Mlb Salary Predictions

Joey Matte

### Examine Data

We are going to analyze Mlb player data. Specifically, we will examine players average annual salary as it relates to their team, position, and the number of years they have been playing. We will start by taking a look at the data.

# load and look at data  
data = read.csv('salaries \_mlb.csv')  
print(head(data))

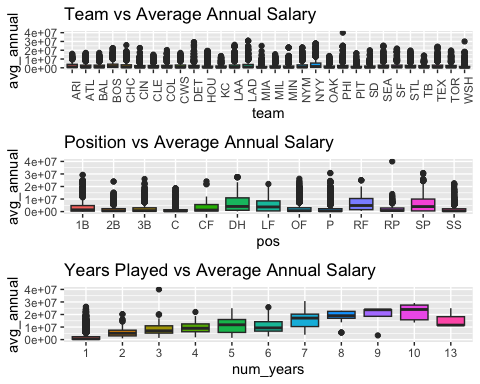
## salary name total\_value pos num\_years years avg\_annual team  
## 1 3800000 Darryl Strawberry 3800000 OF 1 1991 3800000 LAD  
## 2 3750000 Kevin Mitchell 3750000 OF 1 1991 3750000 SF  
## 3 3750000 Will Clark 3750000 1B 1 1991 3750000 SF  
## 4 3625000 Mark Davis 3625000 P 1 1991 3625000 KC  
## 5 3600000 Eric Davis 3600000 OF 1 1991 3600000 CIN  
## 6 3562500 Willie McGee 3562500 OF 1 1991 3562500 SF

# keep columns we are interested in  
data = data[c('avg\_annual','team','pos','num\_years')]  
# convert to factors  
facs = c('team','pos','num\_years')  
data[,facs] = lapply(data[,facs],as.factor)

Lets examine the data graphically.

library(ggplot2)

library(gridExtra)  
t = ggplot(data,aes(team,avg\_annual))+geom\_boxplot(aes(fill=team),show.legend=FALSE) + labs(title='Team vs Average Annual Salary') + theme(axis.text.x = element\_text(angle=90))  
p = ggplot(data,aes(pos,avg\_annual))+geom\_boxplot(aes(fill=pos),show.legend=FALSE) + labs(title='Position vs Average Annual Salary')  
yp = ggplot(data,aes(num\_years,avg\_annual))+geom\_boxplot(aes(fill=num\_years),show.legend=FALSE) + labs(title='Years Played vs Average Annual Salary')  
grid.arrange(t, p, yp)



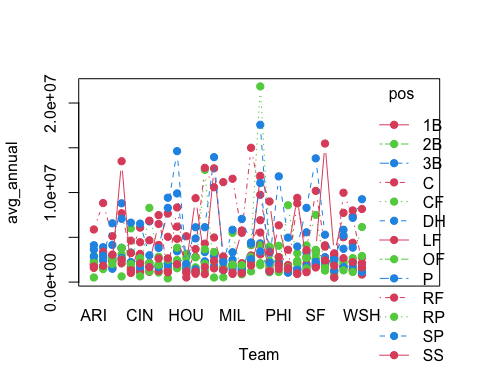
We can see that there are lots of outliers in our data, which we will need to keep in mind. This may be due to teams paying star players much more than others. Since there are quite a bit of them, it doesn’t make sense to drop them from our data set.

### 

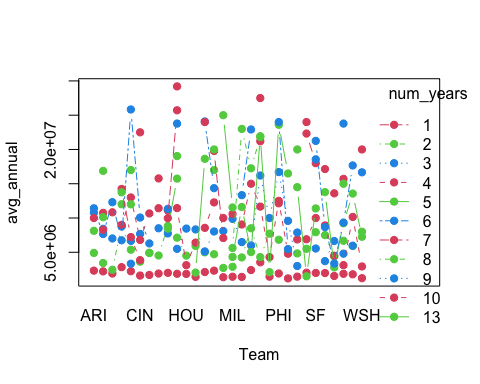
### Interaction Plots

Visualize the interactions between factors.

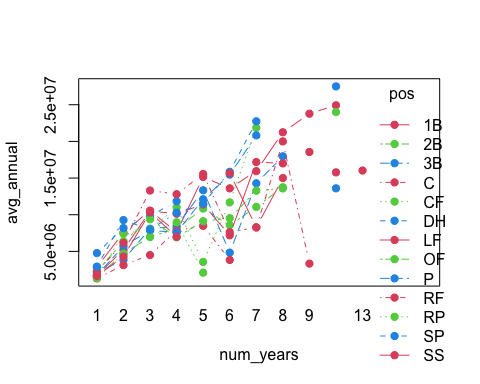
with(data, interaction.plot(team, pos, avg\_annual, type = 'b', col = 10:12, pch = 19, fixed = T, xlab = 'Team', ylab = 'avg\_annual'))



with(data, interaction.plot(team, num\_years, avg\_annual, type = 'b', col = 10:12, pch = 19, fixed = T, xlab = 'Team', ylab = 'avg\_annual'))



with(data, interaction.plot(num\_years, pos, avg\_annual, type = 'b', col = 10:12, pch = 19, fixed = T, xlab = 'num\_years', ylab = 'avg\_annual'))



From our plots, we can see that none of the factors are parallel to another, which tells us there is likely a significant interaction between all of them that affects the average annual salary.

### ANOVA

Next, we will run an ANOVA model including all interactions between factors in order to see which factors and interactions are important in predicting a players annual average salary.

aov\_model = aov(avg\_annual~.^3,data)  
summary(aov\_model)

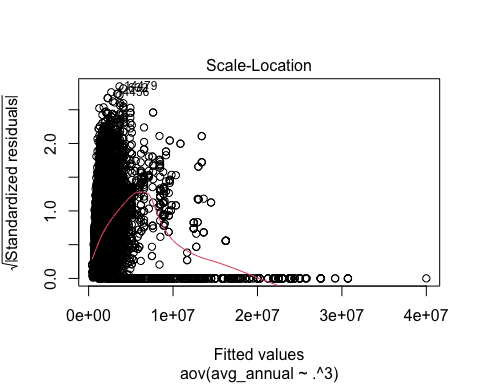
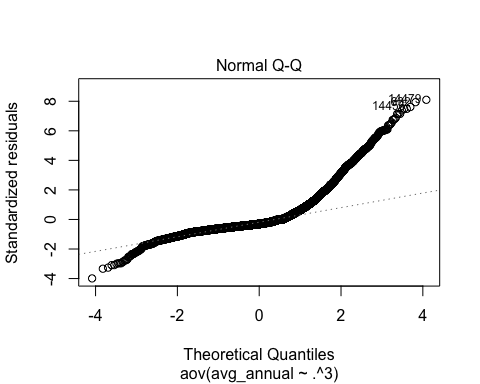
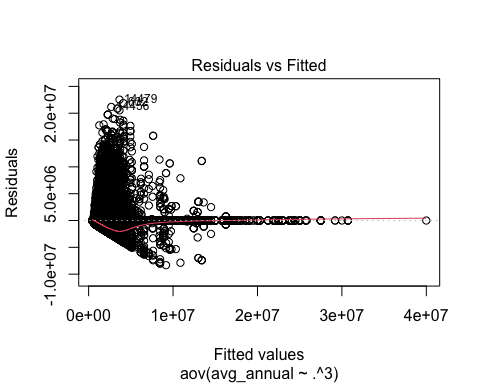
## Df Sum Sq Mean Sq F value Pr(>F)   
## team 29 8.562e+15 2.952e+14 37.883 <2e-16 \*\*\*  
## pos 12 1.963e+16 1.636e+15 209.945 <2e-16 \*\*\*  
## num\_years 10 6.508e+16 6.508e+15 835.098 <2e-16 \*\*\*  
## team:pos 333 8.668e+15 2.603e+13 3.340 <2e-16 \*\*\*  
## team:num\_years 160 1.295e+16 8.095e+13 10.387 <2e-16 \*\*\*  
## pos:num\_years 62 1.703e+15 2.747e+13 3.524 <2e-16 \*\*\*  
## team:pos:num\_years 215 4.370e+15 2.033e+13 2.608 <2e-16 \*\*\*  
## Residuals 22014 1.716e+17 7.793e+12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From our model, all factors and their interactions are significant, meaning a change in each one has a significant effect on the average annual salary, so all should be kept in the model.

### Assumptions

1. Independence: Residuals are roughly evenly spread about the line, showing there is no non-linear relationship.
2. Homoscedasticity: Residuals are roughly evenly spread, showing homoscedasticity.
3. Normality of residuals: The relationship between theoretical and standardized looks to be mostly linear, with skews at the highest and lowest values, however, this is likely due to the outliers, so we will assume that our assumption holds.

plot(aov\_model)

A graph of a number of black and white lines

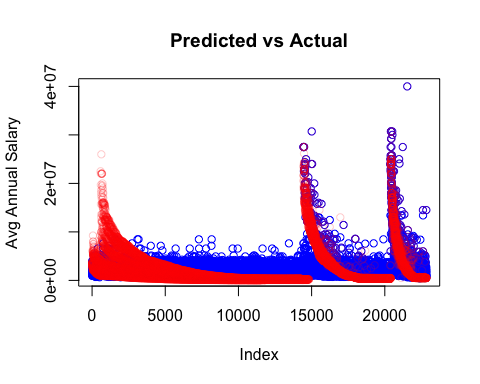
Description automatically generated with medium confidence

### Linear Model and Predictions

# build linear model  
lm\_model = lm(avg\_annual~.^3,data)  
# use model to predict avg annual salary  
predicted = predict(lm\_model, data[c('team','pos','num\_years')])

## Warning in predict.lm(lm\_model, data[c("team", "pos", "num\_years")]):  
## prediction from a rank-deficient fit may be misleading

# graph predicted vs actual avg annual salary  
plot(predicted,col='blue',ylab = 'Avg Annual Salary', main = 'Predicted vs Actual')  
points(data['avg\_annual'],col=rgb(1,0,0,alpha=0.2))



### Conclusions

In analyzing the MLB salary data, we have found that the variables team, position, number of years played, and all of the second and third order interactions between these variables all have a significant effect on the average annual salary of a player. Using this information, and after looking at our model assumptions, we created a linear model to try and predict a players average annual salary based on their team, position, number of years played, and all interactions. The amount of outliers in the data makes this difficult, and our model only explains 42% of the variability in the data largely due to this. Even so, we were able to make a model that predicts the average annual salary decently well, as can be seen in the graph titled ‘Predicted vs Actual’.