DATA 621 Assignment 1

Kai Lukowiak

2018-02-08

1. Data Exploration:

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas. a. Mean / Standard Deviation / Median b. Bar Chart or Box Plot of the data c. Is the data correlated to the target variable (or to other variables?) d. Are any of the variables missing and need to be imputed "fixed"?

Loading the data:

```
library(tidyverse)
library(ggthemes)
library(ggcorrplot)
```

```
dfT <- read_csv('moneyball-training-data.csv')
dfT <- dfT %>% select(-INDEX)
glimpse(dfT)
```

```
## Observations: 2,276
## Variables: 16
## $ TARGET_WINS      <int> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 6...
## $ TEAM_BATTING_H   <int> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 127...
## $ TEAM_BATTING_2B  <int> 194, 219, 232, 209, 186, 200, 179, 171, 197, ...
## $ TEAM_BATTING_3B  <int> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 3...
## $ TEAM_BATTING_HR  <int> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96,...
## $ TEAM_BATTING_BB  <int> 143, 685, 602, 451, 472, 443, 525, 456, 447, ...
## $ TEAM_BATTING_SO  <int> 842, 1075, 917, 922, 920, 973, 1062, 1027, 92...
## $ TEAM_BASERUN_SB  <int> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, ...
## $ TEAM_BASERUN_CS  <int> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 7...
## $ TEAM_BATTING_HBP <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TEAM_PITCHING_H  <int> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 128...
## $ TEAM_PITCHING_HR <int> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96,...
## $ TEAM_PITCHING_BB <int> 927, 689, 602, 454, 472, 443, 525, 459, 447, ...
## $ TEAM_PITCHING_SO <int> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 9...
## $ TEAM_FIELDING_E  <int> 1011, 193, 175, 164, 138, 123, 136, 112, 127,...
## $ TEAM_FIELDING_DP <int> NA, 155, 153, 156, 168, 149, 186, 136, 169, 1...
```

```
dfE <- read_csv('moneyball-evaluation-data.csv')
dfE <- dfE %>% select(-INDEX)
glimpse(dfE)
```

```
## Observations: 259
## Variables: 15
## $ TEAM_BATTING_H   <int> 1209, 1221, 1395, 1539, 1445, 1431, 1430, 138...
## $ TEAM_BATTING_2B  <int> 170, 151, 183, 309, 203, 236, 219, 158, 177, ...
```

```
## $ TEAM_BATTING_3B  <int> 33, 29, 29, 29, 68, 53, 55, 42, 78, 42, 40, 5...
## $ TEAM_BATTING_HR  <int> 83, 88, 93, 159, 5, 10, 37, 33, 23, 58, 50, 1...
## $ TEAM_BATTING_BB  <int> 447, 516, 509, 486, 95, 215, 568, 356, 466, 4...
## $ TEAM_BATTING_SO  <int> 1080, 929, 816, 914, 416, 377, 527, 609, 689,...
## $ TEAM_BASERUN_SB  <int> 62, 54, 59, 148, NA, NA, 365, 185, 150, 52, 6...
## $ TEAM_BASERUN_CS  <int> 50, 39, 47, 57, NA, NA, NA, NA, NA, NA, NA, 2...
## $ TEAM_BATTING_HBP <int> NA, NA, NA, 42, NA, NA, NA, NA, NA, NA, NA, N...
## $ TEAM_PITCHING_H  <int> 1209, 1221, 1395, 1539, 3902, 2793, 1544, 162...
## $ TEAM_PITCHING_HR <int> 83, 88, 93, 159, 14, 20, 40, 39, 25, 62, 53, ...
## $ TEAM_PITCHING_BB <int> 447, 516, 509, 486, 257, 420, 613, 418, 497, ...
## $ TEAM_PITCHING_SO <int> 1080, 929, 816, 914, 1123, 736, 569, 715, 734...
## $ TEAM_FIELDING_E  <int> 140, 135, 156, 124, 616, 572, 490, 328, 226, ...
## $ TEAM_FIELDING_DP <int> 156, 164, 153, 154, 130, 105, NA, 104, 132, 1...
```

Initial Vizualizaton:

Special thanks to this Stack Overflow question

```
ggplot(data = gather(dfT), mapping = aes(x = value)) +
  geom_histogram(bins = 20) + facet_wrap(~key, scales = 'free') +
  theme_tufte()
```

```
## Warning: Removed 3478 rows containing non-finite values (stat_bin).
```
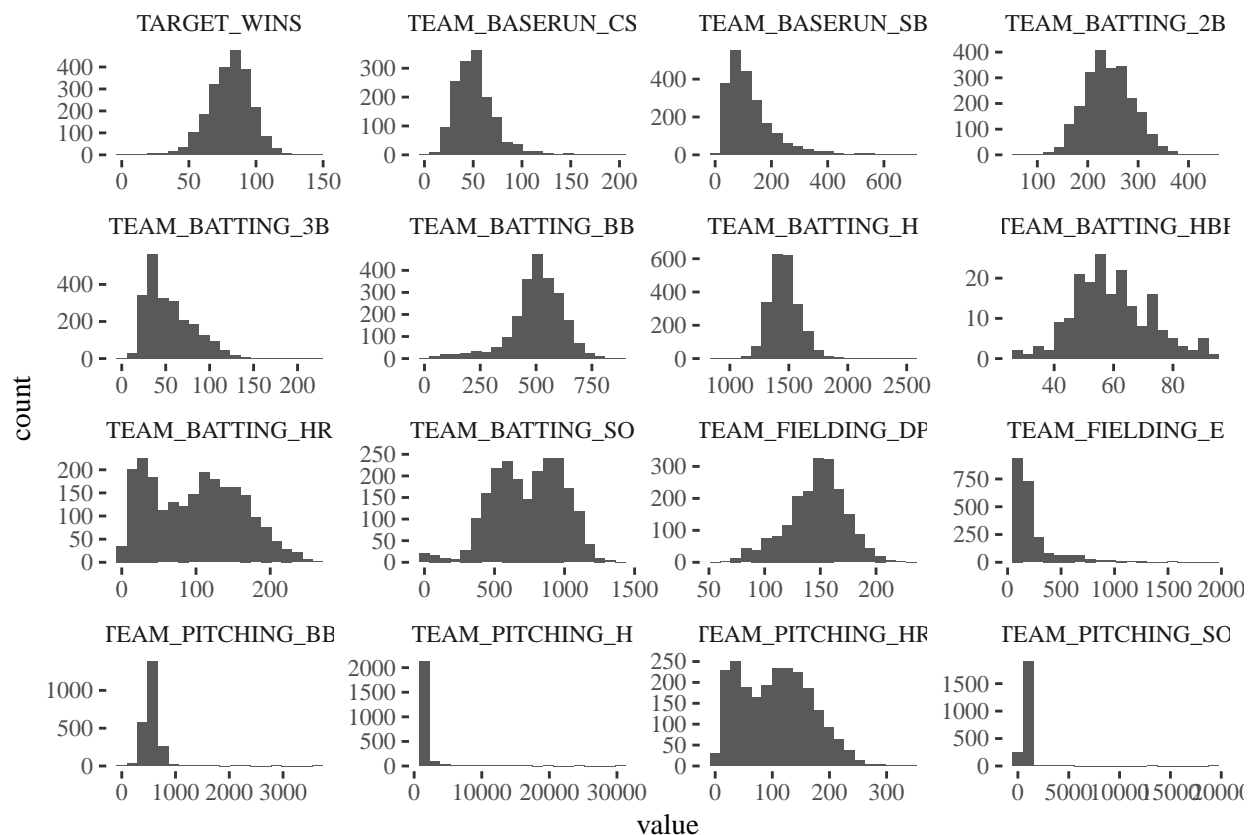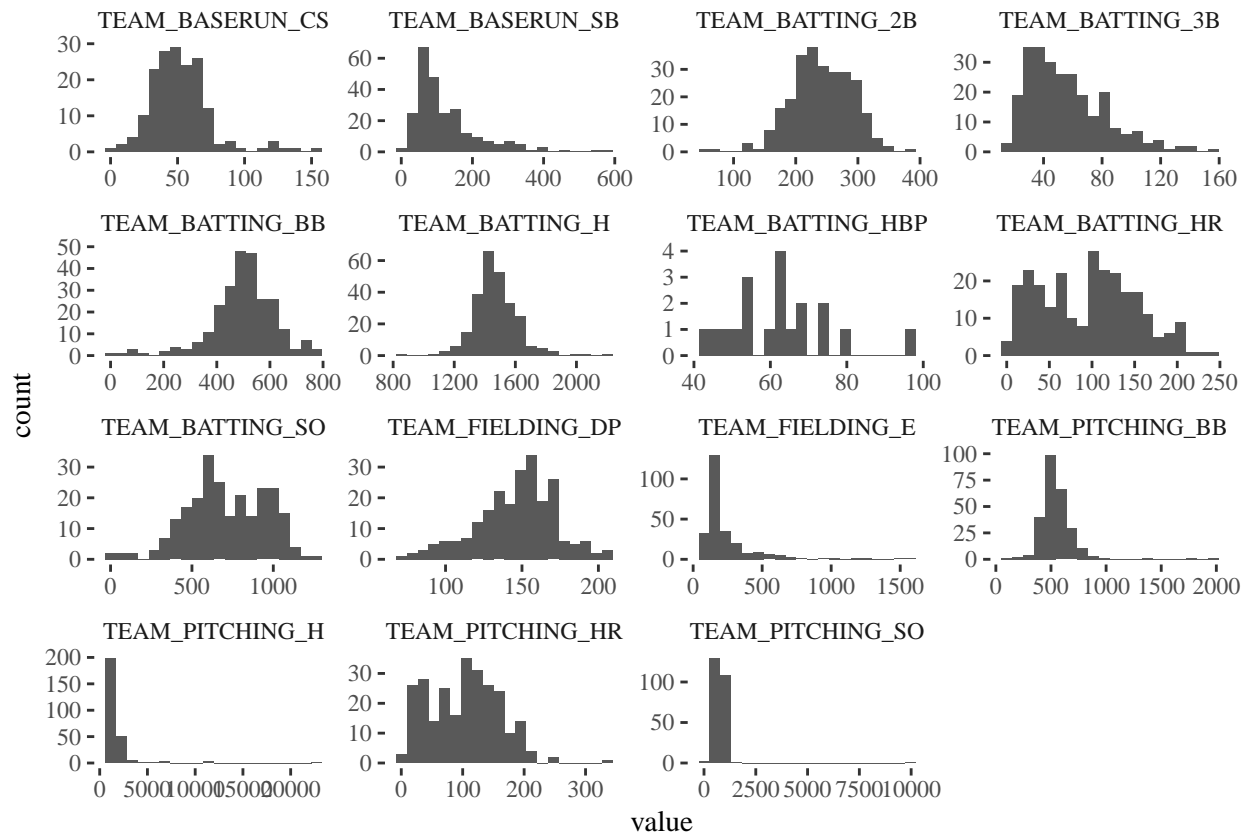


```
ggplot(data = gather(dfE), mapping = aes(x = value)) +
  geom_histogram(bins = 20) + facet_wrap(~key, scales = 'free') +
  theme_tufte()
```

```
## Warning: Removed 407 rows containing non-finite values (stat_bin).
```


From these plots we can see that many variables are aproximetly normally distrubuted. Notable exceptions are `TEAM_BATTING_3B`, `TEAM_BATTING_HR`, `TEAM_PITCHING_H`.

To explore these further: (Also, special thanks to this site)

```r
corr <- round(cor(dfT, use = "complete.obs"), 1) # Complete obs b/c of all NAs
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.1
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```
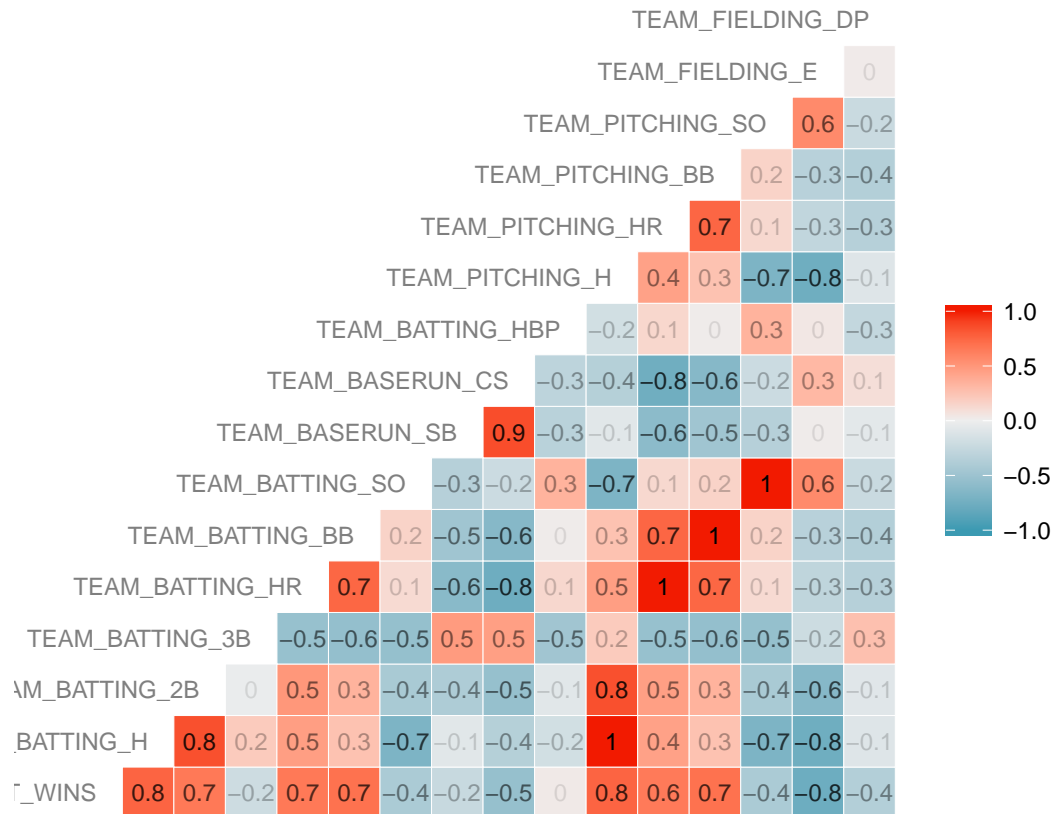
```r
#nba = read.csv("http://datasets.flowingdata.com/ppg2008.csv")
ggcorr(corr, hjust = 1, size = 3, color = "grey50", layout.exp = 1, label = TRUE, label_size = 3, label
```

Correlation matrix (lower-triangular heatmap):

Variable															
TEAM_FIELDING_DP															
TEAM_FIELDING_E	0														
TEAM_PITCHING_SO	0.6	−0.2													
TEAM_PITCHING_BB	0.2	−0.3	−0.4												
TEAM_PITCHING_HR	0.7	0.1	−0.3	−0.3											
TEAM_PITCHING_H	0.4	0.3	−0.7	−0.8	−0.1										
TEAM_BATTING_HBP	−0.2	0.1	0	0.3	0	−0.3									
TEAM_BASERUN_CS	−0.3	−0.4	−0.8	−0.6	−0.2	0.3	0.1								
TEAM_BASERUN_SB	0.9	−0.3	−0.1	−0.6	−0.5	−0.3	0	−0.1							
TEAM_BATTING_SO	−0.3	−0.2	0.3	−0.7	0.1	0.2	1	0.6	−0.2						
TEAM_BATTING_BB	0.2	−0.5	−0.6	0	0.3	0.7	1	0.2	−0.3	−0.4					
TEAM_BATTING_HR	0.7	0.1	−0.6	−0.8	0.1	0.5	1	0.7	0.1	−0.3	−0.3				
TEAM_BATTING_3B	−0.5	−0.6	−0.5	0.5	0.5	−0.5	0.2	−0.5	−0.6	−0.5	−0.2	0.3			
M_BATTING_2B	0	0.5	0.3	−0.4	−0.4	−0.5	−0.1	0.8	0.5	0.3	−0.4	−0.6	−0.1		
BATTING_H	0.8	0.2	0.5	0.3	−0.7	−0.1	−0.4	−0.2	1	0.4	0.3	−0.7	−0.8	−0.1	
T_WINS	0.8	0.7	−0.2	0.7	0.7	−0.4	−0.2	−0.5	0	0.8	0.6	0.7	−0.4	−0.8	−0.4

Legend: 1.0 / 0.5 / 0.0 / −0.5 / −1.0

2. DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set