Anova test on the 4 models –

From the results of anova test we can observe that:

1. Model 2 is better than Model 1 since p-value < 0.05 and RSS value is decreased. So, addition of `region` variable to the model proved to improve the model.
2. Model 3 is not an improvement from Model 2 since p-value > 0.05 and RSS value is unchanged. So, addition of `panels\_flat` variable alone to the model proves to be insignificant.
3. Model 4 is definitely an improvement from Model 3 since p-value < 0.05 and RSS value is decreased significantly. So, addition of all the directional panels to the model has improved the model significantly.

VIF tests –

1. From the vif tests on Model 1 we can observe that all variables have high vif values and is above 10. So the variables are highly correlated and are not acceptable.
2. From the vif tests on Model 2 we can observe that the two variables have high vif values and is above 10. So the variables are highly correlated and are not acceptable. Only one variable `region`, has low vif value at 1.05 which is in between 1 and 5 required to prove it is moderately correlated and is acceptable.
3. From the vif tests on Model 3 we can observe that similarly to model2 all the variables except region` have high vif values and is above 10. So, those variables are highly correlated and are not acceptable.
4. The vif test for Model 4 indicates presence of aliased coefficients in the model which means that the addition of remaining 4 directional panels to the model is resulting in **perfect multicollinearity,** which is not acceptable at all. Since Model 4 has linearly dependent variables we remove `panels\_total` from the model and we do vif test again.
5. From the new VIF test on Model 4 we can observe that three of the directional panels have high multicollinearity and is not acceptable, vif value of ‘panels\_north’ is less than 10 and near the acceptable vif value but it is still above 5. We can consider it acceptable but it is still highly correlated but not as much as other variables.

Conclusion –

Our Findings –

From the above tests we conclude that Model 4 is the best model out of the 4 models to determine effects on Carbon offset. Model 4 p-value < 0.05, Multiple R-squared value is 95% which means the variables in the Model 4 can explain 95% of changes in the dataset and cook’s distance is less than 1 which means that there is no high influence of high leverage points that can alter the model.

From this model we can obtain following observations –

1. By increasing number of solar panels in the south region, we can see positive effects on carbon offset which means we can reduce carbon emissions by adding panels in the south region.
2. By increasing number of solar panels in Northeast and West region, we can see negative effects on carbon offset which means we will increase carbon emissions by adding solar panels in the northeast and west regions.
3. By increasing number of flat facing panels, we can see positive effects on carbon offset which means we can reduce carbon emissions by adding flat facing solar panels.
4. By increasing number of South and East facing panels, we can see the most positive effects on carbon offset which means we can reduce the carbon emissions the most by adding South and East facing solar panels.
5. By increasing number of North facing panels, we can see negative effects on carbon offset which means we can increase carbon emissions by adding North facing solar panels

Caveats –

1. Our models have high multicollinearity. While building early models we found high vif values, had to adjust the models multiple times to obtain descent models. Some variables were even linearly dependent like we have seen in the vif test for Model 4, where we had to remove panels\_total variables to obtain normal vif test results. This means that variables in our model are highly correlated like we have seen in the results of vif test.
2. There is high outlier count in the dataset. We have observed this in the Exploratory Data Analysis in plotting boxplots of Total Potential solar energy generated vs region where we had to rearrange the y-axis to have clearer boxplots. This leads to randomness in our dataset and have wrong effects on models.
3. All the values in the dataset are estimated values, obtained from the google database of imagery and maps. They are not real world recorded values obtained from readings of solar panels directly.