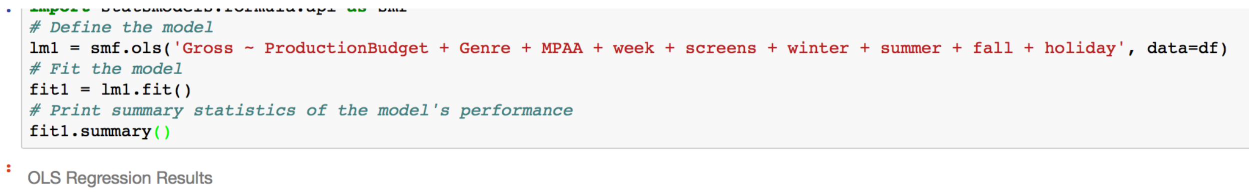
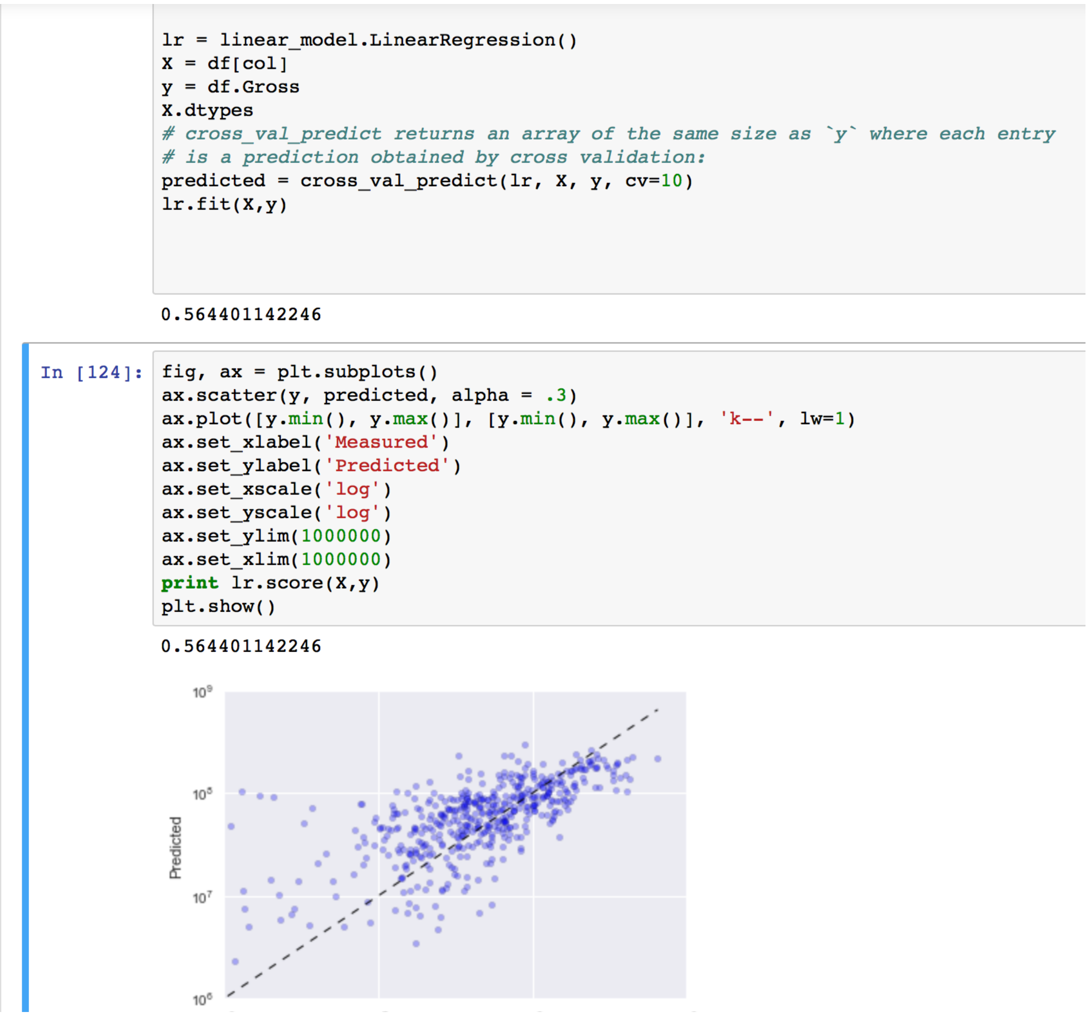
My plan is to predict how much a movie will gross (Domestically only) based on the following features: Genre, Rating, Time of Year, Distributor, Number of Screens, Length of Feature. Data is scraped from BoxofficeMojo.com and numbers.com. My current lr.score is 0.5644.



```
import statsmodels.formula.api as smf
# Define the model
lm1 = smf.ols('Gross ~ ProductionBudget + Genre + MPAA + week + screens + winter + summer + fall + holiday', data=df)
# Fit the model
fit1 = lm1.fit()
# Print summary statistics of the model's performance
fit1.summary()
```

OLS Regression Results



```
lr = linear_model.LinearRegression()
X = df[col]
y = df.Gross
X.dtypes
# cross_val_predict returns an array of the same size as `y` where each entry
# is a prediction obtained by cross validation:
predicted = cross_val_predict(lr, X, y, cv=10)
lr.fit(X,y)
```

```
0.564401142246
```

```
In [124]: fig, ax = plt.subplots()
ax.scatter(y, predicted, alpha = .3)
ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=1)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
ax.set_xscale('log')
ax.set_yscale('log')
ax.set_ylim(1000000)
ax.set_xlim(1000000)
print lr.score(X,y)
plt.show()
```

```
0.564401142246
```