DS6372 Stats 2 Project 2: Kobe Bryant Shot Selection!!!

By Hao Wang, Muchigi Kimari and Billy Nayden

April 2020

1. Introduction

Kobe Bryant marked his retirement from basketball by scoring 60 points in his final game as a member of the Los Angeles Lakers on Wednesday, April 12, 2016. Starting to play professional basketball at the age of 17, Kobe earned the sport's highest accolades throughout his long career.

The original data set contains the location and circumstances of every shot attempted by Bryant during his 20-year career. Our task is to predict whether the shot went in (shot_made_flag = 1) or missed (shot_made_flag = 0).

For this exercise, 5000 of the shot_made_flags have been removed from the original data set. Our goal is to provide the best predictions possible.

2. Data Description

The original data set came to us in CSV format in a file called project2KobeData.csv. We loaded this into R as a data frame called Kobe.

The data has a total of 20,697 observations and 29 variables. The response variable is called shot_made_flag and is a factor with two levels: 1 ("make") and 0 ("miss"). There are 28 explanatory variables, 10 of which are factors and 18 of which are numeric. We summarized these variables in R (**Figure 2.1** and **Figure 2.2**).

3. Exploratory Data Analysis (EDA)

Before conducting our EDA, we first cleaned the data by removing particular variables, and creating new ones to help assist in our analysis and model.

We opted to remove the following variables:

- **Game_event_id:** This is an unnecessary ID field that is covered by multiple other factor variables in the data set.
- **Team_id:** This is a single unique value (1610612474) because Kobe only played with one team over the course of his career.

- **Team_name:** This is a single unique value (Los Angeles Lakers) because Kobe only played with one team over the course of his career.
- **Shot_id:** This duplicates the variable **recId**.

We opted to create the following variables:

- **Home_away:** This is a factor with two levels "home" and "away."
- **Shot_time:** This is the exact time a shot was taken 2) by using **period**, **minutes_remaining**, **seconds_remaining**: **shot_time** = (**period** - 1) * 12 * 60 + (12 - **minutes_remaining**) * 60 + (60 - **seconds_remaining**).
- **Cst_Num:** This assigns a number to each type of shot from **combined_shot_type** (**Figure 3.2**).
- **SeasonID:** This assigns a number to each **season** (**Figure 3.3**).
- **St_Num:** This assigns a number to each **shot_type** (**Figure 3.4**).
- **Sza_Num:** This assigns a number to each **shot_zone_area** (**Figure 3.5**).
- **Szb_Num:** This assigns a number to each **shot_zone_basic** (**Figure 3.6**).
- **Szr_Num:** This assigns a number to each **shot_zone_range** (**Figure 3.7**).

Our new data frame is called kobe_clean and has 32 variables. We maintain the same response variable. We now have 31 explanatory variables, nine factors and 22 explanatory variables (**Figure 3.8** and **Figure 3.9**).

3.1. Checking for possible transformations and outliers

Next, we will check for potential needs for transformation and outliers. We will start with the numeric variables, then use the factor variables.

3.1.1. Numeric variable 1: recID

We checked the histogram with bin width of 1000 (**Table 3.1**) and 100 (**Table 3.2**) and the box plot (**Table 3.3**). The distribution of **recID** is univariate, there is no special pattern in the shape of the data, no outliers, and no need for transformation.

We also checked the scatter plot (**Table 3.4**) with **recID** as the y-axis and **season** on the x-axis. Kobe played in 20 consecutive seasons from 1996-97 through 2015-16. We identified the following while looking for abnormalities in the data:

- In 15 seasons, Kobe made the regular season and the playoffs
- In 5 seasons, Kobe did not qualify for the playoffs and only played in the regular season
- For seasons where Kobe made the playoffs, **recID** will be higher in the playoffs than the regular season

3.1.2. Numeric variable 2: game_id

We checked the histogram with bin width of 1e+07 (**Table 3.5**) and 1e+05 (**Table 3.6**), the boxplot (**Table 3.7**), and the scatter plot (**Table 3.8**).

The two histograms and the box plot demonstrated right skewed distribution with four clusters: regular season 1996-1999, regular season 2000 onward, playoffs 1996-1999, and playoffs 2000 onward. While there appeared to be outliers in the box plot, there were from the playoffs 1996-1999, so we opted to keep the observations in our data set. The scatter plot largely showed the same thing as **recID**, where there are five seasons that Kobe did not make the playoffs.

3.1.3. Numeric variable 3 and 4: lat, lon

We checked both location histograms with bin width of 0.01 (**Table 3.9** and **Table 3.10**) and they indicated that Kobe takes a disproportionate number of shots close to the basket and in the center of the court.

3.1.4. Numeric variable 5 and 6: loc_x, loc_y

We checked both location histograms with bin width of 0.01 (**Table 3.11** and **Table 3.12**) and they indicated the same things as the histograms for **lat** and **lon**.

This led us to consider if shots at location (0,0) should be considered differently in our model. However, the takeaways from exploring histograms (**Table 3.13** and **Table 3.14**) with (0.0) excluded revealed that we should include (0,0) within our existing data set, as well as the following observations:

- Kobe's x distribution showcases a balanced competence to the left and fight sides of the hoop.
- The x distribution is trimodal with peaks at the center area of the hoop and angles at 45 degrees to the left and right of the hoop.
- The distance of three-point shots is 285 inches, and Kobe's number of shots drops off significantly after 285 inches.
- 20.1% of the shots Kobe made were three-point shots.
- Kobe had a 33.1% shooting percentage for three-point shots and 47.8% shooting percentage for two-point shots.

3.1.5. Numeric variable 7, 8 and 9: minutes_remaining, seconds_remaining, shot_time

We looked at the histograms (**Table 3.15, Table 3.16,** and **Table 3.17**) of the three time variables in the data set. These graphs show the following:

- The majority of Kobe's shots happen in quarters 1-4. This is because quarters 5, 6, and 7 correspond to overtime periods that do not happen in every basketball game.
- Kobe takes more shots in the last two seconds of a quarter than at any other time over a given quarter.

Thus, we looked at a histogram of overall shot time (**Table 3.18**) to confirm when these shots happened over the course of a game. This helped us understand that Kobe takes fewer shots in the first half of the second and fourth quarters, likely because he is resting.

3.1.6. Numeric variable 10: period

We looked at the histogram of shots across periods (**Table 3.19**) and found a relatively flat distribution. However, we were able to confirm the rest periods in the first half of the second and fourth quarters we saw in the **shot_time** histogram.

3.1.7. Numeric variable 11: playoffs

We looked at the distribution of shots in the regular season and playoffs (**Table 3.20**) and found unsurprisingly that Kobe took more shots during the regular season, because he played more games. We did not find any outliers.

3.1.8. Numeric variable 12: shot_distance

We investigated the histogram of shot_distance (**Table 3.21**) and found that the majority of shots were taken at a shot distance of 0 (dunks and layups).

Thus, we also investigated the histogram of shot_distance excluding 0 (**Table 3.22**) and found the number of shots increases between 0 and 17 feet, then declines, only to increase again at 25 feet. This is because it is easier for guards like Kobe to get separation for jump shots a little bit removed from the basket where they are less likely to be double-teamed. Additionally, it highlights Kobe's sweet spot shooting distances of 17 and 25 feet.

3.1.9. Numeric variable 13 and 14: game_date, season

We investigate the histograms of shot counts by game_date and season (**Table 3.23** and **Table 3.24**) and found that Kobe's shot counts rise until 2005-2006, where they peak and being to fall. These two variables are highly correlated, however they are sliced differently (game_date provides more detail) so we are going to keep both.

3.1.10. Numeric variable 15: attendance

The histogram for attendance (**Table 3.25**) indicates that attendance is normally distributed.

3.1.11. Numeric variable 16: arena_temp

The histogram for arena_temp (**Table 3.26**) indicates that arena temperature is normally distributed, which makes sense because indoor arenas are climate controlled.

3.1.12. Numeric variable 17: avgnoisedb

The histogram for avgnoisedb (**Table 3.27**) indicates that arena noise is normally distributed. We will investigate if this is correlated to attendance, which is also normally distributed.

3.1.13. Factor variable 1: combined_shot_type

We investigated the histogram of combined_shot_type (**Table 3.28**), which indicated that Kobe's most popular shot was the jump shot, followed by the layup. Kobe made a higher percentage of layups than jump shots.

3.1.14. Factor variable 2: shot_type

The histogram of shot_type (**Table 3.29**) demonstrates that Kobe took more than three times as main two-point shots as he did three-point shots.

3.1.15. Factor variable 3: shot_zone_area

The histogram of shot_zone_area (**Table 3.30**) demonstrates that Kobe took and made more shots from the center than any other zone. He also was more successful from the right than from the left.

3.1.16. Factor variable 4: shot_zone_basic

The histogram of shot_zone_basic (**Table 3.31**) is very consistent with the results we saw from our earlier shot location analysis with the quantitative variables.

3.1.17. Factor variable 5: shot_zone_range

The histogram of shot_zone_range (**Table 3.32**) is very consistent with the results we saw from our earlier shot location analysis with the quantitative variables. It also showcases that Kobe rarely took back court shots.

3.1.18. Factor variable 6: matchup

The histogram of matchup (**Table 3.33**) shows a lot of variance, but it may be due to amount of playing time in each game. It is not evidence of a need for transformation or outliers.

3.1.19. Factor variable 7: opponent

The histogram of opponent (**Table 3.34**) shows a lot of variance, which is due to the change in teams from one season to another, along with many other confounding variables that affect this measure.

3.1.20. Factor variable 8: home_away

The histogram of home_away (**Table 3.35**) indicates that Kobe had a higher shooting percentage at home than away, however the box plot (**Table 3.36**) indicates that the difference is minimal.

3.2. Checking for multicollinearity

We found evidence of multicollinearity between the location variables lon, lat, loc_x, and loc_y by looking at the scatter plots of the variables and establishing the similarity of the shapes (**Table 3.37** and **Table 3.38**).

We then created a Pythagorean distance variable called loc (**Figure 3.10**) using the x and y coordinates from these scatter plots (**Table 3.39**). We then found that this was extremely linearly correlated (**Figure 3.11**) with the shot_distance variable.

Thus, we created a variable called deg (**Figure 3.12**), which measures the angle at which Kobe took a shot from (**Table 3.40**), and removed all the location variables (**Figure 3.13**).

3.3. Major takeaways

We found the following takeaways from our EDA of the numeric variables:

- 17% of total shots were made at location (0,0)
- 20% of total shots were three-point shots
- Kobe made 47.8% of two-point shots
- Kobe made 33.1% of three-point shots
- Kobe made more shots in the last 2 seconds of each quarter than average
- Kobe frequently takes rest for the first half of the second and fourth quarters
- Kobe shows a preference for shots at the distance of 17 feet and 25 feet
- Kobe's peak shot count in the 2005-06 season

We found the following takeaways from our EDA of the factor variables:

- Kobe’s favorite shot type is Jump Shot
- Kobe’s favorite shot zone is center.
- The peak shot count is in season 2008-09

- The three peak odds seasons: 2001-02 (Odds=0.889), 2008-09 (Odds=0.893), 2012-13 (Odds=0.880)
- The farther away from the hoop, the lower shot counts, except for zone range of '16-24 ft', which is even higher than range of '8-16 ft'

Our new data frame Kobe_clean has 20,697 observations and 29 total variables (**Figure 3.14**).

4. Odds Model

We opted to build our model in SAS (**Figure 4.1**). The resulting model is as follows (**Table 4.1**):

Shot_made_odds = 0.3637 - 0.0436(**shot_distance**)

For every foot increase in shot_distance, the odds of a shot made will decrease by a factor of .04, with 95% confidence limits between .036 and .045 (**Table 4.2**).

5. Probability Model

We opted to build our model in SAS (**Figure 5.1**). The resulting model is as follows (**Table 5.1**):

Shot_made_probability = 0.3637 - 0.0436(**shot_distance**)

For every foot increase in shot_distance, the probability of a shot made will decrease by 4.36%, with 95% confidence limits between 4.05% and 4.66% (**Table 5.2**).

6. Playoffs Model

We opted to build our model in SAS (**Figure 6.1**). The resulting model is as follows (**Table 6.1**):

Shot_made_probability = 0.3692 - 0.0358(**playoffs**) - 0.0436(**shot_distance**)

However, the playoffs variable is not significant to the logistic model with a p-value of 0.3751. We are 95% confident that there is very little difference in the odds ratio of shot_made to shot_distance in a regular season game (playoffs=0) and a playoff game (playoffs=1) (**Table 6.2**). Additionally, the plot of predicted shot made probabilities for regular season and playoffs supports this assertion (**Table 6.3**).

7. Logistic Regression Model

For all Logistic Regression models, we used the Kobe_clean data set (**Figure 7.1**). We used a 75/25 split for our train and test sets (**Figure 7.2**). We chose a pprob of .448 because that is Kobe's career shooting percentage.

7.1. Forward Selection, Main Effect, pprob=.448

We chose to build the model in SAS (**Figure 7.3**). The relevant classification values for this model are as follows:

- **AIC:** 19014.523
- **AUC:** 0.7197
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.2. Backward Selection, Main Effect, pprob=.448

We chose to build the model in SAS (**Figure 7.4**). The relevant classification values for this model are as follows:

- **AIC:** 19014.523
- **AUC:** 0.7197
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.3. Stepwise Selection, Main Effect, pprob=.448

We chose to build the model in SAS (**Figure 7.5**). The relevant classification values for this model are as follows:

- **AIC:** 19014.523
- **AUC:** 0.7197
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.4. Forward Selection, Interaction, pprob=.448

We chose to build the model in SAS (**Figure 7.6**). The relevant classification values for this model are as follows:

- **AIC:** 19014.523
- **AUC:** 0.7197
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.5. Backward Selection, Interaction, pprob=.448

We chose to build the model in SAS (**Figure 7.7**). The relevant classification values for this model are as follows:

- **AIC:** 19014.523
- **AUC:** 0.7197
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.6. Stepwise Selection, Interaction, pprob=.448

We chose to build the model in SAS (**Figure 7.8**). The relevant classification values for this model are as follows:

- **AIC:** 19014.523
- **AUC:** 0.7197
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.7. Forward Selection, Polynomial, pprob=.448

We chose to build the model in SAS (**Figure 7.9**). The relevant classification values for this model are as follows:

- **AIC:** 18978.476
- **AUC:** 0.7222
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.8. Backward Selection, Polynomial, pprob=.448

We chose to build the model in SAS (**Figure 7.10**). The relevant classification values for this model are as follows:

- **AIC:** 18978.476
- **AUC:** 0.7222
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

7.9. Stepwise Selection, Polynomial, pprob=.448

We chose to build the model in SAS (**Figure 7.11**). The relevant classification values for this model are as follows:

- **AIC:** 18978.476
- **AUC:** 0.7222
- **Mis-Classification Rate:** 32.0%
- **Sensitivity:** 87.6
- **Specificity:** 44.1
- **Log Loss Function:** 0.269

8. Discriminant Analysis Model

For all DA models we used the Kobe_clean data set (**Figure 7.1**). We used a 75/25 split for our train and test sets (**Figure 7.2**).

8.1. Main Effect, priors '1'=.448

We chose to build the model in SAS (**Figure 8.1**). The relevant classification values for this model are as follows:

- **Mis-Classification Rate:** 36.1%
- **Sensitivity:** 46.0
- **Specificity:** 78.3
- **Log Loss Function:** 0.370

8.2. Polynomial, priors '1'=.448

We chose to build the model in SAS (**Figure 8.2**). The relevant classification values for this model are as follows:

- **Mis-Classification Rate:** 41.1%
- **Sensitivity:** 64.4
- **Specificity:** 54.5
- **Log Loss Function:** 0.392

9. Model Evaluation and Selection

We opted to utilize a logistic regression model utilizing stepwise variable selection and polynomial variables (**Figure 7.11**) for our predictions. We chose this model for the following reasons (**Table 9.1**):

- Among the logistic regression models, this model is tied for the lowest AIC (18978.476)
- Among the logistic regression models, this model is tied for the highest AUC and ranks in the 'Fair' range according to professor's classification table in Session 14 (0.7222)
- This model is tied for the lowest mis-classification rate (32.0%)
- This model is tied for the highest sensitivity (87.6)
- Only the two LDA models have higher specificity than this model (78.3 and 54.5 respectively, compared to 44.1). However, both LDA models have higher mis-classification rates.
- This model is tied for the lowest log loss function (0.269)

Our predictions and the parameter estimates for this model are in the attached Excel sheets.

Prediction file: predlogisticOut.xlsx

Parameter estimates file: Model_9_Parameters_Estimates.xlsx

Appendix 1 (Tables and Charts)

Table 3.1

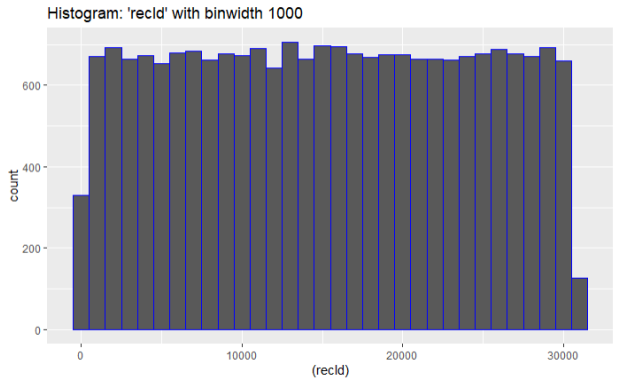

Table 3.2

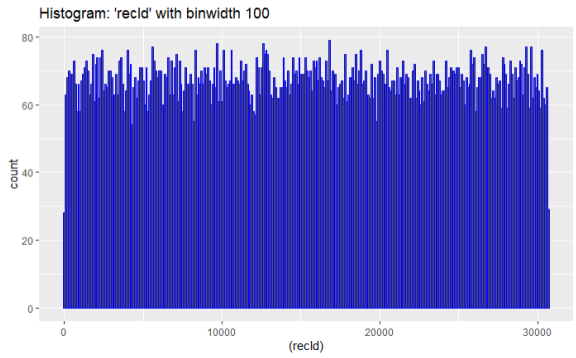

Table 3.3

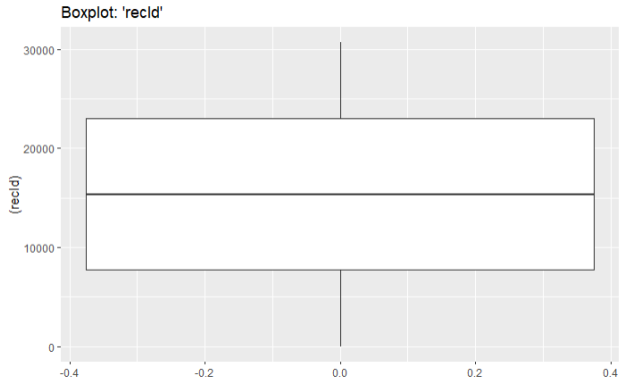

Table 3.4

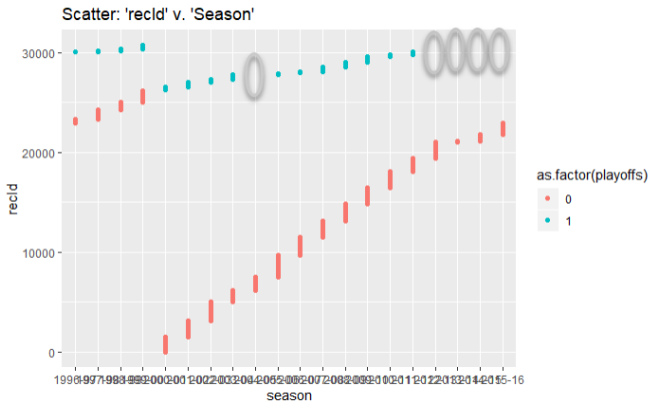

Table 3.5

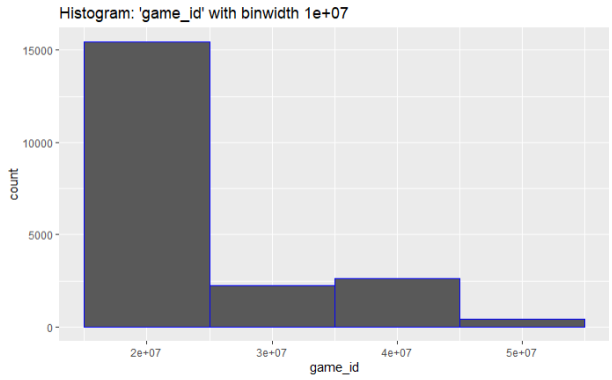

Table 3.6

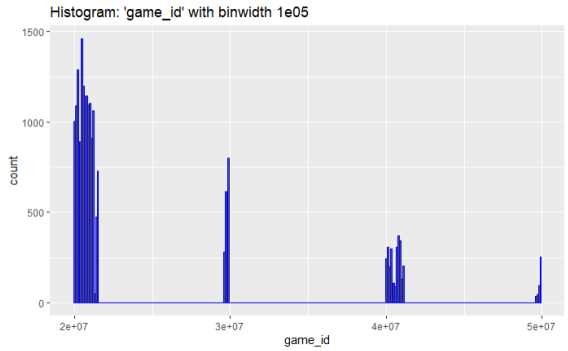

Table 3.7

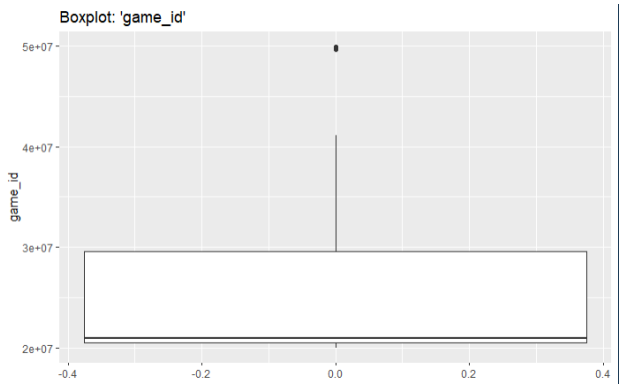

Table 3.8

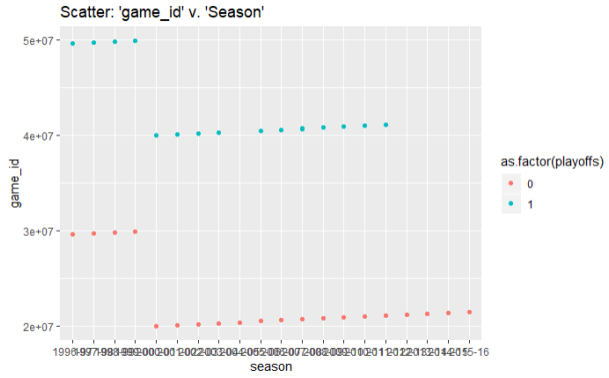

Table 3.9

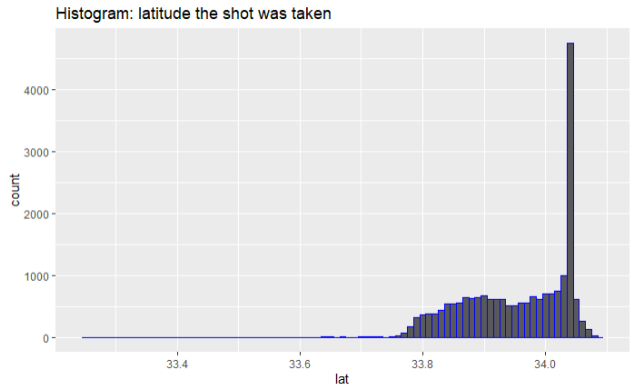

Table 3.10

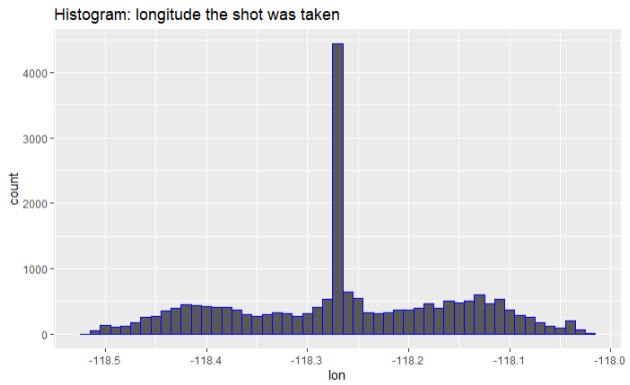

Table 3.11

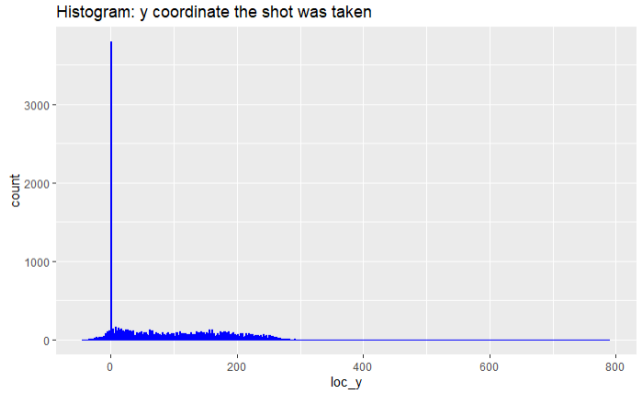

Table 3.12

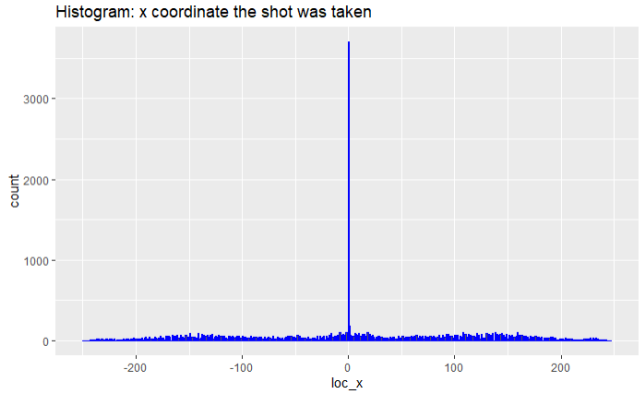

Table 3.13

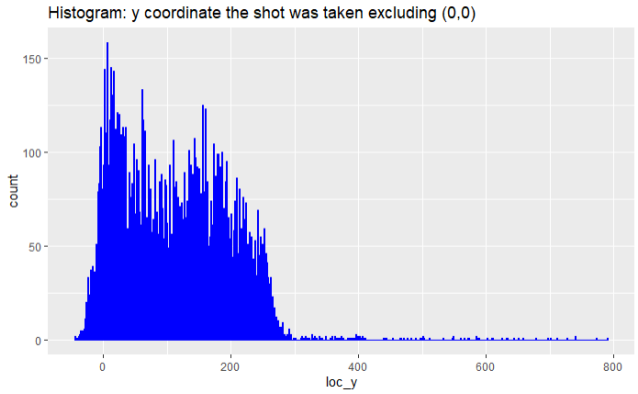

Table 3.14

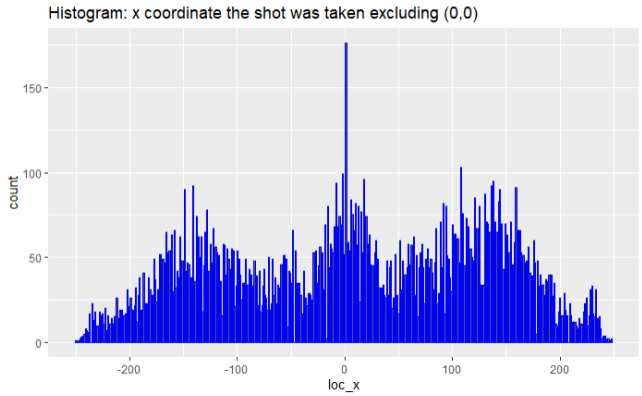

Table 3.15

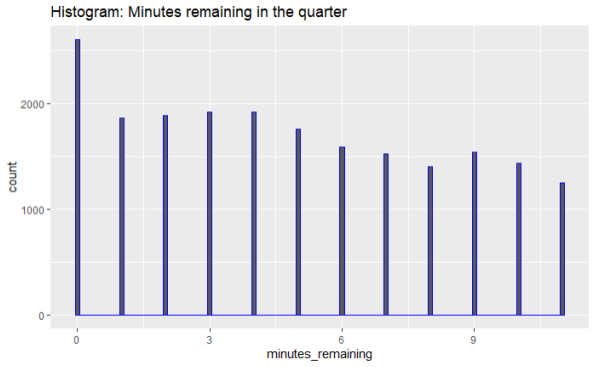

Table 3.16

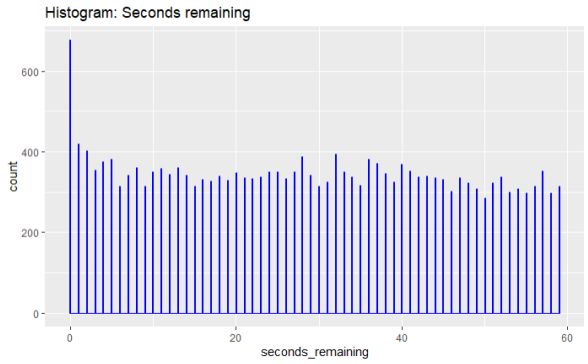

Table 3.17

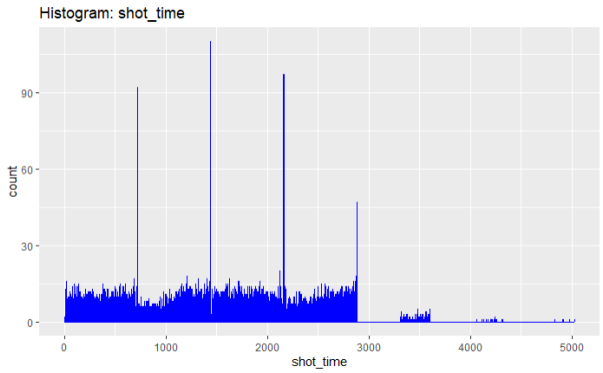

Table 3.18

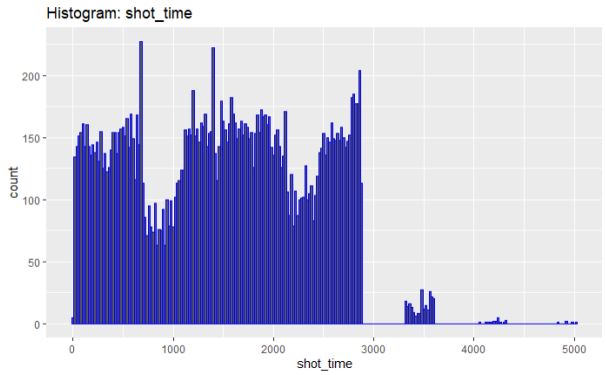

Table 3.19

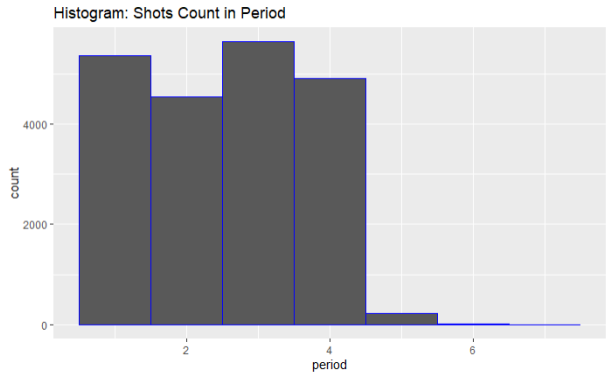

Table 3.20

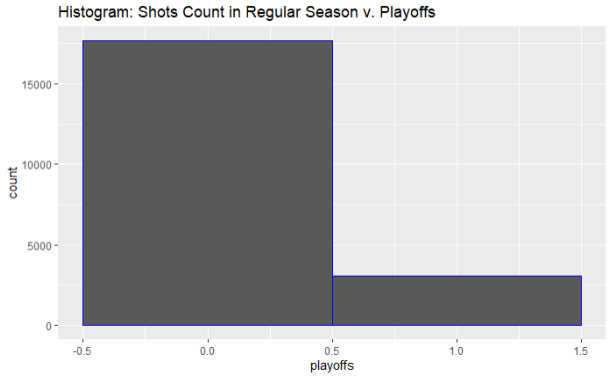

Table 3.21

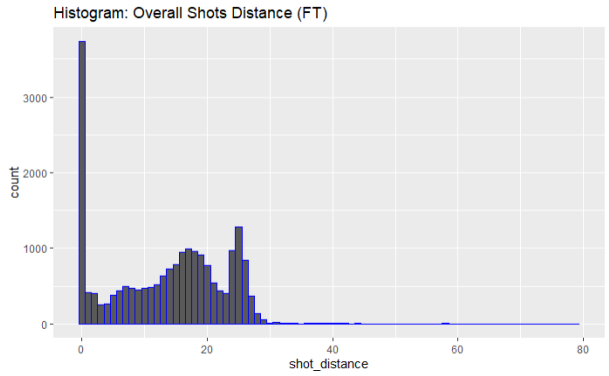

Table 3.22

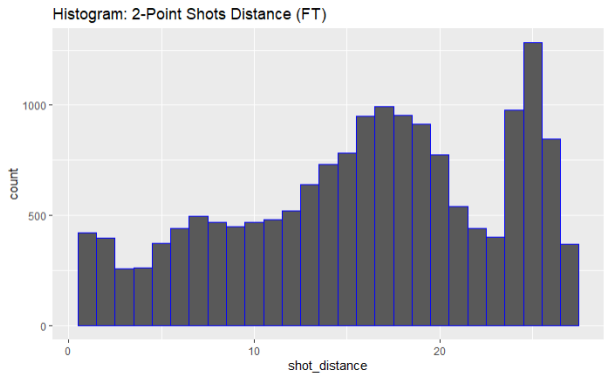

Table 3.23

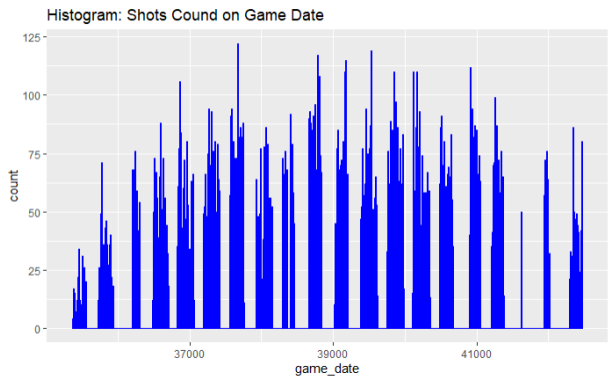

Table 3.24

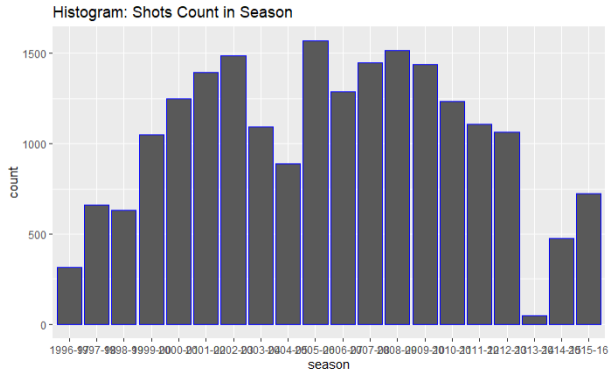

Table 3.25

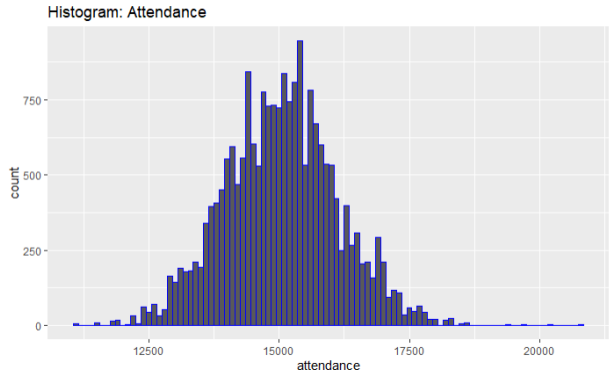

Table 3.26

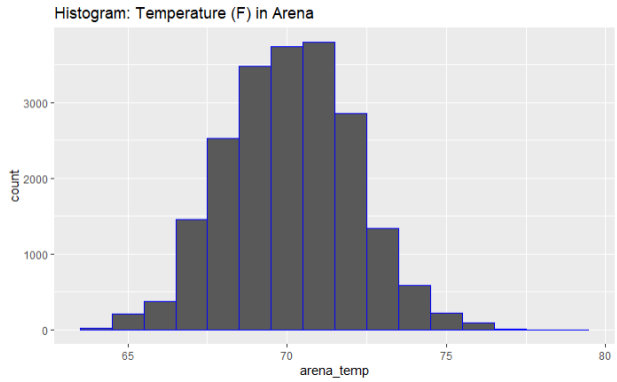

Table 3.27

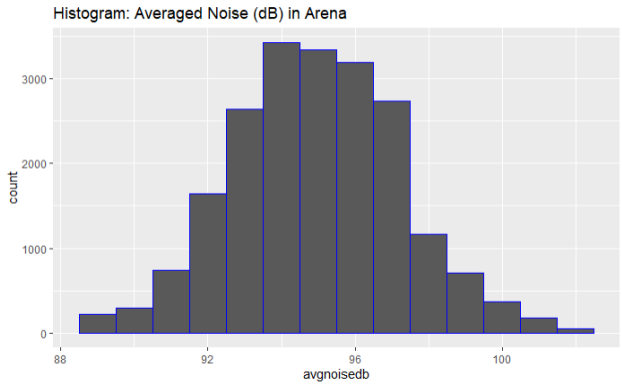

Table 3.28

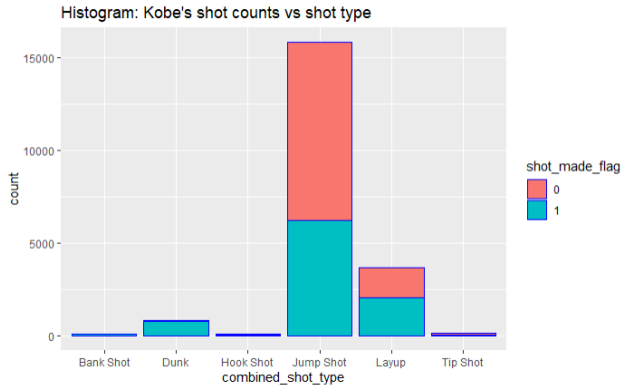

Table 3.29

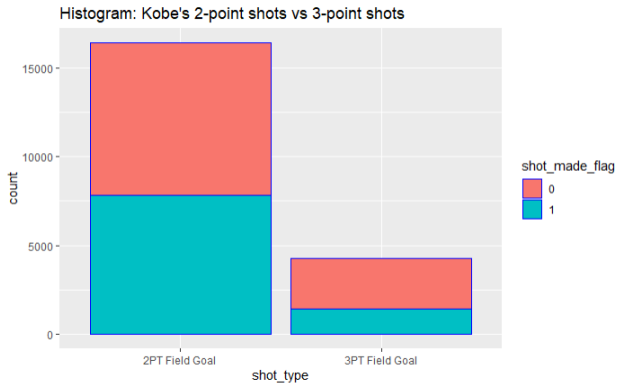

Table 3.30

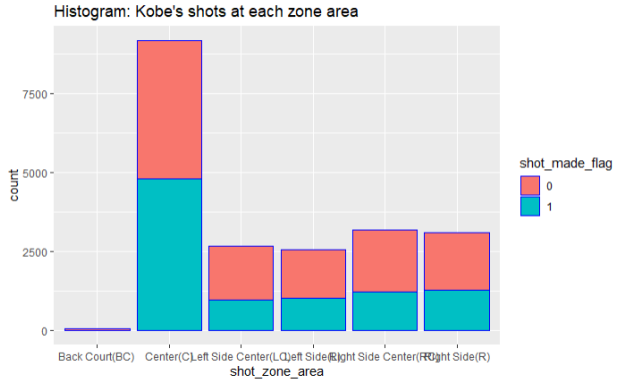

Table 3.31

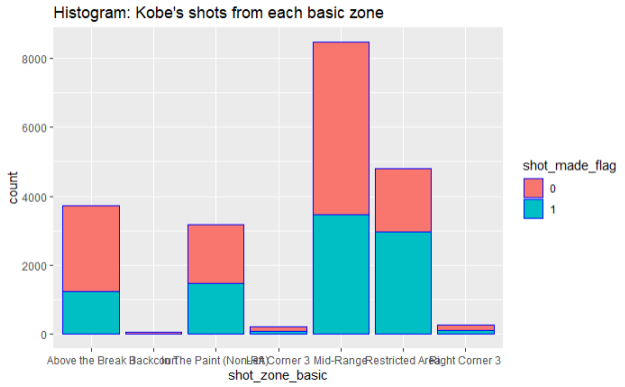

Table 3.32

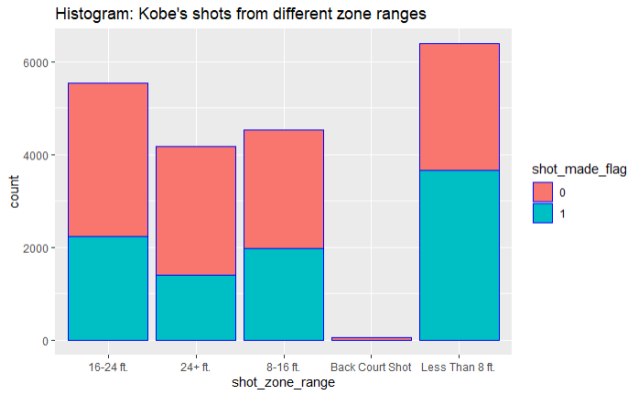

Table 3.33

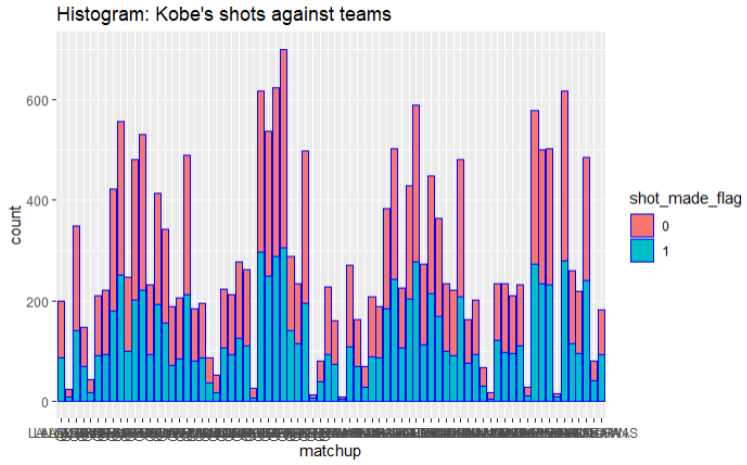

Table 3.34

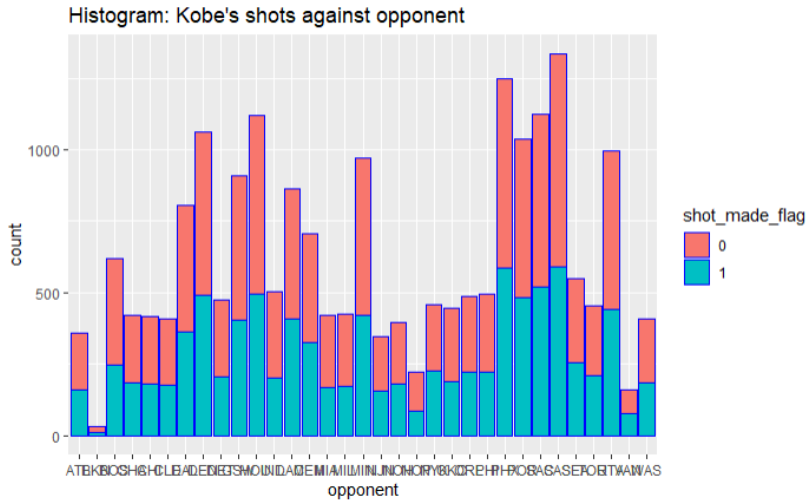

Table 3.35

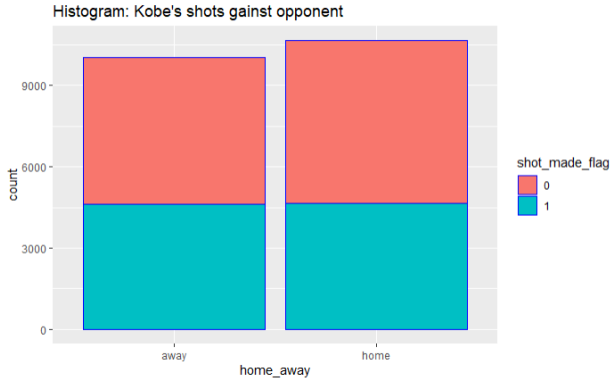

Table 3.36

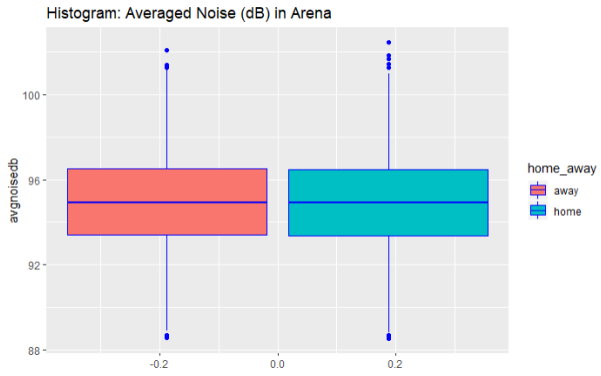

Table 3.37

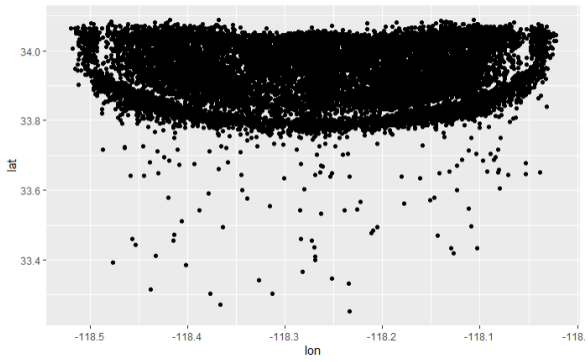

Table 3.38

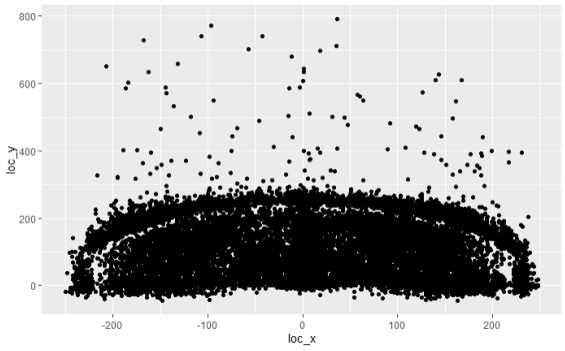

Table 3.39

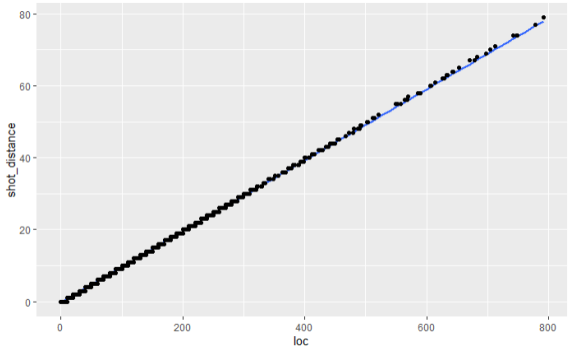

Table 3.40

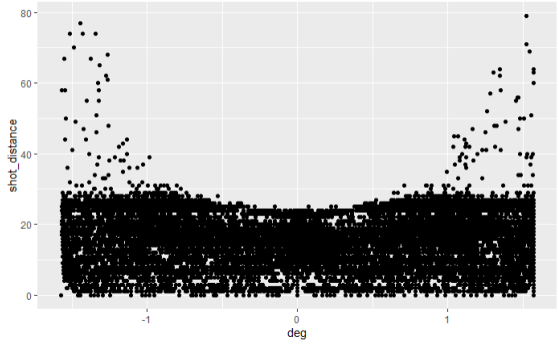

Table 4.1

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3637	0.0248	214.3848	<.0001
shot_distance	1	-0.0436	0.00157	772.7916	<.0001

Table 4.2

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
shot_distance	0.960	0.955	0.964

Table 5.1

Parameter Estimates and Wald Confidence Intervals			
Parameter	Estimate	95% Confidence Limits	
Intercept	0.3637	0.3150	0.4124
shot_distance	-0.0436	-0.0466	-0.0405

Table 5.2

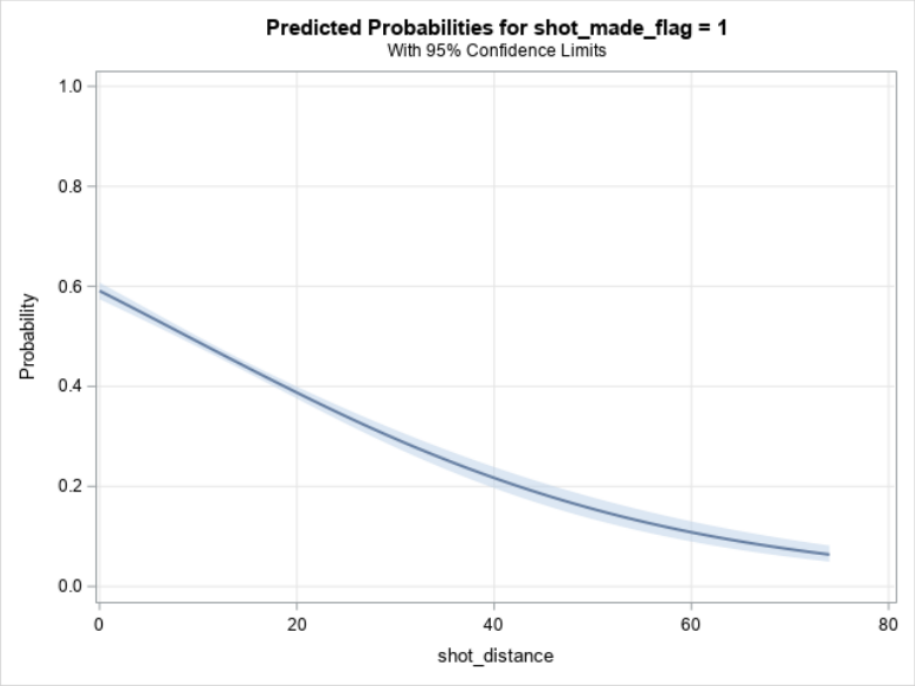

Table 6.1

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3692	0.0256	207.8560	<.0001
playoffs	1	1	-0.0358	0.0403	0.7866	0.3751
shot_distance		1	-0.0436	0.00157	773.2403	<.0001

Table 6.2

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
playoffs 1 vs 0	0.965	0.892	1.044
shot_distance	0.957	0.954	0.960

Table 6.3

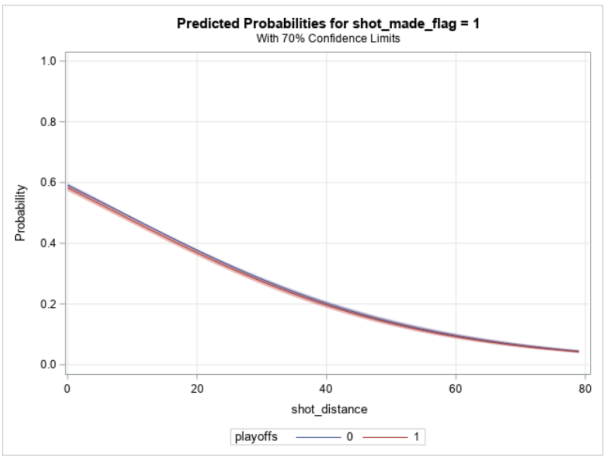

Table 9.1

Prediction Model Selection												
Model #	Model Type	Selection Method	Model Features	pprob	AIC	SC	-2 Log L	AUC	Mis-Classification Rate	Sensitivity	Specificity	Log Loss Function
1	Logistic Regression	Forward	Main Effect	0.448	19014.523	20232.14	18696.523	0.7197	32.0%	87.6	44.1	0.269
2	Logistic Regression	Backward	Main Effect	0.448	19014.523	20232.14	18696.523	0.7197	32.0%	87.6	44.1	0.269
3	Logistic Regression	Stepwise	Main Effect	0.448	19014.523	20232.14	18696.523	0.7197	32.0%	87.6	44.1	0.269
4	Logistic Regression	Forward	Interaction Term	0.448	19014.523	20232.14	18696.523	0.7197	32.0%	87.6	44.1	0.269
5	Logistic Regression	Backward	Interaction Term	0.448	19014.523	20232.14	18696.523	0.7197	32.0%	87.6	44.1	0.269
6	Logistic Regression	Stepwise	Interaction Term	0.448	19014.523	20232.14	18696.523	0.7197	32.0%	87.6	44.1	0.269
7	Logistic Regression	Forward	Polynomial	0.448	18978.476	20196.093	18660.476	0.7222	32.0%	87.6	44	0.269
8	Logistic Regression	Backward	Polynomial	0.448	18978.476	20196.093	18660.476	0.7222	32.0%	87.6	44	0.269
9	Logistic Regression	Stepwise	Polynomial	0.448	18978.476	20196.093	18660.476	0.7222	32.0%	87.6	44	0.269
10	Discriminant Analysis	QDA	Main Effect	priors '1' =.448					36.1%	46.0	78.3	0.370
11	Discriminant Analysis	QDA	Polynomial	priors '1' =.448					41.1%	64.4	54.5	0.392

Appendix 2 (Code)

Figure 2.1

```
## 'data.frame':    20697 obs. of  29 variables:
##  $ recId             : int  1 4 5 6 7 9 12 13 14 16 ...
##  $ action_type       : Factor w/ 54 levels "Alley Oop Dunk Shot",..: 26 5 26 27 26 26 26 41
##  $ combined_shot_type: Factor w/ 6 levels "Bank Shot","Dunk",..: 4 2 4 5 4 4 4 4 4 4 ...
##  $ game_event_id     : int  12 155 244 251 265 309 4 27 66 86 ...
##  $ game_id           : int  20000012 20000012 20000012 20000012 20000012 20000012 20000019
##  $ lat               : num  34 34 34.1 34 33.9 ...
##  $ loc_x             : int  -157 0 -145 0 -65 -94 121 -67 -94 62 ...
##  $ loc_y             : int  0 0 -11 0 108 238 127 110 4 192 ...
##  $ lon               : num  -118 -118 -118 -118 -118 ...
##  $ minutes_remaining : int  10 6 9 8 6 1 11 7 2 0 ...
##  $ period            : int  1 2 3 3 3 3 1 1 1 1 ...
##  $ playoffs          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ season            : Factor w/ 20 levels "1996-97","1997-98",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ seconds_remaining : int  22 19 32 52 12 56 0 9 44 48 ...
##  $ shot_distance     : int  15 0 14 0 12 25 17 12 9 20 ...
##  $ shot_made_flag    : int  0 1 0 1 1 0 1 1 0 0 ...
##  $ shot_type         : Factor w/ 2 levels "2PT Field Goal",..: 1 1 1 1 1 2 1 1 1 1 ...
##  $ shot_zone_area    : Factor w/ 6 levels "Back Court(BC)",..: 4 2 4 2 4 3 5 4 4 2 ...
##  $ shot_zone_basic   : Factor w/ 7 levels "Above the Break 3",..: 5 6 5 6 3 1 5 3 5 5 ...
##  $ shot_zone_range   : Factor w/ 5 levels "16-24 ft.","24+ ft.",..: 3 5 3 5 3 2 1 3 3 1 ...
##  $ team_id           : int  1610612747 1610612747 1610612747 1610612747 1610612747 16106127
##  $ team_name         : Factor w/ 1 level "Los Angeles Lakers": 1 1 1 1 1 1 1 1 1 1 ...
##  $ game_date         : int  36830 36830 36830 36830 36830 36830 36831 36831 36831 36831 ...
##  $ matchup           : Factor w/ 74 levels "LAL @ ATL","LAL @ BKN",..: 29 29 29 29 29 29 72
##  $ opponent          : Factor w/ 33 levels "ATL","BKN","BOS",..: 26 26 26 26 26 26 31 31 31
##  $ shot_id           : int  2 5 6 7 9 11 12 13 14 16 ...
##  $ attendance        : int  14707 14707 14707 14707 14707 14707 15851 15851 15851 15851 ...
##  $ arena_temp        : int  69 69 69 69 69 69 69 69 69 69 ...
##  $ avgnoisedb        : num  94.1 94.1 94.1 94.1 94.1 ...
```

Figure 2.2

```
summary(Kobe)
```

```
##      recId                     action_type    combined_shot_type
##  Min.   :    1   Jump Shot           :12712   Bank Shot:  101
##  1st Qu.: 7700   Layup Shot          : 1734   Dunk     :  844
##  Median :15337   Driving Layup Shot  : 1335   Hook Shot:  110
##  Mean   :15348   Turnaround Jump Shot:  739   Jump Shot:15834
##  3rd Qu.:23033   Fadeaway Jump Shot  :  693   Layup    : 3689
##  Max.   :30691   Running Jump Shot   :  620   Tip Shot :  119
##                  (Other)             : 2864
##  game_event_id      game_id              lat            loc_x
##  Min.   :  2.0   Min.   :20000012   Min.   :33.25   Min.   :-250.000
##  1st Qu.:112.0   1st Qu.:20500095   1st Qu.:33.88   1st Qu.: -67.000
##  Median :254.0   Median :20900337   Median :33.97   Median :   0.000
##  Mean   :249.7   Mean   :24758072   Mean   :33.95   Mean   :   6.674
##  3rd Qu.:368.0   3rd Qu.:29600488   3rd Qu.:34.04   3rd Qu.:  94.000
##  Max.   :653.0   Max.   :49900088   Max.   :34.09   Max.   : 248.000
##
##      loc_y             lon          minutes_remaining     period
##  Min.   :-44.00   Min.   :-118.5   Min.   : 0.000    Min.   :1.000
##  1st Qu.:  3.00   1st Qu.:-118.3   1st Qu.: 2.000    1st Qu.:1.000
##  Median : 72.00   Median :-118.3   Median : 5.000    Median :3.000
##  Mean   : 90.54   Mean   :-118.3   Mean   : 4.893    Mean   :2.525
##  3rd Qu.:160.00   3rd Qu.:-118.2   3rd Qu.: 8.000    3rd Qu.:3.000
##  Max.   :791.00   Max.   :-118.0   Max.   :11.000    Max.   :7.000
##     playoffs          season      seconds_remaining shot_distance
##  Min.   :0.0000   2005-06: 1569   Min.   : 0.00     Min.   : 0.00
##  1st Qu.:0.0000   2008-09: 1516   1st Qu.:13.00     1st Qu.: 5.00
##  Median :0.0000   2002-03: 1487   Median :28.00     Median :15.00
##  Mean   :0.1466   2007-08: 1451   Mean   :28.36     Mean   :13.37
##  3rd Qu.:0.0000   2009-10: 1440   3rd Qu.:43.00     3rd Qu.:21.00
##  Max.   :1.0000   2001-02: 1396   Max.   :59.00     Max.   :79.00
##                   (Other):11838
##  shot_made_flag           shot_type                   shot_zone_area
##  Min.   :0.0000   2PT Field Goal:16414   Back Court(BC)       :  60
##  1st Qu.:0.0000   3PT Field Goal: 4283   Center(C)            :9155
##  Median :0.0000                         Left Side Center(LC) :2668
##  Mean   :0.4477                         Left Side(L)         :2559
##  3rd Qu.:1.0000                         Right Side Center(RC):3162
##  Max.   :1.0000                         Right Side(R)        :3093
##
##              shot_zone_basic        shot_zone_range    team_id
##  Above the Break 3    :3729   16-24 ft.      :5541   Min.   :1.611e+09
##  Backcourt            :  49   24+ ft.        :4174   1st Qu.:1.611e+09
##  In The Paint (Non-RA):3182   8-16 ft.       :4538   Median :1.611e+09
##  Left Corner 3        : 194   Back Court Shot:  60   Mean   :1.611e+09
##  Mid-Range            :8473   Less Than 8 ft.:6384   3rd Qu.:1.611e+09
##  Restricted Area      :4808                          Max.   :1.611e+09
##  Right Corner 3       : 262
##              team_name       game_date            matchup
##  Los Angeles Lakers:20697   Min.   :35374   LAL @ SAS  :  700
##                             1st Qu.:37400   LAL @ SAC  :  624
##                             Median :39024   LAL vs. SAS:  619
##                             Mean   :38907   LAL @ PHX  :  618
##                             3rd Qu.:40246   LAL vs. HOU:  590
##                             Max.   :42473   LAL vs. PHX:  578
##                                             (Other)    :16968
```

```
##     opponent        shot_id        attendance      arena_temp
##  SAS     : 1334   Min.   :    2   Min.   :11065   Min.   :64.0
##  PHX     : 1250   1st Qu.: 7701   1st Qu.:14311   1st Qu.:69.0
##  SAC     : 1126   Median :15340   Median :15058   Median :70.0
##  HOU     : 1122   Mean   :15350   Mean   :15041   Mean   :70.1
##  DEN     : 1061   3rd Qu.:23034   3rd Qu.:15739   3rd Qu.:71.0
##  POR     : 1039   Max.   :30696   Max.   :20845   Max.   :79.0
##  (Other):13765
##    avgnoisedb
##  Min.   : 88.56
##  1st Qu.: 93.40
##  Median : 94.92
##  Mean   : 94.96
##  3rd Qu.: 96.51
##  Max.   :102.43
```

Figure 3.1

```
##  $ action_type       : Factor w/ 54 levels "Alley Oop Dunk Shot",..: 26 5 26 27 26 26 26
##  $ combined_shot_type: Factor w/ 6 levels "Bank Shot","Dunk",..: 4 2 4 5 4 4 4 4 4 4 ...
```

Figure 3.2

```
## 3rd new variable: cst_Num
Kobe_clean <- Kobe_clean %>% mutate(cst_Num = recode(combined_shot_type,
                                                     "Jump Shot" = 1,
                                                     "Layup" = 2,
                                                     "Dunk" = 3,
                                                     "Tip Shot" = 3,
                                                     "Hook Shot" = 3,
                                                     "Bank Shot" = 3))
```

Figure 3.3

```
## 4th new variable: seasonID
Kobe_clean <- Kobe_clean %>% mutate(seasonID = recode(season,
                                                      "1996-97" = 1,
                                                      "1997-98" = 2,
                                                      "1998-99" = 3,
                                                      "1999-00" = 4,
                                                      "2000-01" = 5,
                                                      "2001-02" = 6,
                                                      "2002-03" = 7,
                                                      "2003-04" = 8,
                                                      "2004-05" = 9,
                                                      "2005-06" = 10,
                                                      "2006-07" = 11,
                                                      "2007-08" = 12,
                                                      "2008-09" = 13,
                                                      "2009-10" = 14,
                                                      "2010-11" = 15,
                                                      "2011-12" = 16,
                                                      "2012-13" = 17,
                                                      "2013-14" = 18,
                                                      "2014-15" = 19,
                                                      "2015-16" = 20))
```

Figure 3.4

```
## 5th new variable: st_Num
Kobe_clean <- Kobe_clean %>% mutate(st_Num = recode(shot_type,
                                                    "2PT Field Goal" = 1,
                                                    "3PT Field Goal" = 2))
```

Figure 3.5

```
## 6th new variable: sza_Num
Kobe_clean <- Kobe_clean %>% mutate(sza_Num = recode(shot_zone_area,
                                                     "Center(C)" = 1,
                                                     "Right Side Center(RC)" = 2,
                                                     "Right Side(R)" = 2,
                                                     "Left Side Center(LC)" = 3,
                                                     "Left Side(L)" = 3,
                                                     "Back Court(BC)" = 4))
```

Figure 3.6

```
## 7th new variable: szb_Num
Kobe_clean <- Kobe_clean %>% mutate(szb_Num = recode(shot_zone_basic,
                                                     "Restricted Area" = 1,
                                                     "In The Paint (Non-RA)" = 1,
                                                     "Mid-Range" = 2,
                                                     "Left Corner 3" = 3,
                                                     "Right Corner 3" = 3,
                                                     "Above the Break 3" = 4,
                                                     "Backcourt" = 4))
```

Figure 3.7

```
## 8th new variable: szr_Num
Kobe_clean <- Kobe_clean %>% mutate(szr_Num = recode(shot_zone_range,
                                                     "Less Than 8 ft." = 1,
                                                     "8-16 ft." = 2,
                                                     "16-24 ft." = 3,
                                                     "24+ ft." = 4,
                                                     "Back Court Shot" = 4))
```

Figure 3.8

```
# Check clean data set Kobe_clean
str(Kobe_clean)
```

```
## 'data.frame':    20697 obs. of  32 variables:
##  $ recId             : int  1 4 5 6 7 9 12 13 14 16 ...
##  $ combined_shot_type: Factor w/ 6 levels "Bank Shot","Dunk",..: 4 2 4 5 4 4 4 4 4 4 ...
##  $ game_id           : int  20000012 20000012 20000012 20000012 20000012 20000012 20000019
##  $ lat               : num  34 34 34.1 34 33.9 ...
##  $ loc_x             : int  -157 0 -145 0 -65 -94 121 -67 -94 62 ...
##  $ loc_y             : int  0 0 -11 0 108 238 127 110 4 192 ...
##  $ lon               : num  -118 -118 -118 -118 -118 ...
##  $ minutes_remaining : int  10 6 9 8 6 1 11 7 2 0 ...
##  $ period            : int  1 2 3 3 3 3 1 1 1 1 ...
##  $ playoffs          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ season            : Factor w/ 20 levels "1996-97","1997-98",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ seconds_remaining : int  22 19 32 52 12 56 0 9 44 48 ...
##  $ shot_distance     : int  15 0 14 0 12 25 17 12 9 20 ...
##  $ shot_made_flag    : int  0 1 0 1 1 0 1 1 0 0 ...
##  $ shot_type         : Factor w/ 2 levels "2PT Field Goal",..: 1 1 1 1 1 2 1 1 1 1 ...
##  $ shot_zone_area    : Factor w/ 6 levels "Back Court(BC)",..: 4 2 4 2 4 3 5 4 4 2 ...
##  $ shot_zone_basic   : Factor w/ 7 levels "Above the Break 3",..: 5 6 5 6 3 1 5 3 5 5 ...
##  $ shot_zone_range   : Factor w/ 5 levels "16-24 ft.","24+ ft.",..: 3 5 3 5 3 2 1 3 3 1 ...
##  $ game_date         : int  36830 36830 36830 36830 36830 36830 36831 36831 36831 36831 ...
##  $ matchup           : Factor w/ 74 levels "LAL @ ATL","LAL @ BKN",..: 29 29 29 29 29 29 72
##  $ opponent          : Factor w/ 33 levels "ATL","BKN","BOS",..: 26 26 26 26 26 26 31 31 31
##  $ attendance        : int  14707 14707 14707 14707 14707 14707 15851 15851 15851 15851 ...
##  $ arena_temp        : int  69 69 69 69 69 69 69 69 69 69 ...
##  $ avgnoisedb        : num  94.1 94.1 94.1 94.1 94.1 ...
##  $ home_away         : Factor w/ 2 levels "away","home": 2 2 2 2 2 2 1 1 1 1 ...
##  $ shot_time         : num  98 1061 1588 1628 1788 ...
##  $ cst_Num           : num  1 3 1 2 1 1 1 1 1 1 ...
##  $ seasonID          : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ st_Num            : num  1 1 1 1 1 2 1 1 1 1 ...
##  $ sza_Num           : num  3 1 3 1 3 3 2 3 3 1 ...
##  $ szb_Num           : num  2 1 2 1 1 4 2 1 2 2 ...
##  $ szr_Num           : num  2 1 2 1 2 4 3 2 2 3 ...
```

Figure 3.9

```
summary(Kobe_clean)
```

```
##      recId       combined_shot_type    game_id             lat
##  Min.   :    1   Bank Shot:  101    Min.   :20000012   Min.   :33.25
##  1st Qu.: 7700   Dunk     :  844    1st Qu.:20500095   1st Qu.:33.88
##  Median :15337   Hook Shot:  110    Median :20900337   Median :33.97
##  Mean   :15348   Jump Shot:15834    Mean   :24758072   Mean   :33.95
##  3rd Qu.:23033   Layup    : 3689    3rd Qu.:29600488   3rd Qu.:34.04
##  Max.   :30691   Tip Shot :  119    Max.   :49900088   Max.   :34.09
##      loc_x             loc_y             lon         minutes_remaining
##  Min.   :-250.000   Min.   :-44.00   Min.   :-118.5   Min.   : 0.000
##  1st Qu.: -67.000   1st Qu.:  3.00   1st Qu.:-118.3   1st Qu.: 2.000
##  Median :   0.000   Median : 72.00   Median :-118.3   Median : 5.000
##  Mean   :   6.674   Mean   : 90.54   Mean   :-118.3   Mean   : 4.893
##  3rd Qu.:  94.000   3rd Qu.:160.00   3rd Qu.:-118.2   3rd Qu.: 8.000
##  Max.   : 248.000   Max.   :791.00   Max.   :-118.0   Max.   :11.000
```

```
##      period          playoffs          season      seconds_remaining
##  Min.   :1.000   Min.   :0.0000   2005-06: 1569   Min.   : 0.00
##  1st Qu.:1.000   1st Qu.:0.0000   2008-09: 1516   1st Qu.:13.00
##  Median :3.000   Median :0.0000   2002-03: 1487   Median :28.00
##  Mean   :2.525   Mean   :0.1466   2007-08: 1451   Mean   :28.36
##  3rd Qu.:3.000   3rd Qu.:0.0000   2009-10: 1440   3rd Qu.:43.00
##  Max.   :7.000   Max.   :1.0000   2001-02: 1396   Max.   :59.00
##                                   (Other):11838
##  shot_distance   shot_made_flag           shot_type
##  Min.   : 0.00   Min.   :0.0000   2PT Field Goal:16414
##  1st Qu.: 5.00   1st Qu.:0.0000   3PT Field Goal: 4283
##  Median :15.00   Median :0.0000
##  Mean   :13.37   Mean   :0.4477
##  3rd Qu.:21.00   3rd Qu.:1.0000
##  Max.   :79.00   Max.   :1.0000
##                shot_zone_area              shot_zone_basic
##  Back Court(BC)       :  60   Above the Break 3    :3729
##  Center(C)            :9155   Backcourt            :  49
##  Left Side Center(LC) :2668   In The Paint (Non-RA):3182
##  Left Side(L)         :2559   Left Corner 3        : 194
##  Right Side Center(RC):3162   Mid-Range            :8473
##  Right Side(R)        :3093   Restricted Area      :4808
##                               Right Corner 3       : 262
##        shot_zone_range   game_date            matchup
##  16-24 ft.      :5541   Min.   :35374   LAL @ SAS  :  700
##  24+ ft.        :4174   1st Qu.:37400   LAL @ SAC  :  624
##  8-16 ft.       :4538   Median :39024   LAL vs. SAS:  619
##  Back Court Shot:  60   Mean   :38907   LAL @ PHX  :  618
##  Less Than 8 ft.:6384   3rd Qu.:40246   LAL vs. HOU:  590
##                         Max.   :42473   LAL vs. PHX:  578
##                                         (Other)    :16968
##    opponent       attendance      arena_temp     avgnoisedb
##  SAS    : 1334   Min.   :11065   Min.   :64.0   Min.   : 88.56
##  PHX    : 1250   1st Qu.:14311   1st Qu.:69.0   1st Qu.: 93.40
##  SAC    : 1126   Median :15058   Median :70.0   Median : 94.92
##  HOU    : 1122   Mean   :15041   Mean   :70.1   Mean   : 94.96
##  DEN    : 1061   3rd Qu.:15739   3rd Qu.:71.0   3rd Qu.: 96.51
##  POR    : 1039   Max.   :20845   Max.   :79.0   Max.   :102.43
##  (Other):13765
##  home_away      shot_time       cst_Num         seasonID
##  away:10047   Min.   :   6   Min.   :1.000   Min.   : 1.00
##  home:10650   1st Qu.: 714   1st Qu.:1.000   1st Qu.: 6.00
##               Median :1505   Median :1.000   Median :11.00
##               Mean   :1496   Mean   :1.292   Mean   :10.41
##               3rd Qu.:2160   3rd Qu.:1.000   3rd Qu.:14.00
##               Max.   :5026   Max.   :3.000   Max.   :20.00
##      st_Num          sza_Num          szb_Num          szr_Num
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.207   Mean   :1.816   Mean   :2.001   Mean   :2.368
##  3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :2.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
```

Figure 3.10

```
471 Kobe_clean$loc <- (Kobe_clean$loc_x^2+Kobe_clean$loc_y^2)^0.5
472 Kobe_clean %>% ggplot(aes(x=loc, y=shot_distance)) + geom_smooth(method = lm) + geom_point()
473 cor.test(Kobe_clean$loc, Kobe_clean$shot_distance)
```

Figure 3.11

```
        Pearson's product-moment correlation

data:  Kobe_clean$loc and Kobe_clean$shot_distance
t = 4546, df = 20695, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9994859 0.9995131
sample estimates:
      cor
0.9994997
```

Figure 3.12

```
464 ## We can use trigonometric to create a shot degree variable "deg".
465 Kobe_clean$deg <- atan(Kobe_clean_loc$loc_y/(Kobe_clean_loc$loc_x+.000001))
466 Kobe_clean %>% ggplot(aes(x=deg, y=shot_distance)) + geom_point()
```

Figure 3.13

```
476 ## Now we can remove all the four location variables
477 Kobe_clean <- Kobe_clean %>% select(-c(lat, lon, loc_x, loc_y, loc))
```

Figure 3.14

```
> str(Kobe_clean)
'data.frame':   20697 obs. of  29 variables:
```

Figure 4.1

```
proc logistic data= full descending;
  model shot_made_flag (event='1')= shot_distance / ctable pprob=.5;
  output out=results p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit;
  effectplot / clm alpha=.05 noobs;
run;
```

Figure 5.1

```
 8 /* Proportion test shot_made_flag against shot_distance */
 9 title 'Proportion test shot_made_flag against shot_distance';
10 proc logistic data= full plots=all order=data descending;
11   model shot_made_flag (event='1')= shot_distance / ctable pprob=.5 clparm=wald;
12   output out=results p=prob l=lower u=upper resdev=resdev reschi=pearres
13          xbeta=logit stdxbeta=selogit;
14   effectplot / clm alpha=.05 noobs;
15 run;
```

Figure 6.1

```
13 /* shot_made_flag against playoffs and shot_distance */
14 proc logistic data= full descending;
15   class playoffs (ref="0") / param=ref;
16   model shot_made_flag (event='1')= playoffs shot_distance / ctable pprob=.5;
17   output out=results p=prob l=lower u=upper resdev=resdev reschi=pearres
18          xbeta=logit stdxbeta=selogit;
19   effectplot /*slicefit (sliceby=playoffs)*/ / at (playoffs=all) clm alpha=.3 noobs;
20   effectplot interaction (x=playoffs) / at(shot_distance=7 14 21 28) clm noobs;
21 run;
```

Figure 7.1

```
1 libname xl xlsx 'C:\Edu\GitHub_Stats2\Project2\Kobe_clean.xlsx'; run;
2 data full;
3     set xl.Kobe_clean;
4     shot_dis2 = shot_distance**2;
5     shot_time2 = shot_time**2;
6 run;
```

Figure 7.2

```
/* Prediction with Logistic Regression Model */
data dataIn; set full; randNumber = ranuni(11); run;
data train;  set dataIn; if randNumber <= 1/4 then delete; run;
data test;   set dataIn; if randNumber >  1/4 then delete; run;
proc print data=test;run;
```

Figure 7.3

```
/* Model 1: Logistic Regression, Forward Selection, Main Effect, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season shot_distance shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time
                        / selection=forward ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_1;
run;
proc export
  data=logisticClassified_Model_1
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_1.xlsx"
  replace;
run;
```

Figure 7.4

```
/* Model 2: Logistic Regression, Backward Selection, Main Effect, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season shot_distance shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time
                        / selection=backward ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_2;
run;
proc export
  data=logisticClassified_Model_2
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_2.xlsx"
  replace;
run;
```

Figure 7.5

```
/* Model 3: Logistic Regression, Stepwise Selection, Main Effect, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season shot_distance shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time
                        / selection=stepwise ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_3;
run;
proc export
  data=logisticClassified_Model_3
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_3.xlsx"
  replace;
run;
```

Figure 7.6

```
/* Model 4: Logistic Regression, Forward Selection, Interaction Term, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season | shot_distance shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time
                        / selection=forward ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_4;
run;
proc export
  data=logisticClassified_Model_4
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_4.xlsx"
  replace;
run;
```

Figure 7.7

```
/* Model 5: Logistic Regression, Backward Selection, Interaction Term, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season | shot_distance shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time
                        / selection=backwardward ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_5;
run;
proc export
  data=logisticClassified_Model_5
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_5.xlsx"
  replace;
run;
```

Figure 7.8

```
/* Model 6: Logistic Regression, Stepwise Selection, Interaction Term, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season | shot_distance shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time
                        / selection=stepwise ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_6;
run;
proc export
  data=logisticClassified_Model_6
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_6.xlsx"
  replace;
run;
```

Figure 7.9

```
/* Model 7: Logistic Regression, Forward Selection, Polynomial, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season shot_distance shot_dis2 shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time shot_time2
                        / selection=forward ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_7;
run;
proc export
  data=logisticClassified_Model_7
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_7.xlsx"
  replace;
run;
```

Figure 7.10

```
/* Model 8: Logistic Regression, Backward Selection, Polynomial, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season shot_distance shot_dis2 shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time shot_time2
                        / selection=Backward ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_8;
run;
proc export
  data=logisticClassified_Model_8
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_8.xlsx"
  replace;
run;
```

Figure 7.11

```
/* Model 9: Logistic Regression, Stepwise Selection, Polynomial, pprob=.448 */
proc logistic data= train order=data plots=all descending;
  class action_type combined_shot_type period playoffs season shot_type
                        shot_zone_area shot_zone_basic shot_zone_range
                        matchup opponent home_away;
  model shot_made_flag (event='1') = recId action_type combined_shot_type
                        period playoffs season shot_distance shot_dis2 shot_type shot_zone_area
                        shot_zone_basic shot_zone_range game_date matchup opponent
                        attendance arena_temp avgnoisedb home_away shot_time shot_time2
                        / selection=Stepwise ctable lackfit clparm=wald pprob=.4476977
                         link=glogit;
  output out=logisticOut p=prob l=lower u=upper resdev=resdev reschi=pearres
         xbeta=logit stdxbeta=selogit ;
  score data=test out=logisticClassified_Model_9;
run;
proc export
  data=logisticClassified_Model_9
  dbms=xlsx
  outfile="C:\Edu\GitHub_Stats2\Project2\logisticClassified_Model_9.xlsx"
  replace;
run;
```

Figure 8.1

```
/* Model 10: Discriminant Analysis, pool=test, Main Effect, priors '1' =.448 */
/* pool=test for SAS selects using LDA or QDA */
proc discrim data=train pool=test out=discrimOut_Model_10 crossvalidate testdata=test list;
class shot_made_flag;
var recId shot_distance deg attendance arena_temp avgnoisedb shot_time
    cst_Num seasonID st_Num sza_Num szb_Num szr_Num oppo_Num action_type_Num;
priors '0'=.552 '1'=.448;
run;
```

Figure 8.2

```
/* Model 11: Discriminant Analysis, pool=test, Polynomial, priors '1' =.448 */
/* pool=test for SAS selects using LDA or QDA */
proc discrim data=train pool=test out=discrimOut_Model_11 crossvalidate testdata=test list;
class shot_made_flag;
var recId shot_distance shot_dis2 deg attendance arena_temp avgnoisedb shot_time
    shot_time2 cst_Num seasonID st_Num sza_Num szb_Num szr_Num oppo_Num action_type_Num;
priors '0'=.552 '1'=.448;
run;
```

SAS and R code attachment

project2_Kobe.sas

project2Kobe.Rmd