CUNY SPS DATA 621 - CTG5 - HW1

Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh, Betsy Rosalen

February 27, 2019

1. DATA EXPLORATION

Let's leave instructions in the report for now so that we can easily reference them to make sure we are including everything we need to, cool?

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 and/or Histograms
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"?

Table 1: Summary

	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B
	Min. : 0.0	Min. : 891	Min. : 69	Min. : 0.0
	1st Qu.: 71.0	1st Qu.:1383	1st Qu.:208	1st Qu.: 34.0
	Median : 82.0	Median :1454	Median :238	Median : 47.0
	Mean : 80.8	Mean :1469	Mean :241	Mean : 55.2
	3rd Qu.: 92.0	3rd Qu.:1537	3rd Qu.:273	3rd Qu.: 72.0
	Max. :146.0	Max. :2554	Max. :458	Max. :223.0
	NA	NA	NA	NA

	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB
	Min. : 0.0	Min. : 0	Min. : 0	Min. : 0
	1st Qu.: 42.0	1st Qu.:451	1st Qu.: 548	1st Qu.: 66
	Median :102.0	Median :512	Median : 750	Median :101
	Mean : 99.6	Mean :502	Mean : 736	Mean :125
	3rd Qu.:147.0	3rd Qu.:580	3rd Qu.: 930	3rd Qu.:156
	Max. :264.0	Max. :878	Max. :1399	Max. :697
	NA	NA	NA's :102	NA's :131

	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR
	Min. : 0.0	Min. :29.0	Min. : 1137	Min. : 0
	1st Qu.: 38.0	1st Qu.:50.5	1st Qu.: 1419	1st Qu.: 50
	Median : 49.0	Median :58.0	Median : 1518	Median :107
	Mean : 52.8	Mean :59.4	Mean : 1779	Mean :106
	3rd Qu.: 62.0	3rd Qu.:67.0	3rd Qu.: 1682	3rd Qu.:150
	Max. :201.0	Max. :95.0	Max. :30132	Max. :343
	NA's :772	NA's :2085	NA	NA

	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
	Min. : 0	Min. : 0	Min. : 65	Min. : 52
	1st Qu.: 476	1st Qu.: 615	1st Qu.: 127	1st Qu.:131
	Median : 536	Median : 814	Median : 159	Median :149
	Mean : 553	Mean : 818	Mean : 246	Mean :146
	3rd Qu.: 611	3rd Qu.: 968	3rd Qu.: 249	3rd Qu.:164
	Max. :3645	Max. :19278	Max. :1898	Max. :228
	NA	NA's :102	NA	NA's :286

Subheading here

Put some text in here

Table 2: Standard Deviation

	x
TARGET_WINS	12.1150
TEAM_BATTING_H	76.1479
TEAM_BATTING_2B	26.3293
TEAM_BATTING_3B	9.0439
TEAM_BATTING_HR	32.4132
TEAM_BATTING_BB	74.8421
TEAM_BATTING_SO	104.1564
TEAM_BASERUN_SB	29.9164
TEAM_BASERUN_CS	11.8983
TEAM_BATTING_HBP	12.9671
TEAM_PITCHING_H	75.7886
TEAM_PITCHING_HR	32.3917
TEAM_PITCHING_BB	74.9167
TEAM_PITCHING_SO	104.3472
TEAM_FIELDING_E	16.6322
TEAM_FIELDING_DP	17.6117

Subheading here

Put some text in here

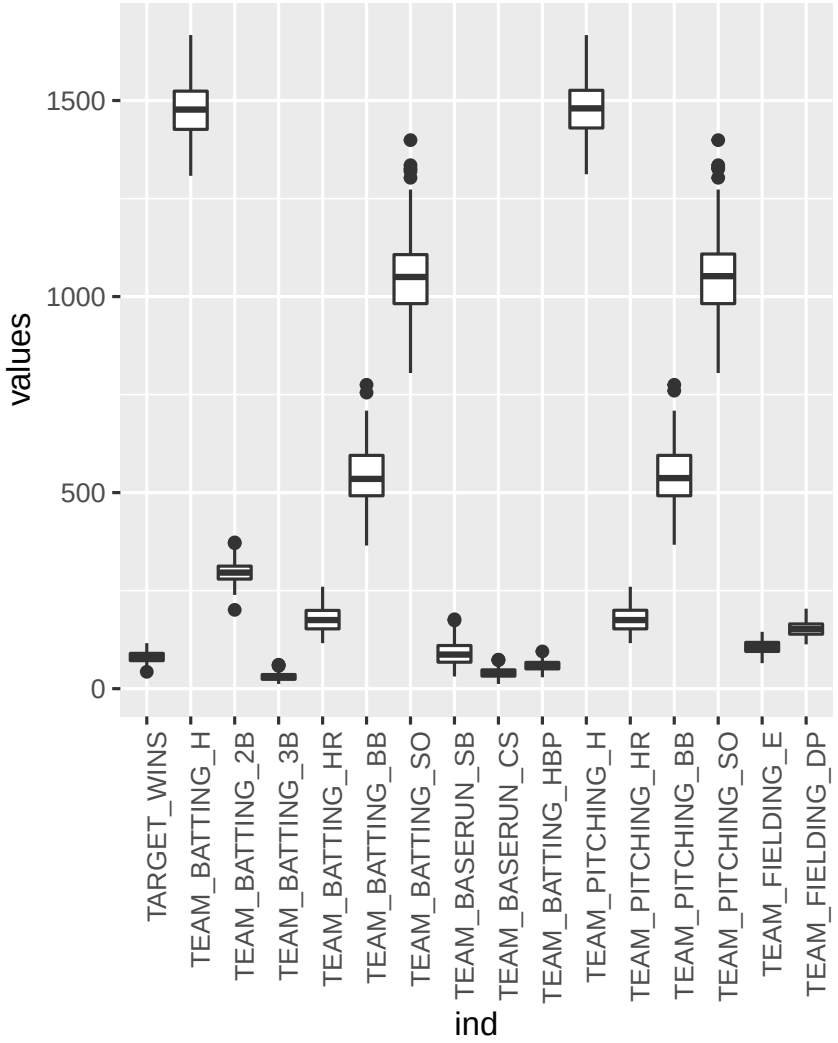

Figure 1: Boxplots

Subheading here

Put some text in here

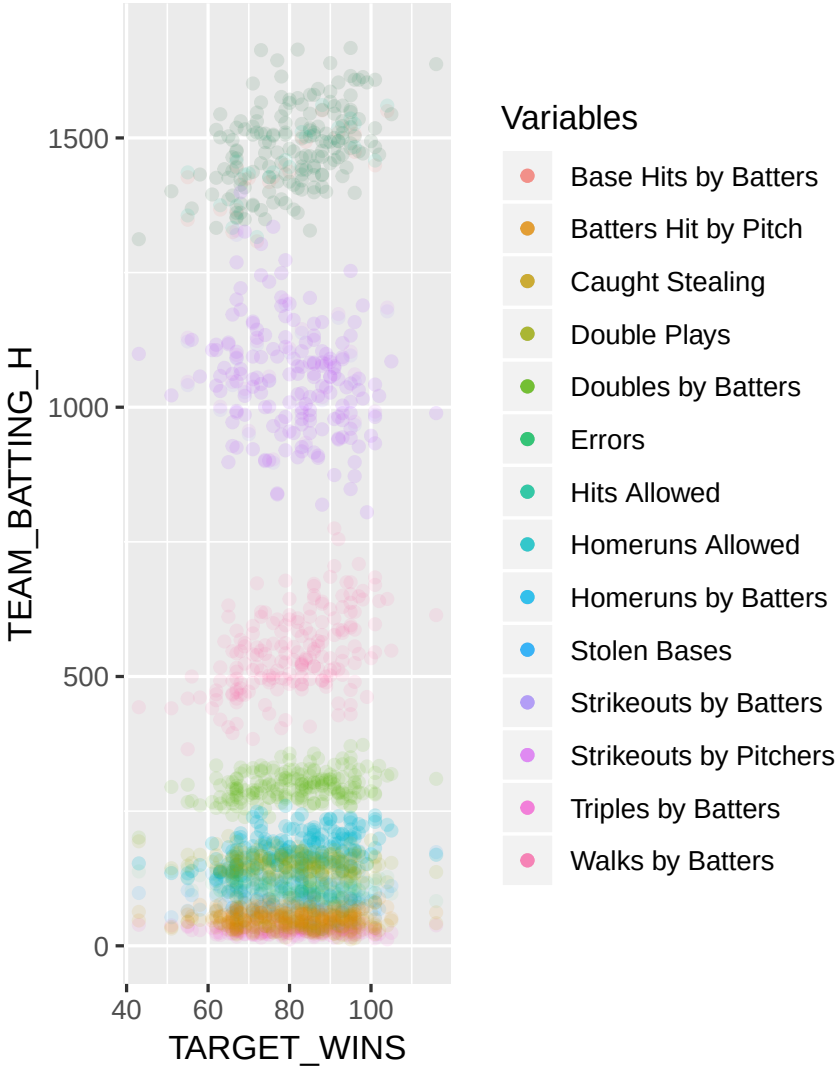

Figure 2: Point Plots

Subheading here

Put some text in here

Subheading here

Put some text in here

2. DATA PREPARATION

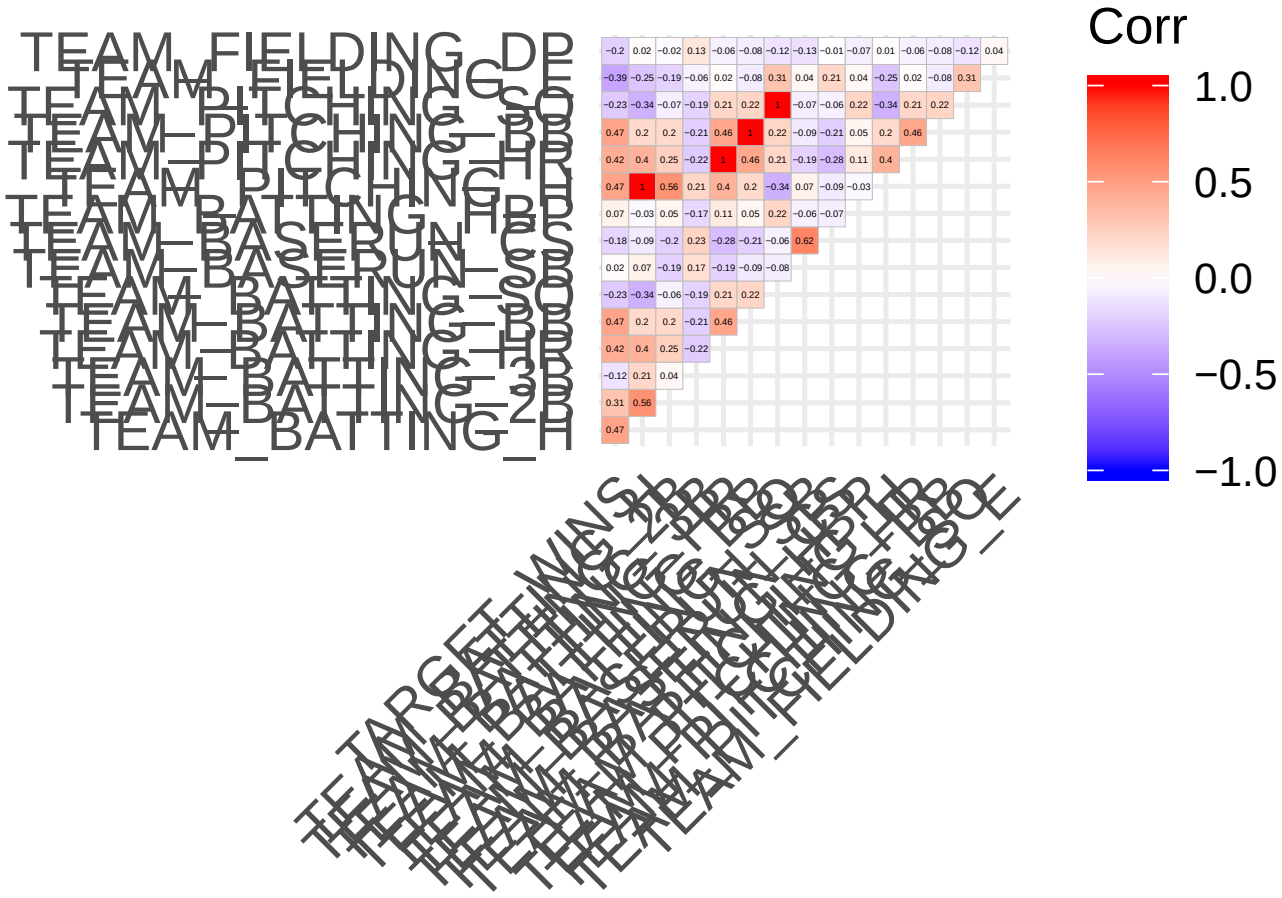

Figure 3: Correlation

Table 3: Missing Values by Variable

	x
TARGET_WINS	0
TEAM_BATTING_H	0
TEAM_BATTING_2B	0
TEAM_BATTING_3B	0
TEAM_BATTING_HR	0
TEAM_BATTING_BB	0
TEAM_BATTING_SO	102
TEAM_BASERUN_SB	131
TEAM_BASERUN_CS	772
TEAM_BATTING_HBP	2085
TEAM_PITCHING_H	0
TEAM_PITCHING_HR	0
TEAM_PITCHING_BB	0
TEAM_PITCHING_SO	102
TEAM_FIELDING_E	0
TEAM_FIELDING_DP	286

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
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.

Appendix

```
# <<<<<<<<<< copy and paste the script file HERE >>>>>>>>>>

# Training Exploration

moneyball_train <- read.csv("./data/moneyball-training-data.csv")[,-1] # use me
moneyball_complete <- moneyball_train[complete.cases(moneyball_train),]

Summary <- summary(moneyball_train)

Standard_Deviation <- sapply(moneyball_complete, sd)

Boxplots <- ggplot(stack(moneyball_complete), aes(x=ind, y=values)) +
    geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

Point_plots <- ggplot(data=moneyball_complete, aes(x=TARGET_WINS)) +
    geom_point(aes(y=TEAM_BATTING_H, color="Base Hits by Batters"), alpha=0.1) +
    geom_point(aes(y=TEAM_BATTING_2B, color="Doubles by Batters"), alpha=0.1) +
    geom_point(aes(y=TEAM_BATTING_3B, color="Triples by Batters"), alpha=0.1) +
    geom_point(aes(y=TEAM_BATTING_HR, color="Homeruns by Batters"), alpha=0.1) +
    geom_point(aes(y=TEAM_BATTING_BB, color="Walks by Batters"), alpha=0.1) +
    geom_point(aes(y=TEAM_BATTING_SO, color="Strikeouts by Batters"), alpha=0.1) +
    geom_point(aes(y=TEAM_BASERUN_SB, color="Stolen Bases"), alpha=0.1) +
    geom_point(aes(y=TEAM_BASERUN_CS, color="Caught Stealing"), alpha=0.1) +
    geom_point(aes(y=TEAM_BATTING_HBP, color="Batters Hit by Pitch"), alpha=0.1) +
    geom_point(aes(y=TEAM_PITCHING_H, color="Hits Allowed"), alpha=0.1) +
    geom_point(aes(y=TEAM_PITCHING_HR, color="Homeruns Allowed"), alpha=0.1) +
    geom_point(aes(y=TEAM_PITCHING_SO, color="Strikeouts by Pitchers"), alpha=0.1) +
    geom_point(aes(y=TEAM_FIELDING_E, color="Errors"), alpha=0.05) +
    geom_point(aes(y=TEAM_FIELDING_DP, color="Double Plays"), alpha=0.1) +
    labs(color="Variables", ylab="Variables")

Correlation <- ggcorrplot(as.data.frame(round(cor(moneyball_complete), 3)),
           type="upper", lab=TRUE, lab_size=.8)

Missing_values <- sapply(moneyball_train, function(x) sum(is.na(x)))
```

Examples to test formatting

footnote/sidenote

Footnotes are on the side!!![1]

[1] Beautiful Evidence