and `not allstar` for the other 7 players to help with your construction of the confusion matrix).

I am using my "best guesses:" other confusion matrices might be a little different depending on how people classified Durant and Fultz.

	Predicted Allstar	Predicted Not
Allstar	1	2
Not	2	5

Exercise 2. From your confusion matrix, calculate the classification rate.

6/10 = 0.6.

Exercise 3. Using $k = 15$ (the max that k can be), construct a confusion matrix using the turnovers variable as the predictor. **Hint**: Look back at Exercise 4 on the previous page.

	Predicted Allstar	Predicted Not
Allstar	3	0
Not	7	0

For this exercise, all 10 players would be predicted to be `allstars` *if* $k = 15$*. Three of these predictions would be correct while 7 of them would not be.*

Exercise 4. From your confusion matrix, calculate the classifciation rate.

3/10 = 0.3.

Exercise 5. Which k is better for the turnovers predictor? $k = 1$ or $k = 15$? Why?

k = 1 is better: we get a higher classification rate!