Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Moneyball OLS Regression

BINGO BONUS:

(20 Pts) Use of the PROC GLM and PROC GENMOD procedures in SAS produced the same coefficient estimates as the PROC REG procedure. I did notice that there were different outputs in terms of the information provided about the selected regressor variables for assessing goodness of fit and maximum likelihood of the coefficient estimates.

Executive Summary & Introduction

Our analysis produced an Ordinary Least Squares (OLS) Regression model that represents 42% of the variability in the estimated number of wins of a 162-game baseball season for a team.

The purpose of this assignment is to develop an ordinary least squares (OLS) regression model to predict the number of wins for a baseball team during a 162-game season. During this project will be emulating the efforts of the Moneyball model developed by Billy Beane and Paul DePodesta of the Oakland Athletics Baseball Club that was used to hire and trade players. Our methodology will consist of different phases to identify the OLS regression model that best represents the variability in the response variable Target_Wins. We will initiate our efforts with exploratory data analysis to understand our data, progress to data manipulation (imputation & transformations) to prepare our data for OLS regression, then finally initiate model development and selection using automated variable selection techniques (forward, backward, stepwise) and criteria (Adjusted R-Square, AIC, BIC) to develop the model that best predicts the number of wins for a baseball team. Once we've selected the best model for our analysis, we will develop a scoring routine that will produce the predicted number of wins for a 162-game season.

Data Exploration

The data consist of 16 variables with 2276 individual records that describe the performance of a baseball team. Each record has been adjusted to match the performance of a team during a regular, 162 game season. An initial observation is that the data does not contain a variable for time, so we must assume that the playing of baseball (rules, athletic ability, supplement use) has been homogenous during the performance of this data. The data dictionary proposes a relationship to the response variable Target_Wins for the given regressor variables that have been provided for our analysis. Of the 16 variables being used for this analysis, six variables (Team_Batting_HBP, Team_Batting_SO, Team_Baserun_SB, Team_Baserun_CS, Team_Fielding_DP, Team_Pitching_SO) are missing a significant amount of records and will require imputation. I intend on dropping Team_Batting_HBP from our analysis due to missing 90% of its values.

Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Further analysis of the regressor variables using the PROC CORR and PROC UNIVARIATE procedures will help us gain additional details about the variables in our dataset and the possible relationships that exist between them. Information gained from the data analysis phase will direct our actions for imputation and transformations as we seek to prepare the data for OLS regression. For each variable in the data that we plan on using in the model we will examine its measures of central tendency (mean, median, mode, standard deviation), adherence to normal distribution (necessary to fulfill the assumptions of OLS regression), and relationships (via Pearson Correlation Coefficients (R)) with other variables.

Target_Wins

The response variable Target_Wins represents the number of wins in a season for a baseball team and has 2276 of 2276 records with a mean of 80.79, median of 82, mode of 83, and standard deviation of 15.75. A frequency plot of the data reveals a normal distribution with minimum kurtosis that is confirmed by its quantiles plot. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds.

Running the PROC CORR procedure for the variables in our data calculates the *Pearson Correlation Coefficients* and provides evidence that the proposed relationship between Target_Wins is contrary to our initial inclinations. Specifically, Team_Baserun_CS, Team_Pitching_BB, and Team_Pitching_HR have positive relationship with then number of wins and Team_Fielding_DP and Team_Pitching_SO have a negative relationship to the number of wins. The results of the PROC CORR procedure run counter to our expectations because we hypothesized that the statistics that reflect a "positive" performance of a team would also represent a winning team and the statistics that result from a "negative" team performance would represent a losing team.

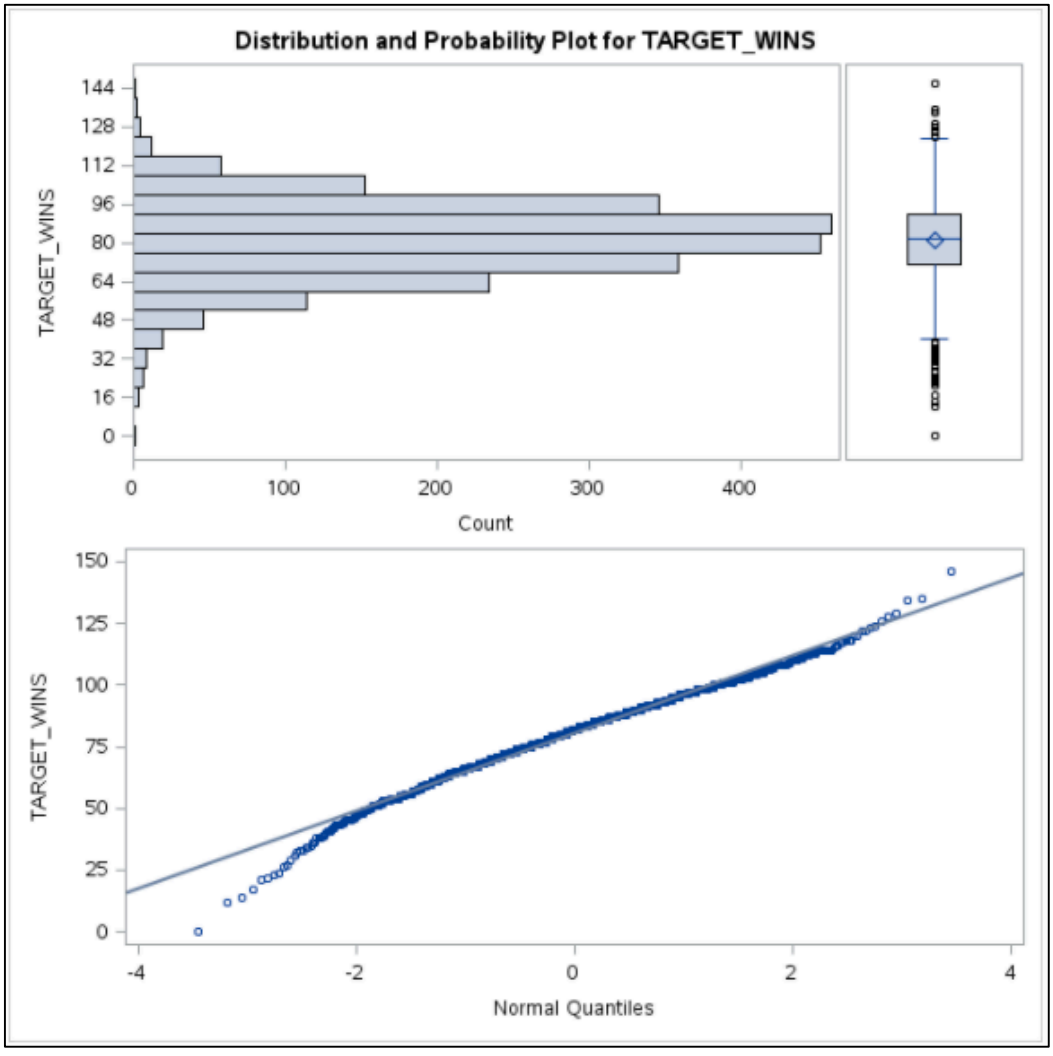
Variable Description & Proposed vs. Actual Relationship with Target_Wins

Variable	**Description**	**Proposed**	**Actual (R value)**
Target_Wins	Number of Wins		
Team_Batting_H	Base Hits by Batters	Positive	Positive (0.38877)
Team_Batting_2B	Doubles by Batters	Positive	Positive (0.28910)
Team_Batting_3B	Triples by Batters	Positive	Positive (0.14261)
Team_Batting_HR	Homeruns by Batters	Positive	Positive (0.17615)
Team_Batting_BB	Walks by Batters	Positive	Positive (0.23256)
Team_Batting_HBP	Batters hit by pitch	Positive	Positive (0.07350)
Team_Batting_SO	Strikeouts by Batters	Negative	Negative (-0.03175)
Team_Baserun_SB	Stolen bases	Positive	Positive (0.13514)
Team_Baserun_CS	Caught Stealing	Negative	Positive (0.02240)

Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Variable	Description	Proposed	Actual (R value)
Team_Fielding_E	Errors	Negative	Negative (-0.17648)
Team_Fielding_DP	Double plays	Positive	Negative (-0.03485)
Team_Pitching_BB	Walks Allowed	Negative	Positive (0.12417)
Team_Pitching_H	Hits Allowed	Negative	Negative (-0.10994)
Team_Pitching_HR	Homeruns Allowed	Negative	Positive (0.18901)
Team_Pitching_SO	Strikeouts by Pitchers	Positive	Negative (-0.07844)

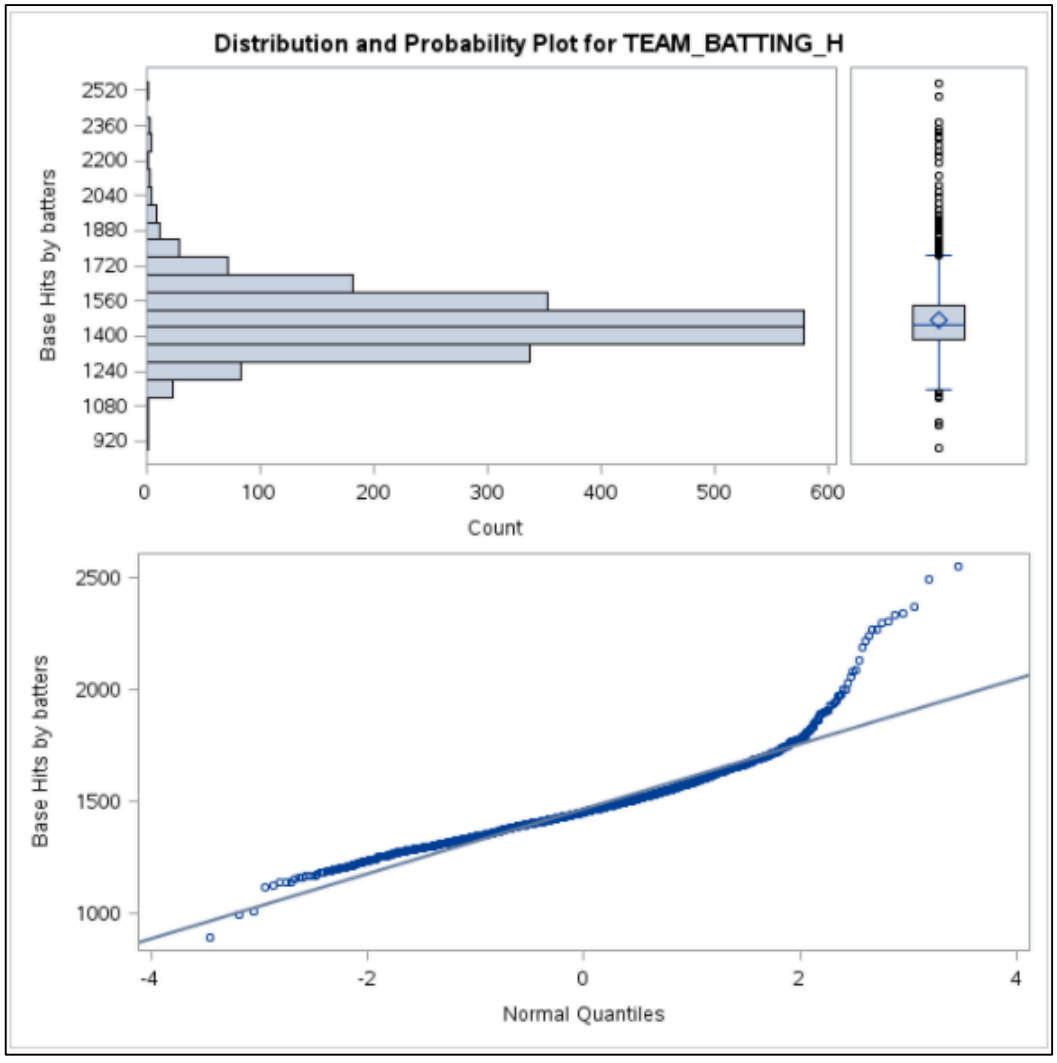
**Red text represents conflicting relationships with variable Target_Wins*



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Batting_H

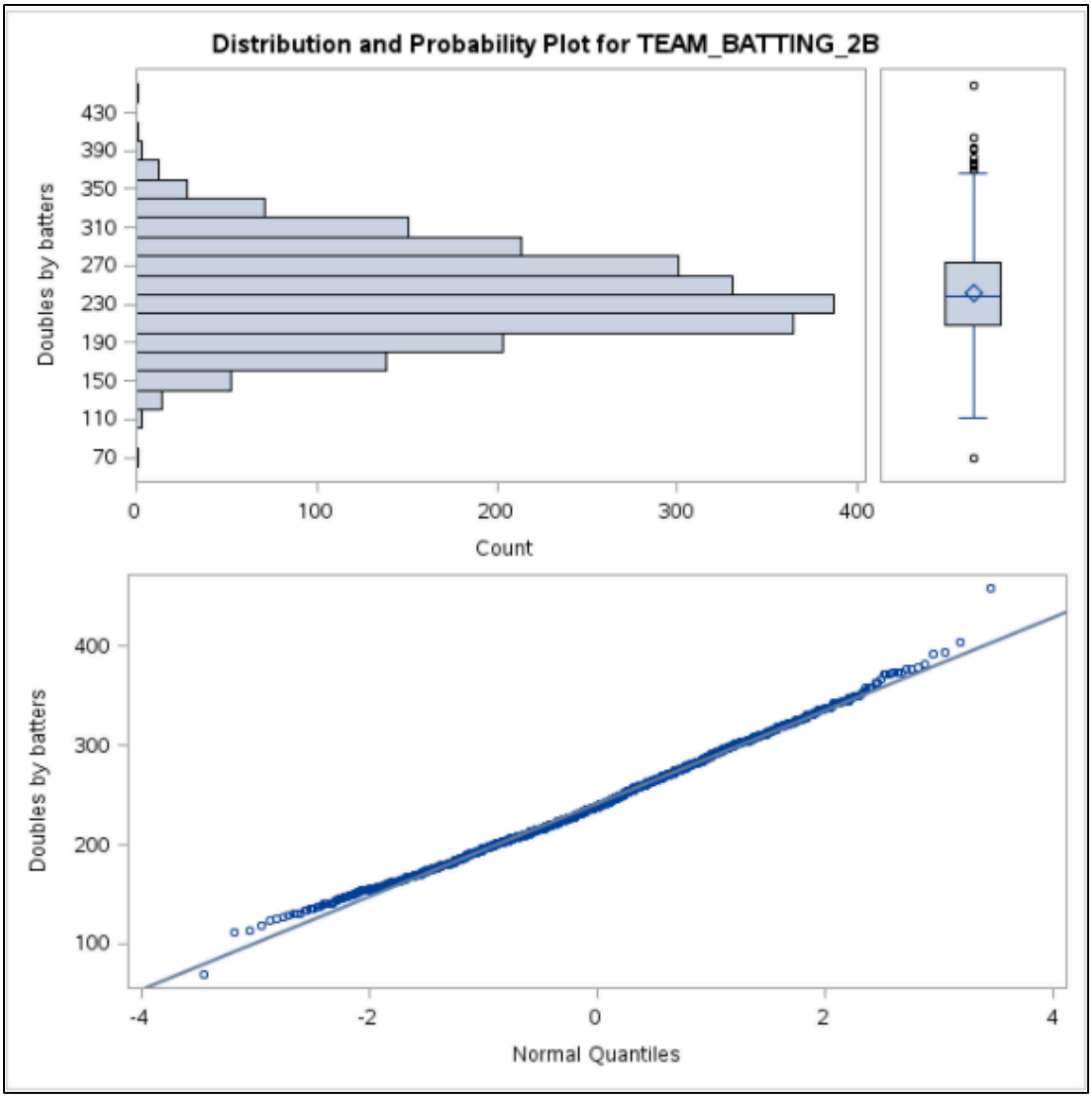
The regressor variable Team_Batting_H represents the number of base hits during a 162-game season and has 2276 of 2276 records with a mean of 1469.26, median of 1454, mode of 1458, and standard deviation of 144.60. A frequency plot of the variable reveals a normal distribution that is slightly left skewed. A review of the quantiles plot shows a departure from normality at the higher values of Team_Batting_H. A box plot of the values shows several outliers beyond the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) bounds. As expected Team_Batting_H has strong positive relationships with Team_Batting_2B (R = 0.56) and Team_Batting_3B (R = 0.43) because they contribute to the number of base hits and has a strong negative relationship with Team_Batting_SO (R = -0.46) that reduces the number of base hits.



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Batting_2B

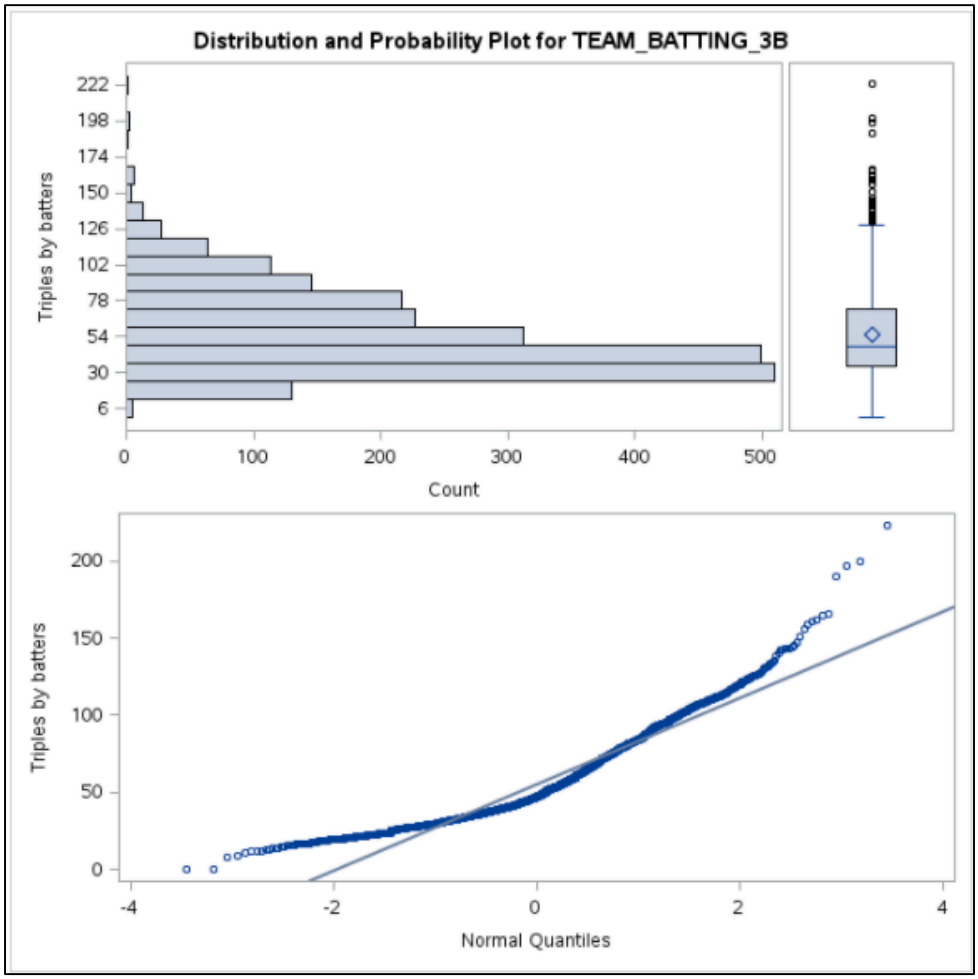
The regressor variable Team_Batting_2B represents the number of doubles by a team in a season. Team_Batting_2B has a mean of 241.25, median of 238, mode of 227, and standard deviation of 46.8. The frequency and quantile plot of the variable reveals a normal distribution of values. A box plot of the values shows a few outliers beyond the Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Batting_2B appears to have strong positive relationships with Team_Batting_H (0.56285) and Team_Batting_HR (0.43540) which indicates that players that hit doubles also hit home runs. Team_Batting_2B also has strong positive relationships with Team_Fielding_DP (0.16) and Team_Pitching_HR (0.45) that I cannot explain.



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Batting_3B

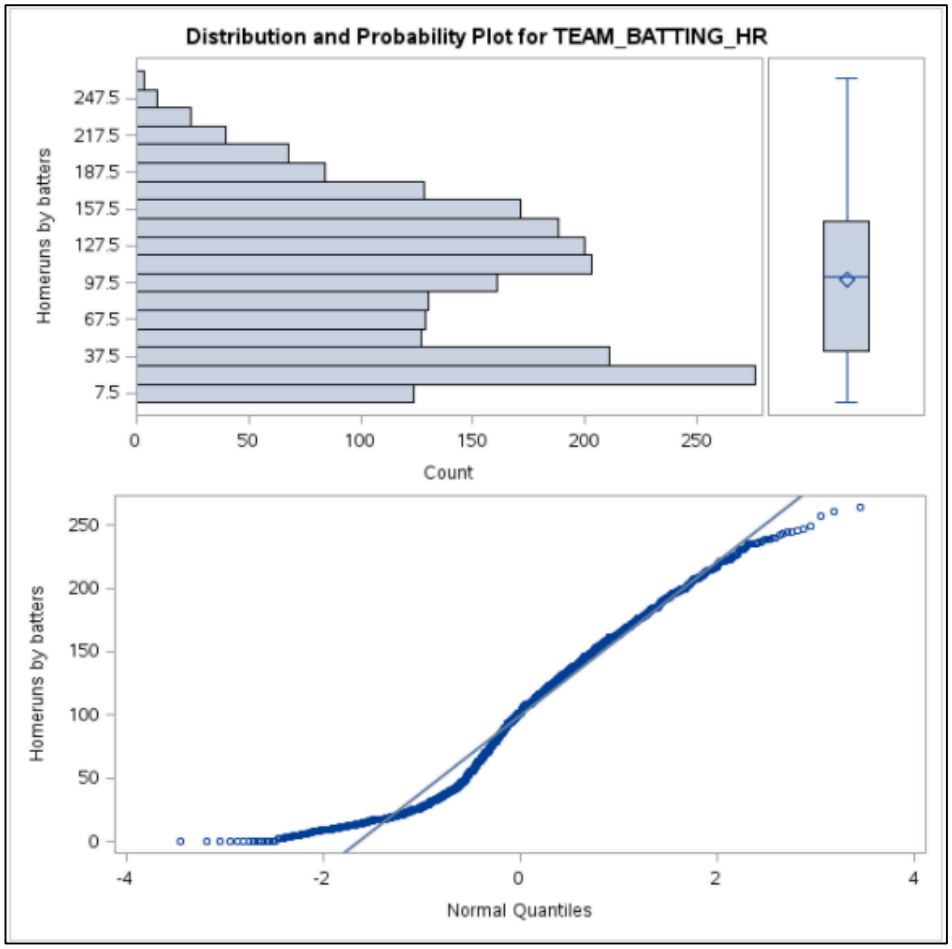
The regressor variable Team_Batting_3B represents the number of triples per team in a season. Team_Batting_3B has a mean of 55.25, median of 47, mode of 35, and standard deviation of 27.94. The frequency and quantile plot of the variable reveals a heavily right-skewed normal distribution of values that depart slightly at the lower and higher values. A box plot of the values shows numerous outliers beyond the Upper (Q3 + 3*(Q3-Q1)) bound. Team_Batting_3B appears to have strong positive relationships with Team_Batting_H (0.43), Team_Baserun_SB (0.53), Team_Baserun_CS (0.35), Team_Fielding_E (0.51). The variable also has some pretty strong negative relationships with Team_Batting_HR (-0.64), Team_Batting_SO (-0.67), Team_Fielding_DP (-0.32), and Team_Pitching_HR (-0.57). There appears to be an underlying structure to the data that I don't fully understand. I believe that the data may represent a tradeoff of players' skills where players who are good at hitting triples are not great fielders or homerun hitters.



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Batting_HR

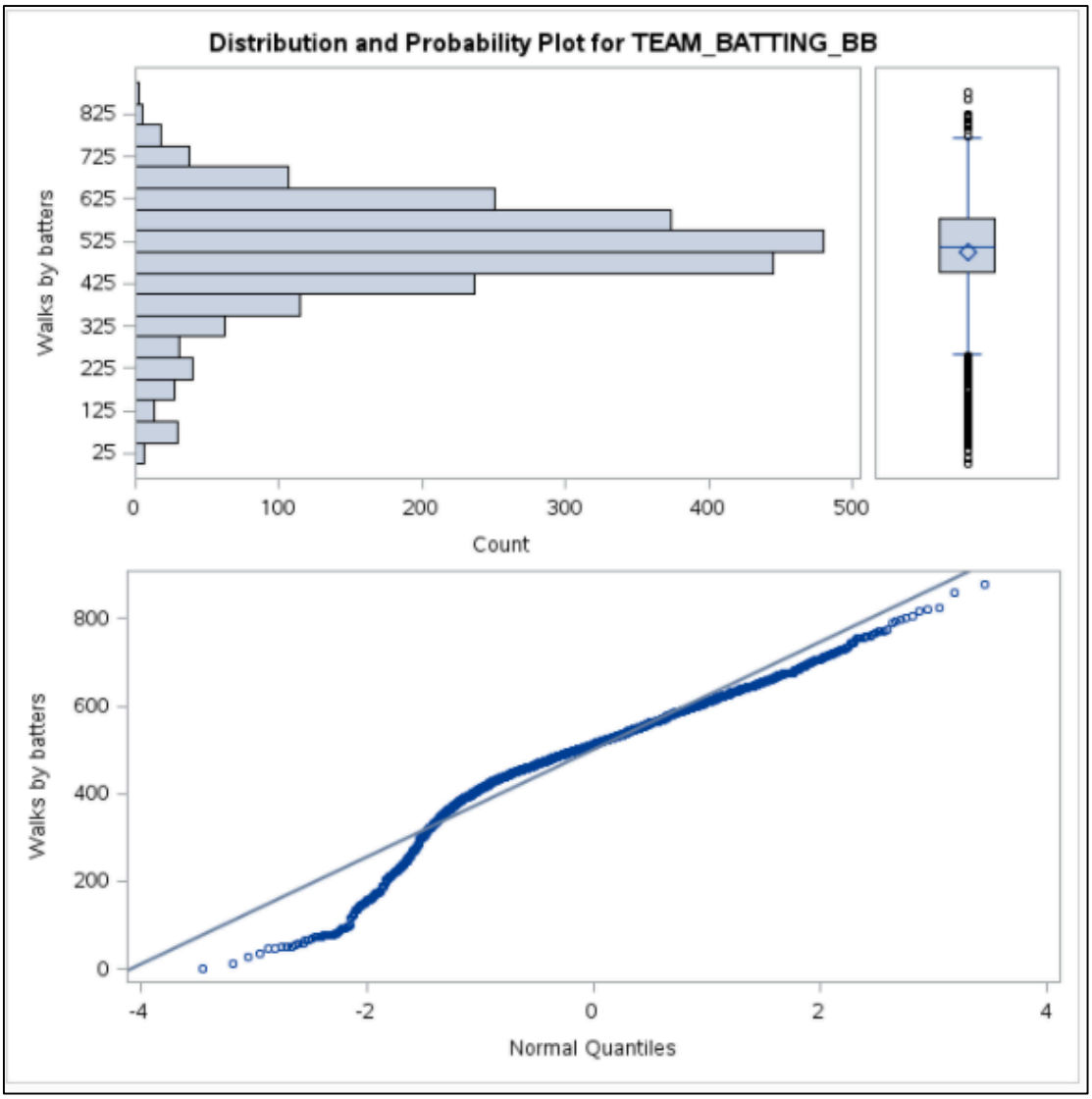
The regressor variable Team_Batting_HR represents the number of homeruns by a team in a season. Team_Batting_HR has a mean of 99.6, median of 102, mode of 21, and standard deviation of 60.55. The frequency and quantile plot of the variable reveals a distribution with twin peak values that occur frequently. The mode value of 21 represents the second peak and marks a significant departure of normality due to the excessive frequency of its occurrence. The box plot of the values shows no outliers beyond the Lower (Q1 - 3*(Q3-Q1)) and Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Batting_HR appears to have strong positive relationships with Team_Batting_2B (0.44), Team_Batting_BB (0.51), Team_Batting_SO (0.73), Team_Fielding_DP (0.45) and Team_Pitching_HR (0.97). The correlation with Team_Pitching_HR is particularly curious because a Pearson Correlation Coefficient value of 0.97 indicates an almost perfect corollary relationship. I cannot begin to understand how the number of homeruns hit by a team corresponds to the number of homeruns their pitching staff allows. Team_Batting_HR also has negative relationships with Team_Batting_3B (-0.63), Team_Baserun_SB (-0.45), Team_Baserun_CS (-0.43), and Team_Fielding_E (-0.58).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Batting_BB

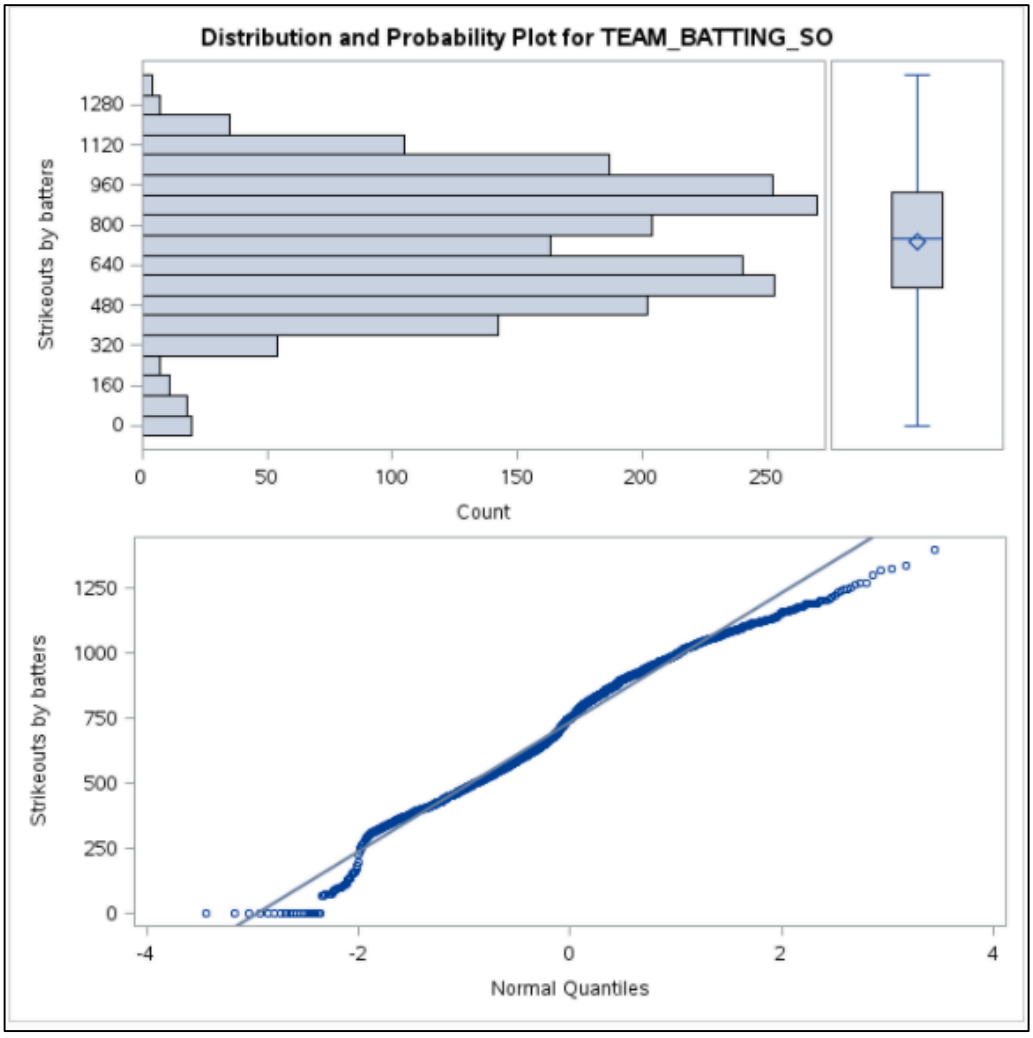
The regressor variable Team_Batting_BB represents the number of walks by batters on a team during a season. Team_Batting_BB has a mean of 501.56, median of 512, mode of 502, and standard deviation of 102.67. A frequency and probability of the variable reveals a normal distribution that is slightly left skewed with a departure from normality occurring at the lower values. A box plot of the values shows several outliers beyond the Lower (Q1 - 3*(Q3-Q1)) and Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Batting_BB has strong positive relationships with Team_Batting_HR (0.51), Team_Batting_SO (0.38), Team_Fielding_DP (0.43), Team_Pitching_BB (0.49), and Team_Pitching_HR (0.46). The regressor variable has a strong negative relationship with Team_Batting_E (-0.66), Team_Pitching_H (-0.45).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Batting_SO

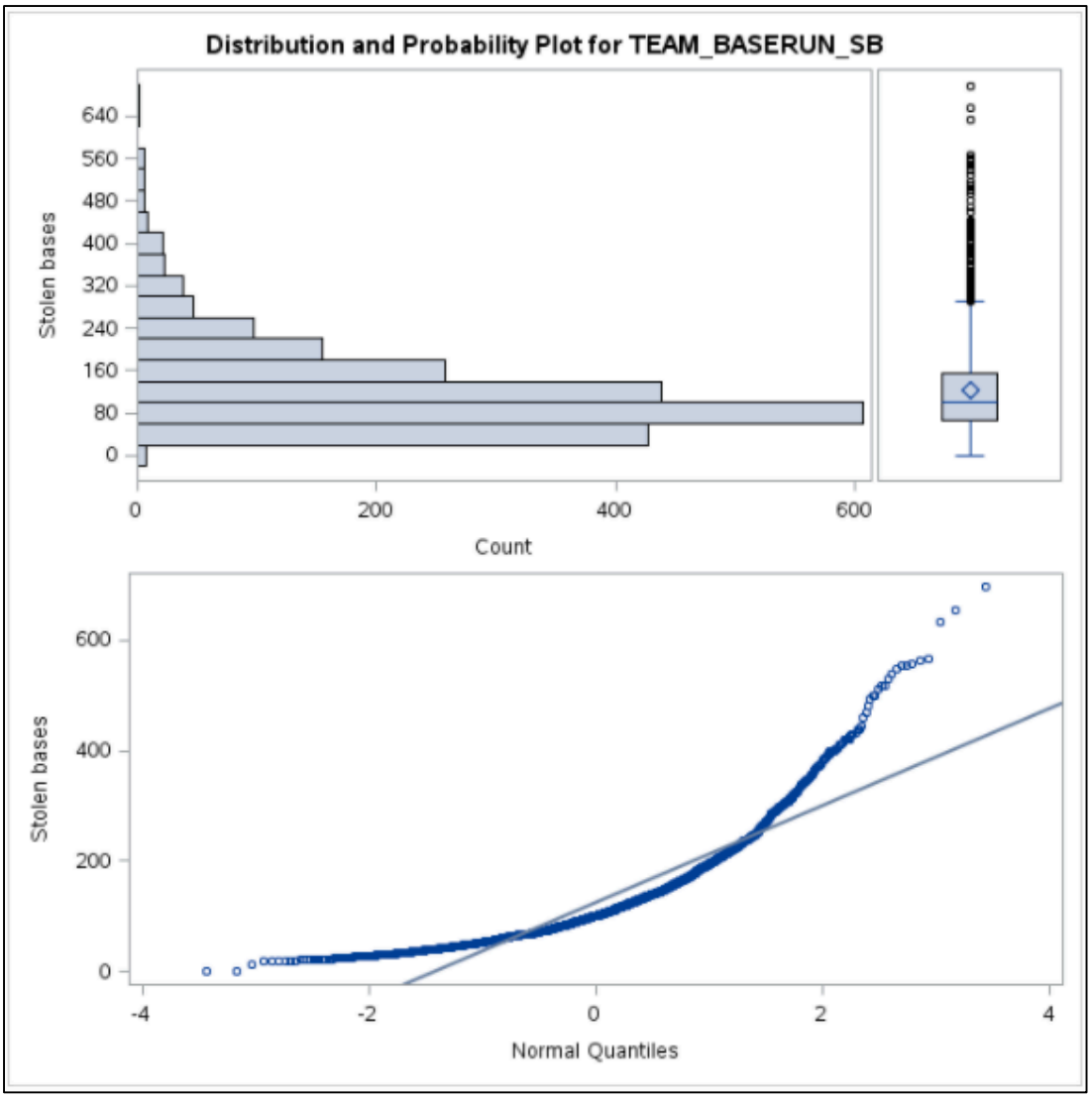
The regressor variable Team_Batting_SO represents the number of strikeouts by batters for a team during a season. Team_Batting_SO has a mean of 735.60, median of 750, and mode of 0, and standard deviation of 248.52. For an unknown reason, the PROC UNIVARIATE procedure is producing a mode of 0.00 despite 960 appearing to be the most frequent value. A frequency and quantile plots reveals a normal distribution that is left skewed. There are two peaks rather than a single peak, indicating that there are two values that occur the most often in the data. A review of the quantiles plot shows a departure from normality at the higher and lower values of Team_Batting_SO. A box plot of the values shows no significant outliers at the upper and lower bounds. The variable has a strong positive relationship with Team_Batting_HR (0.73), Team_Pitching_HR (0.67), and Team_Batting_BB (0.38). Team_Batting_HR has a strong negative relationship with Team_Batting_H (-0.46), Team_Batting_3B (-0.67), Team_Fielding_E (-0.58), and Team_Pitching_H (-0.38).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Baserun_SB

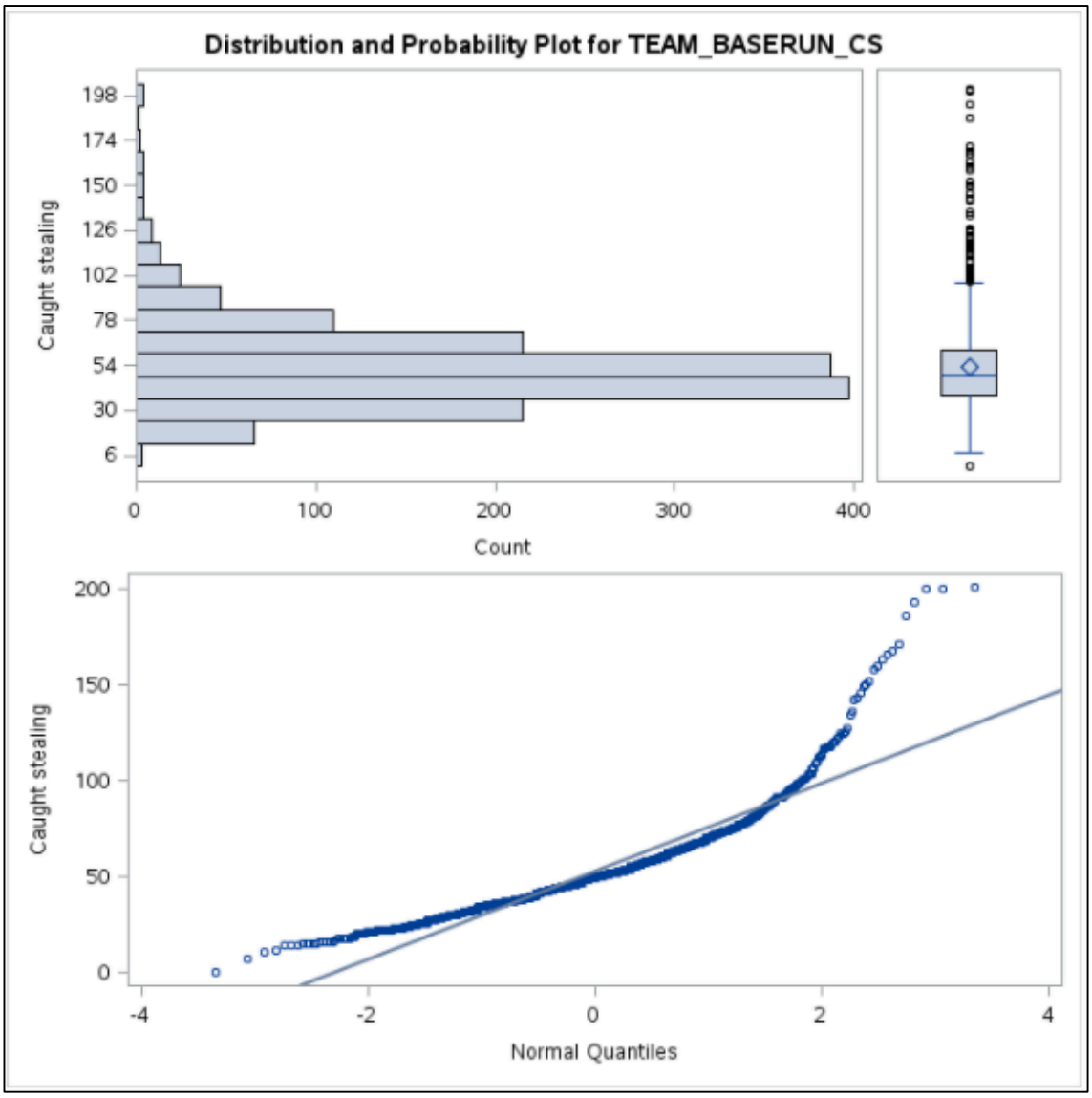
The regressor variable Team_Baserun_SB represents the number of stolen bases by a team during a season. Team_Baserun_SB and has a mean of 124.76, median of 101, mode of 65, and standard deviation of 87.79. A frequency and probability plot of the variable reveals a distribution that is heavily skewed to the right. Further review of the quantiles plot shows a departure from normality for most of the data points. A box plot of the values shows several outliers beyond the Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Baserun_SB has strong positive and negative relationships with Team_Batting_3B (0.53), Team_Baserun_CS (0.66), Team_Fielding_E (0.50), Team_Batting_HR (-0.45), Team_Fielding_DP (-0.50), and Team_Pitching_HR (-0.41).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Baserun_CS

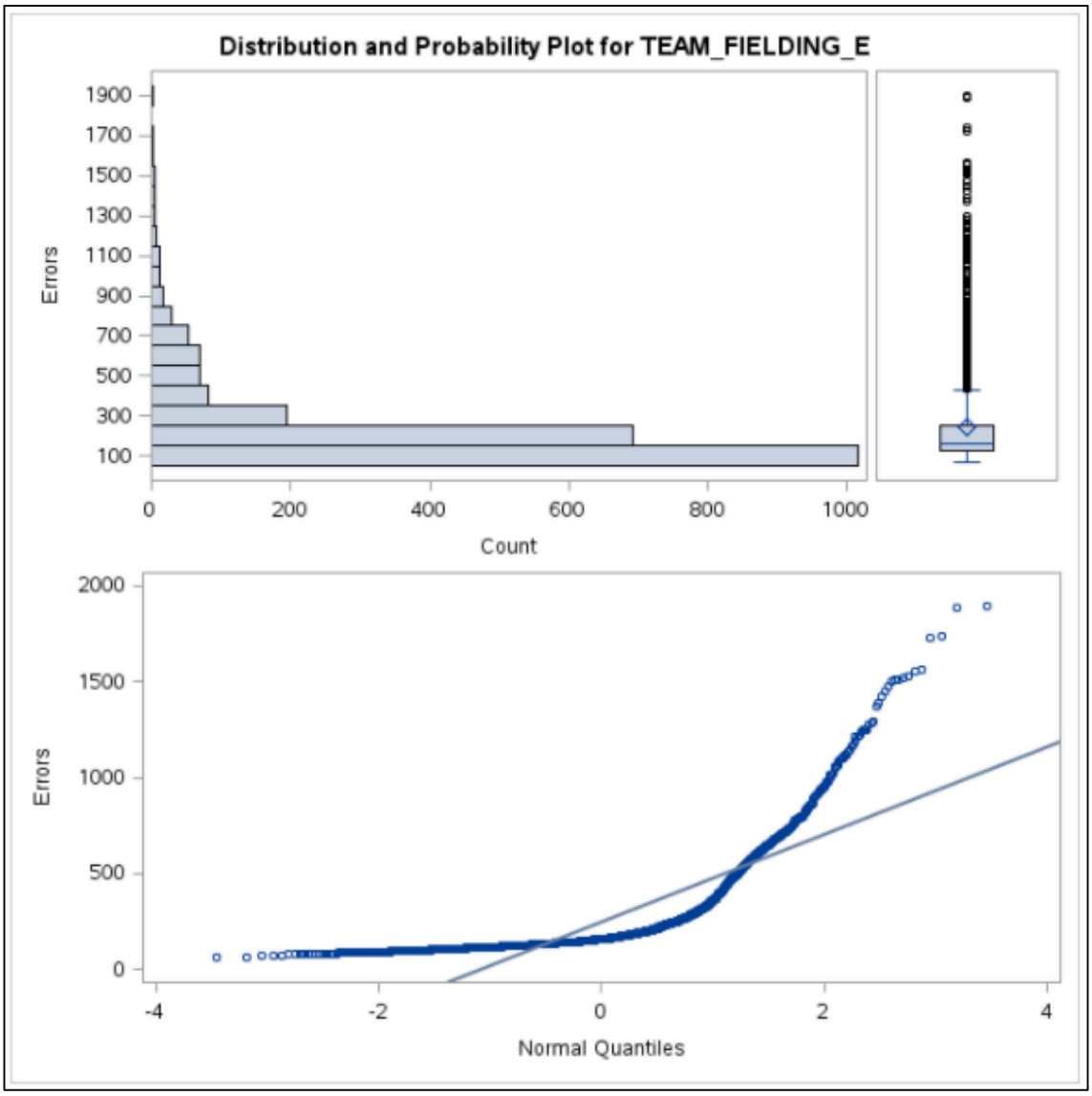
The regressor variable Team_Baserun_CS represents the number of batters caught stealing bases during a season and has a mean of 52.80, median of 49, mode of 52, and standard deviation of 22.96. The distribution, probability, and box plot of the variable reveals data that follows a right-skewed normal distribution and has several outliers above the Upper bound (Q3 + 3*(Q3-Q1)). A review of the quantiles plot shows a departure from normality both the higher and lower values. Team_Baserun_CS has strong positive correlation with Team_Batting_3B (0.35) Team_Baserun_SB (0.66) (those who steal get caught stealing more). Additionally, the variable has negative relationships with Team_Batting_HR (-0.43) and Team_Pitching_HR (0.42).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Fielding_E

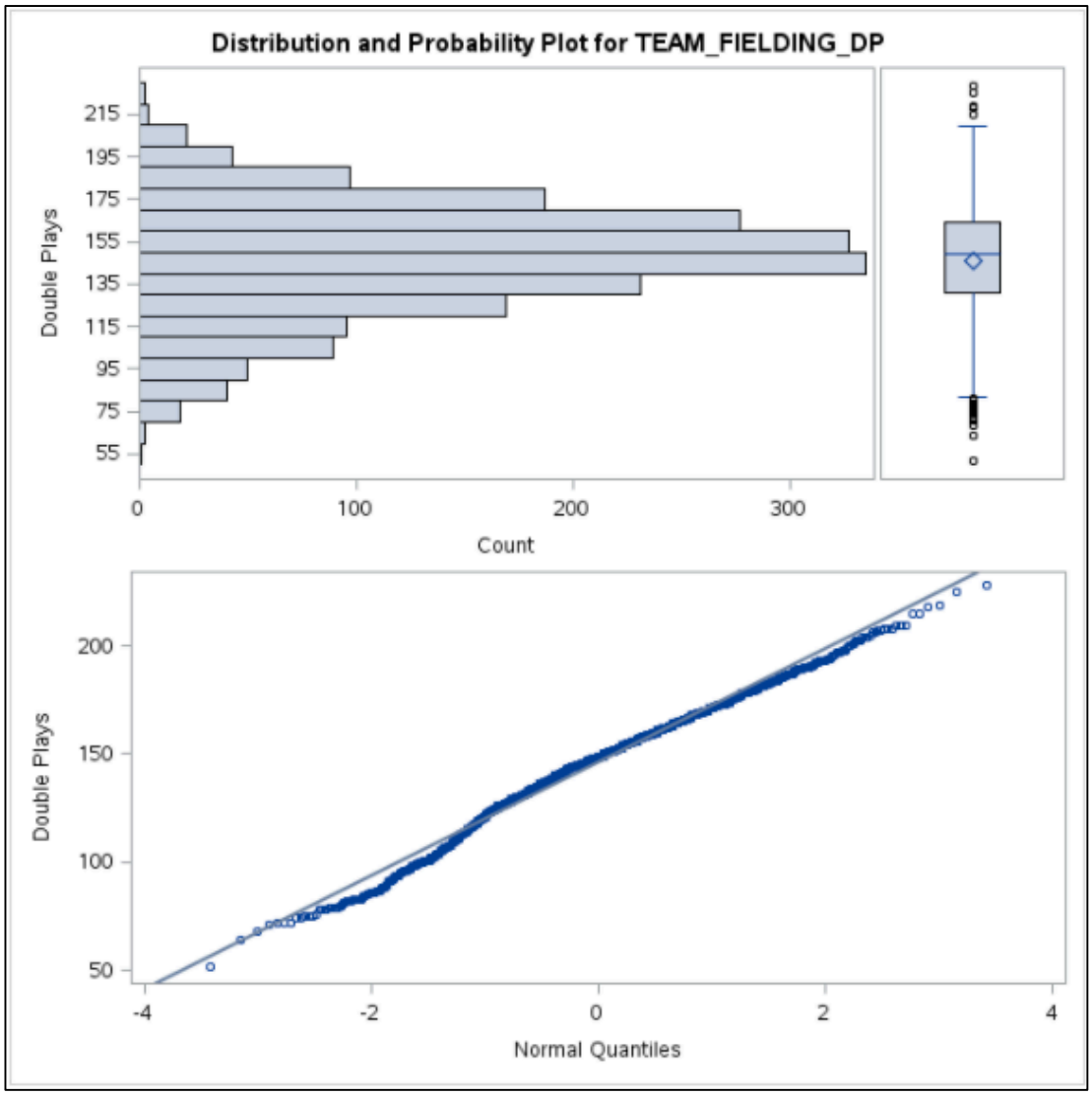
The regressor variable Team_Fielding_E represents the number of fielding errors during a season and has a mean of 246.48, median of 159, mode of 122, and standard deviation of 227.27. The distribution, probability, and box plot of the variable reveals data is heavily right-skewed and has several outliers above the Upper bound (Q3 + 3*(Q3-Q1)). A review of the quantiles plot shows a departure from normality both the higher and lower values. Team_Fielding_E has strong positive correlation with Team_Batting_3B (0.50) Team_Baserun_SB (0.50) and Team_Pitching_H (0.67). Additionally, the variable has negative relationships with Team_Batting_HR (-0.59), Team_Batting_BB (-0.66), Team_Batting_SO (-0.58), Team_Fielding_DP (-0.50), and Team_Pitching_HR (-0.49).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Fielding_DP

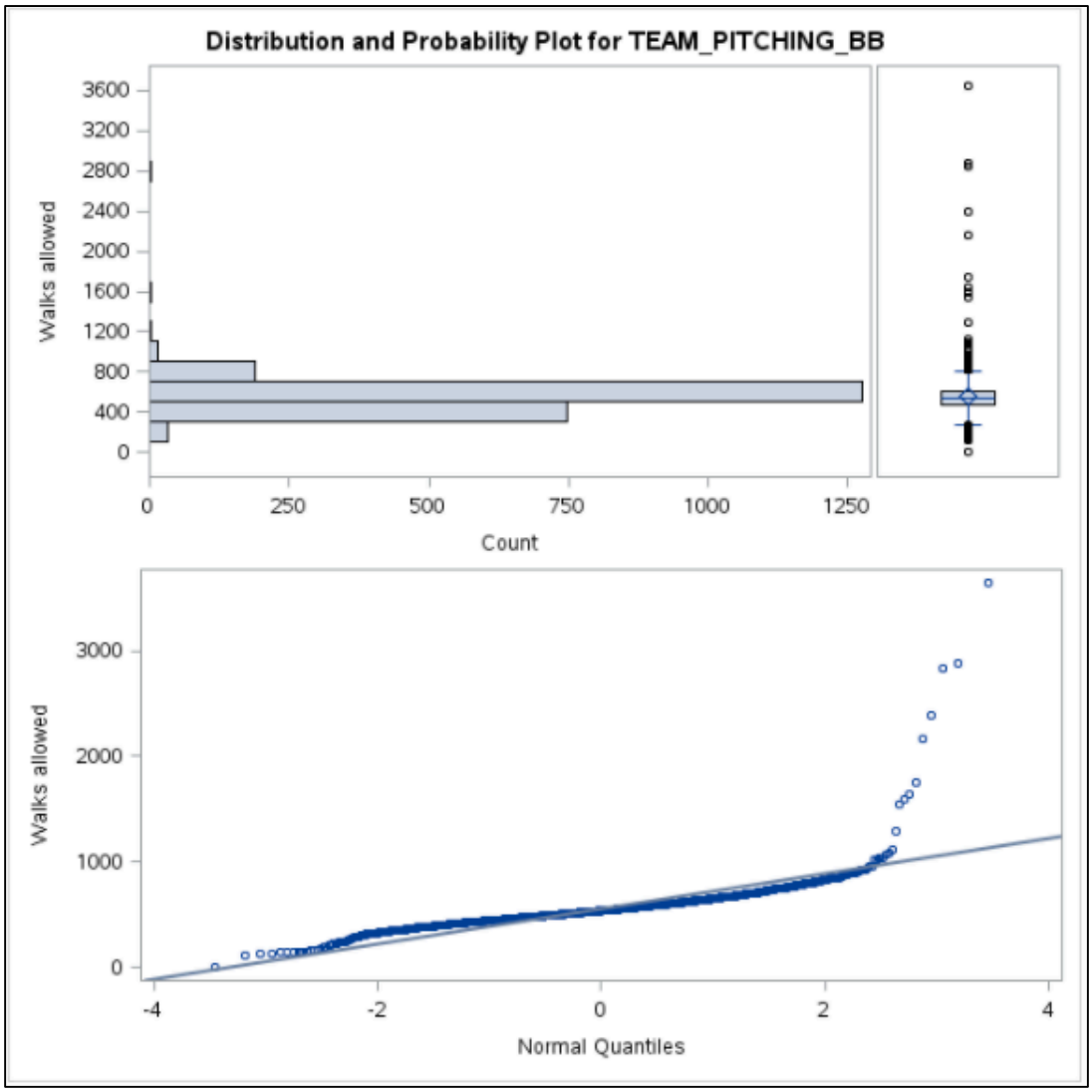
The regressor variable Team_Batting_DP represents the number of double plays during a season and has a mean of 146.39, median of 149, mode of 148, and standard deviation of 26.23. A review of the distribution and probability plot of the variable reveals a normal distribution that is nearly perfect. A review of the quantiles plot shows minimal departure from normality. A box plot of the values shows several outliers beyond the Lower (Q1 - 3*(Q3-Q1)) and Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Fielding_DP has strong positive relationships with Team_Batting_HR (0.45), Team_Batting_BB (0.43), Team_Pitching_BB (0.32) and Team_Pitching_HR (0.44). The regressor has negative corollary relationships with Team_Baserun_SB (-0.50) and Team_Fielding_E (-0.50).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Pitching_BB

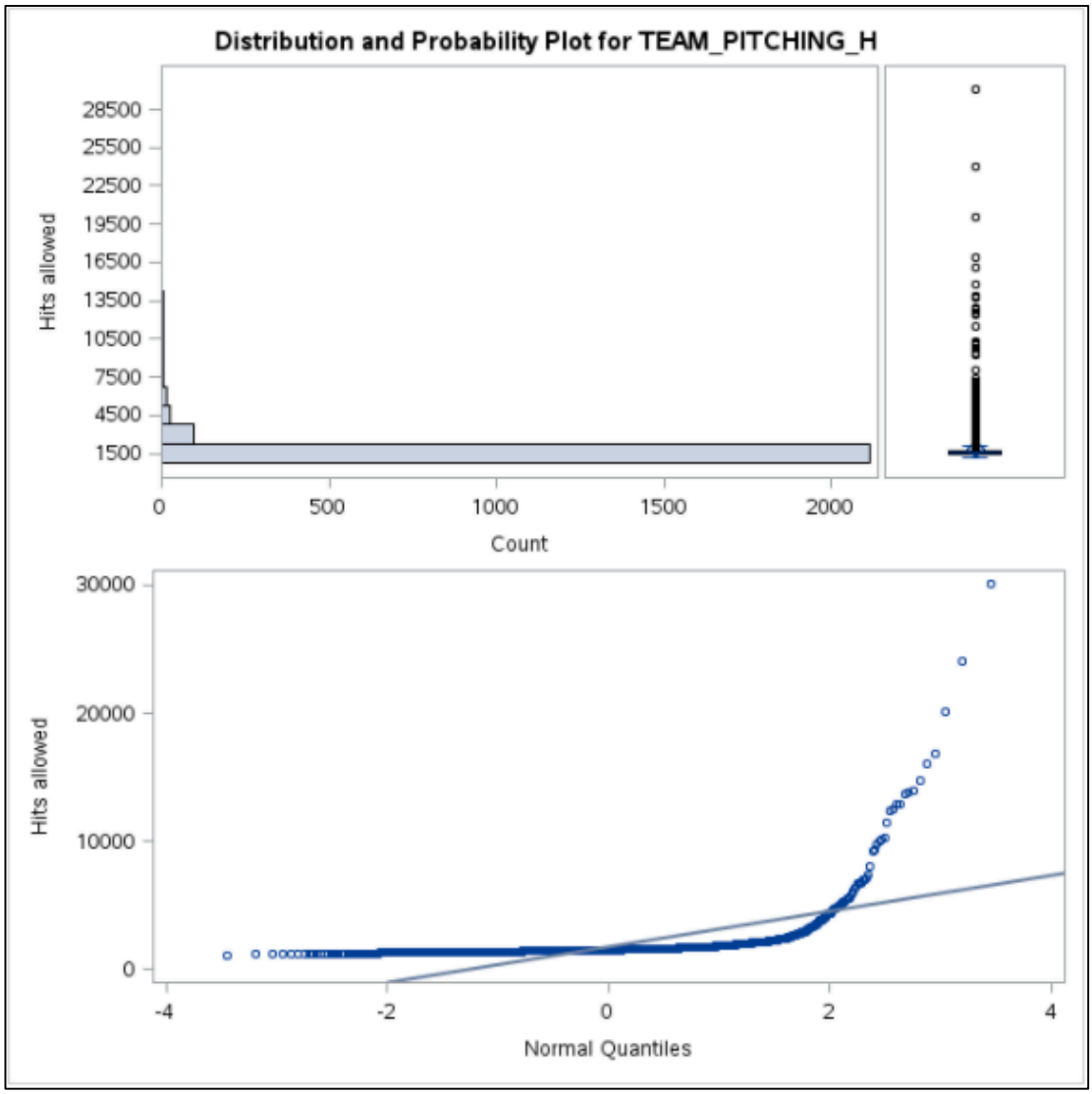
The regressor variable Team_Pitching_BB represents the number of walks allowed during a season and has a mean of 553, median of 536.5, mode of 536, and standard deviation of 166.36. A review of the distribution and probability plot of the variable reveals a normal distribution that is appears nearly perfect due to its scale. A review of the quantiles plot shows minimal departure from normality. A box plot of the values shows several outliers beyond the Lower (Q1 - 3*(Q3-Q1)) and Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Pitching_BB has strong positive relationships with Team_Batting_BB (0.49), Team_Fielding_DP (0.32), Team_Pitching_H (0.32) and Team_Pitching_SO (0.49). The regressor has no strong negative corollary relationships with the other variables in the data set.



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Pitching_H

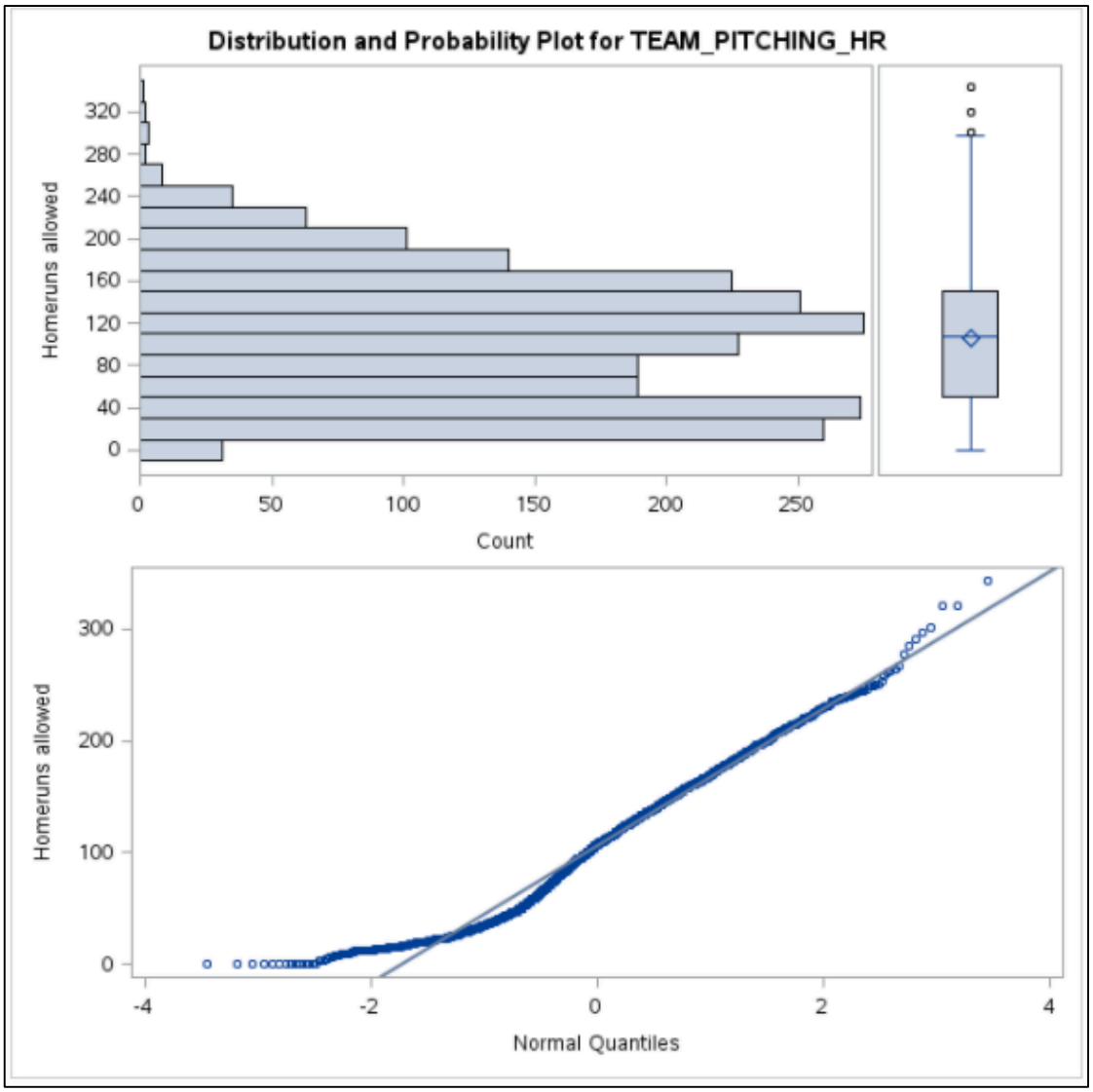
The regressor variable Team_Pitching_H represents the number of hits allowed during a season and has a mean of 1779.21, median of 1518, mode of 1494, and standard deviation of 1406.84. Distribution and probability plots of the variable reveals a heavily right skewed distribution. A box plot of the values shows several outliers beyond the Lower (Q1 - 3*(Q3-Q1)) and Upper (Q3 + 3*(Q3-Q1)) bounds. Team_Pitching_H has strong positive relationships with Team_Fielding_E (0.67) & Team_Pitching_BB (0.32) and has a strong negative relationship with Team_Batting_BB (-0.45) & Team_Batting_SO (-0.38).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Pitching_HR

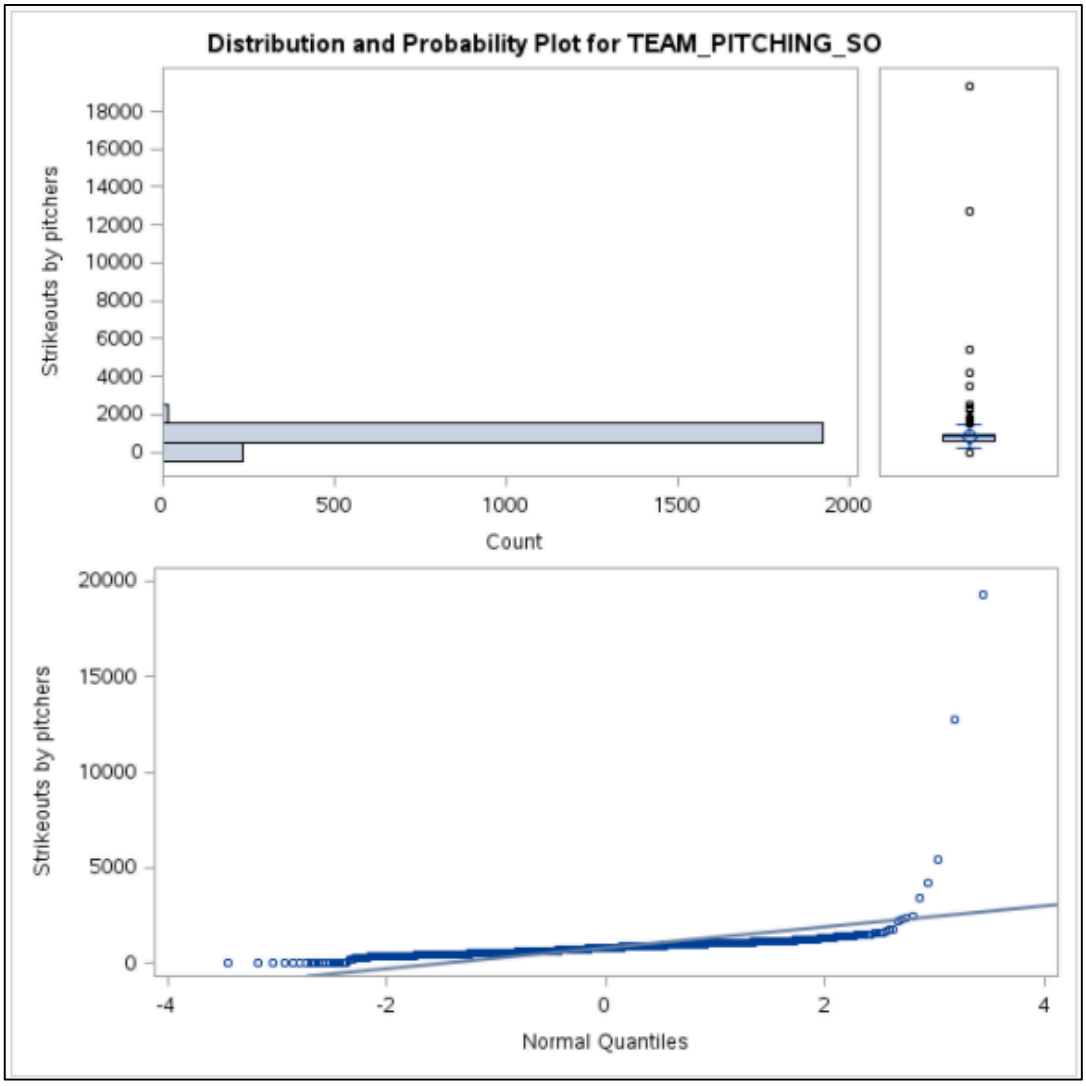
The regressor variable Team_Pitching_HR represents the number of homeruns allowed during a season with mean of 105.7, median 107, mode 114, and standard deviation 61.3. Distribution plot of the variable reveals data that some what follows a normal distribution with twin peak values. A review of the quantiles plot shows a departure from normality mostly at the lower values of Team_Pitching_HR. A box plot of the values shows a few outliers beyond Upper (Q3 + 3*(Q3-Q1)) bound. Team_Pitching_HR has strong positive relationships with Team_Batting_HR (0.97), Team_Batting_2B (0.45), and Team_Batting_SO (0.67). The variable has significant negative relationships with Team_Batting_3B (-0.57), Team_Baserun_SB (-0.41), Team_Baserun_CS (-0.42), and Team_Fielding_E (-0.49).



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Pitching_SO

The regressor variable Team_Pitching_SO represents the number of strikeouts by pitchers during a season with mean 817.73, median 813, mode 0, and standard deviation 144.60. A frequency plot of the variable reveals al distribution that is heavily right skewed due to the scale of the plot. A review of the quantiles plot shows a departure from normality at the higher values of Team_Batting_H. A box plot of the values shows several outliers beyond the upper (Q3 + 3*(Q3-Q1)) bound. Team_Pitching_SO has strong positive relationships with Team_Batting_SO (0.42) and Team_Pitching_BB (0.49) and no strong negative relationships with other variables.



Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Data Preparation

The Data Preparation phase will consist of actions taken to prep the data for OLS regression. Specifically, we will impute the missing values of the variables with a measure of central tendency, adjust the data set for outliers that may exhibit excessive influence upon the variable (pull it one way or the other), and transform the variable using mathematical functions to adhere to the assumptions of OLS regression.

Reviewing the output of the earlier PROC MEANS procedure we can see that we have six variables that are missing values. Team_Batting_HBP represents the number of batters hit by a pitch and is missing an excessive 91% of its values. I've chosen to drop this variable from the data because any effort to replace the missing values with a measure of central tendency would unnaturally push the data towards my own preferences rather than what occurred. Additionally, Team_Batting_HBP has an R value of 0.07 with the response variable Target_Wins so I feel that its removal will have minimal impact upon our model. For the remaining five variables with missing values (Team_Batting_SO, Team_Baserun_SB, Team_Baserun_CS, Team_Fielding_DP-DP, and Team_Pitching_SO) I've reviewed their distribution plots and selected the measure of central tendency that best reflects the data. For the variables that follow a normal distribution (Team_Batting_SO, Team_Baserun_CS, Team_Fielding_DP, & Team_Pitching_SO) I've chosen to impute the missing values with the mean value for that variable. For the remaining variable, Team_Baserun_SB, that is positively-skewed (right skewed) I've chosen to replace the missing values with the variable's mode. A separate flag variable was created for each regressor that required imputation to indicate that a missing value was replaced. A flag of one indicates a replacement value; a flag of zero indicates a preexisting value.

The MEANS Procedure

Variable	Label	N	N Miss
TARGET_WINS		2276	0
TEAM_BATTING_H	Base Hits by batters	2276	0
TEAM_BATTING_2B	Doubles by batters	2276	0
TEAM_BATTING_3B	Triples by batters	2276	0
TEAM_BATTING_HR	Homeruns by batters	2276	0
TEAM_BATTING_BB	Walks by batters	2276	0
TEAM_BATTING_HBP	Batters hit by pitch	191	2085
TEAM_BATTING_SO	Strikeouts by batters	2174	102
TEAM_BASERUN_SB	Stolen bases	2145	131
TEAM_BASERUN_CS	Caught stealing	1504	772
TEAM_FIELDING_E	Errors	2276	0
TEAM_FIELDING_DP	Double Plays	1990	286
TEAM_PITCHING_BB	Walks allowed	2276	0
TEAM_PITCHING_H	Hits allowed	2276	0
TEAM_PITCHING_HR	Homeruns allowed	2276	0
TEAM_PITCHING_SO	Strikeouts by pitchers	2174	102

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	TARGET_WINS
TEAM_BATTING_HBP	0.07350
Batters hit by pitch	0.3122
	191

Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

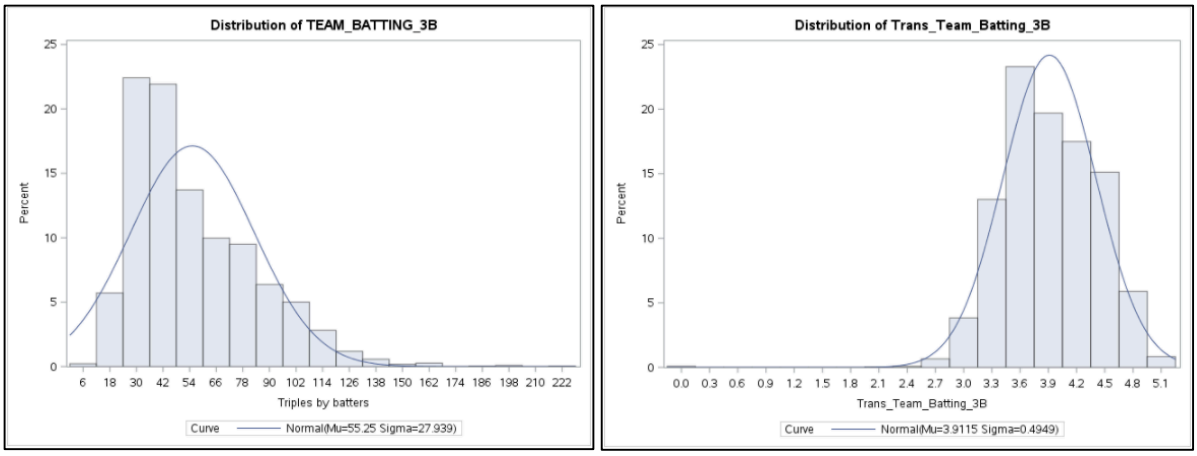
Box plots of the data have identified 12 variables with outliers beyond the beyond the Lower (Q1 - 3*(Q3-Q1)) and Upper (Q3 + 3*(Q3-Q1)) bounds. The concern with outliers in the data is that extreme values will pull the model towards that value, minimizing the impact of the rest of the variable's data upon the model. While it may be convenient to remove the extreme outliers, I want to take care to not remove data that reflects actual phenomena that occurred within the data set though it may degrade the accuracy of the model. It appears that the variables Team_Batting_H, Team_Batting_2B, Team_Batting_3B, Team_Batting_BB, Team_Baserun_SB, Team_Baserun_CS, Team_Fielding_E, Team_Fielding_DP, Team_Pitching_BB, Team_Pitching_H, Team_Pitching_HR, and Team_Pitching_SO have significant outliers. After doing several iterations where I capped variables at the lower (Q1 - 3*(Q3-Q1)) and upper (Q3 + 3*(Q3-Q1)) only Team_Pitching_H increased the Adjusted R-Square value for the model. Team_Pitching_H upper values were capped at 2475 and lower values were capped at 627.

Data transformations were focused on altering variables so they would conform to the normality assumptions necessary for OLS regression. We altered the variables Team_Batting_3B, Team_Baserun_SB, Team_Fielding_E, Team_Pitching_BB, Team_Pitching_H and Team_Pitching_SO through systematic trial and error until we had a distribution/frequency plot that more closely aligned with a normal distribution and increased the Adjusted R^2 value of our transformation test model.

Summary of Variable Transformations

Team_Batting_3B (Before/After Transformation)

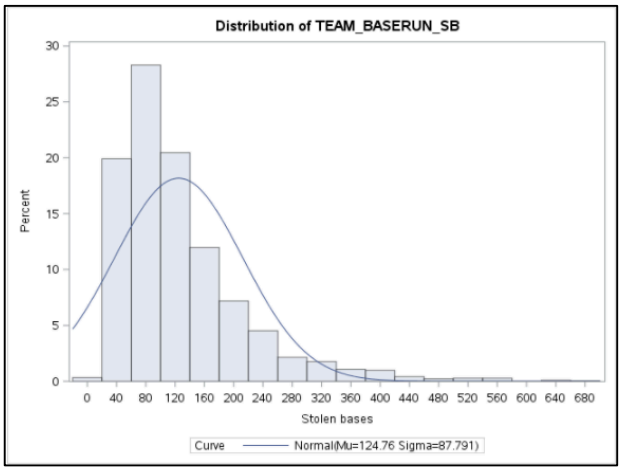
- Performed logarithmic transformation of *Team_Batting_3B* to correct for skewness.

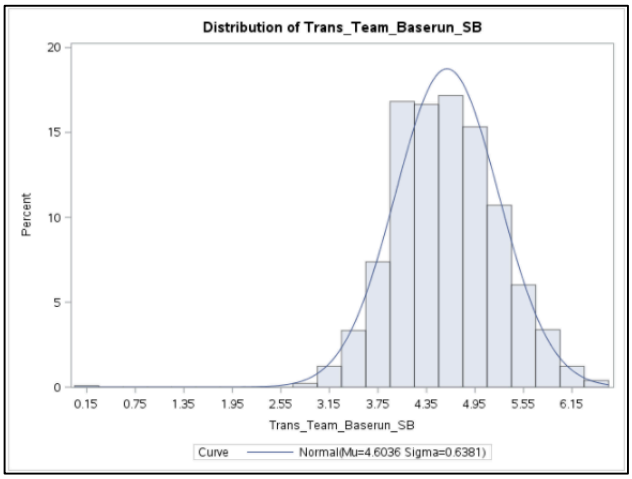


Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Baserun_SB (Before/After Transformation)

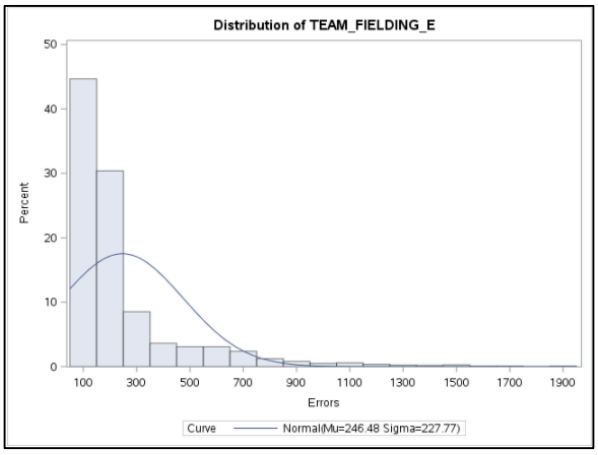
- Performed logarithmic transformation of *Team_Baserun_SB* to correct for skewness.

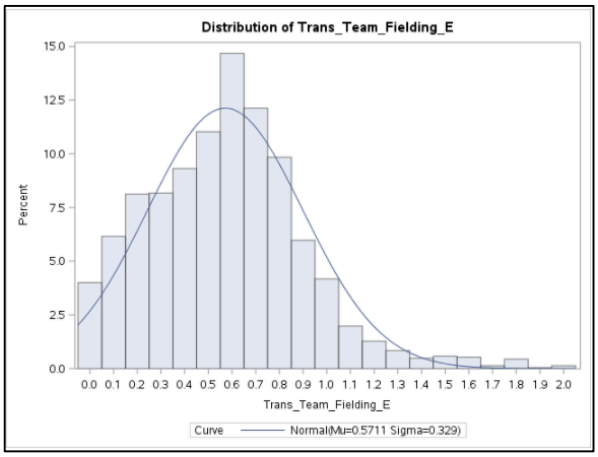




Team_Fielding_E (Before/After Transformation)

- Performed logarithmic transformation of *Team_Fielding_E* to correct for skewness.
- Divided by the mean to break the peak value into sub-values (standardization).
- Took the absolute value of the transformation results to correct for negative values.

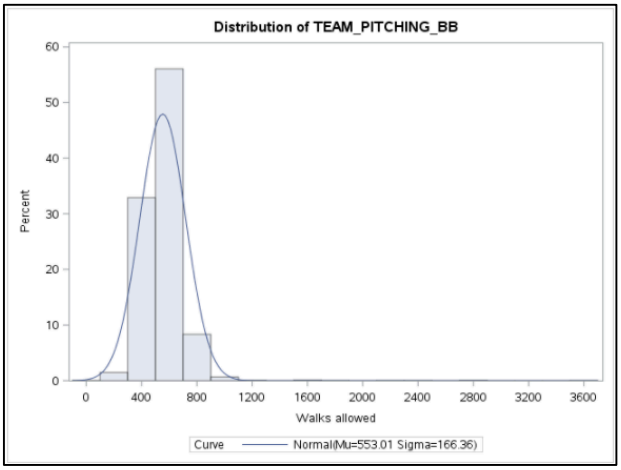


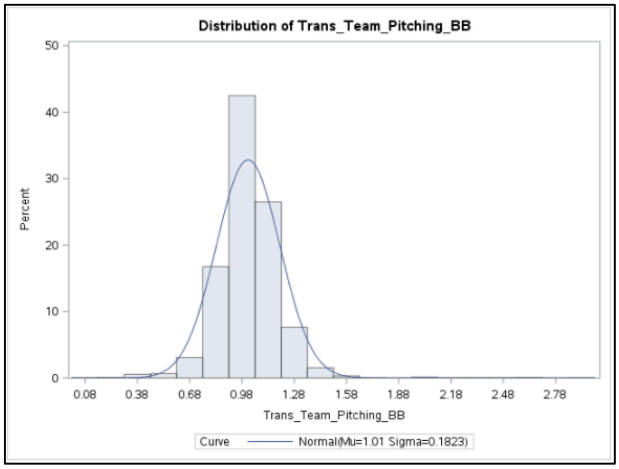


Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Pitching_BB (Before/After Transformation)

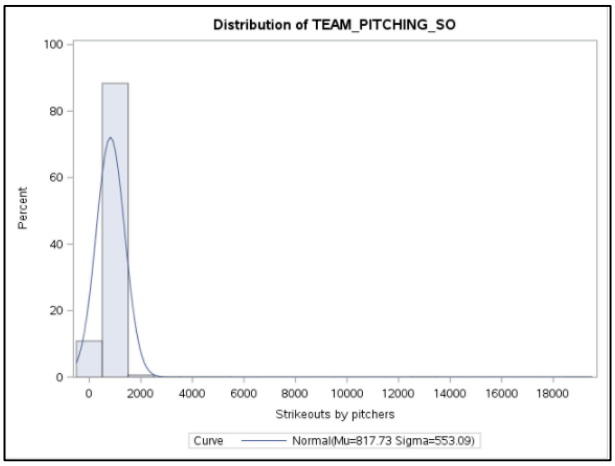
- Performed logarithmic transformation of *Team_Pitching_BB* to correct for skewness.
- Divided by the mean to break the peak value into sub-values (standardization).

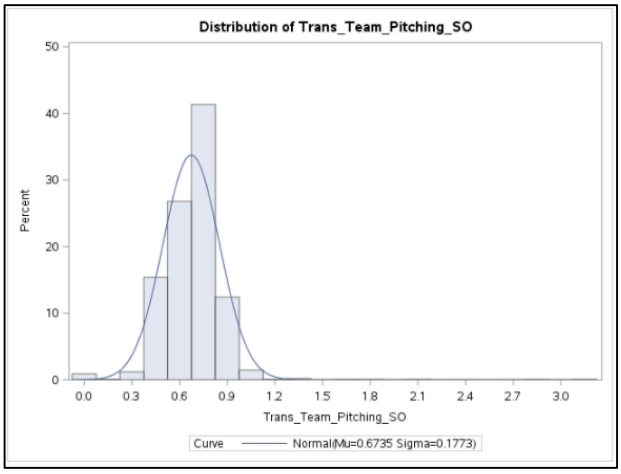




Team_Pitching_SO *(Before/After Transformation)*

- Performed logarithmic transformation of *Team_Pitching_SO* to correct for skewness.
- Divided by the mean to break the peak value into sub-values (standardization).

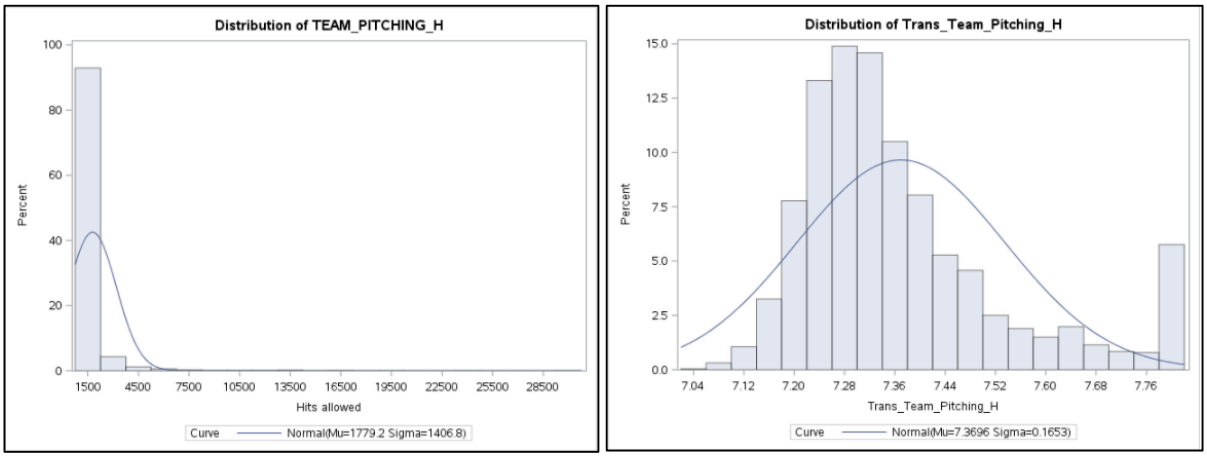




Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Team_Pitching_H *(Before/After Transformation)*

- Performed logarithmic transformation of *Team_Pitching_H* to correct for skewness.



Model Development & Selection

In the process of our analysis we built seven models for comparison: one base model with the provided data set, a model with imputed values and flag value, a model with imputed & transformed values, a rate based model (all variables divided by 162), and three models using forward, backward, and stepwise automated variable selection techniques (Adjusted R-Square Criteria, AIC, BIC). The six later models went through an iterative trail & error process to find which variations of variables, based off imputations, outlier manipulation, and transformations, to produce the best model. After performing the functional transformations, I also experimented with converting all my variables to rates by dividing each metric by 162 (the number of games in the season). While this effort did seem to minimize the magnitude of the peaks, the highest Adjusted R-Square value I was able to achieve was 0.3442, which meant that the rate transformation model (combined with the previous transformations) encompassed ~ 34% of the variability in the response Target_Wins. The forward and backward automated techniques produced models with superior Adjusted R-Square values, of 0.4216 and 0.4212 respectively, but had excessively high variance inflation factors (VIF) > 10 that warned of multicollinearity (inflation). The stepwise model was selected based off adhering to most of OLS regression assumptions, highest Adjusted R-squared values, VIFs < 10, lowest AIC & BIC, lowest root mean square error, comparative number of terms, and results of the scoring data set.

Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

Our selected stepwise model for computing the predicted amount of Target_Wins for a 162 game-season takes the equation form of:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \mathcal{E}$$

Where:

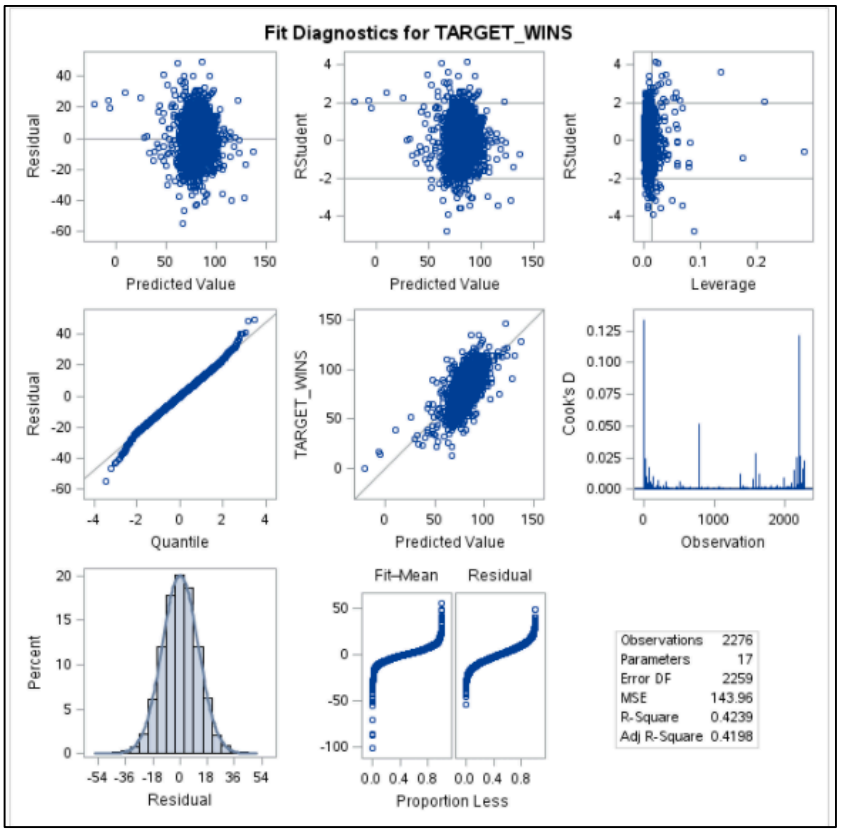
In Model	In Data	Beta	Value
Y	Predicted Target Wins	β_0	165.02689
X1	Team_Batting_H	β_1	0.05473
X2	Team_Batting_2B	β_3	-0.03031
X3	Team_Batting_3B	β_4	0.07642
X4	Team_Batting_BB	β_5	0.02652
X5	Team_Pitching_H	β_6	0.00154
X6	Team_Pitching_HR	β_7	0.05785
X7	Team_Fielding_E	β_8	-0.05941
X8	Imp_Team_Batting_SO	β_9	-0.01650
X9	Imp_Team_Baserun_SB	β_{10}	0.06180
X10	F_Team_Baserun_SB	β_{11}	39.42668
X11	Imp_Team_Fielding_DP	β_{12}	-0.10891
X13	F_Team_Fielding_DP	β_{13}	2.92487
X14	F_Team_Pitching_SO	β_{14}	9.08696
X15	Trans_Team_Baserun_SB	β_{15}	-1.60493
X16	Trans_Team_Fielding_E	β_{16}	6.84819
X17	Trans_Team_Pitching_H	β_{17}	-20.09114
E	Error Term		

With Goodness-of-Fit Diagnostics:

Metric	Value
Root MSE	11.99
R-Square	0.4239
Adjusted R-Square	0.4198
F-Value	103.9

Assignment #1
Daren Purnell
Predict_411 Section 60
KAGGLE NAME: DJPURNELL

The Adjusted R-square value for this model is at 0.4198, which means about 42% of the variability in the response variable Target_Wins is represented by the model. Using the automatically generated ODS output to assess the goodness-of-fit of this model, I can observe that the residuals, the error between the estimated and actual value of the response variable, seem to be small. The residuals seem to spread out evenly with no significant structure (homoscedastic) and follow a normal distribution as seen in the QQ and distribution plot. The Cook's D plot shows three main outliers that warrant further analysis to better understand their effects upon the model.



There are couple of things in this model that require explanation; while certainly not intuitive, this series of imputations and transformations produced the best model.

- Team_Batting_2B has negative coefficient which is counter-intuitive for a variable thought to increase the number of Target_Wins.
- The coefficients for the imputed flags for Team_Baserun_SB and Team_Pitching_SO (indicated by the prefix F_) are much larger than the other variables. It's not clear why a flag for the replacement of a variable would have such a large coefficient.

- Team_Fielding_Errors shows up twice in the selected equation with opposite signs. When I was trying different iterations of transformations, I accidently kept the original of Team_Fielding_Errors and it improved the quality of the model.
- Team_Baserun_SB shows up twice in the selected equation with opposite signs. When I was trying different iterations of transformations, I accidently kept the original of Team_Baserun_SB and it improved the quality of the model.
- Team_Pitching_H shows up twice in the selected equation with opposite signs. When I was trying different iterations of transformations, I accidently kept the original of Team_Pitching_H and it improved the quality of the model.
- The imputed version of Team_Fielding_DP (double plays) has a negative coefficient despite double plays helping the team win.

Conclusion

Our efforts to use OLS Regression to predict the number of wins for baseball team in a 162-game season cumulated with the selection of model that represents ~ 42% of the variability in the response variable Target_Wins. The analysis focused on understanding the underlying structure and nuances of each variable in the data set. Admittedly, there are still a few things that we don't understand such as the strong correlation between distinct, seemingly unrelated variables. If we had more time, we would perform factor analysis to better understand the structure of the data. For variables that were found to be missing an excessive amount of their records, imputation (data replacement) was conducted using that variables measures of central tendency (mean, mode, median) as replacement values. Later, we focused on manipulating the variables to adhere to a normal distribution to fit the assumptions of OLS regression. I also must concede, that while exploratory data analysis did reveal information about baseball, the selected model does not truly represent the mechanism of winning baseball games. We also experimented with using a rate based model that divided each of the variables by 162 to minimize the magnitude of the peaks and adhere to a normal distribution. In several instances, we used multiple variations of a variable in one model to produce the best outcomes. This decision was the price that was paid to adhere to the independence, linearity, multicollinearity, and error assumptions required to use OLS regression. Our selected model represents a black box that produces an estimate of baseball wins when provided with the proper inputs.

"Linear Regression is simple and powerful. If you can get away with using it as a predictive model, then do so. Even if it means taking a few liberties."

Donald Wedding