DATASET

Dataset shows per capita state and local public expenditures and associated state demographic and economic characteristics for 48 states (USA) during the year of 1960.

Variables:

- EX – per capita state and local public expenditures
- ECAB – economic ability index (income, retail sales, value of output per capita are equally measured)
- MET – percentage of population living in metropolitan areas
- GROW – percent change in population (1950-1960)
- YOUNG – percent of population between 5-19 years of age
- OLD – percent of population over 65 years of age
- STATE – state name

Example 15 rows:

	ex	ecab	met	grow	young	old	state
0	256	85.5	19.7	6.9	29.6	11.0	'ME'
1	275	94.3	17.7	14.7	26.4	11.2	'NH'
2	327	87.0	0.0	3.7	28.5	11.2	'VT'
3	297	107.5	85.2	10.2	25.1	11.1	'MA'
4	256	94.9	86.2	1.0	25.3	10.4	'RI'
5	312	121.6	77.6	25.4	25.2	9.6	'CT'
6	374	111.5	85.5	12.9	24.0	10.1	'NY'
7	257	117.9	78.9	25.5	24.8	9.2	'NJ'
8	257	103.1	77.9	7.8	25.7	10.0	'PA'
9	336	116.1	68.8	39.9	26.4	8.0	'DE'
10	269	93.4	78.2	31.1	27.5	7.3	'MD'
11	213	77.2	50.9	21.9	28.8	7.3	'VA'
12	308	108.4	73.1	22.2	28.0	8.2	'MI'
13	273	111.8	69.5	21.8	26.9	9.2	'OH'
14	256	110.8	48.1	18.3	27.5	9.6	'IN'

The dataset was fairly clean. Just small adjustments were made, such as modifying the type of a few instances from string to double (negative values were strings in the initial dataset). We decided to drop the State column as it will not be used in the analysis.

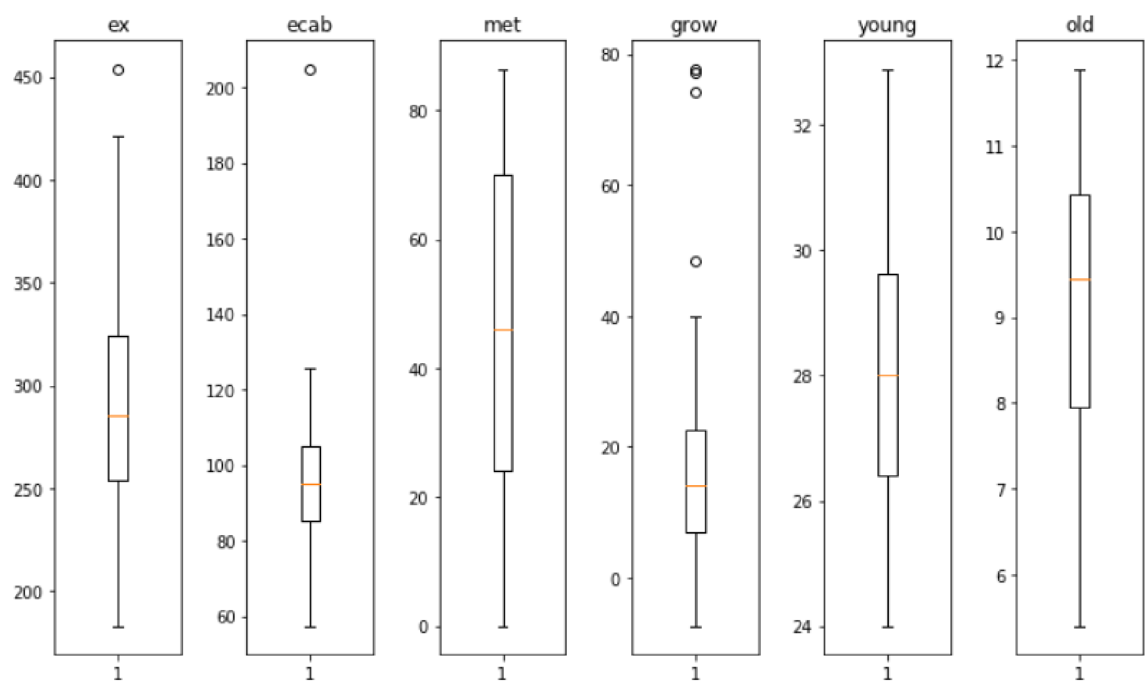
ANALYSIS

The purpose of the analysis is to build a regression model where public expenditures is the dependant variable. We want to understand how public expenditures change based on different characteristics.

Summary statistics for our variables.

	ex	ecab	met	grow	young	old
count	48.000000	48.000000	48.000000	48.000000	48.000000	48.00000
mean	286.645833	96.754167	46.168750	18.729167	28.114583	9.21250
std	58.794807	22.252831	26.938797	18.874749	2.148526	1.63936
min	183.000000	57.400000	0.000000	-7.400000	24.000000	5.40000
25%	253.500000	85.400000	24.100000	6.975000	26.400000	7.95000
50%	285.500000	95.300000	46.150000	14.050000	28.000000	9.45000
75%	324.000000	105.100000	69.975000	22.675000	29.625000	10.42500
max	454.000000	205.000000	86.500000	77.800000	32.900000	11.90000

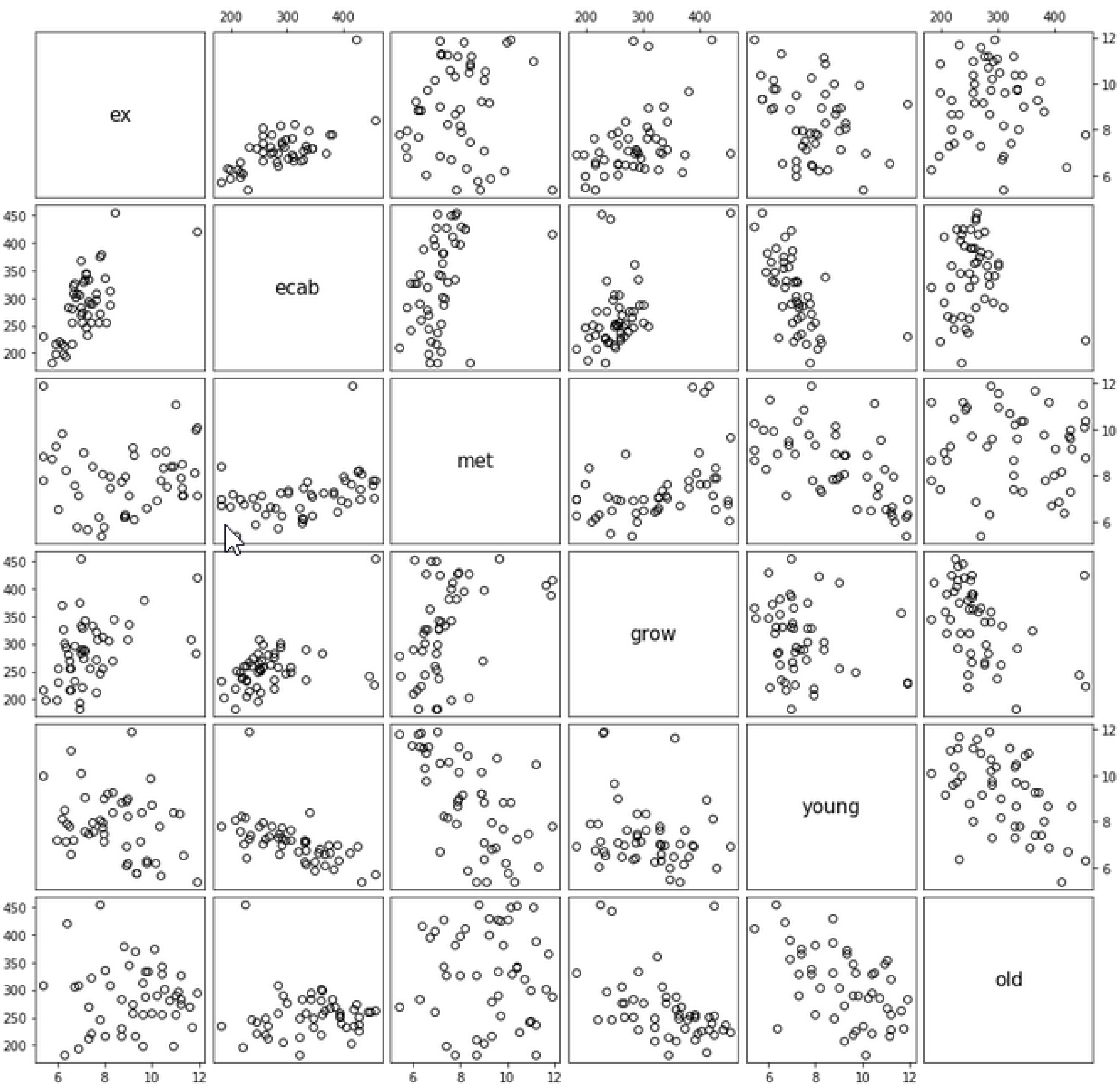
We start by plotting box plots for each of the variables in our dataset.



We notice that some variables have outliers (ex, ecab, grow). Also, some variables have asymmetric whiskers. These variables could potentially benefit from transformations if we want to achieve more normally distributed values for such variable.

Next we're going to look at a Scatterplot Matrix for our variables.

Scatterplot Matrix



The biggest correlation seems to be between ex and ecab variables.

We confirm this fact by looking at a correlation table which shows correlations between our variables.

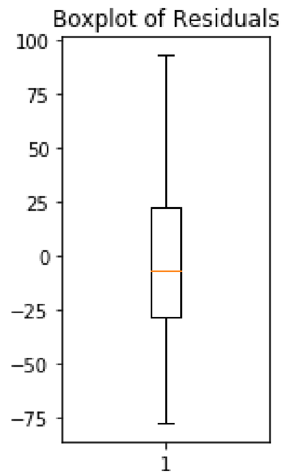
	ex	ecab	met	grow	young	old
ex	1.000000	0.655863	0.045235	0.405287	-0.293197	-0.023396
ecab	0.655863	1.000000	0.408926	0.460072	-0.589468	-0.044496
met	0.045235	0.408926	1.000000	0.404023	-0.626280	-0.041053
grow	0.405287	0.460072	0.404023	1.000000	-0.204488	-0.412582
young	-0.293197	-0.589468	-0.626280	-0.204488	1.000000	-0.524929
old	-0.023396	-0.044496	-0.041053	-0.412582	-0.524929	1.000000

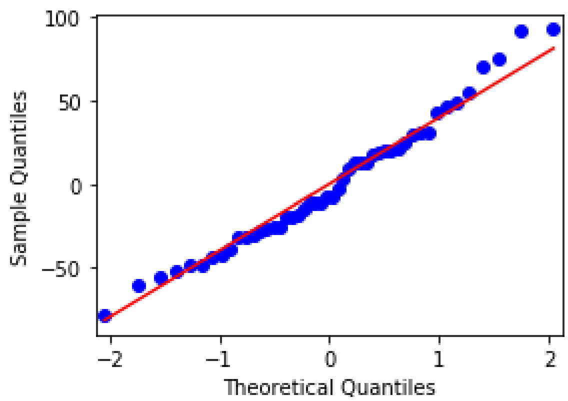
FITTING INITIAL LINEAR REGRESSION MODEL

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     ex   R-squared:                       0.535
Model:                            OLS   Adj. R-squared:                  0.479
Method:                 Least Squares   F-statistic:                     9.647
Date:                Tue, 04 Dec 2018   Prob (F-statistic):           3.47e-06
Time:                        17:15:44   Log-Likelihood:                -244.80
No. Observations:                  48   AIC:                             501.6
Df Residuals:                      42   BIC:                             512.8
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         67.2588    303.756      0.221      0.826    -545.745     680.263
ecab           1.8007      0.430      4.189      0.000       0.933       2.668
met           -0.7074      0.375     -1.886      0.066      -1.464       0.049
grow           0.8690      0.436      1.994      0.053      -0.010       1.748
young          0.7436      7.353      0.101      0.920     -14.094      15.582
old            4.4110      7.145      0.617      0.540     -10.009      18.831
==============================================================================
Omnibus:                        1.850   Durbin-Watson:                   2.086
Prob(Omnibus):                  0.396   Jarque-Bera (JB):                1.741
Skew:                           0.441   Prob(JB):                        0.419
Kurtosis:                       2.697   Cond. No.                     5.75e+03
==============================================================================
```

Model is relatively well fit. Adjusted R squared is decently sized. The worrying fact is that only one variable seems to be statistically significant.

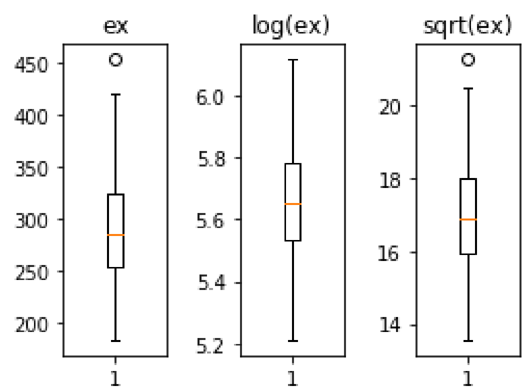
Let's look at the box plot, Normal Q-Q plot of residuals to confirm that residuals are nearly normally distributed.

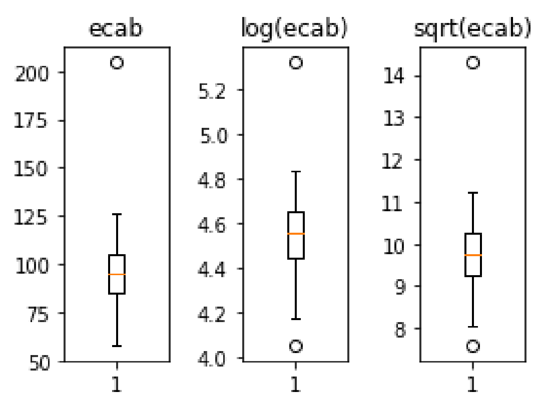


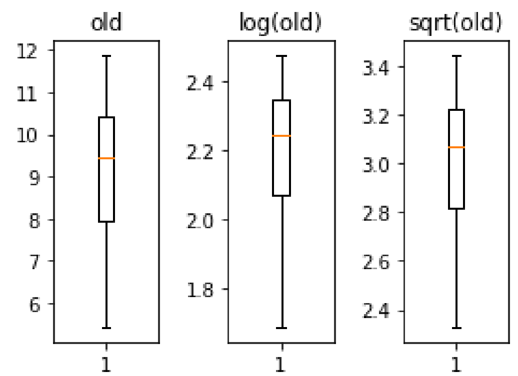


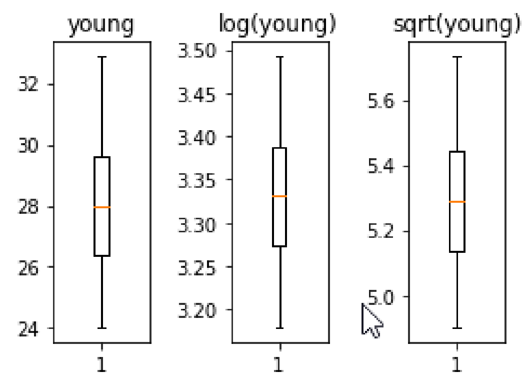
Looks like the residuals are nearly normally distributed. So the assumptions of the regression model are satisfied. Just to make sure, we ran a Shapiro-Wilk normality test. P-value was equal to 0.365, which is higher than 0.05, so we do not have a reason to reject the null hypothesis about the normality of residuals.

Next we will see if transformation of certain variables may remove outliers and improve the distribution of said variables.









Transforming ex variable into log(ex) seems to have made the box plot look very close to a box plot for a normal distribution. We don't see any improvements for the rest of the variables.

We'll train the regression model after transforming the ex variable.

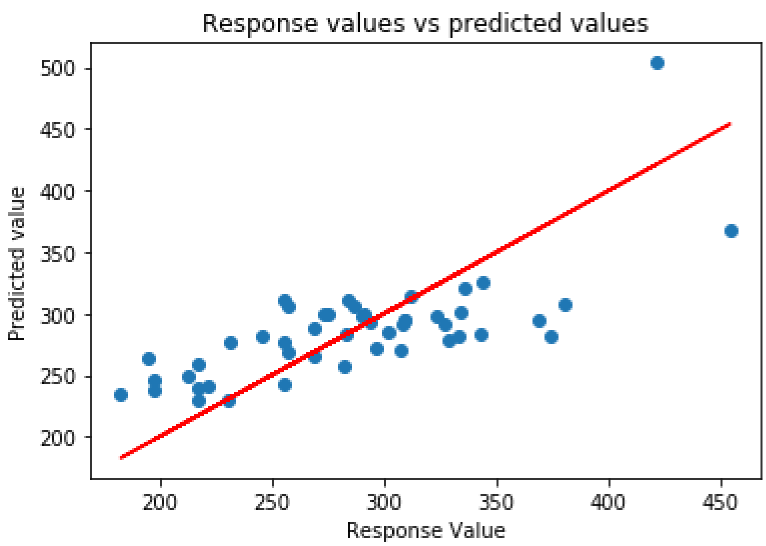
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     ex   R-squared:                       0.515
Model:                            OLS   Adj. R-squared:                  0.458
Method:                 Least Squares   F-statistic:                     8.936
Date:                Tue, 11 Dec 2018   Prob (F-statistic):           7.67e-06
Time:                        15:32:10   Log-Likelihood:                 25.883
No. Observations:                  48   AIC:                            -39.77
Df Residuals:                      42   BIC:                            -28.54
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.7671      1.080      4.414      0.000       2.588       6.947
ecab           0.0059      0.002      3.891      0.000       0.003       0.009
met           -0.0022      0.001     -1.626      0.112      -0.005       0.001
grow           0.0034      0.002      2.213      0.032       0.000       0.007
young          0.0038      0.026      0.145      0.886      -0.049       0.057
old            0.0244      0.025      0.961      0.342      -0.027       0.076
==============================================================================
Omnibus:                        1.139   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.566   Jarque-Bera (JB):                0.949
Skew:                           0.064   Prob(JB):                        0.622
Kurtosis:                       2.323   Cond. No.                     5.75e+03
==============================================================================
```

Adjusted R squared is lower than the one from the initial model, we decide to go back to the initial model and remove variables that are statistically insignificant based on t-vaue (young old).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     ex   R-squared:                       0.525
Model:                            OLS   Adj. R-squared:                  0.493
Method:                 Least Squares   F-statistic:                     16.24
Date:                Tue, 04 Dec 2018   Prob (F-statistic):           3.01e-07
Time:                        17:10:58   Log-Likelihood:                -245.27
No. Observations:                  48   AIC:                             498.5
Df Residuals:                      44   BIC:                             506.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        130.8154     28.136      4.649      0.000      74.111     187.520
ecab           1.8161      0.321      5.650      0.000       1.168       2.464
met           -0.7090      0.258     -2.751      0.009      -1.228      -0.190
grow           0.6862      0.378      1.815      0.076      -0.076       1.448
==============================================================================
Omnibus:                        1.069   Durbin-Watson:                   2.036
Prob(Omnibus):                  0.586   Jarque-Bera (JB):                1.080
Skew:                           0.326   Prob(JB):                        0.583
Kurtosis:                       2.661   Cond. No.                         524.
==============================================================================
```

Adjusted R squared has improved slightly to 0.493.

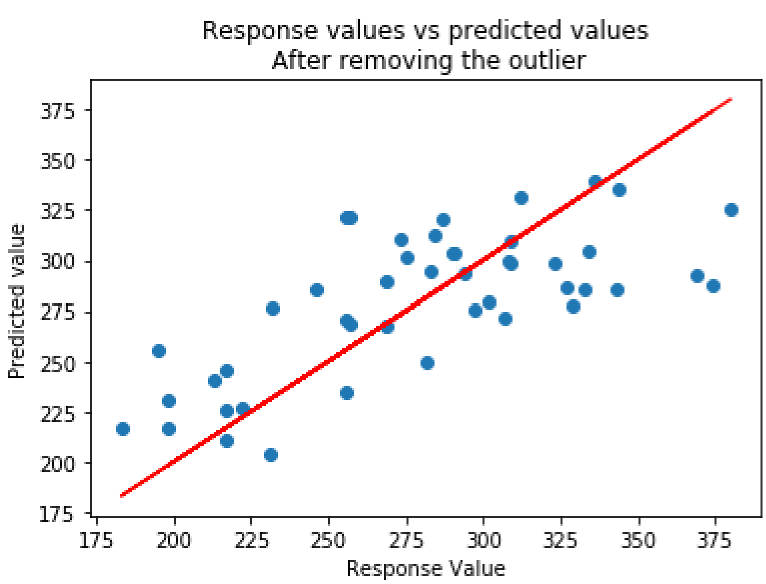
Next we look at the plot of predicted values based on response values.



We notice that one of the points in an outlier that may be pulling the line upwards, which is most likely skewing the model. Let's try to train the model again after removing this point.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     ex   R-squared:                       0.580
Model:                            OLS   Adj. R-squared:                  0.550
Method:                 Least Squares   F-statistic:                     19.76
Date:                Tue, 04 Dec 2018   Prob (F-statistic):           3.34e-08
Time:                        17:11:39   Log-Likelihood:                -234.98
No. Observations:                  47   AIC:                             478.0
Df Residuals:                      43   BIC:                             485.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         51.3560     34.256      1.499      0.141     -17.727     120.439
ecab           2.7286      0.392      6.956      0.000       1.938       3.520
met           -0.9521      0.242     -3.940      0.000      -1.439      -0.465
grow           1.0359      0.354      2.927      0.005       0.322       1.750
==============================================================================
Omnibus:                        0.738   Durbin-Watson:                   1.817
Prob(Omnibus):                  0.691   Jarque-Bera (JB):                0.840
Skew:                           0.226   Prob(JB):                        0.657
Kurtosis:                       2.525   Cond. No.                         680.
==============================================================================
```

Adjusted R squared has increased to 0.55 and is the highest yet.



Plot of the Response values versus predicted values has improved as well. We decide to keep this as our final model.

CONCLUSIONS

Final Model

EX = 51.36 + 2.72 * **ECAB** – 0.95 **MET** + 1.03 * **GROW**

- ECAB has the biggest influence on public expenditures. The higher the economic ability of a state, the better economical shape of a state (higher income/sales) the more money a state will have to spend.
- Increase in percentage of people in metropolitan areas decreases the public expenditures. Better infrastructure within the city, where there's higher density of population may require lower overall cost per capita.
- Higher percent change in population (last 10 years) causes higher public expenditures. Sudden population increase may cause temporary increases in public expenditures (until various public services are adjusted to a higher amount of people).