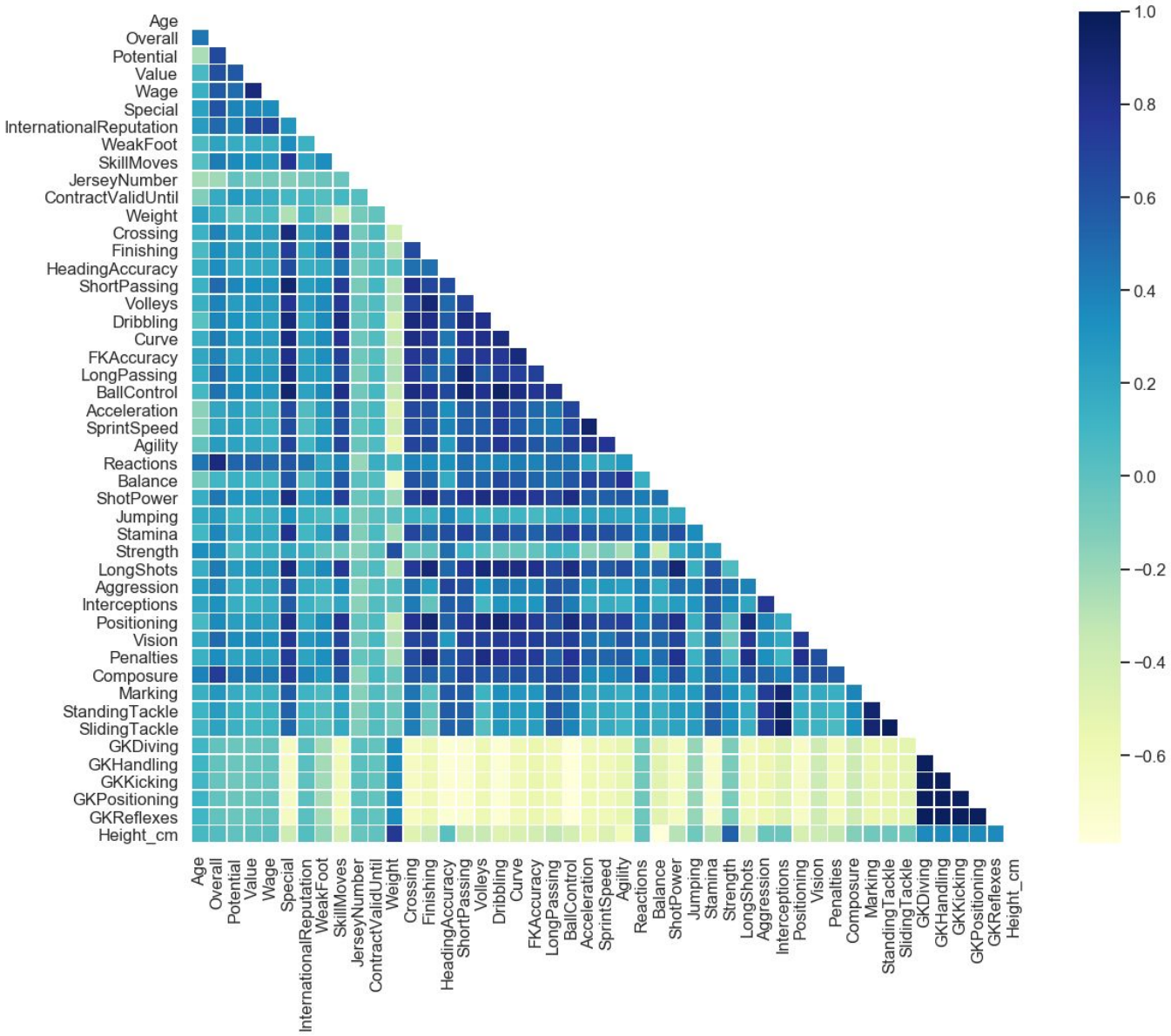
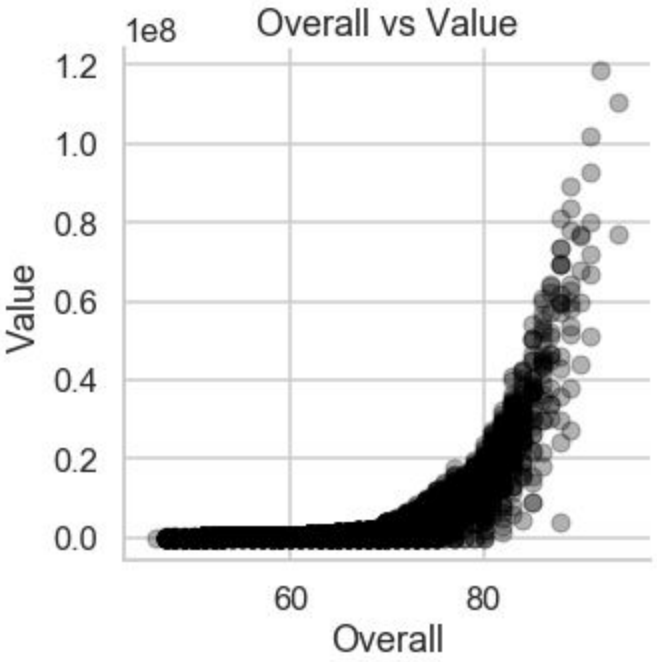
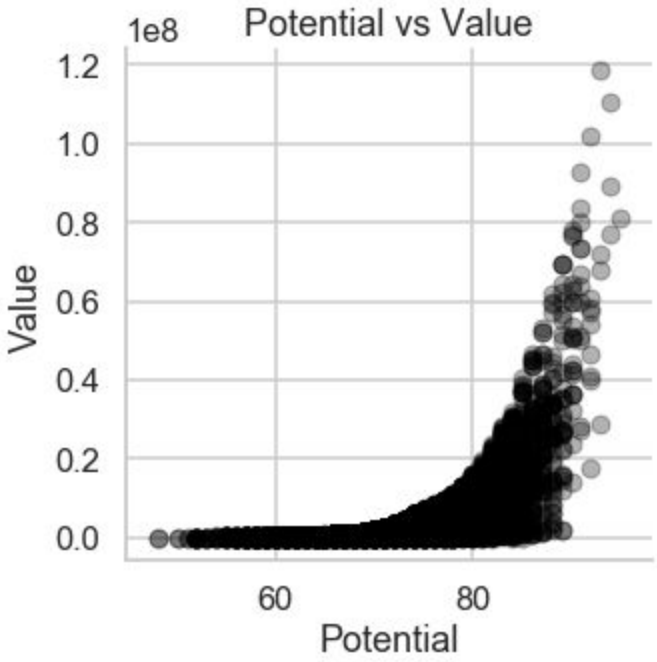
FIFA19 Data Analysis

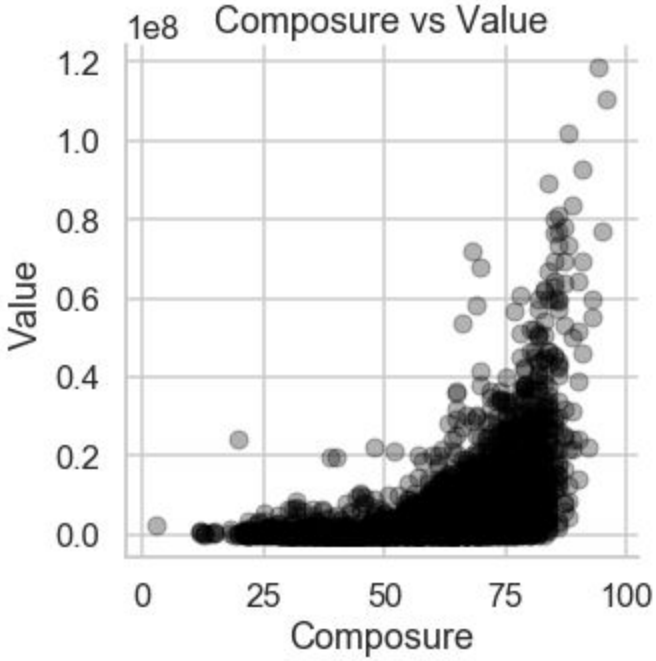
Here we shall examine how certain variables relate to our target variable: Player Value. A reasonable place to start is to check whether there are any strong correlations between any of the variables as well as our target variable. This will provide insight regarding any “linear” relationship; however, it will not be a good tool for non-linear relationships. I will consider any Pearson correlation coefficient value (r) above 0.5(absolute value), to be of significance, and will require further exploration when selecting a model for predicting player’s value.

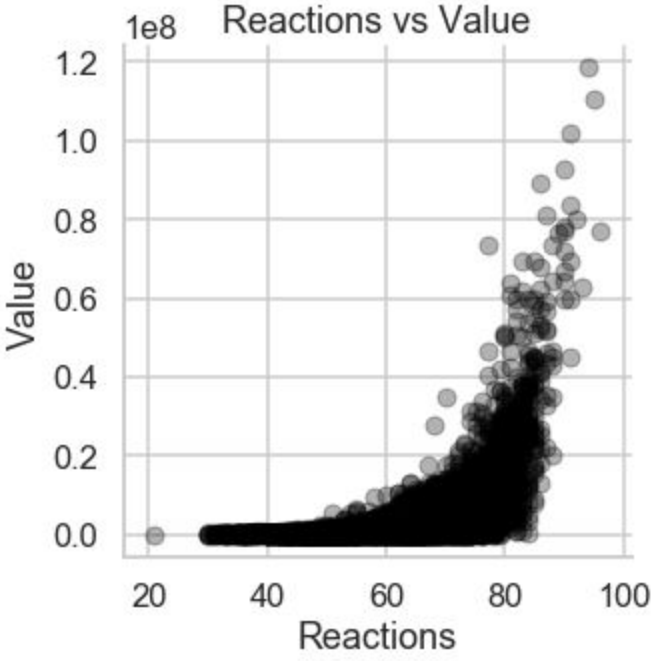


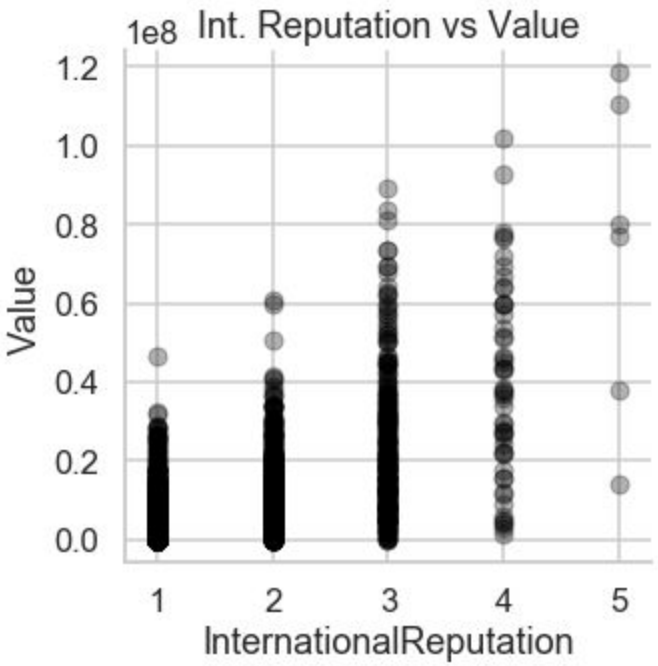
Zooming in at the value column, the notable features with the highest correlations are as follows : Overall, Potential, Wage, International Reputation, Reactions, and Composure. All of which have correlation coefficient of greater than equal to 0.5.

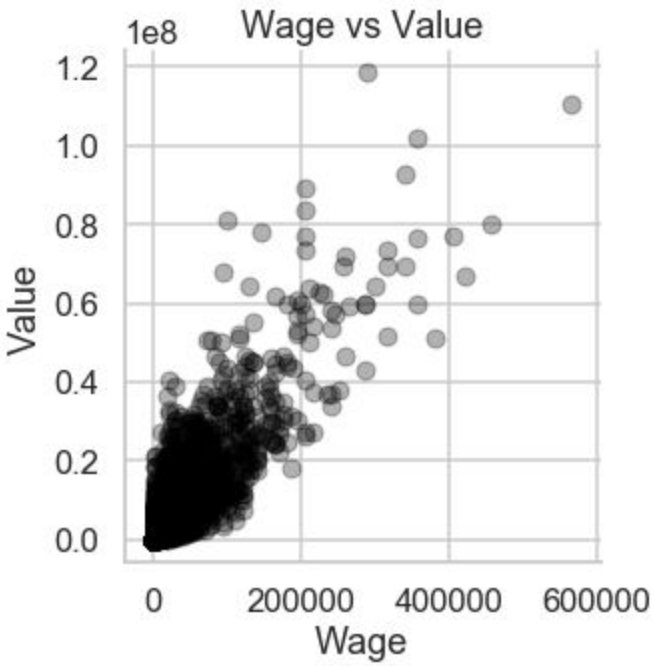

Overall vs Value
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Value
60
80
Overall


Potential vs Value
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Value
60
80
Potential


Composure vs Value
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Value
0
25
50
75
100
Composure

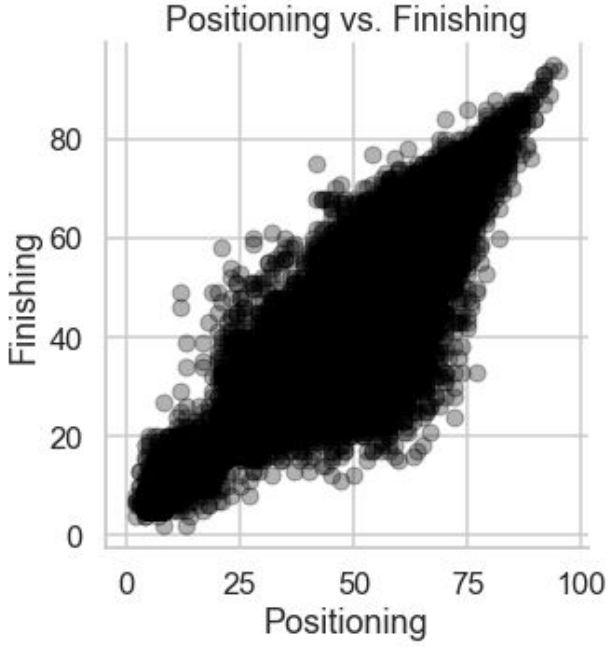

Reactions vs Value
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Value
20
40
60
80
100
Reactions

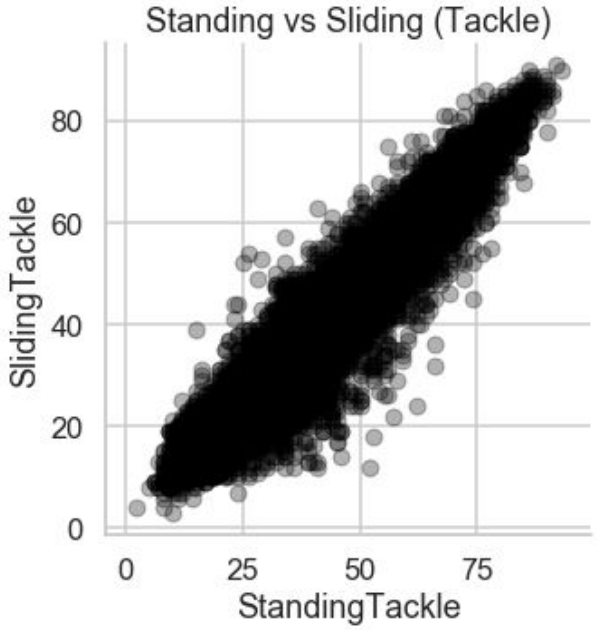




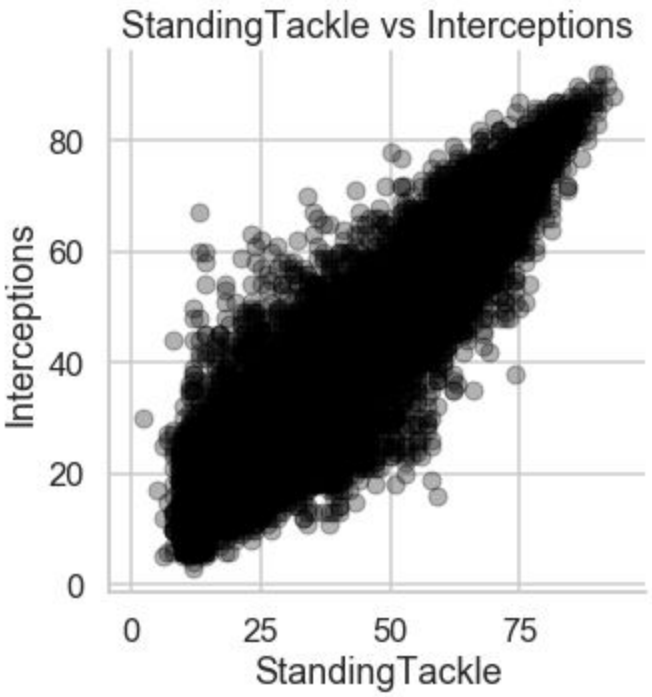
The scatter plots further support the linear relationship between the variables mentioned.

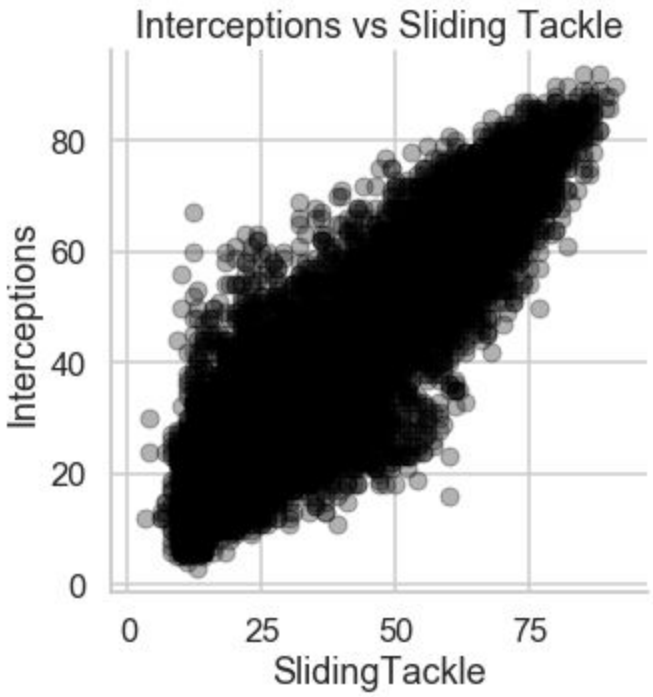
Looking at the correlation between other variables we can see that there are strong correlations between certain variables as expected. This is because these values are related to each other. For example, Finishing and Positioning have correlation values of above 0.9. This is expected because it would be normal for someone with a good positioning to have a good finishing score on average. This is true for many of the variables, thus I will need to use a method that will minimize any multicollinearity from my chosen model. I may need to cluster certain skills that are related to each other to one general skill. Examples, shown below.





The same is true for Tackling of either kind and Interceptions.





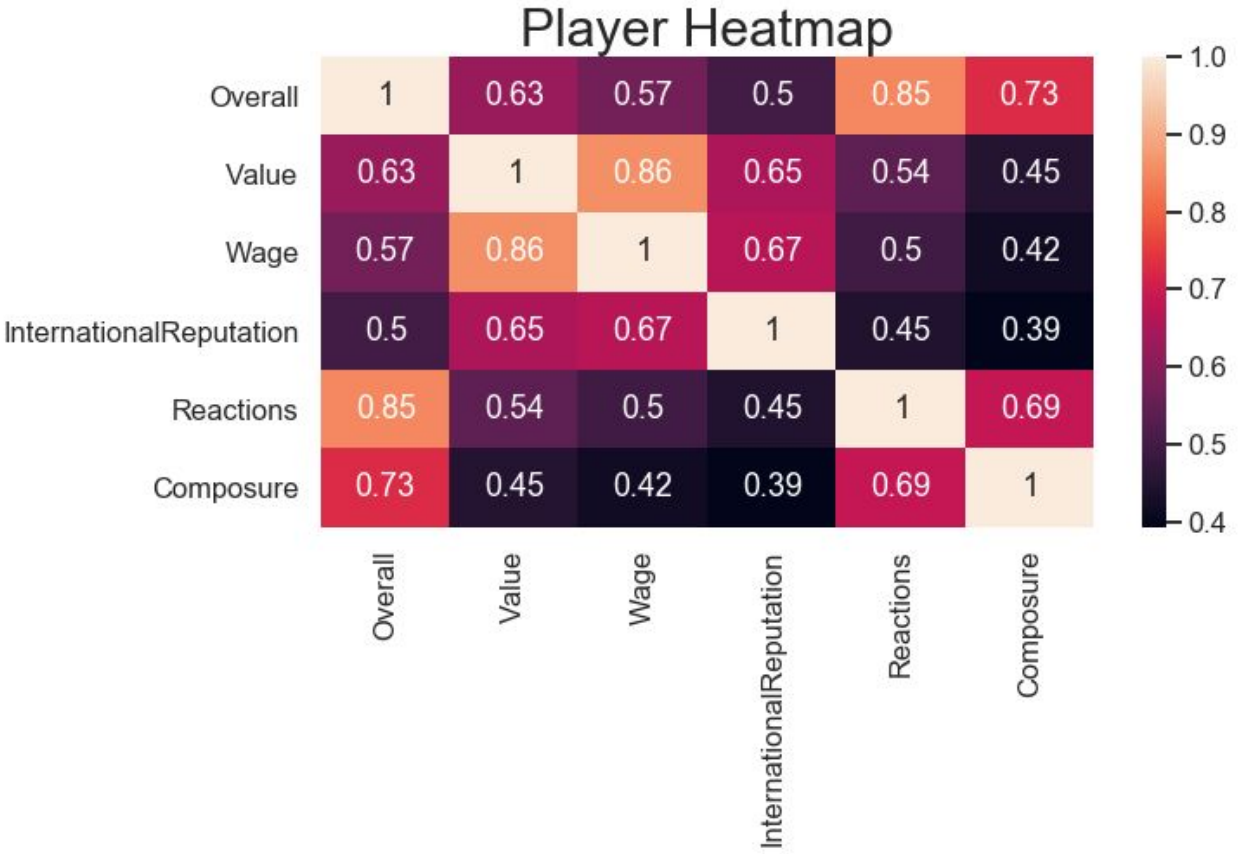
Hypothesis Testing (ANOVA)

We have the following ANOVA results for the model Value ~ Wage + InternationalReputation + Reactions + Overall + Composure, based on the highest correlation values.

	df	sum_sq	mean_sq	F
Wage	1.0	4.199774e+17	4.199774e+17	58447.534094
InternationalReputation	1.0	6.468290e+15	6.468290e+15	900.180882
Reactions	1.0	7.604239e+15	7.604239e+15	1058.268916
Overall	1.0	4.920042e+15	4.920042e+15	684.713770
Composure	1.0	1.533055e+14	1.533055e+14	21.335266
Residual	18201.0	1.307841e+17	7.185545e+12	NaN

	PR(>F)
Wage	0.000000e+00
InternationalReputation	4.396260e-193
Reactions	1.081045e-225
Overall	3.438348e-148
Composure	3.882125e-06
Residual	NaN

We can see that all of the p-values for the 2-way ANOVA are close to 0. Thus, if our null Hypothesis is that at least one of the estimators is 0, we would reject that hypothesis. Meaning, all of these features add significant value to this initial Ordinary Least Squares model. Let's also take a look at the heatmap zoomed in at the features above.



We can see that Composure has the lowest correlation with our target "Value". Also, we can check for multicollinearity here between any of our features. I will only consider correlation coefficient values above 0.7 for multicollinearity. We can see that Overall has high correlation with Reactions and Composure. There is also moderately strong correlation between Wage and Reputation, as well as Reactions and Composure. We may consider dropping Composure or Reactions or both, if we see that the model performs better.