Overview:

The purpose of the is research is to create a model that will allow us to predict players scoring production at 5v5 into the future and leverage that against their market value to find play driving players at superb value. This will mostly impact middle of the lineup players as top-end talent is signed for long periods of time and are outliers in the data. Additionally, these predictions can help internally justify an overpayment against market value.

Data & Contents:

- Data from HockeyReference.com queried via HockeyR package
- Filtered to forwards only, 5v5 stats, > 10 GP in each season, and < 0.80 5v5 points per game (outlier remove). Players must have two seasons of data to be included in model.

Model & Model Eval:

- Stepwise regression to model the first season into the future.
- Training: 17-18, 19-20 to 22-23 season

```
Call:
lm(formula = p_pg_target ~ p_pg_prev1 + p_pg_prev2 + mean_toi_minutes.x_prev1 +
    icf_per60_prev1 + age__prev1, data = training_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36814 -0.06703 -0.00192  0.06258  0.34756

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.0394775  0.0267724   1.475 0.140518
p_pg_prev1               0.2508241  0.0272125   9.217  < 2e-16 ***
p_pg_prev2               0.2262377  0.0247196   9.152  < 2e-16 ***
mean_toi_minutes.x_prev1 0.0160535  0.0015352  10.457  < 2e-16 ***
icf_per60_prev1          0.0041892  0.0011428   3.666 0.000254 ***
age__prev1              -0.0056350  0.0006434  -8.758  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1019 on 1691 degrees of freedom
Multiple R-squared:  0.5174,    Adjusted R-squared:  0.516
F-statistic: 362.6 on 5 and 1691 DF,  p-value: < 2.2e-16
```

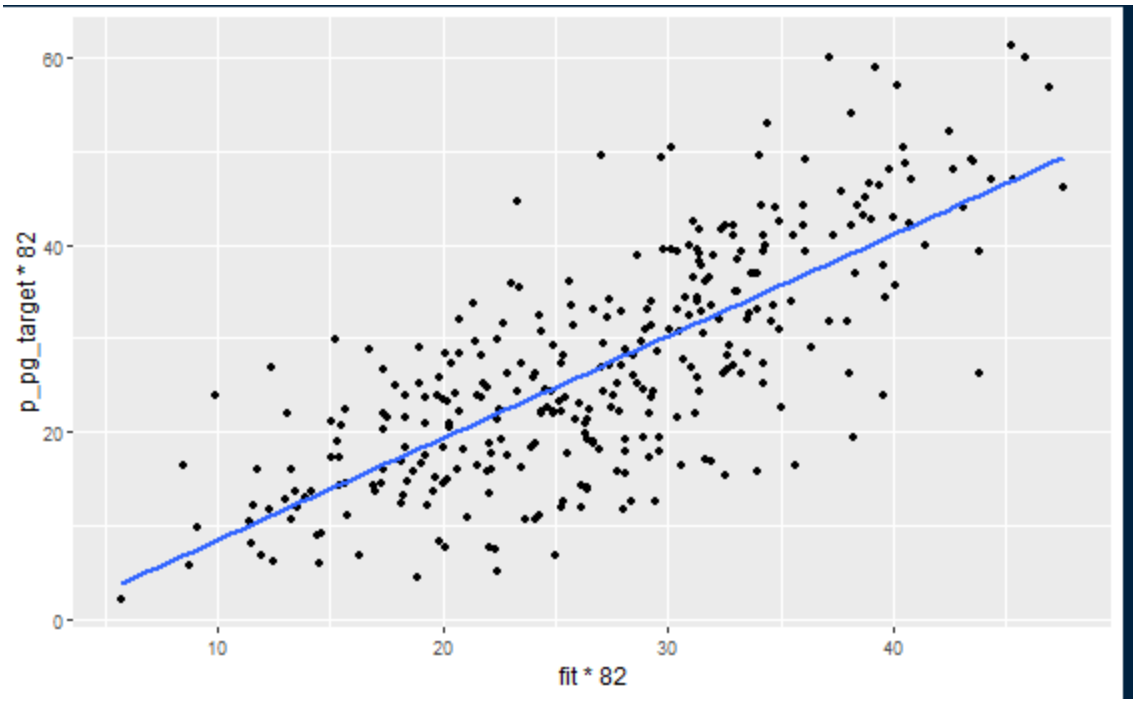
- Test: 18-19 season

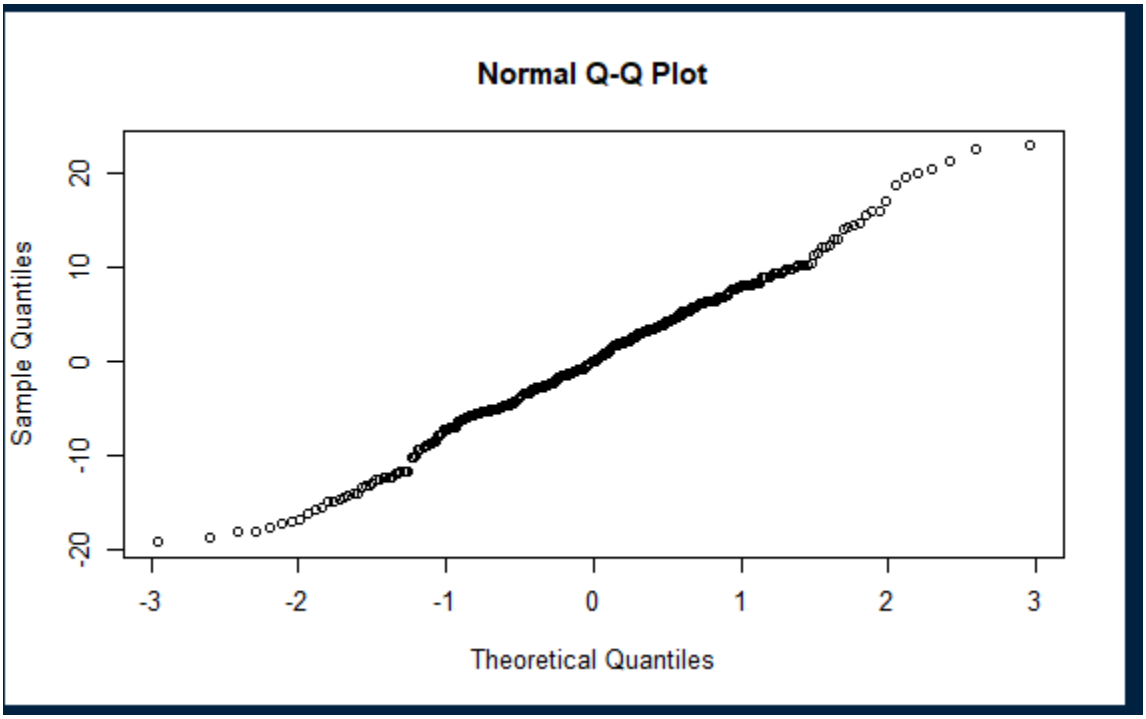
```
Call:
lm(formula = p_pg_target ~ p_pg_prev1 + p_pg_prev2 + mean_toi_minutes.x_prev1 +
    icf_per60_prev1 + age__prev1, data = testing_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.248839 -0.058127  0.001455  0.061814  0.267177

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.0159508  0.0604785  -0.264 0.792150
p_pg_prev1               0.3019991  0.0669136   4.513 9.04e-06 ***
p_pg_prev2               0.1960930  0.0572433   3.426 0.000695 ***
mean_toi_minutes.x_prev1 0.0195671  0.0035256   5.550 6.08e-08 ***
icf_per60_prev1          0.0005021  0.0026055   0.193 0.847303
age__prev1              -0.0044257  0.0014480  -3.056 0.002433 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09774 on 314 degrees of freedom
Multiple R-squared:  0.5682,    Adjusted R-squared:  0.5613
F-statistic: 82.64 on 5 and 314 DF,  p-value: < 2.2e-16

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06976 0.25239 0.33008 0.32981 0.39978 0.57999
```

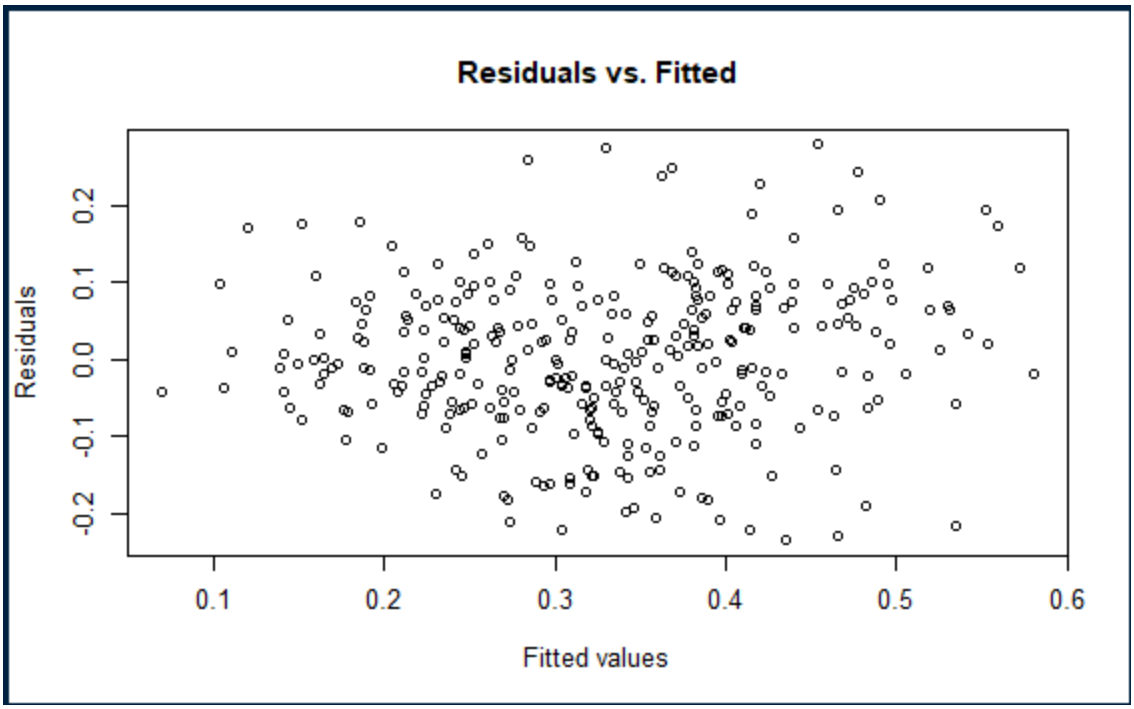
- Assumptions made on model features when predicting 2+ seasons into the future. Leveraged research on the age curve and point production to create age buckets of '26u', '27-29', '30+' and rank each player into a quartile based on their point per game production from the most recent season and used the average change from each model feature in each quartile to create the updated model feature. See code below:

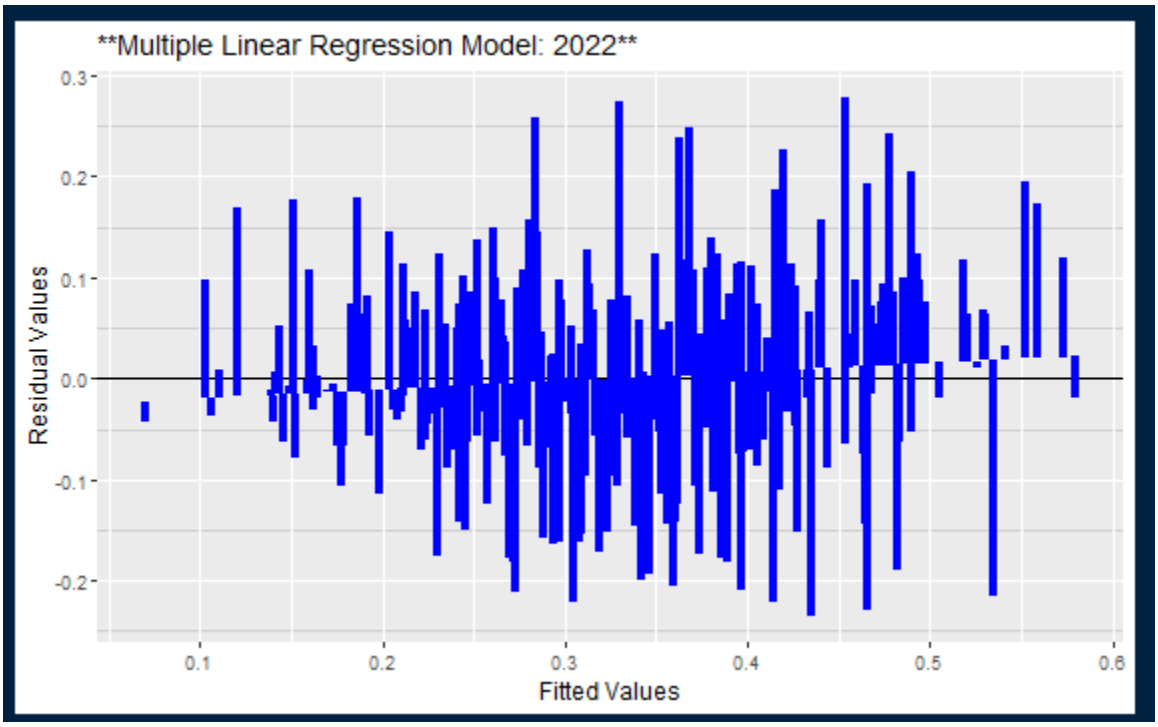
```
age_buckets_test <- two_szn_model_df %>%
  select(player_name, age__prev1, p_pg_prev1, p_pg_prev2, mean_toi_minutes.x_prev1, icf_per60_prev1
         ,mean_toi_minutes.x_prev2, icf_per60_prev2) %>%
  mutate(age_bucket=ifelse(age__prev1>=30,"30+",
                           ifelse(age__prev1>=27,"27-29",
                                  ifelse(age__prev1<=26,"26u",NA)))) %>%
  group_by(age_bucket) %>%
  mutate(quartile = ntile(p_pg_prev1, 4)) %>%
  ungroup() %>%
  group_by(age_bucket, quartile) %>%
        summarise(avg_pts_delta=mean(p_pg_prev1-p_pg_prev2),
           avg_toi_delta=mean(mean_toi_minutes.x_prev1-mean_toi_minutes.x_prev2),
           avg_icf_delta=mean(icf_per60_prev1-icf_per60_prev2)) %>%
  ungroup() %>% |
  mutate(id=paste0(age_bucket,"_",quartile))
```

- **Evaluations:**


p_pg_target * 82
fit * 82
0
20
40
60
10
20
30
40


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-20
-10
0
10
20
-3
-2
-1
0
1
2
3

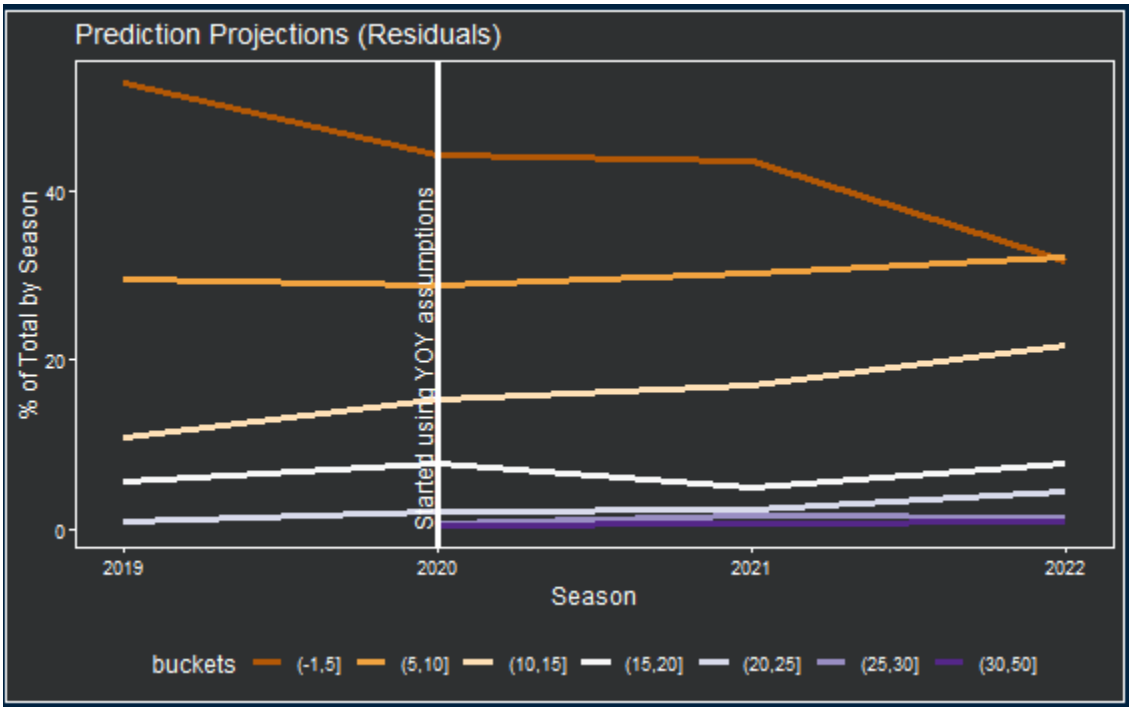


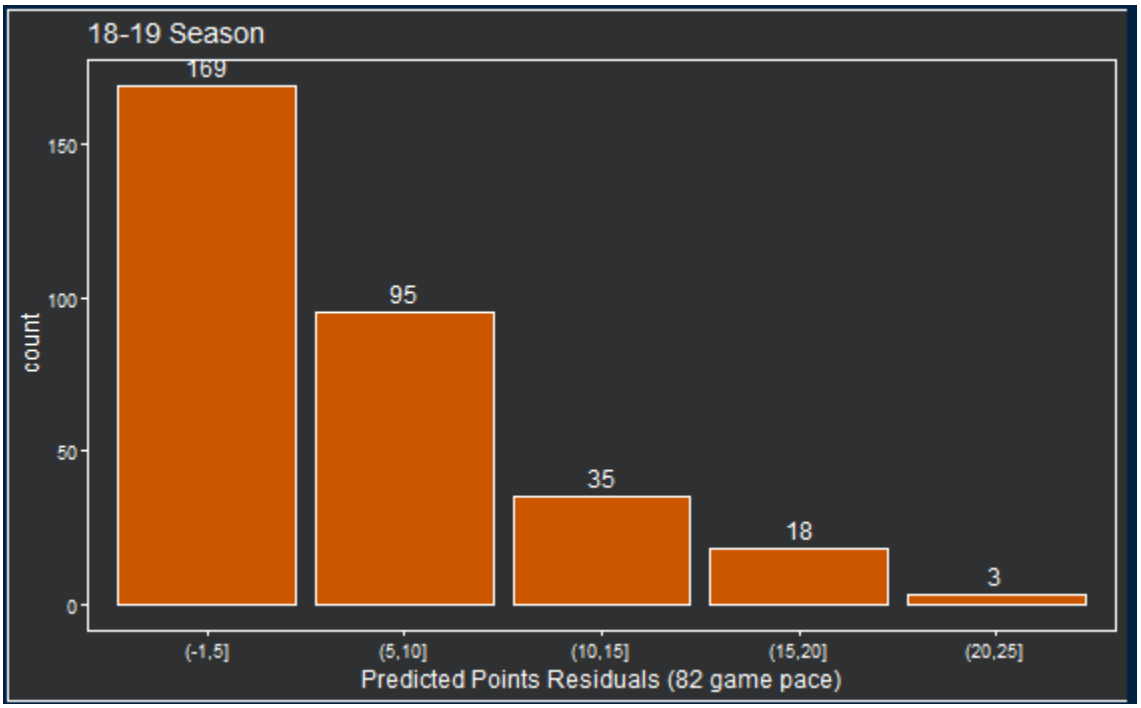


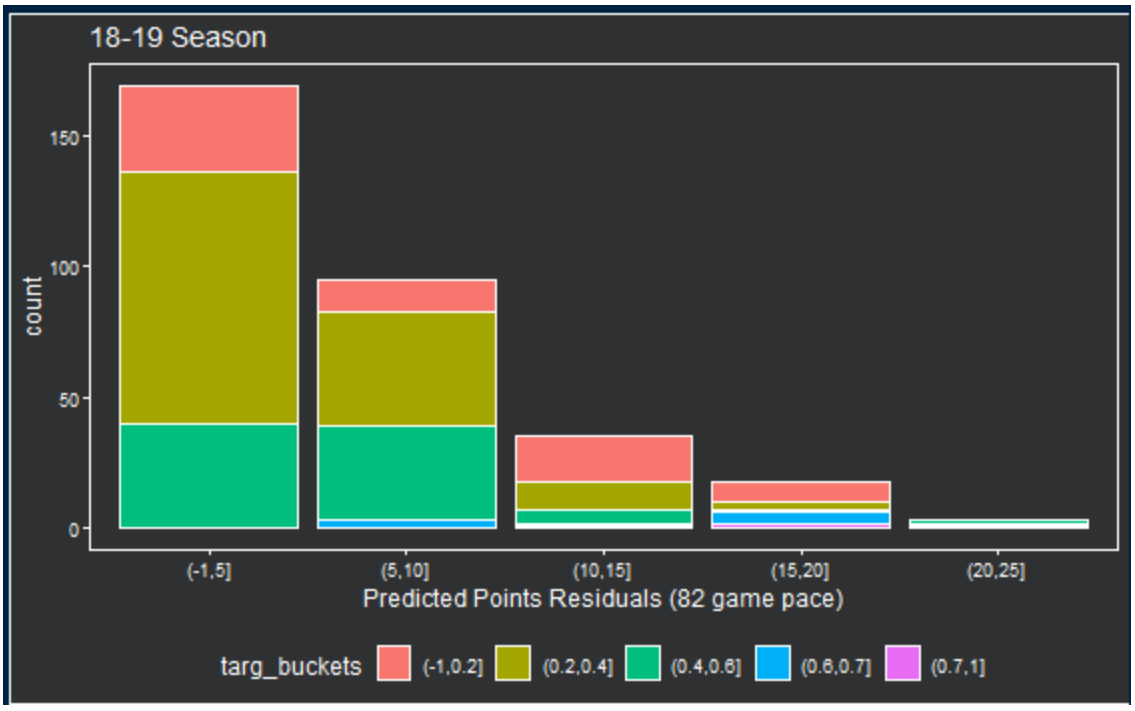
Outcome:

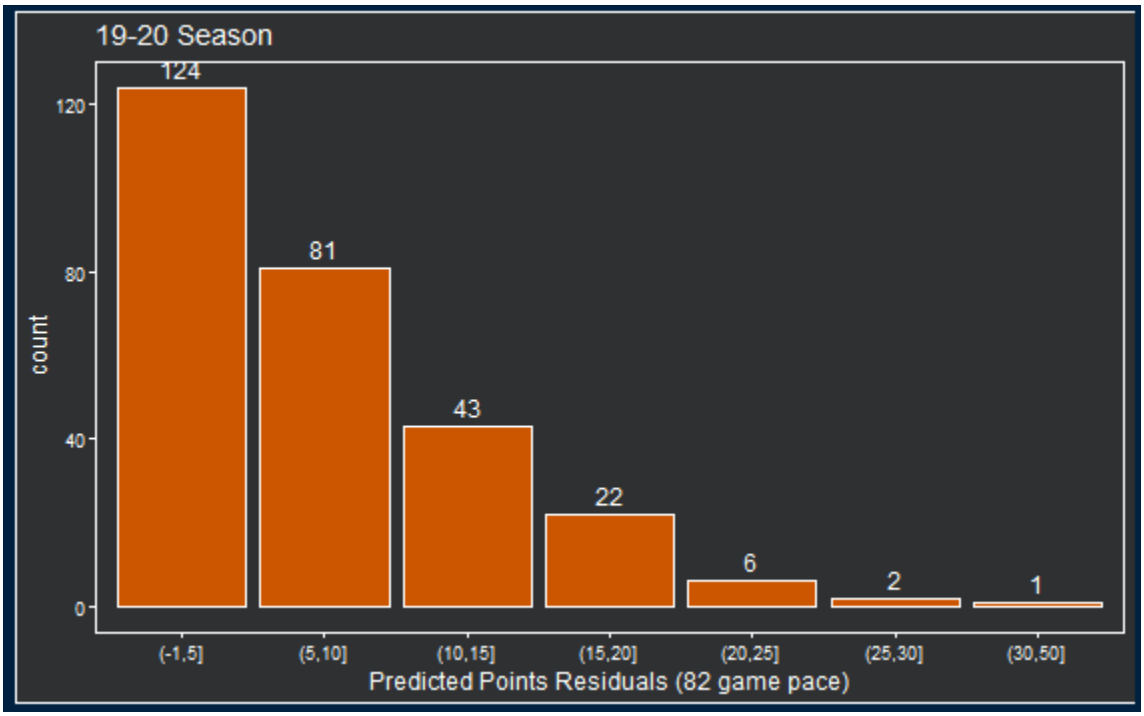
- Although the r^2 scores weren't spectacular and only account for ~50% of the variability the models did generalize well from train to test. And as evidenced by the residuals of each season the model does a pretty good job predicting both +1 season ahead using strictly regression techniques while +2 seasons and beyond predict well using both

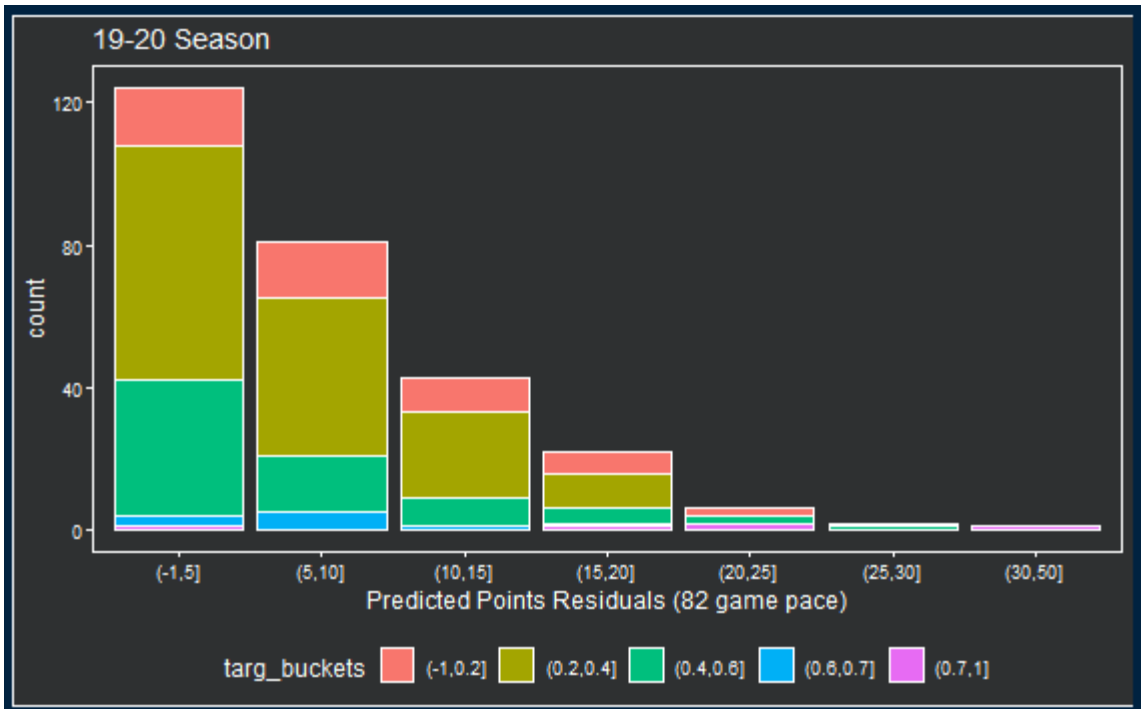
regression techniques and assumptions. Also as we hoped the model does a better job predicting middle of the lineup player than others.

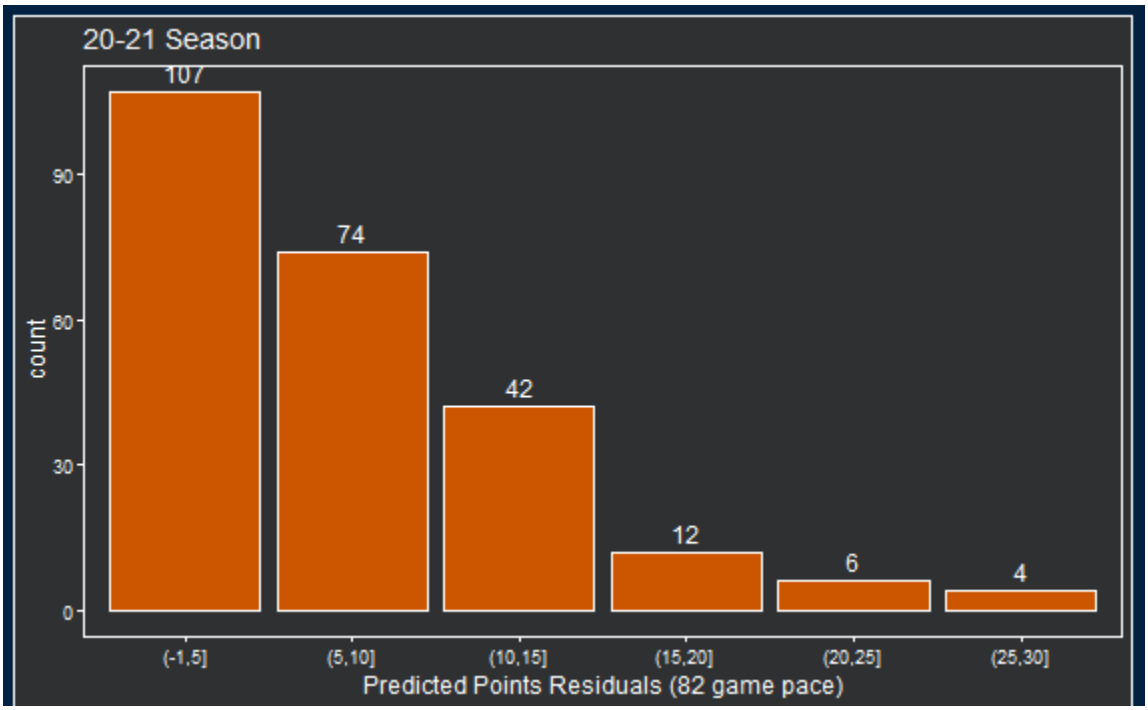


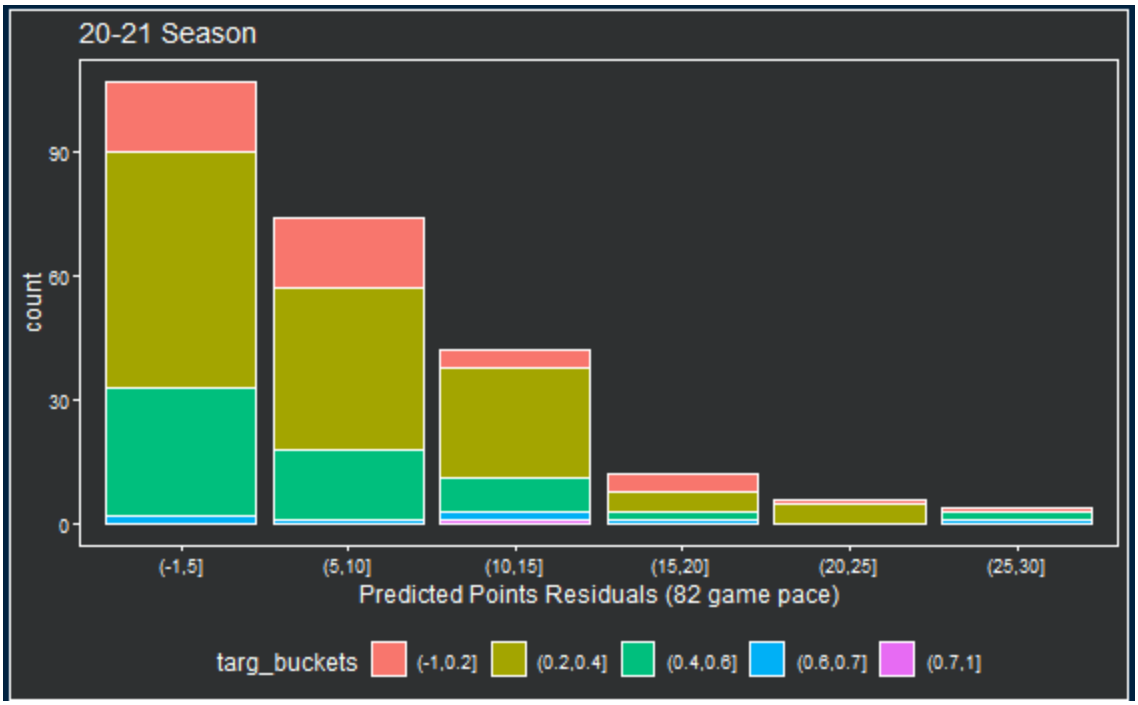


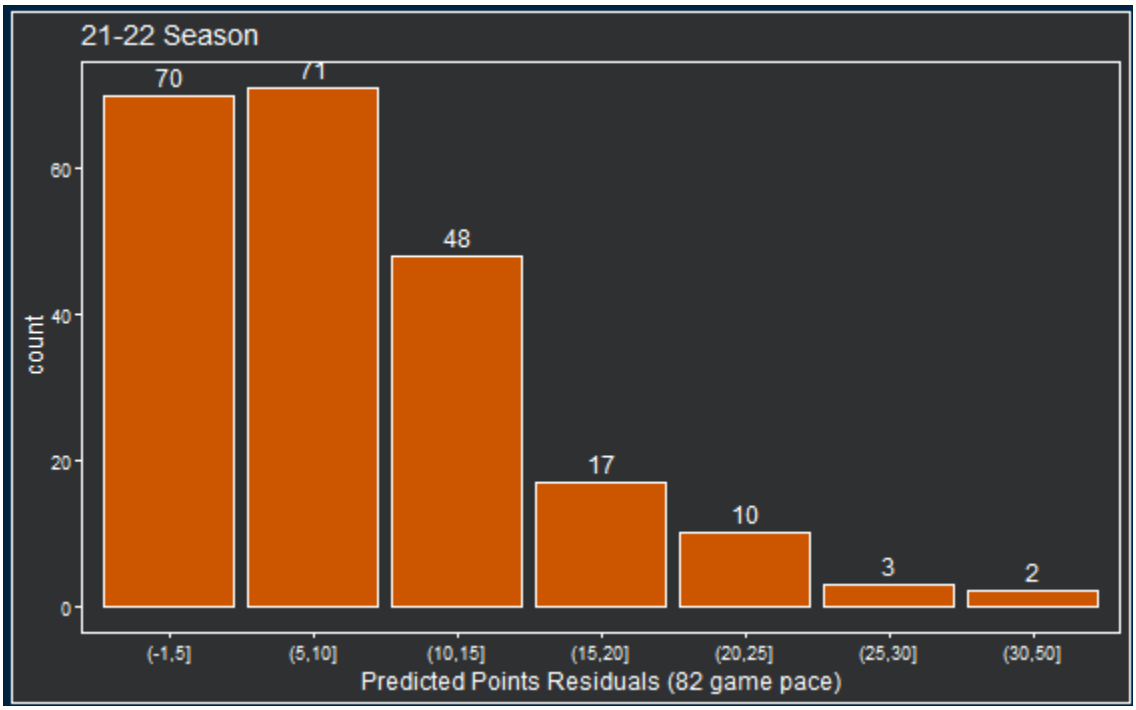

18-19 Season
count
150
100
50
0
(-1,5]
(5,10]
(10,15]
(15,20]
(20,25]
Predicted Points Residuals (82 game pace)
targ_buckets
(-1,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.7]
(0.7,1]

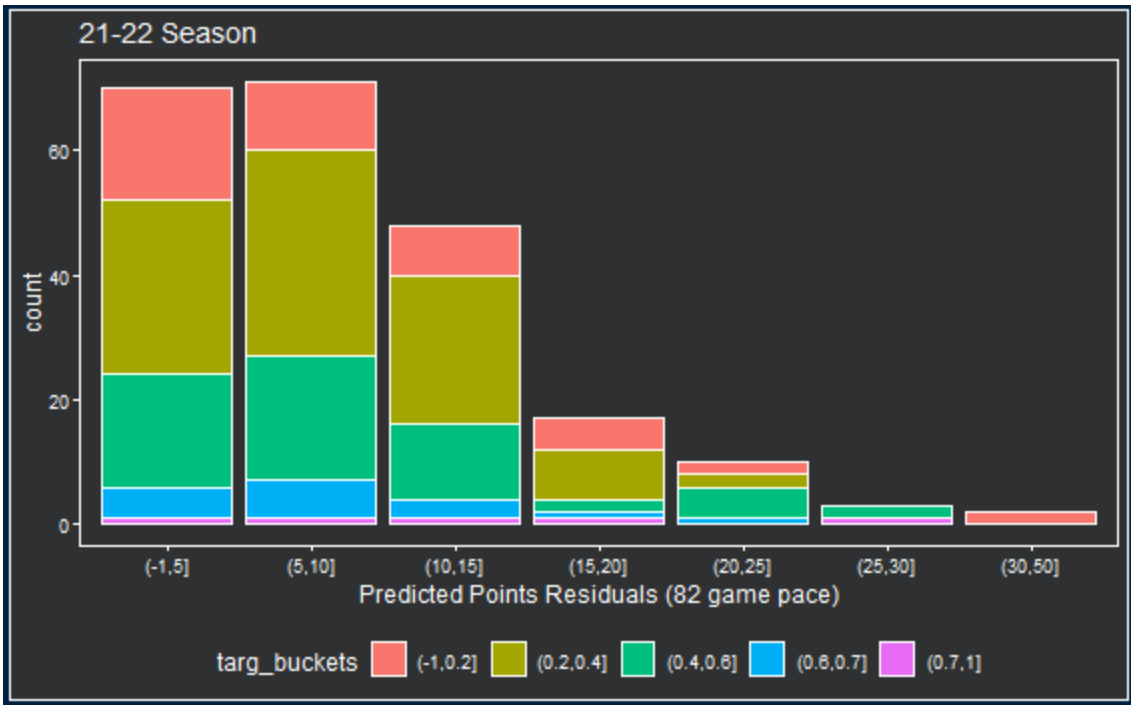

19-20 Season
count
120
80
40
0
124
81
43
22
6
2
1
(-1,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
(30,50]
Predicted Points Residuals (82 game pace)


19-20 Season
120
80
40
0
count
(-1,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
(30,50]
Predicted Points Residuals (82 game pace)
targ_buckets
(-1,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.7]
(0.7,1]


20-21 Season
107
74
42
12
6
4
90
60
30
0
count
(-1,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
Predicted Points Residuals (82 game pace)


20-21 Season
count
90
60
30
0
(-1,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
Predicted Points Residuals (82 game pace)
targ_buckets
(-1,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.7]
(0.7,1]


21-22 Season
count
60
40
20
0
70
71
48
17
10
3
2
(-1,5]
(5,10]
(10,15]
(15,20]
(20,25]
(25,30]
(30,50]
Predicted Points Residuals (82 game pace)



Next Steps:

With access to better data I would like to include both team level and career to date features into the model to see if it would help it perform more effectively. As I progress in my modelling capabilities, possibly testing out this data using a quadratic model, considering the Age Curve in professional sports, or a Lasso Regression to draw out important features rather than just using a Stepwise technique.