ENE 434 Lab 8 Assignment

#Pre-Assignment Necessesities (included in lab) ##Loading Packages

In the lab, we did the following necessary activites: Loaded data, created incentive per kw column,changed to MW instead of kw, and estimated and predicted with first gam model.

#Assignment ##Question 1 Can you create the variables zip_year_total_mw, that is, the cumulative amount of capacity in each zip code (a finer geographic division than county). Chart the relationship between total installed capacity in a year and costs at the zip level. Include zip_year_total_mw in the linear model instead of county_year_total_mw. Are the estimated results substantially different? ###creating zipsum columns

```
zip_sum <- pv_df %>% group_by(zip, year) %>%
  summarize(
    zip_year_total_mw = sum(nameplate)
  )
```

```
## `summarise()` has grouped output by 'zip'. You can override using the `.groups` argument.
```

```
pv_df_zip <- left_join(pv_df,zip_sum,by=c('zip','year'))
##changing to MW instead of kw

pv_df_zip$zip_year_total_mw = pv_df$county_year_total/1000
pv_df_zip$contractor_year_total_mw = pv_df$contractor_year_total/1000
pv_df_zip$contractor_market_share_perc = pv_df$contractor_market_share*100

lm_mod2 = lm(cost_per_kw ~ date +
  county +
    sector +
    nameplate +
    zip_year_total_mw +
    incentive_per_kw +
    contractor_year_total_mw +
    lease +
  china,
    data=pv_df_zip)

summary(lm_mod2)
```

```
##
## Call:
## lm(formula = cost_per_kw ~ date + county + sector + nameplate +
##     zip_year_total_mw + incentive_per_kw + contractor_year_total_mw +
```

```
##     lease + china, data = pv_df_zip)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10097   -870   -219    530  98113
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.508e+04  3.117e+02  80.464  < 2e-16 ***
## date                     -1.247e+00  1.987e-02 -62.764  < 2e-16 ***
## countyAmador             -6.246e+02  1.422e+02  -4.393 1.12e-05 ***
## countyButte              -8.038e+02  6.802e+01 -11.818  < 2e-16 ***
## countyCalaveras          -6.498e+02  1.019e+02  -6.379 1.79e-10 ***
## countyColusa             -6.895e+02  2.055e+02  -3.355 0.000794 ***
## countyContra Costa        2.842e+01  3.703e+01   0.767 0.442796
## countyEl Dorado          -3.401e+02  5.508e+01  -6.174 6.69e-10 ***
## countyFresno             -4.257e+02  3.642e+01 -11.688  < 2e-16 ***
## countyGlenn              -5.988e+02  1.371e+02  -4.368 1.25e-05 ***
## countyHumboldt           -4.429e+01  1.451e+02  -0.305 0.760213
## countyImperial           -1.730e+03  7.236e+02  -2.391 0.016799 *
## countyInyo               -7.940e+02  2.605e+02  -3.048 0.002302 **
## countyKern               -2.710e+02  3.770e+01  -7.189 6.56e-13 ***
## countyKings              -6.525e+02  7.219e+01  -9.038  < 2e-16 ***
## countyLake               -2.912e+02  1.187e+02  -2.453 0.014187 *
## countyLassen             -1.304e+03  5.396e+02  -2.416 0.015712 *
## countyLos Angeles         3.433e+02  3.529e+01   9.725  < 2e-16 ***
## countyMadera             -5.231e+02  7.083e+01  -7.385 1.54e-13 ***
## countyMarin              -5.182e+02  5.253e+01  -9.865  < 2e-16 ***
## countyMariposa           -7.491e+02  2.237e+02  -3.349 0.000811 ***
## countyMendocino          -1.257e+02  9.955e+01  -1.263 0.206663
## countyMerced             -7.444e+02  7.933e+01  -9.384  < 2e-16 ***
## countyMono               -7.730e+02  1.747e+02  -4.426 9.63e-06 ***
## countyMonterey           -3.512e+02  7.438e+01  -4.722 2.34e-06 ***
## countyNapa               -1.563e+02  6.392e+01  -2.446 0.014456 *
## countyNevada             -7.622e+02  7.733e+01  -9.857  < 2e-16 ***
## countyOrange              2.753e+02  3.245e+01   8.483  < 2e-16 ***
## countyPlacer             -3.658e+02  4.323e+01  -8.463  < 2e-16 ***
## countyPlumas             -5.784e+02  1.519e+02  -3.809 0.000140 ***
## countyRiverside           2.606e+02  3.597e+01   7.243 4.42e-13 ***
## countySacramento         -4.581e+02  3.183e+02  -1.439 0.150032
## countySan Benito         -5.071e+02  1.470e+02  -3.449 0.000563 ***
## countySan Bernardino      2.583e+02  3.431e+01   7.530 5.10e-14 ***
## countySan Diego           1.660e+02  3.256e+01   5.098 3.44e-07 ***
## countySan Francisco       1.242e+03  4.380e+01  28.346  < 2e-16 ***
## countySan Joaquin        -5.243e+02  5.331e+01  -9.836  < 2e-16 ***
## countySan Luis Obispo    -4.125e+02  5.009e+01  -8.235  < 2e-16 ***
## countySan Mateo          -2.941e+01  4.652e+01  -0.632 0.527268
## countySanta Barbara      -5.090e+02  5.615e+01  -9.064  < 2e-16 ***
## countySanta Clara        -3.261e+02  3.303e+01  -9.871  < 2e-16 ***
## countySanta Cruz         -6.205e+02  5.723e+01 -10.843  < 2e-16 ***
## countyShasta             -6.998e+02  9.990e+01  -7.005 2.49e-12 ***
## countySolano             -1.808e+01  6.149e+01  -0.294 0.768767
## countySonoma             -3.638e+02  3.914e+01  -9.296  < 2e-16 ***
## countyStanislaus         -8.378e+02  1.084e+02  -7.730 1.09e-14 ***
```

```
## countySutter             -5.551e+02  9.587e+01  -5.790 7.06e-09 ***
## countyTehama             -4.439e+02  1.557e+02  -2.851 0.004364 **
## countyTrinity             1.630e+03  1.617e+03   1.008 0.313346
## countyTulare             -3.954e+02  4.452e+01  -8.882  < 2e-16 ***
## countyTuolumne           -5.340e+02  1.442e+02  -3.703 0.000213 ***
## countyVentura            -1.951e+02  3.784e+01  -5.157 2.52e-07 ***
## countyYolo               -1.743e+02  5.900e+01  -2.954 0.003136 **
## countyYuba                4.108e+01  1.163e+02   0.353 0.723801
## sectorGovernment          1.310e+03  9.643e+01  13.580  < 2e-16 ***
## sectorNon-Profit         -6.267e+02  9.469e+01  -6.618 3.65e-11 ***
## sectorResidential         1.851e+02  4.239e+01   4.366 1.27e-05 ***
## nameplate                -1.217e+00  2.113e-01  -5.761 8.40e-09 ***
## zip_year_total_mw        -1.958e+01  1.028e+00 -19.050  < 2e-16 ***
## incentive_per_kw          5.610e-01  1.894e-02  29.615  < 2e-16 ***
## contractor_year_total_mw -7.238e-01  6.595e-01  -1.098 0.272374
## lease                     1.680e+02  1.295e+01  12.975  < 2e-16 ***
## china                    -3.222e+02  1.363e+01 -23.637  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1617 on 104998 degrees of freedom
##   (1562 observations deleted due to missingness)
## Multiple R-squared:  0.3747, Adjusted R-squared:  0.3743
## F-statistic:  1015 on 62 and 104998 DF,  p-value: < 2.2e-16
```

###creating zipdata for ggmaps to use

```
zip_data = pv_df %>% group_by(zip, year) %>% summarise(
  avg_zip_cost = mean(cost_per_kw, na.rm=TRUE),
  avg_zip_cost_less_sub = mean(cost_ex_subsid_per_kw, na.rm=TRUE),
  zip_year_total = sum(nameplate, na.rm=TRUE),
  zip_lat = mean(latitude, na.rm=TRUE),
  zip_long = mean(longitude, na.rm=TRUE),
  cost_per_capacity = avg_zip_cost/zip_year_total
)
```

```
## `summarise()` has grouped output by 'zip'. You can override using the `.groups` argument.
```

###setting up ggmaps

excluding "maps_api_key = '_____' ^^

```
p_load(ggmap, gganimate, gifski)
register_google(key = maps_api_key)
```

###loading in calfornia map, animating to show change over time This part is excluded and the gif is attached so that the file would knit!

```
#print(animate_object)
```

##Question 2 In lab 7 we used cumulative capacity to model learning curves. In this lab we used days/dates to model change of prices over time. Run regressions (linear and semi-parametric (GAM)) where you

use cumulative capacity instead of days/date to model the effects over time. For the semi-parametric model, create a prediction and compare with the results in the lab. Are the results substantially different? ###creating a cumulative capacity value and using this in a linear model

```
pv_df = pv_df %>% arrange(date) %>% mutate(
  cum_cap = cumsum(nameplate)
)

lm_mod3 = lm(cost_per_kw ~ cum_cap +
  county +
    sector +
    nameplate +
    county_year_total_mw +
     contractor_year_total_mw +
    incentive_per_kw +
    lease +
  china,
    data=pv_df)

summary(lm_mod2)
```

```
##
## Call:
## lm(formula = cost_per_kw ~ date + county + sector + nameplate +
##     zip_year_total_mw + incentive_per_kw + contractor_year_total_mw +
##     lease + china, data = pv_df_zip)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10097   -870   -219    530  98113
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.508e+04  3.117e+02  80.464  < 2e-16 ***
## date                     -1.247e+00  1.987e-02 -62.764  < 2e-16 ***
## countyAmador             -6.246e+02  1.422e+02  -4.393 1.12e-05 ***
## countyButte              -8.038e+02  6.802e+01 -11.818  < 2e-16 ***
## countyCalaveras          -6.498e+02  1.019e+02  -6.379 1.79e-10 ***
## countyColusa             -6.895e+02  2.055e+02  -3.355 0.000794 ***
## countyContra Costa        2.842e+01  3.703e+01   0.767 0.442796
## countyEl Dorado          -3.401e+02  5.508e+01  -6.174 6.69e-10 ***
## countyFresno             -4.257e+02  3.642e+01 -11.688  < 2e-16 ***
## countyGlenn              -5.988e+02  1.371e+02  -4.368 1.25e-05 ***
## countyHumboldt           -4.429e+01  1.451e+02  -0.305 0.760213
## countyImperial           -1.730e+03  7.236e+02  -2.391 0.016799 *
## countyInyo               -7.940e+02  2.605e+02  -3.048 0.002302 **
## countyKern               -2.710e+02  3.770e+01  -7.189 6.56e-13 ***
## countyKings              -6.525e+02  7.219e+01  -9.038  < 2e-16 ***
## countyLake               -2.912e+02  1.187e+02  -2.453 0.014187 *
## countyLassen             -1.304e+03  5.396e+02  -2.416 0.015712 *
## countyLos Angeles         3.433e+02  3.529e+01   9.725  < 2e-16 ***
## countyMadera             -5.231e+02  7.083e+01  -7.385 1.54e-13 ***
## countyMarin              -5.182e+02  5.253e+01  -9.865  < 2e-16 ***
## countyMariposa           -7.491e+02  2.237e+02  -3.349 0.000811 ***
```

```
## countyMendocino          -1.257e+02  9.955e+01  -1.263 0.206663
## countyMerced             -7.444e+02  7.933e+01  -9.384  < 2e-16 ***
## countyMono               -7.730e+02  1.747e+02  -4.426 9.63e-06 ***
## countyMonterey           -3.512e+02  7.438e+01  -4.722 2.34e-06 ***
## countyNapa               -1.563e+02  6.392e+01  -2.446 0.014456 *
## countyNevada             -7.622e+02  7.733e+01  -9.857  < 2e-16 ***
## countyOrange              2.753e+02  3.245e+01   8.483  < 2e-16 ***
## countyPlacer             -3.658e+02  4.323e+01  -8.463  < 2e-16 ***
## countyPlumas             -5.784e+02  1.519e+02  -3.809 0.000140 ***
## countyRiverside           2.606e+02  3.597e+01   7.243 4.42e-13 ***
## countySacramento         -4.581e+02  3.183e+02  -1.439 0.150032
## countySan Benito         -5.071e+02  1.470e+02  -3.449 0.000563 ***
## countySan Bernardino      2.583e+02  3.431e+01   7.530 5.10e-14 ***
## countySan Diego           1.660e+02  3.256e+01   5.098 3.44e-07 ***
## countySan Francisco       1.242e+03  4.380e+01  28.346  < 2e-16 ***
## countySan Joaquin        -5.243e+02  5.331e+01  -9.836  < 2e-16 ***
## countySan Luis Obispo    -4.125e+02  5.009e+01  -8.235  < 2e-16 ***
## countySan Mateo          -2.941e+01  4.652e+01  -0.632 0.527268
## countySanta Barbara      -5.090e+02  5.615e+01  -9.064  < 2e-16 ***
## countySanta Clara        -3.261e+02  3.303e+01  -9.871  < 2e-16 ***
## countySanta Cruz         -6.205e+02  5.723e+01 -10.843  < 2e-16 ***
## countyShasta             -6.998e+02  9.990e+01  -7.005 2.49e-12 ***
## countySolano             -1.808e+01  6.149e+01  -0.294 0.768767
## countySonoma             -3.638e+02  3.914e+01  -9.296  < 2e-16 ***
## countyStanislaus         -8.378e+02  1.084e+02  -7.730 1.09e-14 ***
## countySutter             -5.551e+02  9.587e+01  -5.790 7.06e-09 ***
## countyTehama             -4.439e+02  1.557e+02  -2.851 0.004364 **
## countyTrinity             1.630e+03  1.617e+03   1.008 0.313346
## countyTulare             -3.954e+02  4.452e+01  -8.882  < 2e-16 ***
## countyTuolumne           -5.340e+02  1.442e+02  -3.703 0.000213 ***
## countyVentura            -1.951e+02  3.784e+01  -5.157 2.52e-07 ***
## countyYolo               -1.743e+02  5.900e+01  -2.954 0.003136 **
## countyYuba                4.108e+01  1.163e+02   0.353 0.723801
## sectorGovernment          1.310e+03  9.643e+01  13.580  < 2e-16 ***
## sectorNon-Profit         -6.267e+02  9.469e+01  -6.618 3.65e-11 ***
## sectorResidential         1.851e+02  4.239e+01   4.366 1.27e-05 ***
## nameplate                -1.217e+00  2.113e-01  -5.761 8.40e-09 ***
## zip_year_total_mw        -1.958e+01  1.028e+00 -19.050  < 2e-16 ***
## incentive_per_kw          5.610e-01  1.894e-02  29.615  < 2e-16 ***
## contractor_year_total_mw -7.238e-01  6.595e-01  -1.098 0.272374
## lease                     1.680e+02  1.295e+01  12.975  < 2e-16 ***
## china                    -3.222e+02  1.363e+01 -23.637  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1617 on 104998 degrees of freedom
##   (1562 observations deleted due to missingness)
## Multiple R-squared:  0.3747, Adjusted R-squared:  0.3743
## F-statistic:  1015 on 62 and 104998 DF,  p-value: < 2.2e-16
```

###doing a gam model with the cumulative capacity value

```
gam_mod2=gam(cost_per_kw ~ s(cum_cap) +
    sector +
    nameplate +
    lease +
    county_year_total_mw +
     contractor_year_total_mw +
    incentive_per_kw +
    china,
    family=gaussian,
    data=pv_df)
summary(gam_mod2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## cost_per_kw ~ s(cum_cap) + sector + nameplate + lease + county_year_total_mw +
##     contractor_year_total_mw + incentive_per_kw + china
##
## Parametric coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5418.85101   43.60120 124.282  < 2e-16 ***
## sectorGovernment          974.61481   95.92139  10.161  < 2e-16 ***
## sectorNon-Profit         -660.74675   94.27181  -7.009 2.42e-12 ***
## sectorResidential         175.97218   42.16034   4.174 3.00e-05 ***
## nameplate                  -2.39796    0.20821 -11.517  < 2e-16 ***
## lease                     128.86319   12.78333  10.081  < 2e-16 ***
## county_year_total_mw        3.42414    0.60678   5.643 1.67e-08 ***
## contractor_year_total_mw    2.84948    0.66341   4.295 1.75e-05 ***
## incentive_per_kw            0.89384    0.01791  49.904  < 2e-16 ***
## china                    -302.90341   13.64405 -22.200  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(cum_cap) 8.983      9 949.5  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.377   Deviance explained = 37.7%
## GCV = 2.6028e+06  Scale est. = 2.6023e+06  n = 105061
```

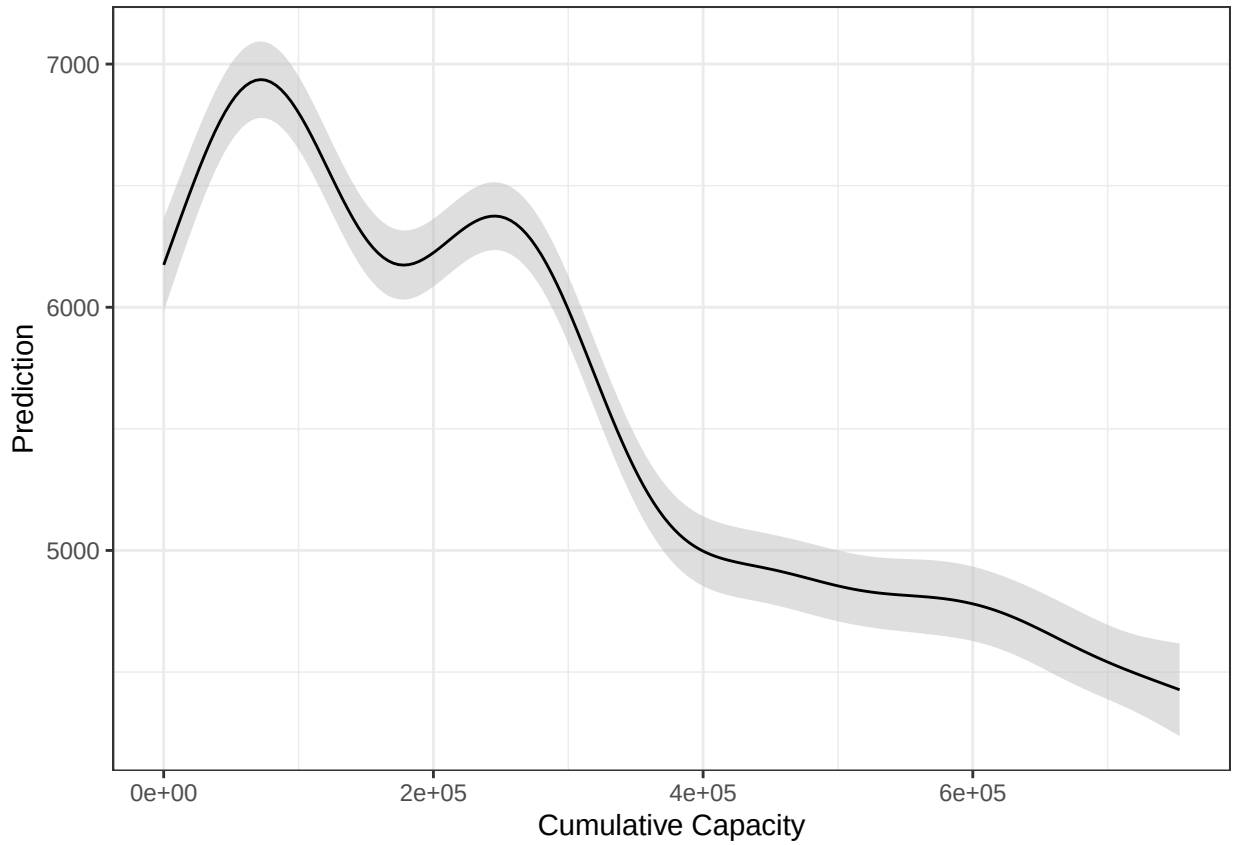
###setting up test dataset and predicting with previous gam model

```
pdat2 =with(pv_df,
list(
cum_cap = round(seq(min(cum_cap), max(cum_cap), length = 200)),
sector = rep("Residential",200),
nameplate = rep(mean(nameplate), 200),
county_year_total_mw = rep(mean(county_year_total_mw), 200),
contractor_year_total_mw = rep(mean(contractor_year_total_mw), 200),
```

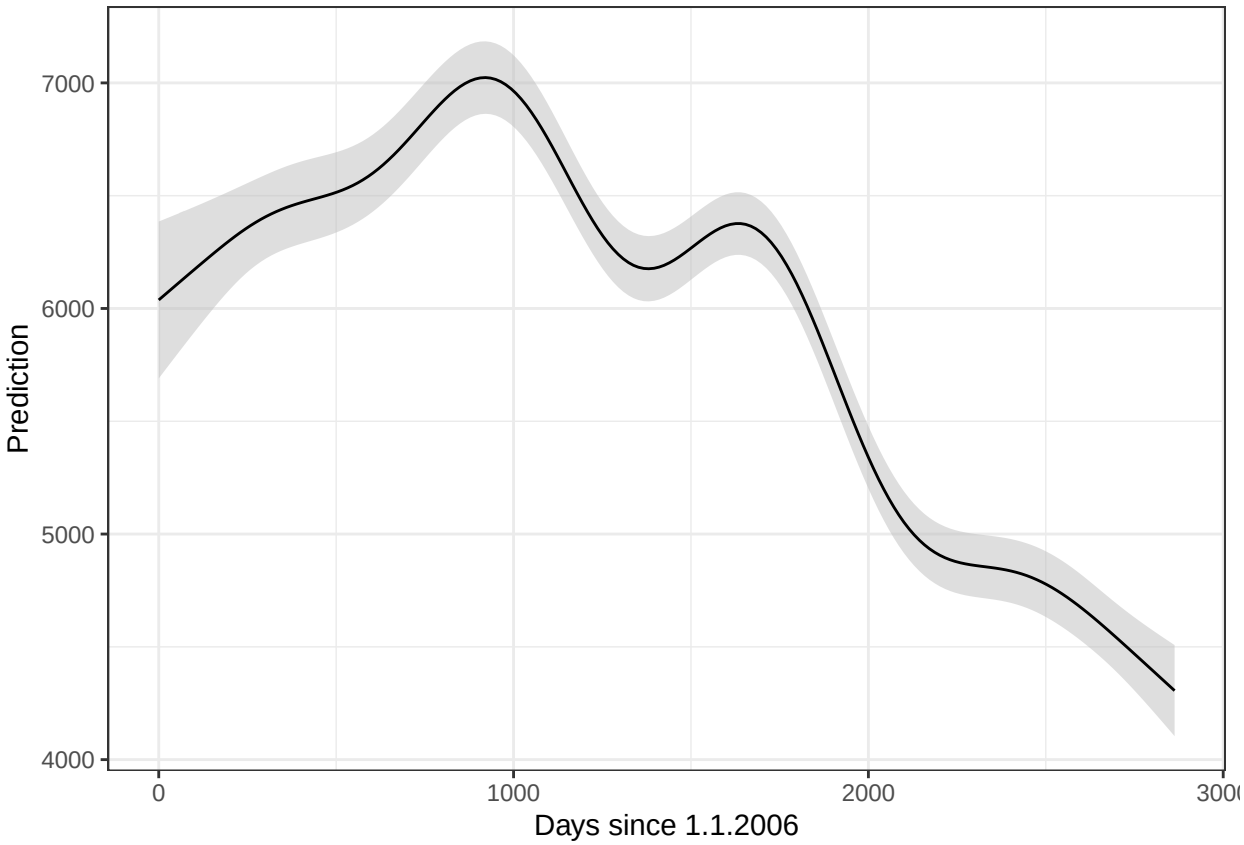
```
incentive_per_kw = rep(mean(incentive_per_kw/1000), 200),
lease = rep(0, 200),
china = rep(0, 200)
))

pred2 = predict(gam_mod2, pdat2, type = "terms", se.fit = TRUE)

pred2_fit = as_tibble(pred2$fit)
pred2_fit["intercept"] = coef(gam_mod2)[1]

pred2_fit = pred2_fit %>% mutate(prediction = rowSums(.))

pred2_fit["cum_cap"] = with(pv_df, round(seq(min(cum_cap), max(cum_cap), length = 200)))

pred2_se= as_tibble(pred2$se)
pred2_fit["prediction_se"] = rowSums(pred2_se)
pred2_fit["upper"] = pred2_fit$prediction +  2 * pred2_fit$prediction_se
pred2_fit["lower"] = pred2_fit$prediction - 2* pred2_fit$prediction_se
```

###plotting cumulative capacity gam model as well as time gam model and comparing

```
ggplot(pred2_fit, aes(x=cum_cap, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Cumulative Capacity", y="Prediction") +
    theme_bw()
```

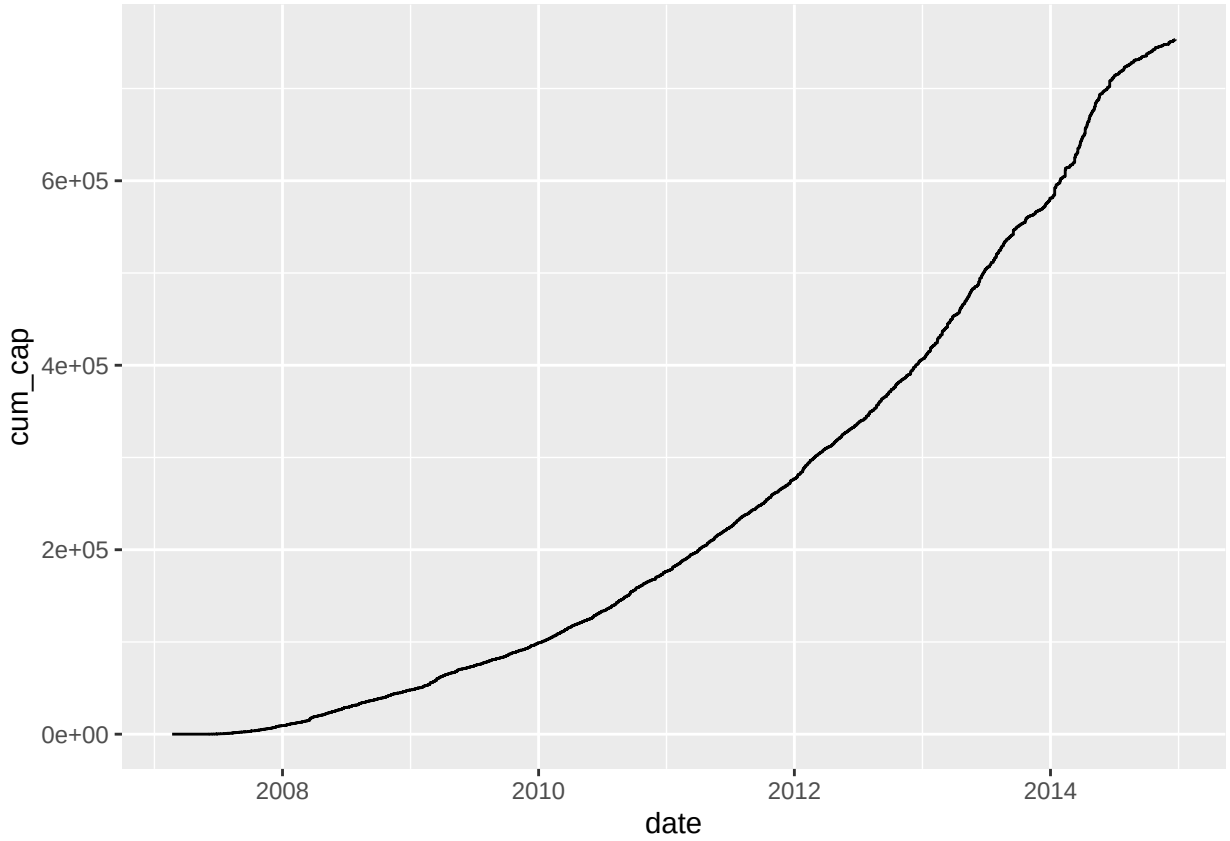



```
ggplot(pred1_fit, aes(x=days, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Days since 1.1.2006", y="Prediction") +
    theme_bw()
```



The results appear different, but when taken into account with exponential increase in cumulative capacity over time, the graphs are saying the same thing! The graph below depicts capacity over time ###plotting cumulative capacity over time

```
ggplot(data = pv_df, mapping = aes(x = date, y = cum_cap)) +
  geom_line()
```



##Question 3 Now create a prediction model from a GAM estimation for Chinese panels over time. In addition, create a new variable that indicates the share (from 0-1) of Chinese panels among installed solar panels per month. Use this series to create a prediction model (so instead of inputting a column of 0s or 1s to indicate Chinese, or non-Chinese panels, you would put in the series representing the share of panels.) Compare the predictive curve of cost over time with the curve representing non-Chinese panels. What does this tell you about the influence of Chinese panels on the average cost of solar panel systems during this time period?

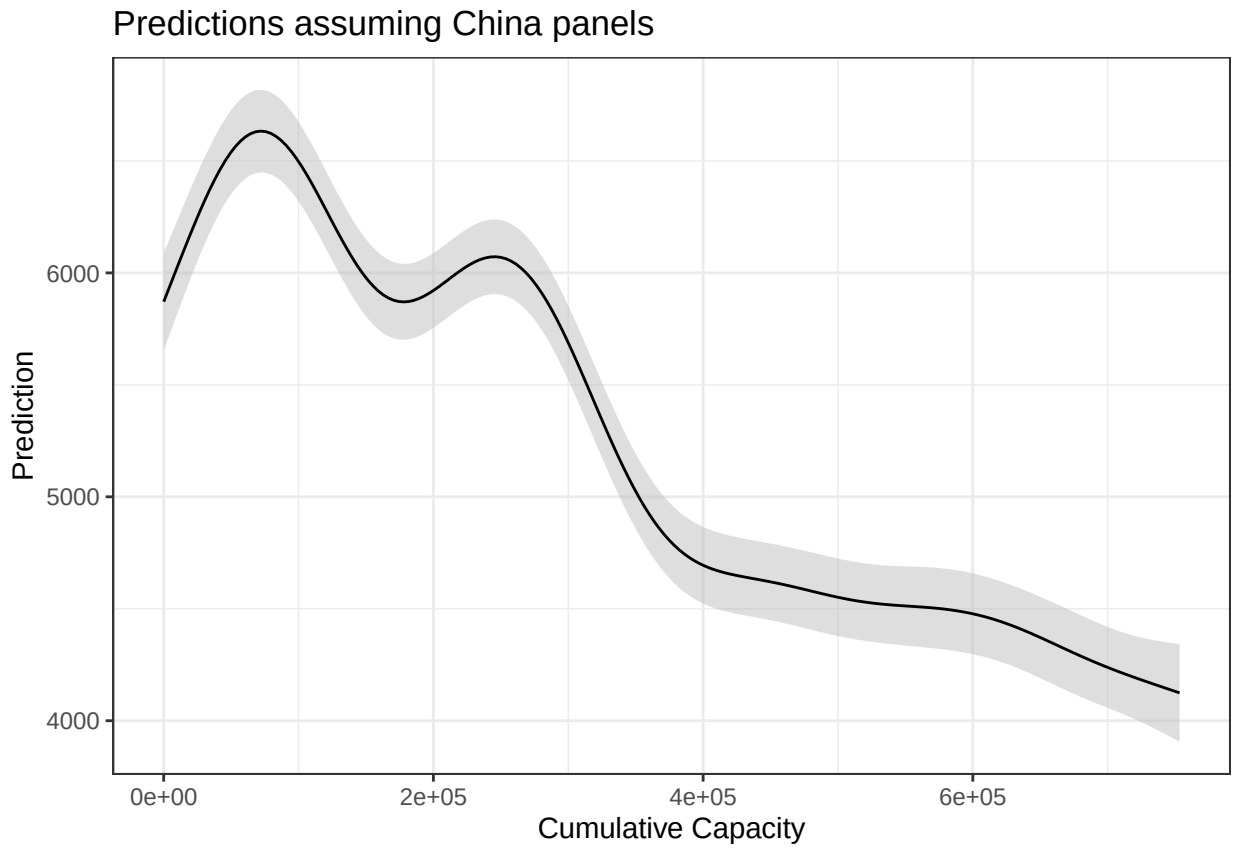
###setting up prediction for model with only chinese panels on cumulative capacity

```
pdat3 =with(pv_df,
list(
cum_cap = round(seq(min(cum_cap), max(cum_cap), length = 200)),
sector = rep("Residential",200),
nameplate = rep(mean(nameplate), 200),
county_year_total_mw = rep(mean(county_year_total_mw), 200),
contractor_year_total_mw = rep(mean(contractor_year_total_mw), 200),
incentive_per_kw = rep(mean(incentive_per_kw/1000), 200),
lease = rep(0, 200),
china = rep(1, 200)
))

pred3 = predict(gam_mod2, pdat3, type = "terms", se.fit = TRUE)

pred3_fit = as_tibble(pred3$fit)
pred3_fit["intercept"] = coef(gam_mod2)[1]
```

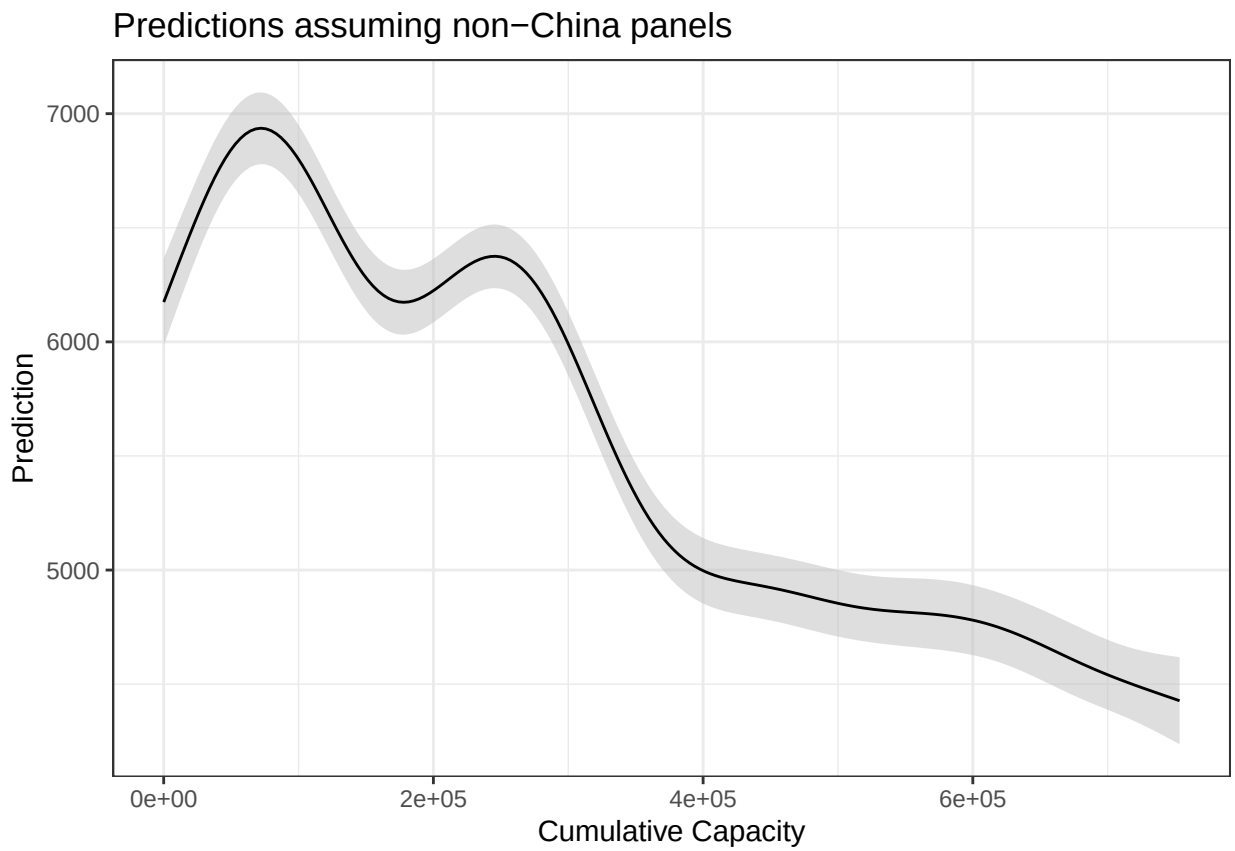
```
pred3_fit = pred3_fit %>% mutate(prediction = rowSums(.))

pred3_fit["cum_cap"] = with(pv_df, round(seq(min(cum_cap), max(cum_cap), length = 200)))

pred3_se= as_tibble(pred3$se)
pred3_fit["prediction_se"] = rowSums(pred3_se)
pred3_fit["upper"] = pred3_fit$prediction +  2 * pred3_fit$prediction_se
pred3_fit["lower"] = pred3_fit$prediction - 2* pred3_fit$prediction_se
```

###plotting first only china, then plotting only nonchina to compare the difference

```
ggplot(pred3_fit, aes(x=cum_cap, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Cumulative Capacity", y="Prediction") +
  ggtitle('Predictions assuming China panels') +
    theme_bw()
```




```
ggplot(pred2_fit, aes(x=cum_cap, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Cumulative Capacity", y="Prediction") +
  ggtitle('Predictions assuming non-China panels') +
    theme_bw()
```



If we predict the panels to be only chinese, then they cost less across the whole distribution of time.

###Appending china ratio column (the hard way)

```
pv_df_alter <- pv_df %>%
  mutate(yearz = year-2007) %>%
  mutate(cum_month = (yearz)*12 + month)
pv_df_alter2 <- pv_df_alter %>%
  group_by(cum_month) %>%
  summarize(
    total_capacity = sum(cum_cap, na.rm = TRUE)
  ) %>%
  arrange(cum_month)
#pv_df_alter_simple <- pv_df_alter %>%
  #select(cum_month) %>%
  #distinct()
pv_df_alter2[nrow(pv_df_alter2)+1,] <- 1
pv_df_alter2[nrow(pv_df_alter2)+1,] <- 3
pv_df_alter2[nrow(pv_df_alter2)+1,] <- 4
pv_df_alter2 <- pv_df_alter2 %>%
  add_column(cum_cap_month_china = 0)
#for (i in 1:max(pv_df_alter$cum_month)) {
#  dataset <- pv_df_alter %>%
#    filter(cum_month==i)
#  pv_df_alter2$cum_cap_month[i] = sum(dataset$nameplate, na.rm=TRUE)
#}
```

```
for (i in 1:max(pv_df_alter$cum_month)) {
  dataset <- pv_df_alter %>%
    dplyr::filter(cum_month==i) %>%
    dplyr::filter(china == 1)
  pv_df_alter2$cum_cap_month_china[i] = sum(dataset$nameplate, na.rm = TRUE)
}
pv_df_ratio <- pv_df_alter2 %>%
  mutate(china_ratio = cum_cap_month_china/total_capacity) %>%
  dplyr::filter(!is.na(china_ratio))
pv_df_final <- left_join(pv_df_alter, pv_df_ratio, by = 'cum_month')
```

###creating another gam model using the new china ratio value

```
gam_mod3=gam(cost_per_kw ~ s(cum_cap) +
    sector +
    nameplate +
    lease +
    county_year_total_mw +
     contractor_year_total_mw +
    incentive_per_kw +
    china_ratio,
    family=gaussian,
    data=pv_df_final)
summary(gam_mod3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## cost_per_kw ~ s(cum_cap) + sector + nameplate + lease + county_year_total_mw +
##     contractor_year_total_mw + incentive_per_kw + china_ratio
##
## Parametric coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5362.62350   42.66682 125.686  < 2e-16 ***
## sectorGovernment          944.90627   94.22405  10.028  < 2e-16 ***
## sectorNon-Profit         -656.74757   90.83462  -7.230 4.86e-13 ***
## sectorResidential         186.17806   41.28619   4.509 6.51e-06 ***
## nameplate                  -2.25364    0.20496 -10.995  < 2e-16 ***
## lease                     104.40409   12.64259   8.258  < 2e-16 ***
## county_year_total_mw        3.18568    0.60449   5.270 1.37e-07 ***
## contractor_year_total_mw   -2.19128    0.62495  -3.506 0.000455 ***
## incentive_per_kw            0.90971    0.01777  51.191  < 2e-16 ***
## china_ratio                -0.38435    0.02634 -14.594  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df    F p-value
## s(cum_cap) 8.983      9 1013  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

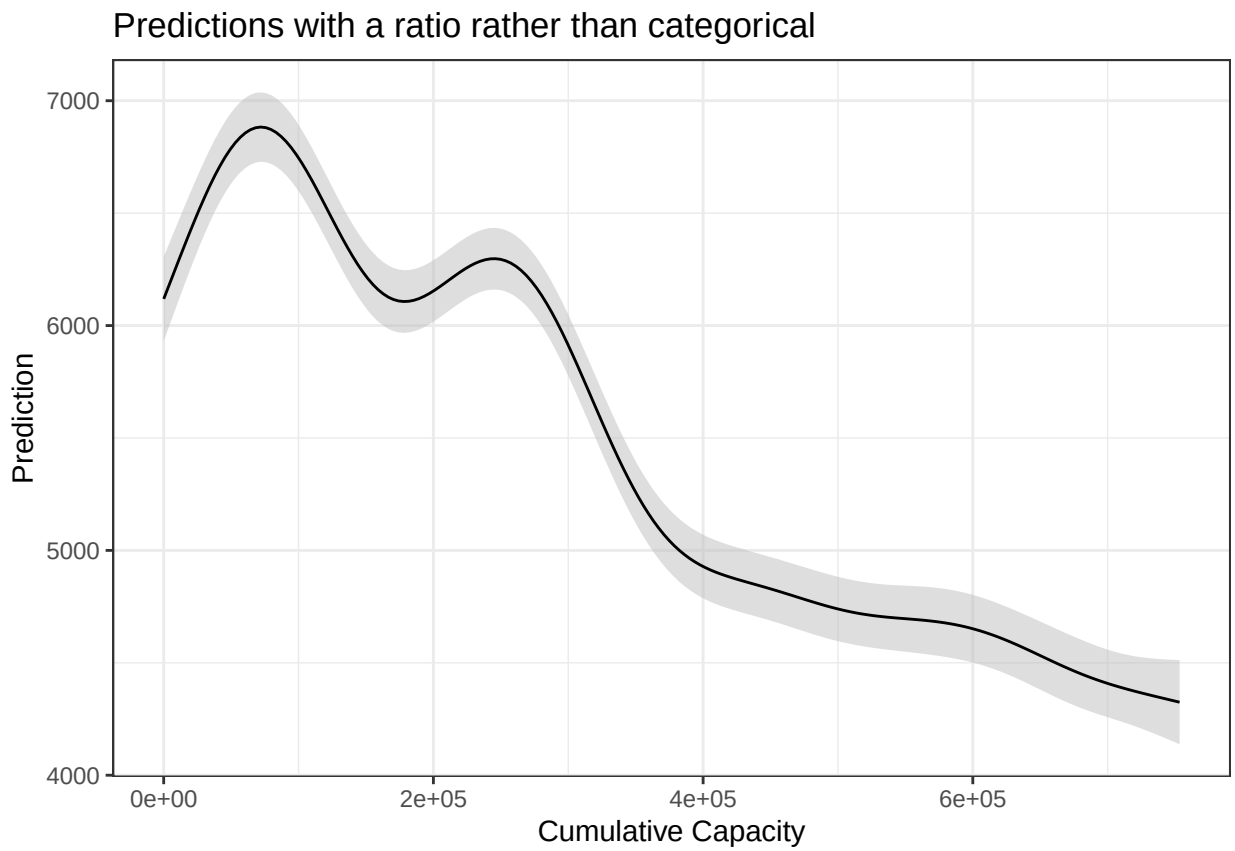
```
##
## Rank: 18/19
## R-sq.(adj) =  0.373   Deviance explained = 37.3%
## GCV = 2.6221e+06  Scale est. = 2.6217e+06  n = 106623
```

###setting up prediction with china ratio value rather than china ratio dummy

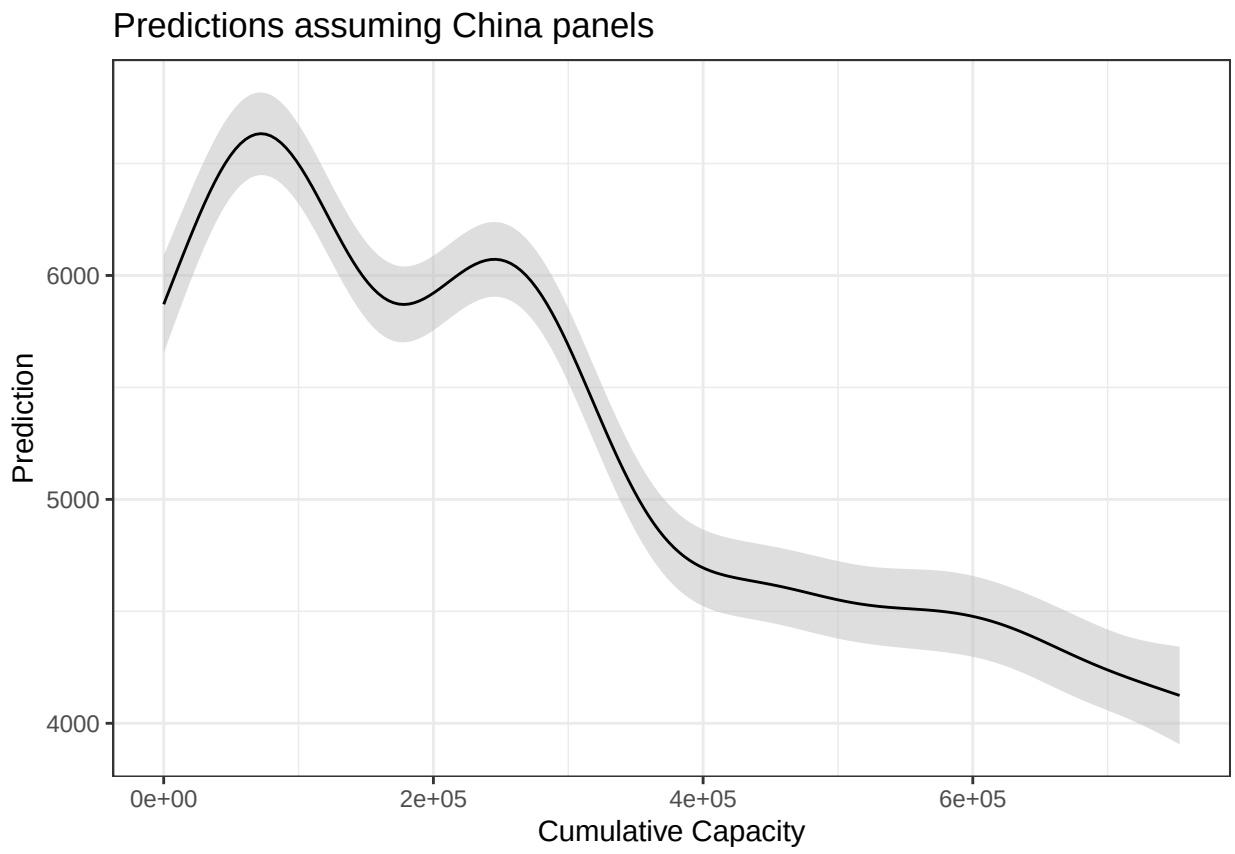
```
pdat4=with(pv_df_final,
list(
cum_cap = round(seq(min(cum_cap), max(cum_cap), length = 200)),
sector = rep("Residential",200),
nameplate = rep(mean(nameplate), 200),
county_year_total_mw = rep(mean(county_year_total_mw), 200),
contractor_year_total_mw = rep(mean(contractor_year_total_mw), 200),
incentive_per_kw = rep(mean(incentive_per_kw/1000), 200),
lease = rep(0, 200),
china_ratio = round(seq(min(china_ratio), max(china_ratio), length = 200))
))

pred4 = predict(gam_mod3, pdat4, type = "terms", se.fit = TRUE)

pred4_fit = as_tibble(pred4$fit)
pred4_fit["intercept"] = coef(gam_mod3)[1]

pred4_fit = pred4_fit %>% mutate(prediction = rowSums(.))

pred4_fit["cum_cap"] = with(pv_df, round(seq(min(cum_cap), max(cum_cap), length = 200)))

pred4_se= as_tibble(pred4$se)
pred4_fit["prediction_se"] = rowSums(pred4_se)
pred4_fit["upper"] = pred4_fit$prediction +  2 * pred4_fit$prediction_se
pred4_fit["lower"] = pred4_fit$prediction - 2* pred4_fit$prediction_se
```

###graphing first our china ratio prediction, then all china prediction, then nochina prediction to compare the three

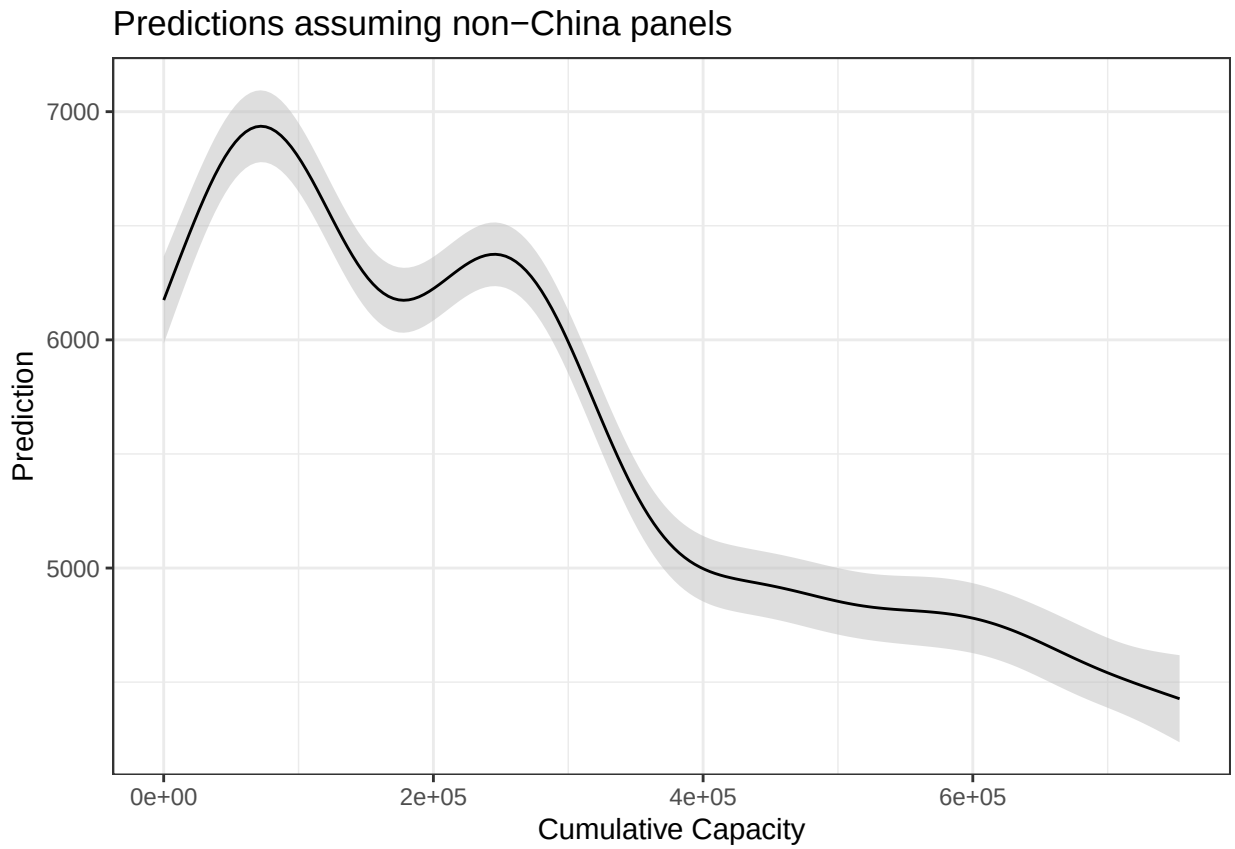
```
ggplot(pred4_fit, aes(x=cum_cap, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Cumulative Capacity", y="Prediction") +
  ggtitle('Predictions with a ratio rather than categorical') +
    theme_bw()
```




```
ggplot(pred3_fit, aes(x=cum_cap, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Cumulative Capacity", y="Prediction") +
  ggtitle('Predictions assuming China panels') +
    theme_bw()
```



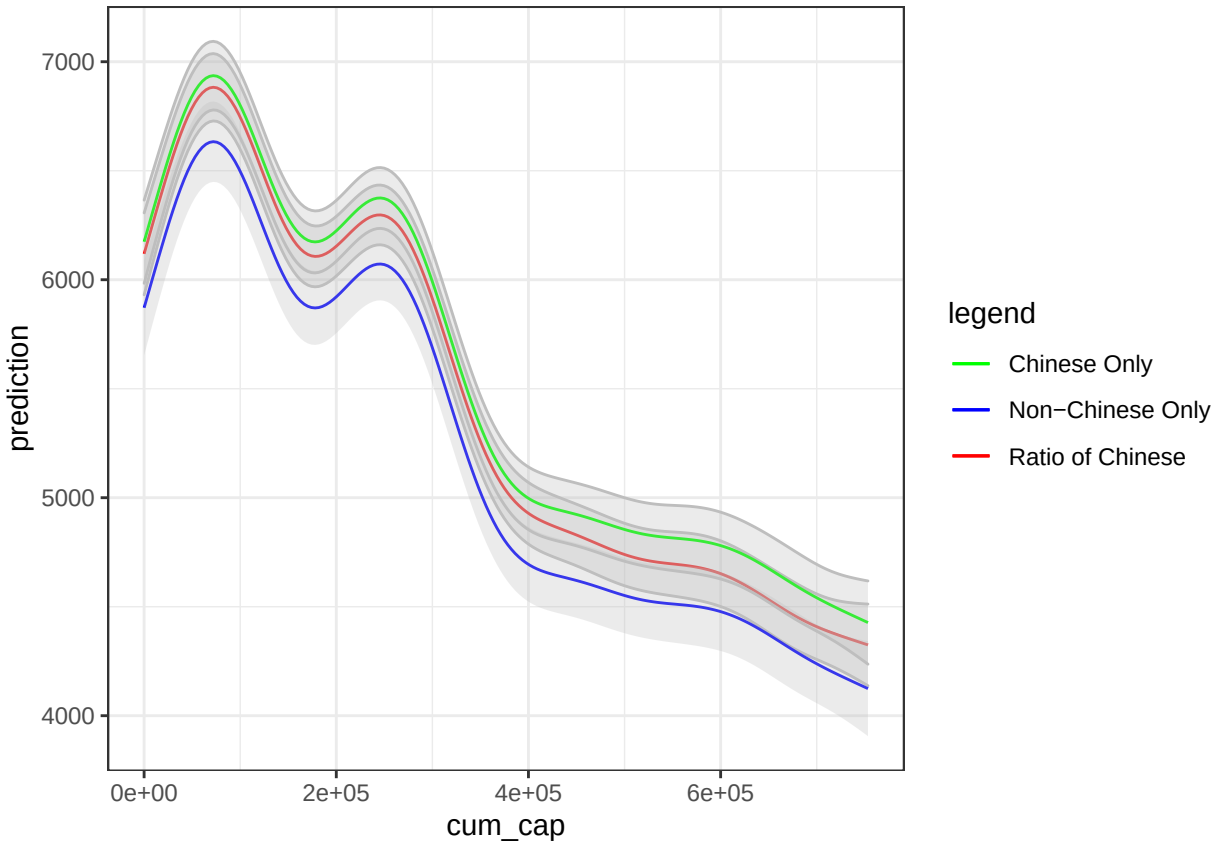

```
ggplot(pred2_fit, aes(x=cum_cap, y=prediction, ymin=upper, ymax=lower))+
    geom_ribbon(alpha=.5, fill="grey") +
    geom_line() +
    labs(x="Cumulative Capacity", y="Prediction") +
  ggtitle('Predictions assuming non-China panels') +
    theme_bw()
```



###All three plotted together

```
colors <- c("Non-Chinese Only" = "blue", "Ratio of Chinese" = "red", "Chinese Only" = "green")
ggplot() +
  geom_line(aes(x=cum_cap, y = prediction, ymin = upper, ymax = lower, color = 'Ratio of Chinese'), data
  geom_ribbon(aes(ymin=lower, ymax=upper, x=cum_cap, fill = "band"), alpha = 0.3, data = pred4_fit, colo
  geom_line(aes(x=cum_cap, y = prediction, ymin = upper, ymax = lower, color = 'Non-Chinese Only'), data
  geom_ribbon(aes(ymin=lower, ymax=upper, x=cum_cap, fill = "band"), alpha = 0.3, data = pred3_fit, fill
  geom_line(aes(x=cum_cap, y = prediction, ymin = upper, ymax = lower, color = 'Chinese Only'), data = p
  geom_ribbon(aes(ymin=lower, ymax=upper, x=cum_cap, fill = "band"), alpha = 0.3, data = pred2_fit, colo
    theme_bw() +
  labs(color = "legend") +
  scale_color_manual(values = colors)
```

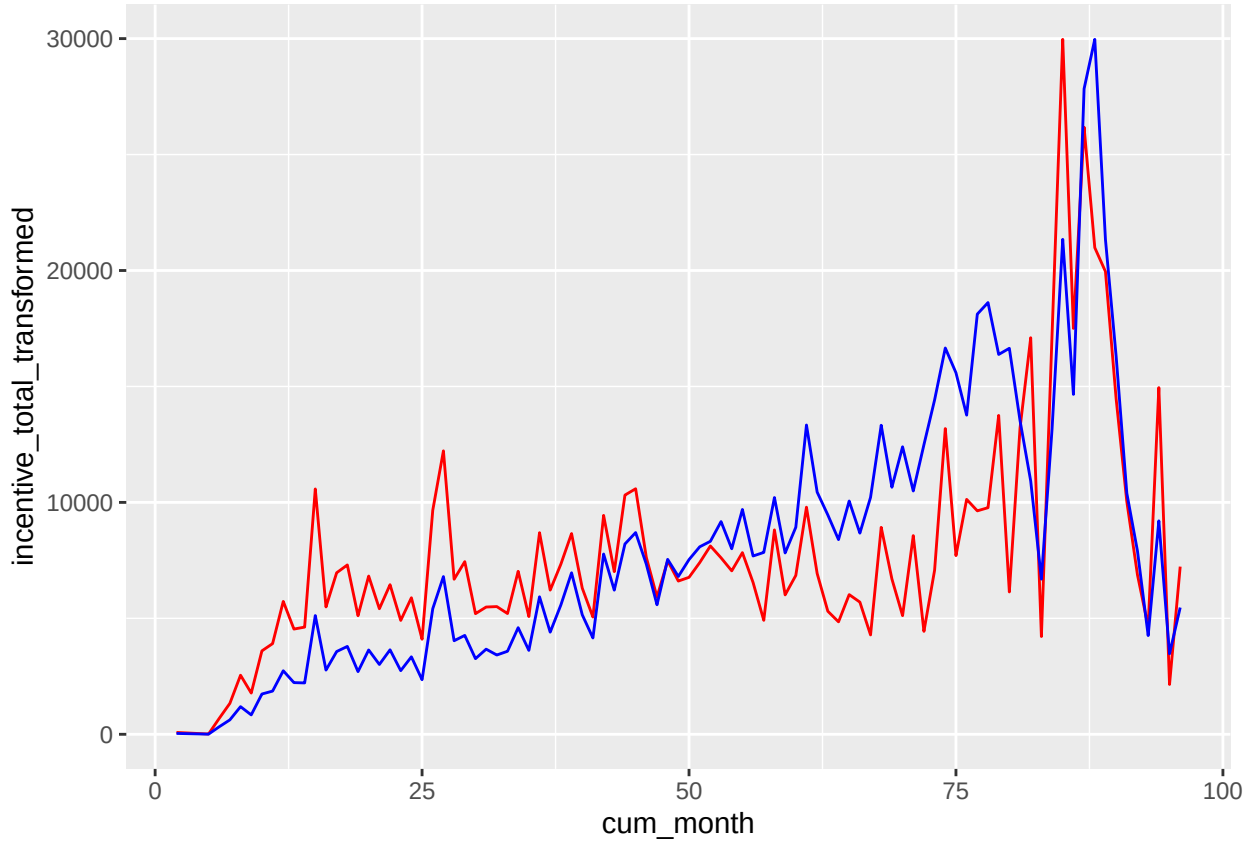
```
## Warning: Ignoring unknown aesthetics: ymin, ymax

## Warning: Ignoring unknown aesthetics: ymin, ymax

## Warning: Ignoring unknown aesthetics: ymin, ymax
```



All this indicates that chinese panels decreased the cost of soalr panels across the time period. As their market penetration increased, the average price of solar panel installation decreased. ##Question 4 Open ended question: What other questions could you answer with this data set? Show in the form figures, regressions, or other estimations. You could also consider downloading updated data here.

###Answer There's a lot of information on incentives in this dataset, so it'd be interesting to look at some of this data. A super prude look would be:

```
pv_df_final %>% group_by(cum_month) %>%
  summarize(
    incentive_total = sum(incentive_amount, na.rm=TRUE),
    new_nameplate = sum(nameplate, na.rm = TRUE)
  ) %>%
  mutate(incentive_total_transformed = max(new_nameplate)/max(incentive_total)*incentive_total) %>%
  ggplot() +
  geom_line(aes(x = cum_month, y = incentive_total_transformed), color = 'red') +
  geom_line(aes(x = cum_month, y = new_nameplate), color = 'blue')
```



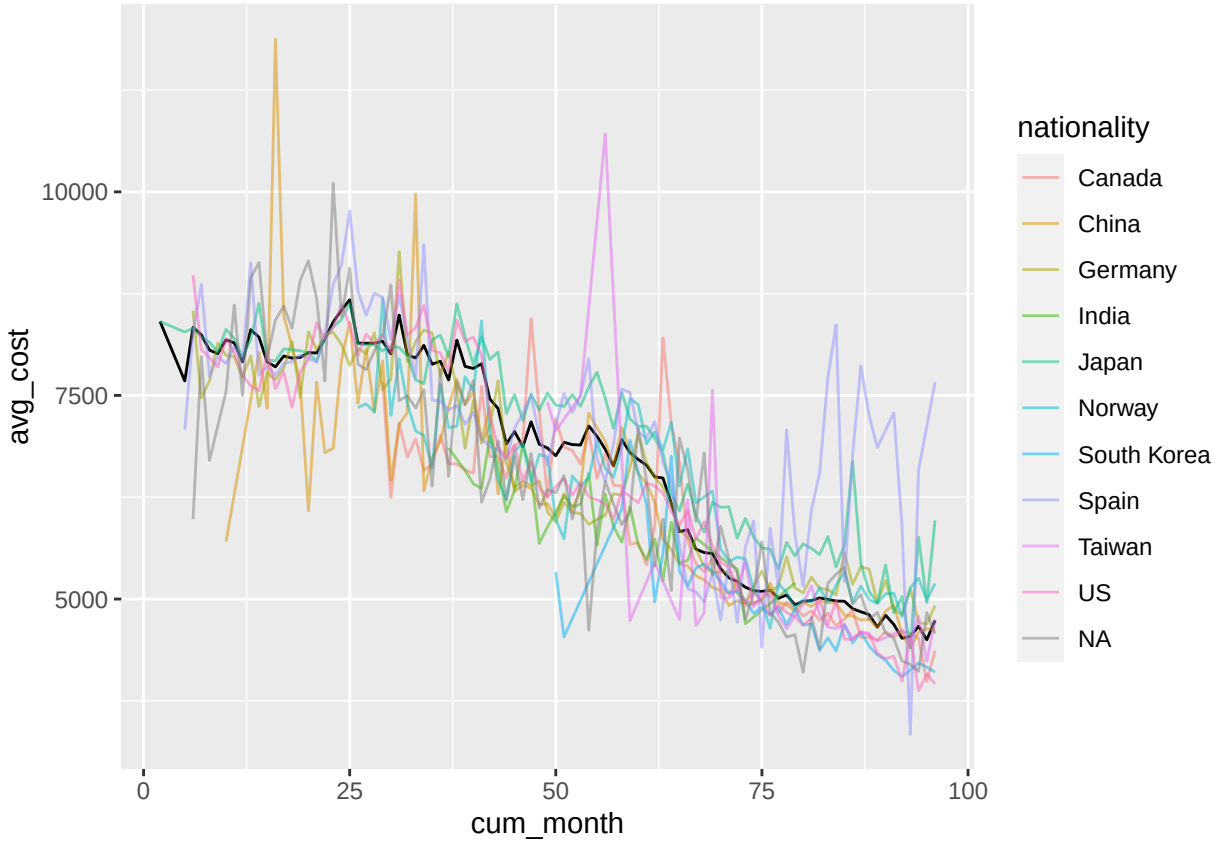
Here we looked at monthly incentives (transformed to be in the same range as nameplate capacity) and nameplate capacity. We can see there is quite precise correlation, as one would expect. However, there are some interesting deviations that 50 months in where people are installing without incentives, then a huge spike ~80 months in.

There's also manufacturer and contractor data that would be interesting to look at!

```
average_cost <- pv_df_final %>% group_by(cum_month) %>%
  summarize(avg_cost = mean(cost_per_kw))
nation_data <- pv_df_final %>% group_by(nationality, cum_month) %>%
  summarize(
    avg_cost = mean(cost_per_kw)
  )
```

```
## `summarise()` has grouped output by 'nationality'. You can override using the `.groups` argument.
```

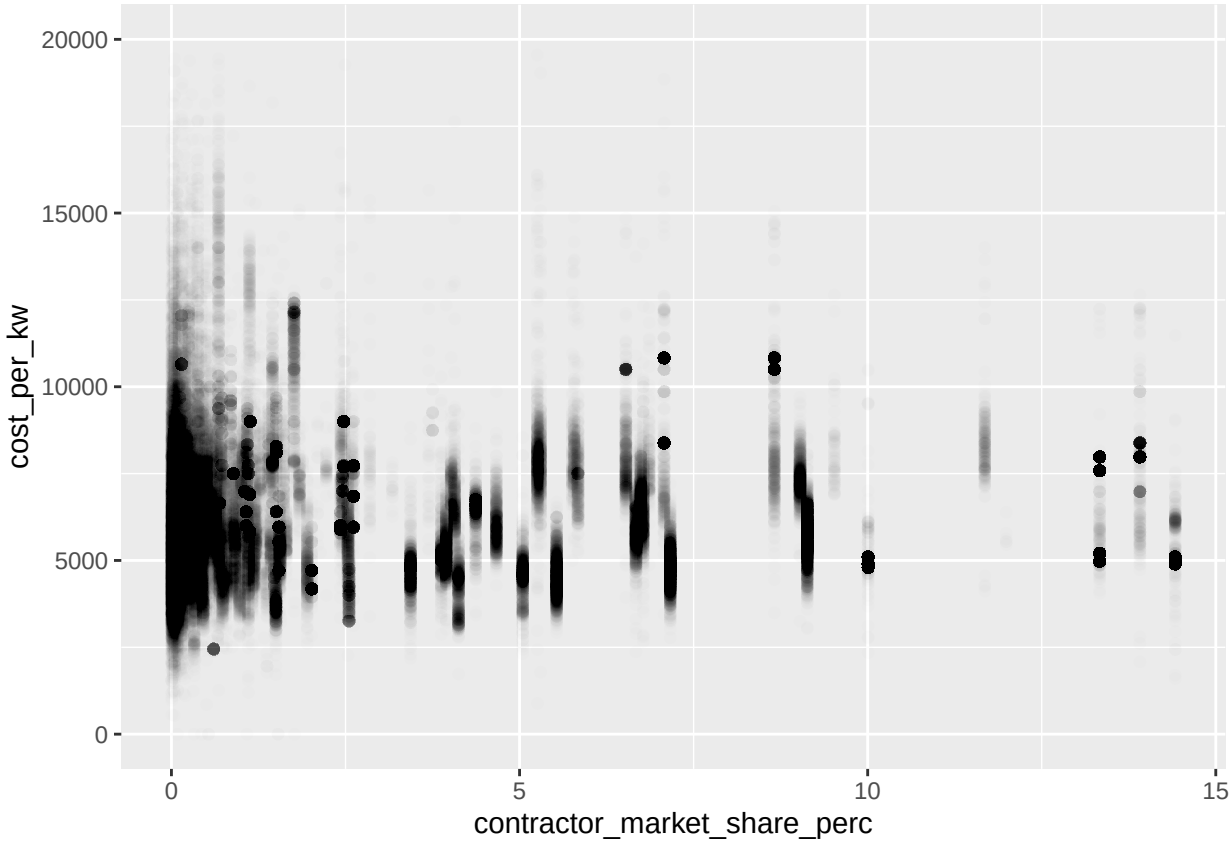
```
ggplot() +
geom_line(aes(x = cum_month, y = avg_cost), data = average_cost) +
geom_line(mapping = aes(x = cum_month, y = avg_cost, color = nationality), data = nation_data, alpha =
```



There's a bunch of fun stuff to look at here, but I have a feeling some of the high's and lows are due to limited data in those periods.

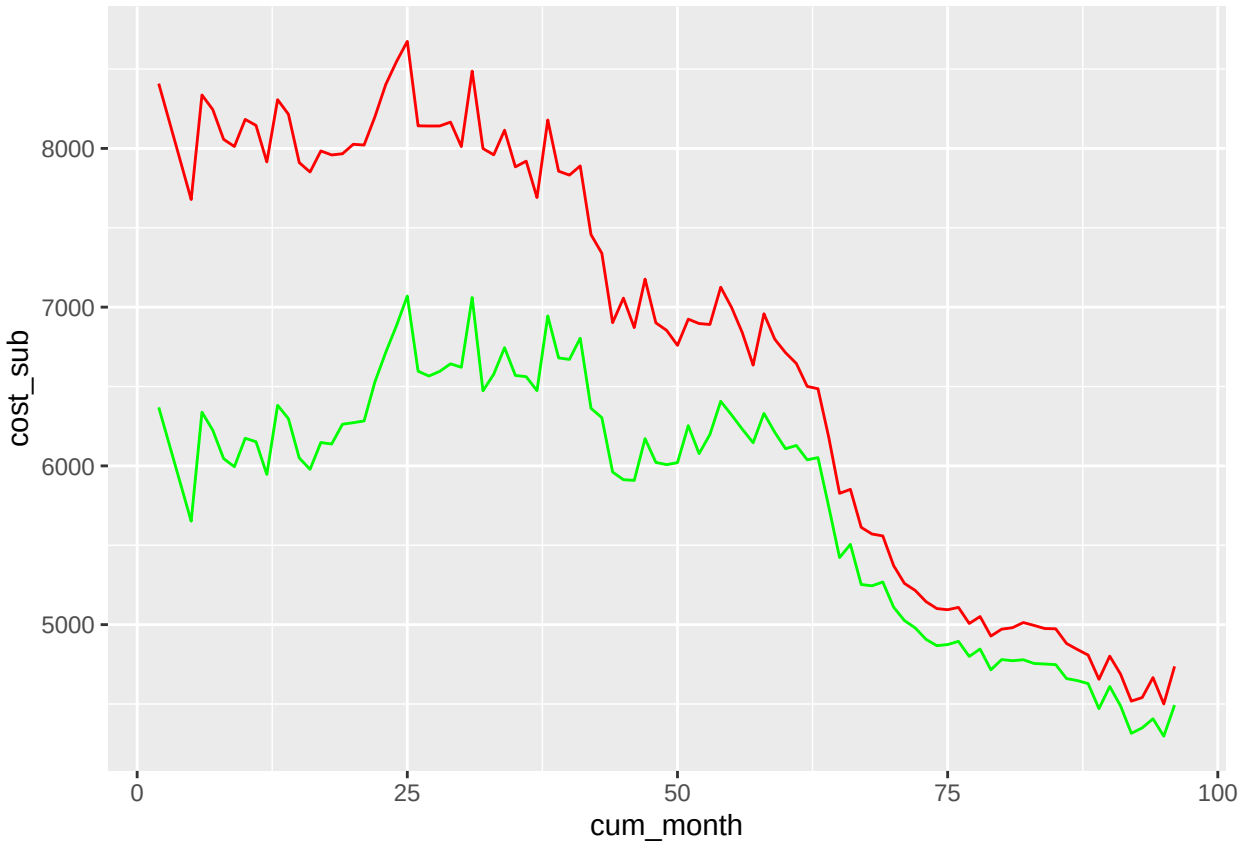
```
pv_df %>%
  ggplot(mapping = aes(x = contractor_market_share_perc, y = cost_per_kw)) +
  geom_point(alpha = .007) +
  scale_y_continuous(limits = c(0,20000))
```

```
## Warning: Removed 31 rows containing missing values (geom_point).
```



This is not that interesting, but could warrant a little extra investigation: are there no economies of scale in solar panel installation?

```
pv_df_final %>%
  group_by(cum_month) %>%
  summarize(
    cost_sub = mean(cost_ex_subsid_per_kw, na.rm=TRUE),
    cost = mean(cost_per_kw, na.rm=TRUE)
  ) %>%
  ggplot() +
  geom_line(aes(x = cum_month, y = cost_sub), color = 'green') +
  geom_line(aes(x = cum_month, y = cost), color = 'red')
```



This one is pretty straightforward, just looking at the significance of subsidies over time. They're not as large as they once were!

##End Note

Thank you for a great class, it was a lot of fun and the assignments were some of the most relevant assignments I've ever done. HAGS -Ethan