Predicting Economic Growth

Bruce Stoutenburg

December 14, 2018

Introduction

In this data project, I am attempting to predict future economic growth with current economic indicators. There is currently an immense amount of time and money spent by governments, organizations and individuals in an attempt to accurately predict future economic growth. The predictions that current methods produce are often wrong and they rarely even agree across methodologies; therefore, finding a good predictor of economic activity will be of tremendous worth. The methodologies currently used include projection onto historical growth, regression, qualitative assessments etc., but no method has emerged as being consistently accurate (we can clearly see this is true due to the fact that many methods still exist).

Data

For my analysis I will use data obtained from The World Bank (GDP growth is the metric I will use to measure economic growth). The World Bank is well regarded as an accurate source of data, and it measures and makes public over 1500 features for every country in the world. These features include GDP growth (my variable of interest), foreign investments, education levels etc. The following code snippet gives a brief overview of the steps I took to obtain the data. Anyone can follow the link in the code to download the zip file containing the csv files I need, but to automate the task, I will use selenium, and I will also use Python's zipfile package to unpack the files.

```
from selenium import webdriver
#followinng this link automatically downloads a zipped folder
driver = webdriver.Chrome(r'C:\Users\bstoutenburg\Downloads\chromedriver.exe')
driver.get('http://databank.worldbank.org/data/download/WDI_csv.zip')

#extract the files from the downloads folder
import zipfile
zip_ref = zipfile.ZipFile(r'C:\Users\bstoutenburg\Downloads\WDI_csv.zip', 'r')
zip_ref.extractall(r'C:\Users\bstoutenburg\Downloads\ACME\Math403\Data_Project')
zip_ref.close()
```

Once the data is downloaded I will make several groups of countries that are likely to have similar drivers of growth. These groups will be created based on government type. It's not always obvious how to classify the government of a country (for example some people will classify China as communist while others will disagree); therefore, to avoid any personal bias (although bias may still be present in the outside data sources I choose to use), I will categorize countries based

on the classification given by the CIA at https://www.cia.gov/library/publications/the-world-factbook/fields/2128.html. By using the CIA data I hope to have accurate data that reflects a well thought out methodology for classification.

In the following code snippet, I get the source html from the web-page listing government types by country, put it into a BeautifulSoup format, and use the built in functionality of BeautifulSoup as well as regular expressions to build a dictionary mapping countries to government types. In the snippet I also reference the variable named 'index', which I defined previously (see auxiliary files) and it is a list of all country names in the data set from The World Bank. I use an abbreviation scheme to name my governments where a 'pr' at the beginning indicates presidential, 'pa' indicates parliamentary, 'de' and 're' indicate democracy and republic respectively, 'co_st' indicates a communist state, 'co_mo' indicates a constitutional monarchy, 'ab_mo' indicates an absolute monarchy, and 'other' indicates all government types not accounted for.

```
resp=requests.get('https://www.cia.gov/library/publications/...')
soup=BeautifulSoup(resp.text,'html.parser')
table=soup.findAll('td')
pres_rep=re.compile('president[\s\S]*republic')
pres_dem=re.compile('president[\s\S]*democ')
parl_dem=re.compile('parliam[\s\S]*democ')
parl_rep=re.compile('parliam[\s\S]*republic')
cons_mon=re.compile('consti[\s\S]*monarchy')
comm_sta=re.compile('communist')
abs_mona=re.compile('absolute monarchy')
gov_dict={}
for i in range(0,len(table),2):
    if table[i].text.strip('\n') in index:
        #In this if statement I iterate through the regular expressions
        #and build the dictionary entry by entry see auxiliary files for details
```

```
In [8]: gov_dict

Out[8]: {'Afghanistan': 'pr_re',
         'Albania': 'pa_re',
         'Algeria': 'pr_re',
         'American Samoa': 'pr_de',
            ...
         'Vietnam': 'co_st',
         'Zambia': 'pr_re',
         'Zimbabwe': 'pr_re'}
```

The actual government types are a little more nuanced than the seven types I categorize countries into, for example The Ukraine is listed as a semi-presidential republic, but the regular expressions I use to identify government types will map The Ukraine to a presidential republic.

Cleaning and Engineering

Initially, the data is far from ready for analysis. I have to break up the countries into government groups based on the dictionary I defined previously, reformat the data so that columns correspond to features and rows correspond to observations, then I need to eliminate the large amounts of NaNs. Below is a snippet of the data before cleaning.

```
In [4]: data.head()
```

```
Out[4]:   Country Name      Country Code   Indicator Name  ...  2016       2017
        0   Arab World          ARB     2005 PPP conversion...  NaN        NaN
        1   Arab World          ARB     2005 PPP conversion...  NaN        NaN
        2   Arab World          ARB     Access to clean fue...  84.570425  NaN
        3   Arab World          ARB     Access to electrici...  88.768654  NaN
        4   Arab World          ARB     Access to electrici...  78.958780  NaN
```

I will use the pandas.DataFrames.Groupby method to group the World Bank Data by country name, then iterate through these groups and, using gov_dict defined above, make eight lists of DataFrames (one for each government type plus anything not included in the seven main types), then I will concatenate the DataFrames lists which gives me eight DataFrames. Once I have the groups, I will eliminate features that are similar (for example one feature gives the percent of the population that is literate, and then there are other features that give the percent of various age categories that are literate, percentage of genders that are literate etc.). I will then go through the groups and drop features that have abnormally high amounts of NaNs. I also drop all years before 2004 since older years tend to have more sparse data. Following is a snippet of data that comes from the other group.

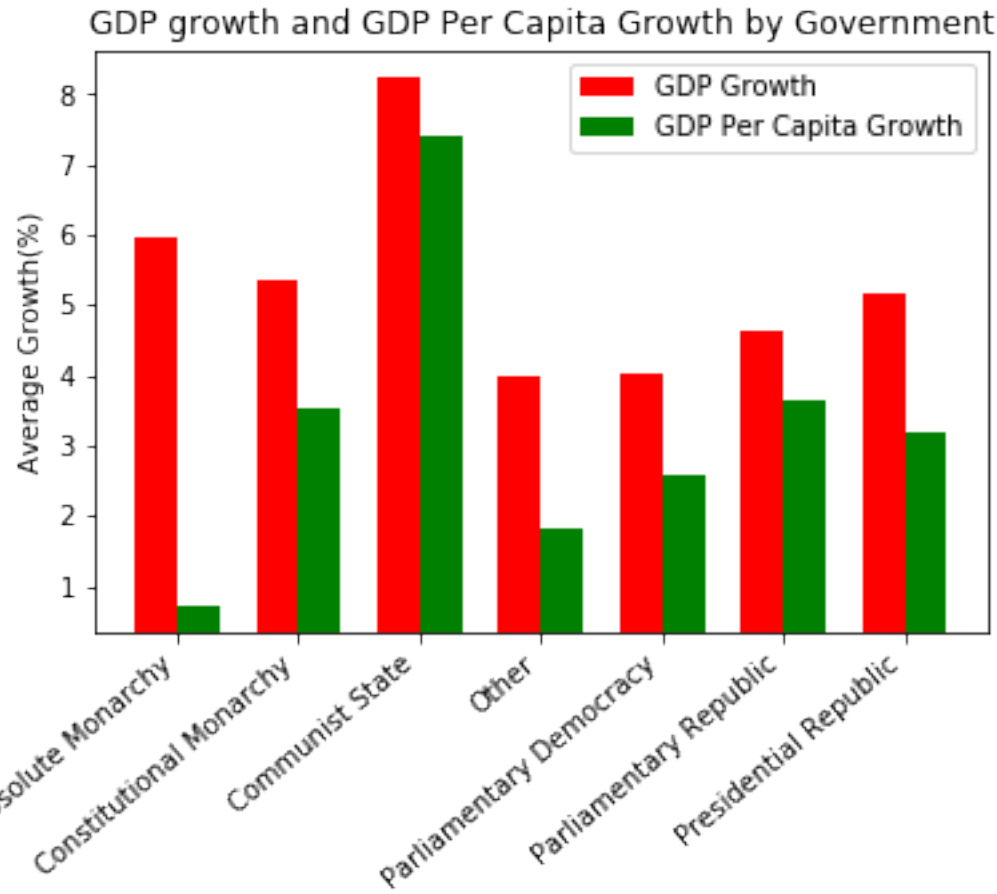
```
In [4]: other.head()
```

```
Out[4]:         Year    GDP growth  GDP per capita growth ... Merchandise trade (% of GDP)
        0        2004   4.461630    2.835457              ...        80.719013
        1        2005   11.870729   10.172609             ...        79.090891
        2        2006   6.500547    4.893524              ...        84.241930
        3        2007   6.352317    4.768227              ...        79.540921
        4        2008   2.667356    1.264395              ...        81.764244
```

At this point the data has come down from originally having over 26 million points (including NaNs), down to about 50,000 points now. Every single country has no NaNs, and every feature is shared by all countries in its group. Also note, I made a decision to combine the 'presidential democracy' group with 'other' because after making the eliminations there were very few observations in the 'presidential democracy' group, so I now have seven groups.

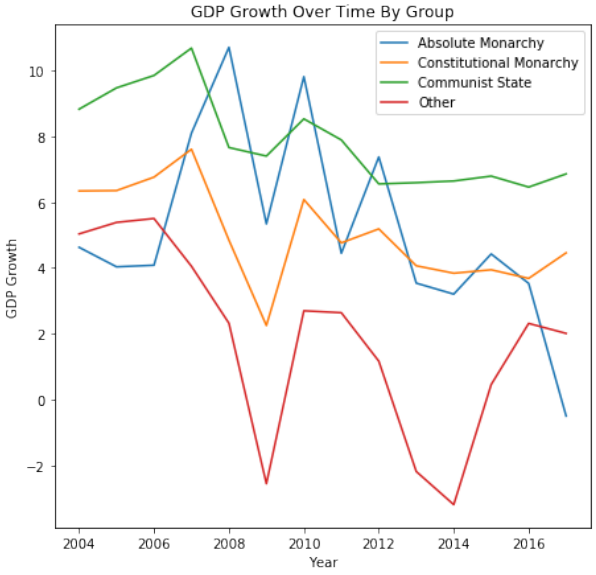
Analysis and Visualization

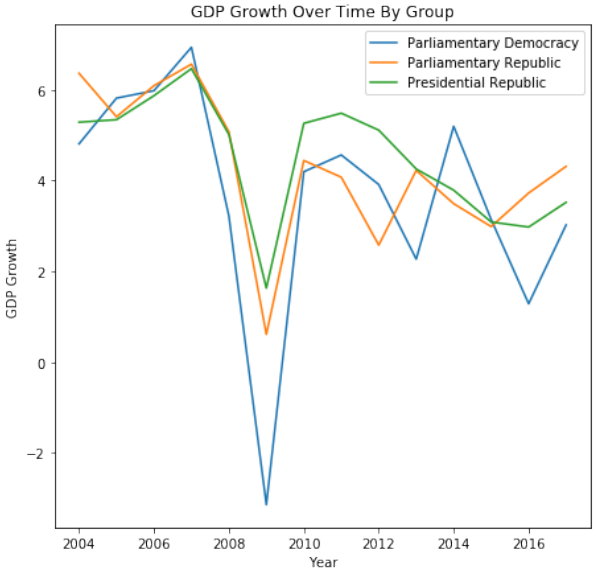
I am ready to analyze the data. First, I want to see how GDP growth looks for each group. I can easily visualize average GDP growth and GDP Per Capita growth by group through a grouped bar chart. I want to know if some groups are growing more than other groups. If so, can I identify any reasons why a group is growing more or less than the others?

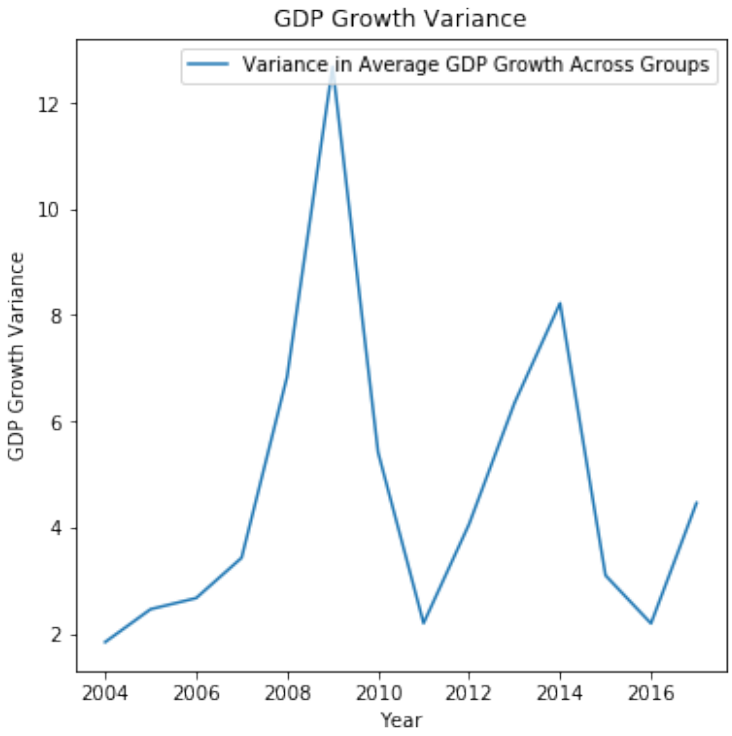


Immediately, it is noticeable that all the groups have positive growth for both GDP and GDP Per Capita. The Absolute Monarchy group provides one interesting feature. This group has the second highest Average GDP growth but the lowest GDP Per Capita growth. Does this indicate that a country can greatly expand its GDP by encouraging population growth at the expense of individuals not making significant economic gains? Is there a confounding factor that causes this. Reverse causality? Or is this just a statistical anomaly?

Now that I know how the economies grow on average, I would like to see how the growth from year to year compares across groups. Is growth correlated? To help answer this question, I will plot average growth for each group by year. To help with visualization I will split the results into two graphs (a single graph results in an unreadable mess of lines), and to further help understand the relationship across groups, I will plot the variance in average GDP growth across groups for each year.





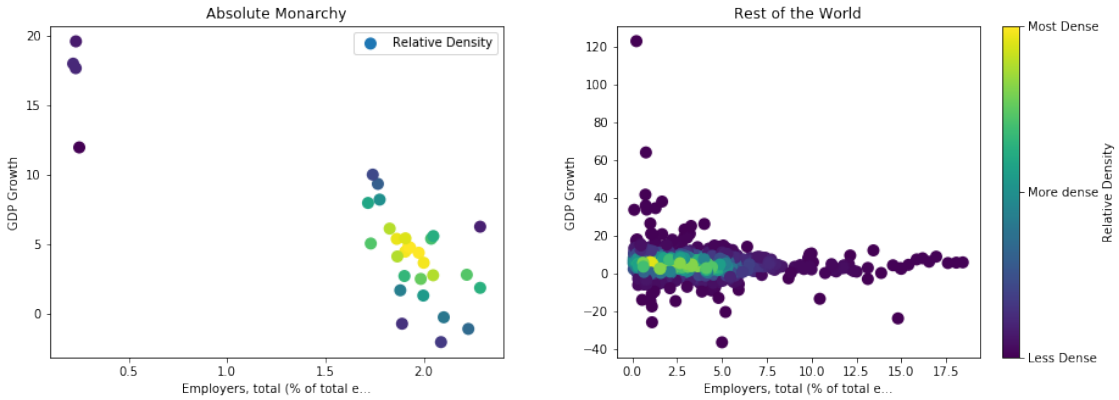


We can see here that there is some correlation, but change in GDP growth is not constant across groups. There are some times when there is greater variance in the data; for example, during the financial crisis in 2008-2009, average GDP growth did tend to go down in all groups, but variance across groups shot way up, which indicates that some groups were hit harder by the financial crisis compared to others.

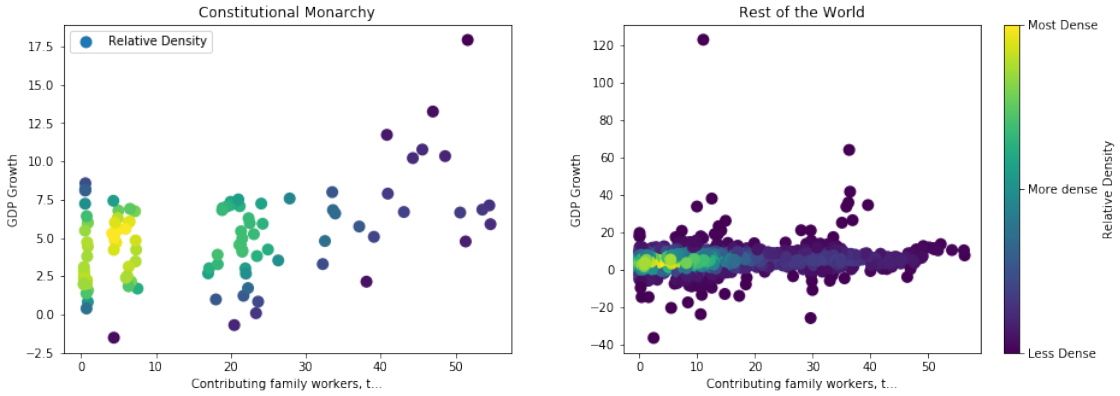
After this initial visualization, I want to find some of the features that can best predict future GDP growth. I will attempt to identify single variables that have a high correlation with future GDP growth. I want to initially focus my attention on identifying single variables that may predict future GDP growth because I want to know if any variables seem likely to be valuable in predicting GDP growth, and a single variable is much easier to visualize and gain intuitive understanding into. To do this, I will run a single regression on every feature in each group, (each regression will also include a constant) and I will measure the R^2 value in each regression (Statsmodels will be used to run the regressions). Since I am constraining my regression to a single variable, selecting the highest R^2 value is equivalent to selecting both the lowest BIC and AIC scores; therefore, I will select the feature that gives the highest R^2 value, and plot that feature against the next year's GDP growth. In addition, I want to see if any of these features seems to have a different distribution in its respective group compared to the rest of the world, so beside each group, I will also plot the feature against future GDP growth for the rest of the world. All observations for each group shares exactly the same features, but not all observations share exactly the same features across groups, thus the plots for the rest of the world may not always include all other countries.

Since the data is particularly dense for some groups, a scatter plot can quickly become unreadable as dots get piled on top of one another; therefore I will use the scipy.stats.gaussian_kde function to approximate a pdf for the distribution and plot a heat map of the scatter plot to help with visualization.

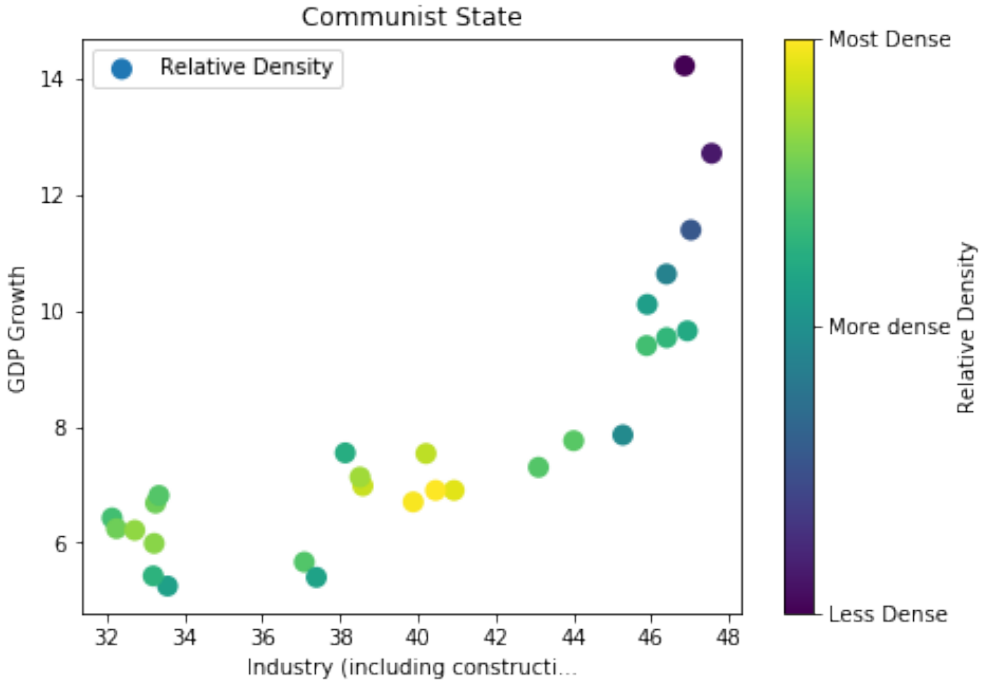
```
In Countries whose government was Absolute Monarchy, the single variable that
gave the highest R squared value was
Employers, total (% of total employment) (modeled ILO estimate).
It had an R squared value of 0.7484442777733069
```



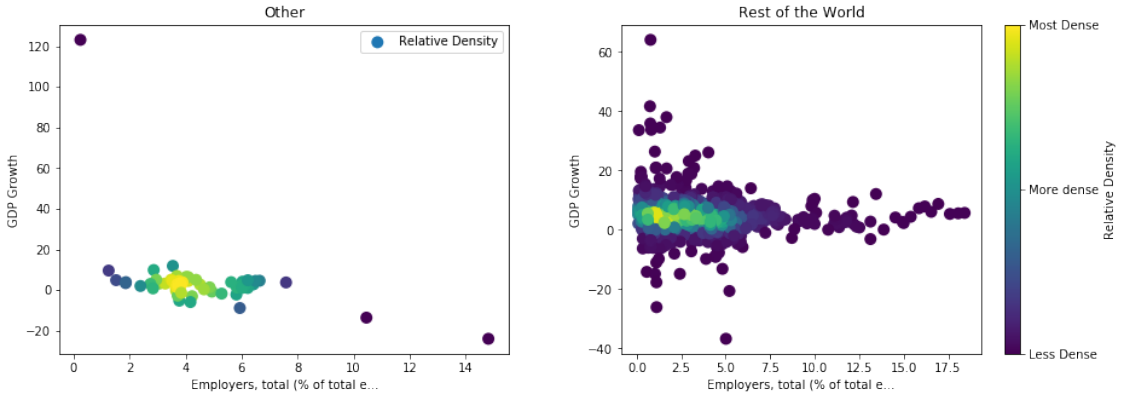
In Countries whose government was Constitutional Monarchy, the single variable that gave the highest R squared value was
Contributing family workers, total (% of total employment) (modeled ILO estimate).
It had an R squared value of 0.18912450359908173



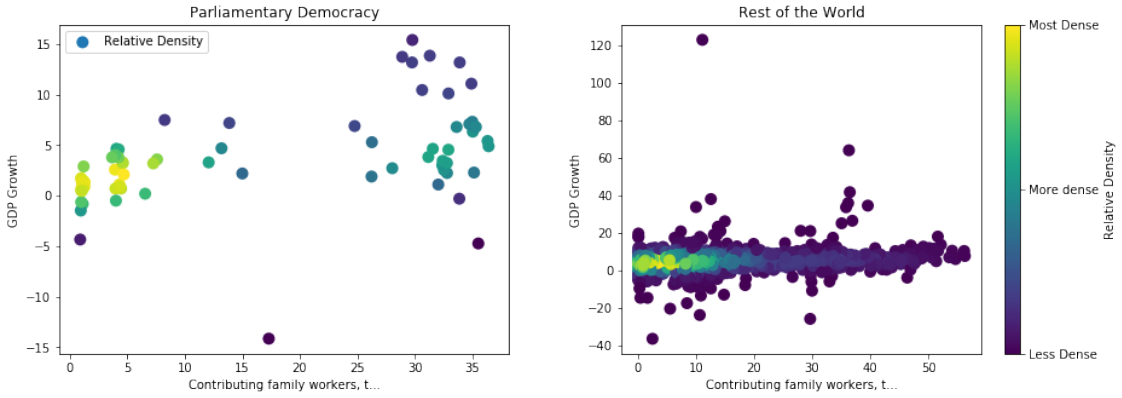
In Countries whose government was Communist State, the single variable that gave the highest R squared value was:
Industry (including construction), value added (% of GDP).
It had an R squared value of 0.6519688139816965



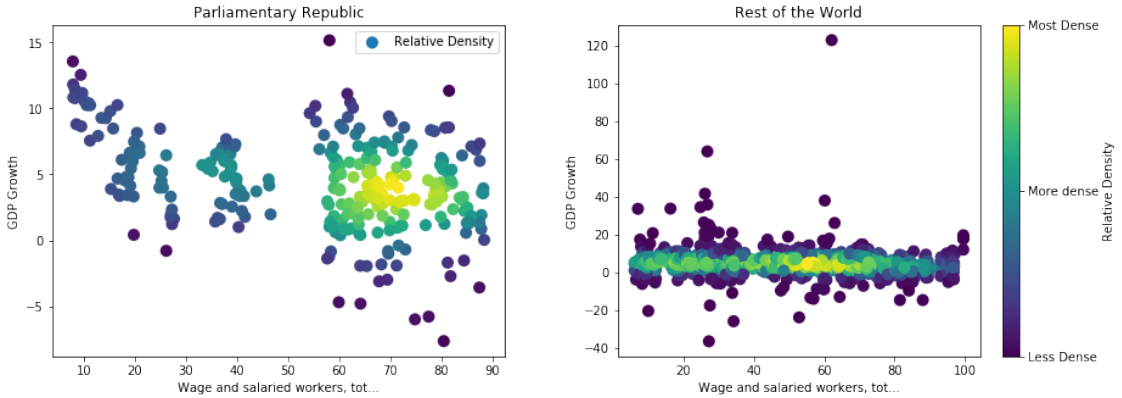
In Countries whose government was Other, the single variable that gave the highest R squared value was
Employers, total (% of total employment) (modeled ILO estimate).
It had an R squared value of 0.18311028191520318



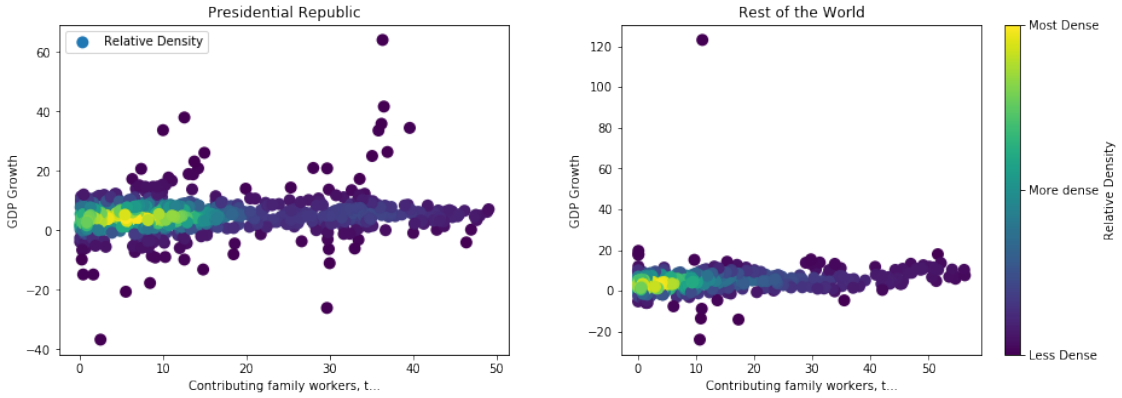
In Countries whose government was Parliamentary Democracy, the single variable that gave the highest R squared value was
Contributing family workers, total (% of total employment) (modeled ILO estimate).
It had an R squared value of 0.19658127165438977



In Countries whose government was Parliamentary Republic, the single variable that gave the highest R squared value was
Wage and salaried workers, total (% of total employment) (modeled ILO estimate).
It had an R squared value of 0.1361585553986805



In Countries whose government was Presidential Republic, the single variable that gave the highest R squared value was
Contributing family workers, total (% of total employment) (modeled ILO estimate).
It had an R squared value of 0.025894515982961774



Note, I skipped the graph for the rest of the world associated with Communist State since the best single variable for Communist State was 'Industry (including construction), value added (% of GDP)' and no other country had this variable in the data set.

After skimming the graphs, all the distributions seem to be at least somewhat similar across groups, but in order to determine if the distributions are truly similar, more analysis is necessary. I will use scipy.stats.ks_2samp (which calculates the Kolmogorov-Smirnov statistic on 2 samples), to compare the distributions, and determine which, if any, has a statistically different distribution. The following code snippet shows how this is calculated. The snippet uses the variable 'best_vals' which is defined previously (see auxiliary files), this is a list of the feature names that gave the highest R^2 value for each group. I also reference the variable 'groups', which is a list of DataFrames (one for each group).

```
p_vals=[]
for i in range(len(groups)):
    group_vals=np.asarray(groups[i][best_vals[i]]).astype(float)
    rest_vals=None
    for j in groups:
        if i==j:
            break
        if rest_vals is None:
            rest_vals=groups[j].asarray()
        else:
            rest_vals=np.hstack((rest_vals,groups[j][best_vars[i]].asarray()))
    p_vals.append(ks_2samp(group_vals,rest_vals[1]))
```

Test Results:

The p-value for countries whose government is Absolute Monarchy:
2.44485565481206e-06

The p-value for countries whose government is Constitutional Monarchy :
1.0953436660973378e-06

The p-value for countries whose government is Communist State::
2.020121156562933e-19

The p-value for countries whose government is Other:
9.615351426592739e-11

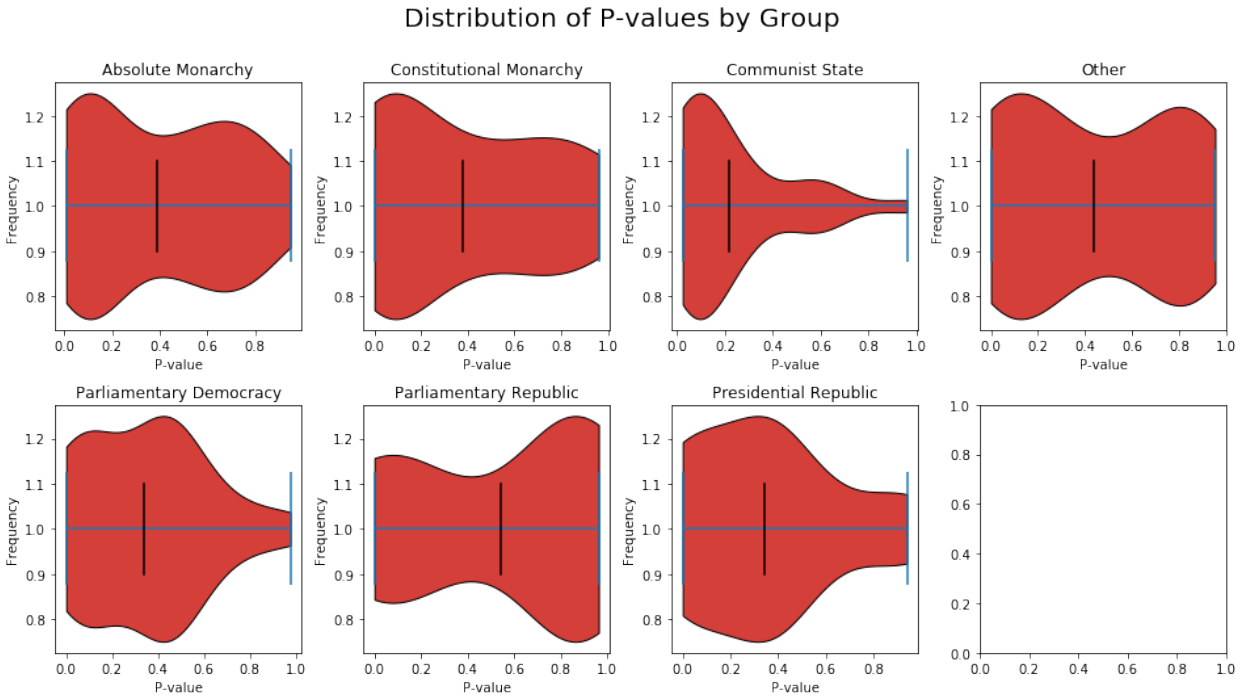
The p-value for countries whose government is Parliamentary Democracy:
4.2355219362894694e-05

The p-value for countries whose government is Parliamentary Republic:
9.965244511366208e-16

The p-value for countries whose government is Presidential Republic:
6.946934487890843e-08

The default null hypothesis is that the samples come from the same distribution. None of the p-values are even close to failing .05 confidence level (the largest p-value is about 0.0000424), so we can say with near certainty that the distribution of the selected feature in each group comes from a different distribution compared to the rest of the world.

To finish my analysis, I want to find out how significant the features in each group are in predicting GDP growth. To do this I will perform a linear regression for each group using all the features in the group, and I will then examine the p-values for each feature. I want to know if the features in the groups tend to be significant. To help gain an understanding of how significant the features are on average, I will plot the distribution of the P-values using a violin plot, and included in each plot will be a line segment showing the mean for each group.



By looking at these plots, I can easily see that some groups have many features that are relatively insignificant. For example, parliamentary republics have a dense distribution of p-values close to one, and a comparatively low density close to zero with a mean of nearly .6, so in the future I will likely need to eliminate some of the features in this group. On the other hand, Communist States have a mean of about .2 with a very large distribution near zero, and a very small distribution near one, thus I probably will not need to eliminate as many features in this group compared to the other groups. All the groups have at least some p-values close to zero, so each group has some features that seem significant. These results show that there may be a way to create a model with predictive power.

Conclusion

It is tremendously valuable, yet extremely difficult, to predict future economic growth. Corporations will need this information when deciding whether to invest in growth, governments will need the information when considering policies, central banks need the information when trying determine how to adjust interest rates and so on. This is an important topic that needs to be addressed.

The initial analysis of the data obtained from the world bank shows that many features are insignificant, but there are also many features that seem significant. When I looked at some of the most significant features in each group compared to the rest to the world, I found that there was significant differences in how the features were distributed. The implications of this observation is unclear as of yet, but this observation supports the intuition that there are fundamental differences from one group to another. The analysis I performed only looked at one variable, and did not compare the relationship between multiple variables. It is possible that the relationships across groups are also very different.

More research will be done to uncover how economic growth can be modeled in different countries. Once good models are developed, I can make more comparisons between groups of countries. Will a model from one group give descent predictions for another group, or is each group different enough to make a model trained on one group nearly useless in another? More importantly, can a model trained on historical data give accurate predictions for the future? This is what I hope to accomplish in the future by fitting the data to other machine learning models including non-linear models. Machine learning and economic modeling are both areas with on-going research, so future analysis will be exciting and rewarding, and I hope to uncover valuable insights to help further the research in this area.