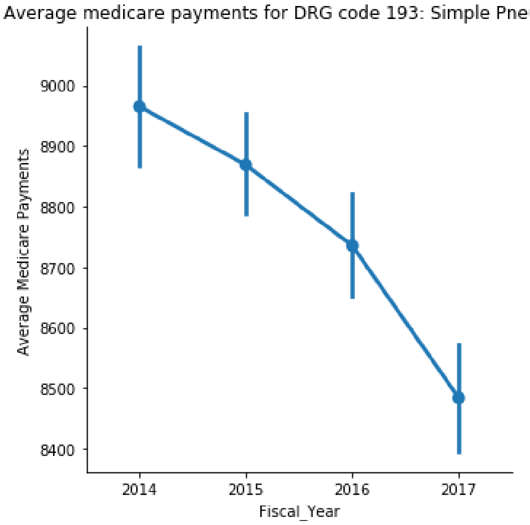
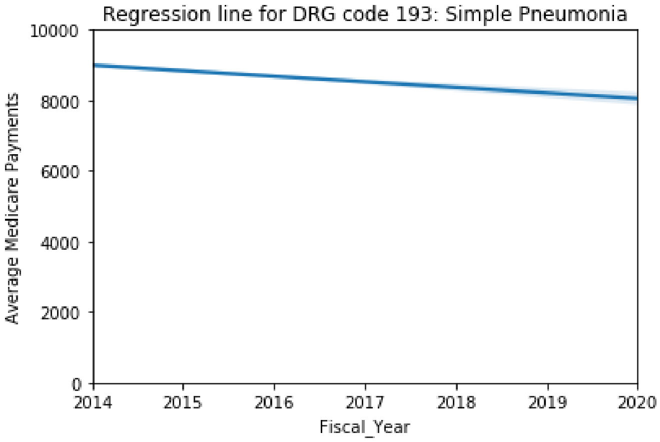
Capstone 1: In depth analysis

I decided to complete a linear regression because I am attempting to predict a value in the future (rather than a category that would have preferenced logistic regression). I figured that a linear regression would be a fairly simple, straightforward way to get useful data like the rate of change for payments as a whole or for an individual diagnosis. I do appreciate that typically, you would want more data points to create a more accurate regression, and that 4 data points is not ideal.

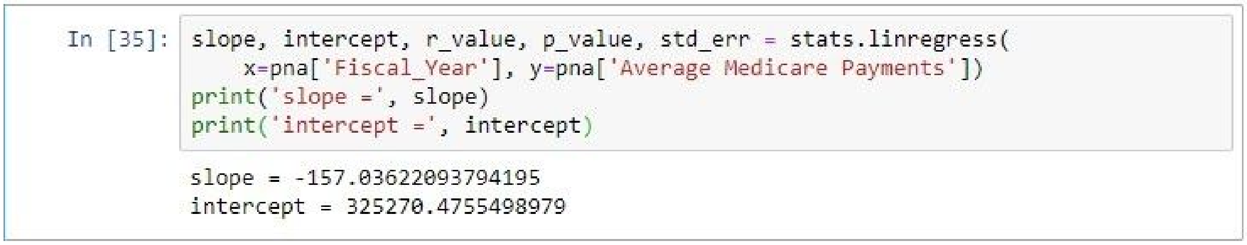
I started by looking at a catplot of the data generated using seaborn to better visualize where I thought a regression line might go. Here is the one for my simple pneumonia plot:



I then plotted a regression line for the data using seaborn to see if they would visually appear consistent:



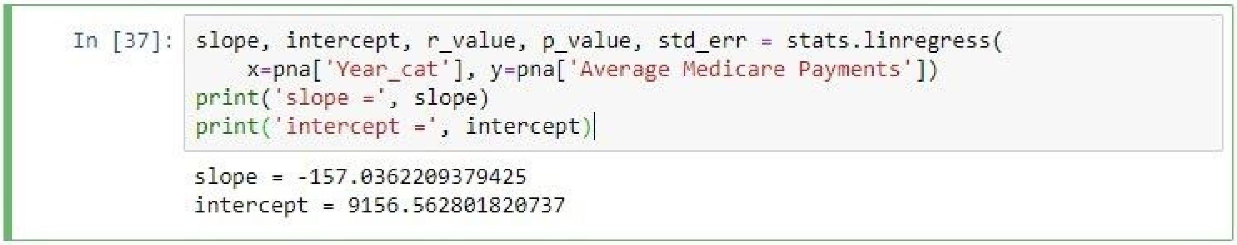
A problem I ran into was realizing that seaborn does not allow you to extract some of the statistical data, such as a slope or intercept. So I calculated those points using scipy.stats.



In [35]:
```
slope, intercept, r_value, p_value, std_err = stats.linregress(
    x=pna['Fiscal_Year'], y=pna['Average Medicare Payments'])
print('slope =', slope)
print('intercept =', intercept)
```

```
slope = -157.03622093794195
intercept = 325270.4755498979
```

I determined that using fiscal year was messing up my calculations, so I created a column to treat the year as a categorical value of 1-4. After I made that change, the results were consistent with the graphical analysis I completed earlier:



In [37]:
```
slope, intercept, r_value, p_value, std_err = stats.linregress(
    x=pna['Year_cat'], y=pna['Average Medicare Payments'])
print('slope =', slope)
print('intercept =', intercept)
```

```
slope = -157.0362209379425
intercept = 9156.562801820737
```

I utilized that same technique for 2 other diagnoses of interest as well as the average medicare payment as a whole. I found that medicare payments are increasing at an average rate of $223 per year. Of the three diagnoses I looked at, Simple pneumonia is decreasing at a rate of ~$160, major joint replacement is decreasing at a rate of ~$164, and heart transplant is increasing at a rate of ~$7,800.

For future exploration, I could utilize a regularization like lasso regression to attempt to determine which one or few diagnoses are most contributing to overall cost or which state(s)/regions/zipcodes are most impacting the average cost. I could also tailor this approach to a given diagnosis or set of diagnoses of interest to a particular hospital or system.