Performing a shot analysis on Kobe Bryant

STAT167 Final Project Presentation

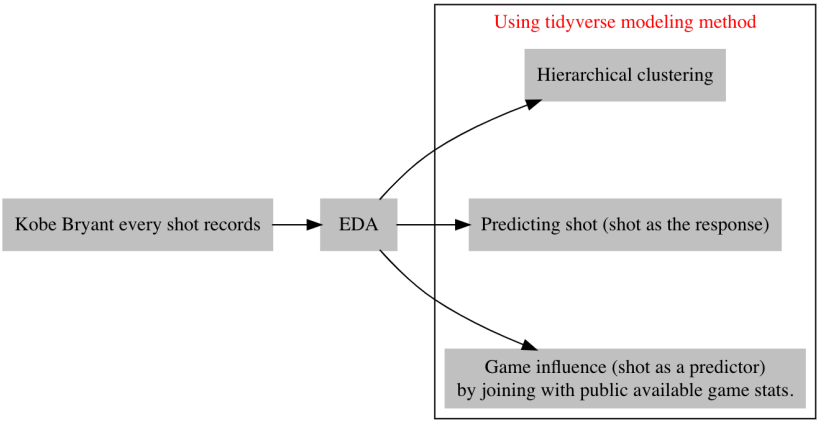
Oceans 4
(Lauren, Natasha, Patrick and Shiyuan)

06/02/2020

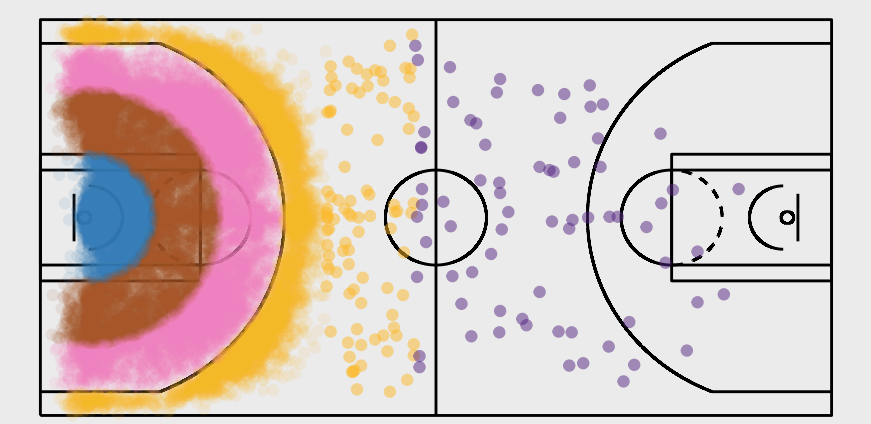
Rationale of the project

- Main topic: How accurate is Kobe and what variables contribute to his success as a player?
  - What types of shots (layups, jump shots, etc.) does Kobe do best (highest accuracy)?
  - What are his strengths and weaknesses as a basketball player?
  - What parts of the court is he the most accurate in? (hotspots on the court)
  - What season(s) was Kobe's prime? Season(s) he won MVP? Seasons he was injured? Or How Kobe improved/ receded over time? (also using 2pt, 3pt, ft, season, etc as variable)?

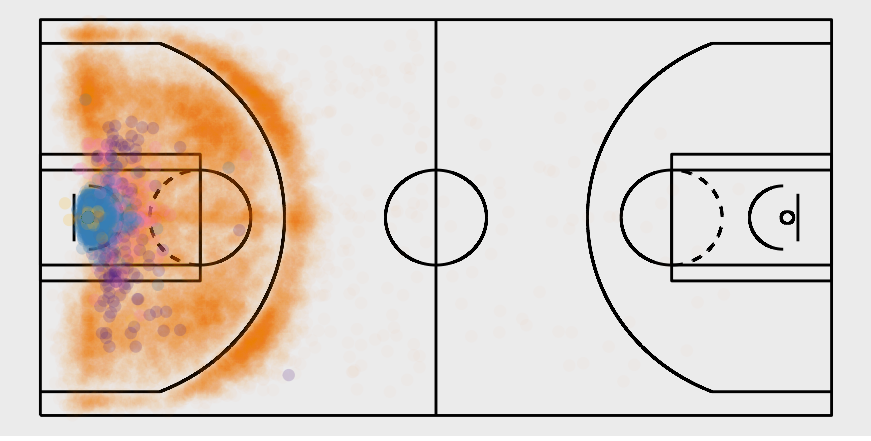
Workflow


Using tidyverse modeling method
Hierarchical clustering
Kobe Bryant every shot records
EDA
Predicting shot (shot as the response)
Game influence (shot as a predictor)
by joining with public available game stats.

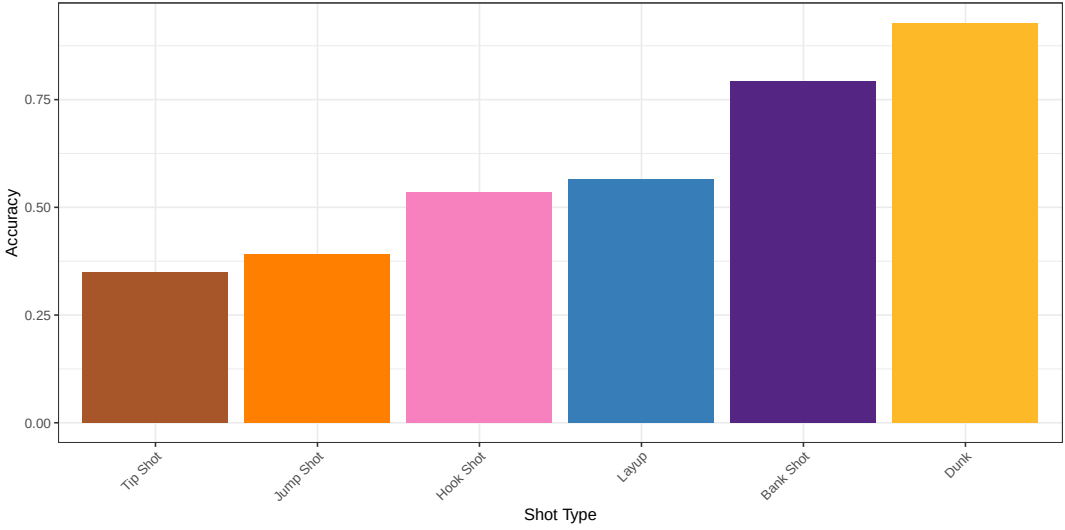
Shot Distance



Shot Types



What types of shots did Kobe do best?



Modeling in tidyverse

```
## Documented in a vignette from broom package
set.seed(167)
df <- kobe %>%
   select(shot_made_flag, shot_distance, t_sec, home, opponent, season) %>%
   modelr::crossv_kfold(k = 10) %>%   # produce training and test column
   mutate(glm = map(train, ~glm(shot_made_flag ~ ., data = .x, family = binomial)), # mapping training to glm
          pred = map2(glm, test, ~predict.glm(object = .x, newdata = .y, type = "response")), # predict by test
          pred_class = map(pred, ~if_else(.x > 0.5, 1, 0)), # apply a cutoff and obtain prediction class
          true_class = map(test, ~{as_tibble(.x)$shot_made_flag}), # extract original label
          misclass_error = map2_dbl(pred_class, mis_class, ~mean(.x != .y))) # calculating misclassification error
```

```
# A tibble: 10 x 8
   train        test         .id   glm          pred          pred_class    true_class    misclass_error
   <named list> <named list> <chr> <named list> <named list>  <named list>  <named list>           <dbl>
 1 <resample>   <resample>   01    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.396
 2 <resample>   <resample>   02    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.405
 3 <resample>   <resample>   03    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.403
 4 <resample>   <resample>   04    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.4
 5 <resample>   <resample>   05    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.400
 6 <resample>   <resample>   06    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.411
 7 <resample>   <resample>   07    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.405
 8 <resample>   <resample>   08    <glm>        <dbl [2,569]> <dbl [2,569]> <dbl [2,569]>          0.390
 9 <resample>   <resample>   09    <glm>        <dbl [2,569]> <dbl [2,569]> <dbl [2,569]>          0.417
10 <resample>   <resample>   10    <glm>        <dbl [2,569]> <dbl [2,569]> <dbl [2,569]>          0.396
```

Detailed examination of the output

```
# A tibble: 10 x 8
   train        test         .id   glm          pred          pred_class    true_class    misclass_error
   <named list> <named list> <chr> <named list> <named list>  <named list>  <named list>           <dbl>
 1 <resample>   <resample>   01    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.396
 2 <resample>   <resample>   02    <glm>        <dbl [2,570]> <dbl [2,570]> <dbl [2,570]>          0.405
  ...
 9 <resample>   <resample>   09    <glm>        <dbl [2,569]> <dbl [2,569]> <dbl [2,569]>          0.417
10 <resample>   <resample>   10    <glm>        <dbl [2,569]> <dbl [2,569]> <dbl [2,569]>          0.396
```

```
str(df$test[[1]])
List of 2
 $ data: tibble [25,697 × 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
  ..$ shot_made_flag: num [1:25697] 0 1 0 1 0 1 1 0 0 1 ...
  ..$ shot_distance : num [1:25697] 15 16 22 0 14 0 12 12 25 17 ...
  ..$ t_sec         : num [1:25697] 622 465 412 379 572 532 372 216 116 660 ...
  ..$ home          : chr [1:25697] "away" "away" "away" "away" ...
  ..$ opponent      : chr [1:25697] "POR" "POR" "POR" "POR" ...
  ..$ season        : chr [1:25697] "2000-01" "2000-01" "2000-01" "2000-01" ...
 $ idx : int [1:2570] 2 4 24 35 55 56 61 63 64 74 ...
 - attr(*, "class")= chr "resample"
```

```
str(df$pred[[1]])
 Named num [1:2570] 0.44 0.611 0.313 0.617 0.306 ...
 - attr(*, "names")= chr [1:2570] "1" "2" "3" "4" ...
```

The advantages of implementing modeling in table

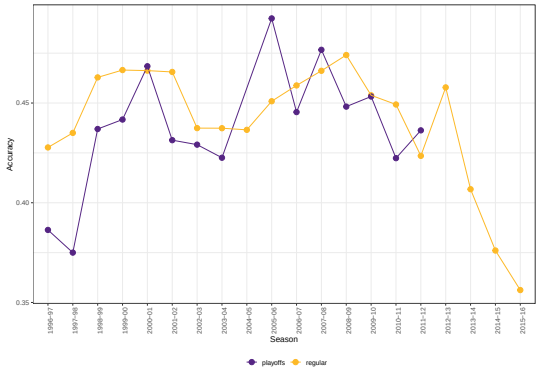
- "Vectorizing" the code and it is thought to be more efficient (Wickham, 2014).
- Easier to track input/output of each step, thus ensuring reproducibility.

Kobe accuracy varies by different season/playoffs

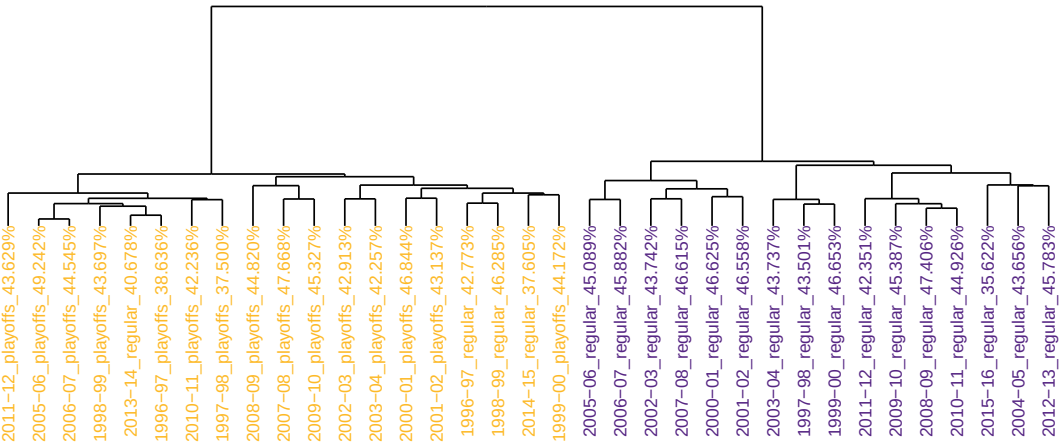
- 20 years of career, some of them have regular and playoffs seasons.
- Actual season x playoffs = 35 categories in total.

Kobe accuracy varies by different season/playoffs

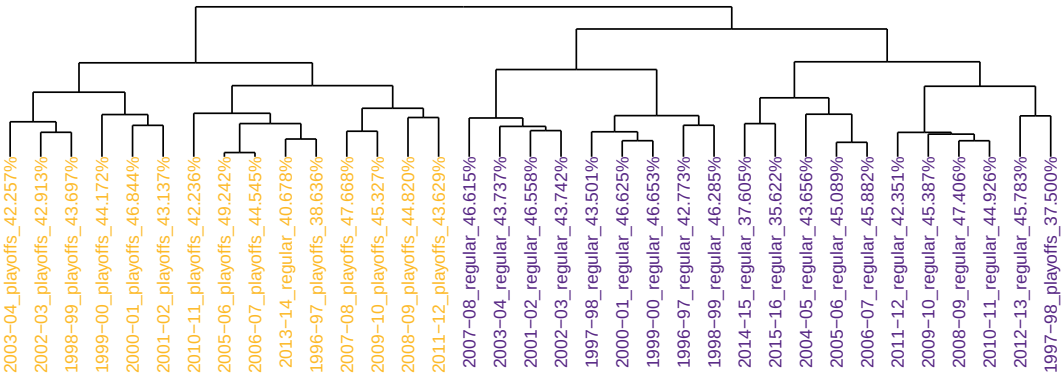
- 20 years of career, some of them have regular and playoffs seasons.
- Actual season x playoffs = 35 categories in total.
- Acuracy varies



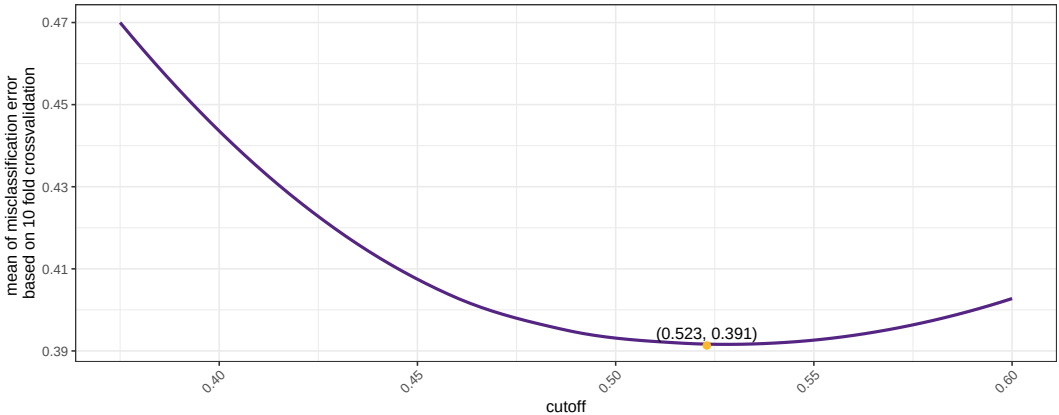
Hierarchical clustering based on distance



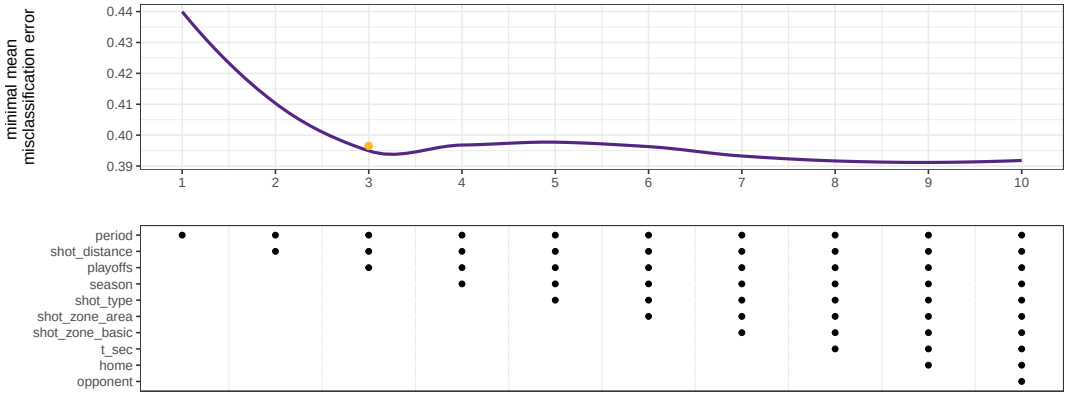
Hierarchical clustering based correlation



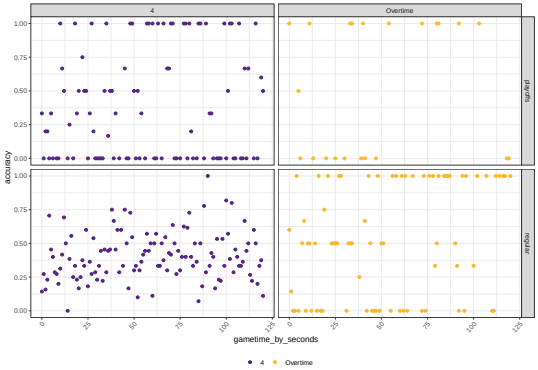
Choosing a cutoff for full GLM model



Selecting variables that best predict his accuracy



Shot accuracy towards the end of the game



Playoffs

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.312496   0.076956   4.061 8.84e-05 ***
QuarterOvertime    0.145024   0.100721   1.440    0.153
gametime_by_seconds 0.001350   0.001097   1.230    0.221

Residual standard error: 0.4248 on 118 degrees of freedom
Multiple R-squared:  0.0267,    Adjusted R-squared:  0.01021
F-statistic: 1.619 on 2 and 118 DF,  p-value: 0.2025
```

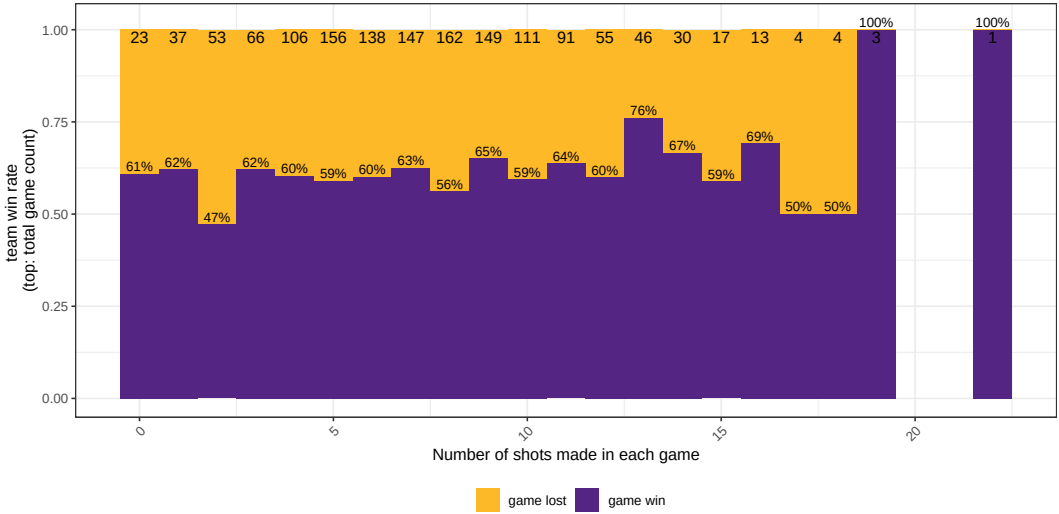
Regular season

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.2964657  0.0441531   6.715 1.99e-10 ***
QuarterOvertime    0.1324562  0.0423870   3.125  0.00205 **
gametime_by_seconds 0.0018200  0.0005894   3.088  0.00231 **

Residual standard error: 0.2908 on 196 degrees of freedom
Multiple R-squared:  0.083, Adjusted R-squared:  0.07364
F-statistic:  8.87 on 2 and 196 DF,  p-value: 0.0002052
---
```

Does Kobe have a positive influence on overall game?

```
glm(WL~accuracyByDay, data = wl, family = binomial) %>% summary()

Call:
glm(formula = WL ~ accuracyByDay, family = binomial, data = wl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1330  -1.2744   0.8276   0.9900   1.5683

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8839     0.1869  -4.730 2.25e-06 ***
accuracyByDay  3.0502     0.4140   7.367 1.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1888.8  on 1411  degrees of freedom
Residual deviance: 1829.7  on 1410  degrees of freedom
AIC: 1833.7
```

How much Kobe need to score for the team to win?



What we have done

- From this data we were able to build multiple models.
- Clustering model for shot accuracy for regular season and playoffs.
- GLM for shot accuracy based on period, time left in game, and type of game (regular season/playoffs).
- GLM for overall "game influence".

We used the Kaggle data set for our data visualization and modeling. We had considered other data sets that include injury information but we discovered that the data points were either insufficient to make conclusions or didn't answer our questions.

How has the work been split?

- Patrick: **The basketball expert** to make sure we are using the correct variables and making reasonable predictions.
- Lauren: **Statistical modeling** and testing variables to make sure our data makes sense.
- Shiyuan: **Coding** and dealing with any difficulties we have with coding.
- Natasha: Data Visualization and **overall clean up**. Typing up all project related info.

Thank you for your attention!

References

Grolemund, G., and Wickham, H. R for Data Science:: Import, Tidy, Transform, Visualize, and Model Data (O'REILLY).

Wickham, H. (2014). Advanced R (Taylor & Francis).