Hypothesis 1

On average, minutes played is positively correlated to a player's three point score, two point score, free throw score, offensive rebound score, defensive rebound score, block score, and steal score.

Reasoning

The more minutes you play the more chances you have to either score a point or prevent the other team from scoring a point.

Data

In order to run this regression smoothly, there were many steps I had to take to clean my data frame. To start off I created a subset data from my original that contained all basketball statistics between the year 1986 and the year 2017. After this I created yet another subset including only the variables described in the hypothesis. Then I created a function that would drop all outliers using quantiles. Once this was completed I decided to log transform every variable in order to compress the data and make my model linear. This arose a problem within the data as some of these log transformed variables contained inf and na values. I modified the dataset to drop all rows with either a na or inf value inorder to get my working data frame.

Main regression

My dependent variable for this regression analysis will be minutes played as I wish to see the effects it has on three point scores, two point scores, free throw scores, offensive rebounds, defensive rebounds blocks, and steals. These will be considered my independent variables.

Regression

```
Call:
lm(formula = hp1_working_df$ln_mp ~ hp1_working_df$ln_blk + hp1_working_df$ln_drb +
    hp1_working_df$ln_stl + hp1_working_df$ln_ft + hp1_working_df$ln_orb +
    hp1_working_df$ln_3p + hp1_working_df$ln_2p)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87313 -0.10922 -0.00076  0.11182  0.99673

Coefficients:
                       Estimate Std. Error t value             Pr(>|t|)
(Intercept)            3.082050   0.011541 267.044 < 0.0000000000000002 ***
hp1_working_df$ln_blk -0.030931   0.002594 -11.926 < 0.0000000000000002 ***
hp1_working_df$ln_drb  0.361515   0.005373  67.285 < 0.0000000000000002 ***
hp1_working_df$ln_stl  0.249645   0.003840  65.020 < 0.0000000000000002 ***
hp1_working_df$ln_ft   0.011161   0.003973   2.809              0.00498 **
hp1_working_df$ln_orb -0.025363   0.003796  -6.682      0.0000000000249 ***
hp1_working_df$ln_3p   0.062698   0.001486  42.187 < 0.0000000000000002 ***
hp1_working_df$ln_2p   0.264683   0.005238  50.529 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1784 on 9777 degrees of freedom
Multiple R-squared:  0.9546,    Adjusted R-squared:  0.9545
F-statistic: 2.934e+04 on 7 and 9777 DF,  p-value: < 0.00000000000000022
```

Regression Analysis

After close analysis of this regression I am able to confirm that all of the variables I have chosen are positively correlated except for blocking stats and offensive rebounds. Since all of these variables are log transformed, I am able to tell that a ten percent increase in minutes played results in a 0.3% decrease in a players blocking stats. Even though this is a negative correlation, It has a very small correlation which tells me that even though this is statistically significant (low p value) minutes played does not affect a players blocking stats in a largely impactful way. The other negative coefficient is similar in that minutes played has a very small impact on it. For a ten percent increase in minutes played, a player's offensive rebounds decrease by 0.2%. The independent variables with the strongest correlation to minutes played are defensive rebounds, steals, and two point shots. Respectively, a ten percent increase in minutes played yields a 3.6% increase in defensive rebounds, a 2.5% increase in steals and a 2.6% increase in 2 pointers.

My insight as to why these are more highly correlated is that these aspects of a basketball game are more likely to happen than the negatively correlated aspects. For example, a defensive rebound is much more likely to occur vs an offensive rebound since when a team is on defense, they usually have more players under the basket than the opposing team. Similarly, blocking a shot in the NBA is considered very hard to do which makes sense as to why on average, the more a person plays the less blocked shots he has. The R squared value sits high at 0.9546 meaning that my data is highly correlated. Also with a very low p value my data is considered statistically significant.

With all that being said I am forced to reject my null hypothesis since there are negatively correlated variables however, my logic in creating that hypothesis was correct since the majority

of variables are highly correlated and the variables that are negatively correlated are so in a very minor manner.