The Colonial Origins of Comparative Development - Project by Sweta Gangopadhyay and Tianxiong Hu

Introduction

Source of Data and Paper Description

For our research project, we have decided to look at a paper by Daron Acemoglu, Simon Johnson and James A. Robinson called "The Colonial Origins of Comparative Development". The paper is a study about what factors have led to differences in economic development across countries that were former colonies. The paper looks at primarily British and French colonies and compares their GDP per capita in 1995. The authors of the paper hypothesized that the GDP per capita of the former colonies were closely tied to the quality of institutions in the countries. If institutions like governmental bodies and educational bodies were more extractive (i.e. took away resources/power from the citizens of the country), Acemoglu et al. hypothesized that these extractive institutions would lead to fewer property rights, fewer investments in human capital, and hence lower GDPs. They also looked at what caused institutions to be extractive vs. non-extractive. They hypothesized that colonizers only set up good institutions if they themselves settled in the colony, making the European settler mortality rate an effective proxy for institution quality.

This is the link to the data and tables used by the authors in their paper: https://economics.mit.edu/faculty/acemoglu/data/ajr2001

For the purpose of this project we used their data set to recreate some of their findings and then extend the paper by adding some of our hypotheses and tests as well. The data set is a collection of GDP per capita, avg. protection against expropriation (their measure of institution quality with higher numbers corresponding to better institutions), natural resource reserves (like oil, gold, zinc, etc.), geography, climate, colonizer details (like colonizer country, years since independence, duration of colonization etc.), and ethnographic factors like religion, language and ethnicity. The remaining variables in the data set used by the paper are not used in this project, and so are not mentioned.

Our Take on the Data

In this project, we aim to explore the data set a little more and look at the trends in the data. With our collection of former colony/country data, we looked at the biggest colonizers, wealth by continent and even correlations between variables to test through regressions later. We also recreated the base ordinary least squares regressions found in the paper and extended their analysis by adding and testing against many other variables, based on our own hypotheses (found below).

Hypotheses

1. We want to look at the effect of:
    - Religion
    - Ethnicity
    - Language

On the main variables of interest in the paper - insititution quality and log GDP. We think that with greater ethnographic diversity it will be more difficult to coordinate economic policies that will satisfy all groups - which may mean lower economic performance.

However, we are also open to the idea that with greater diversity (like immigration), economic activity may actually be very high. We hope to run correlations and regressions to test these relationships and see whether our hypotheses are correct.

Importing and cleaning the dataset

After importing the data set, we cleaned it up so that there were no null values. We also dropped all the columns of variables we will not be using to make the data set look more succinct. We then set the index to the country name and our final, cleaned data set looked like this:

In [36]:

```
df.set_index('name')
df
```

Out[36]:

	name	yr.col	colonizer	continent	s.america	mortality	logem4	logpgp12	lat_abst	avexpr	...	zinc	oilres	y
3	Angola	1483.0	Portugal	Africa	0.0	280.00	5.634789	8.716929	0.136667	5.363636	...	0.0	146000.0	
7	Argentina	1512.0	Spain	South America	1.0	68.90	4.232656	NaN	0.377778	6.386364	...	0.0	46900.0	
11	Australia	1788.0	England	Other	0.0	8.55	2.145931	10.705441	0.300000	9.318182	...	12.0	99100.0	
23	Burkina Faso	1896.0	France	Africa	0.0	280.00	5.634789	7.321831	0.144445	4.454545	...	0.0	0.0	
25	Bangladesh	1537.0	Portugal	Asia	0.0	71.41	4.268438	7.540647	0.266667	5.136364	...	0.0	0.0	
31	Bahamas, The	1718.0	England	Other	0.0	85.00	4.442651	10.361836	0.268333	7.500000	...	0.0	0.0	
39	Bolivia	1524.0	Spain	South America	1.0	71.00	4.262680	8.571013	0.188889	5.636364	...	0.0	14300.0	
41	Brazil	1500.0	Portugal	South America	1.0	71.00	4.262680	9.385043	0.111111	7.909091	...	1.0	19400.0	
51	Canada	1535.0	France	Other	0.0	16.10	2.778819	10.658045	0.666667	9.727273	...	15.0	185000.0	
59	Côte d'Ivoire	1843.0	France	Africa	0.0	668.00	6.504288	7.620299	0.088889	7.000000	...	0.0	7460.0	
61	Cameroon	1884.0	Germany	Africa	0.0	280.00	5.634789	7.758745	0.066667	6.454545	...	0.0	31900.0	
63	Congo, Dem. Rep.	1885.0	Belgium	Africa	0.0	240.00	5.480639	8.395268	0.011111	4.681818	...	0.0	338000.0	
65	Colombia	1499.0	Spain	South America	1.0	71.00	4.262680	9.266961	0.044444	7.318182	...	0.0	57.0	
71	Costa Rica	1502.0	Spain	South America	1.0	NaN	4.357990	9.468531	0.111111	7.045455	...	0.0	0.0	
81	Dominican Republic	1492.0	Spain	South America	1.0	130.00	4.867535	9.230497	0.211111	6.181818	...	0.0	0.0	
83	Algeria	1830.0	France	Africa	0.0	NaN	4.359270	9.049626	0.311111	6.500000	...	0.0	339000.0	
85	Ecuador	1563.0	Spain	South America	1.0	71.00	4.262680	9.189732	0.022222	6.545455	...	0.0	141.0	
87	Egypt, Arab Rep.	1798.0	France	Africa	0.0	67.80	4.216562	8.813373	0.300000	6.772727	...	0.0	111000.0	
95	Ethiopia	1896.0	Italy	Africa	0.0	26.00	3.258096	7.037666	0.088889	5.727273	...	0.0	0.0	
103	Gabon	1875.0	France	Africa	0.0	280.00	5.634789	9.685675	0.011111	7.818182	...	0.0	579000.0	
109	Ghana	1874.0	England	Africa	0.0	668.00	6.504288	7.624223	0.088889	6.272727	...	0.0	30.0	
111	Guinea	1890.0	France	Africa	0.0	483.00	6.180017	6.974252	0.122222	6.545455	...	0.0	0.0	
113	Gambia, The	1618.0	England	Africa	0.0	1470.00	7.293018	7.574620	0.147556	8.272727	...	0.0	0.0	
119	Guatemala	1519.0	Spain	South America	1.0	71.00	4.262680	8.537425	0.170000	5.136364	...	0.0	2700.0	
121	Guyana	1616.0	Netherlands	Other	0.0	32.18	3.471345	8.131310	0.055556	5.886364	...	0.0	0.0	
123	Hong Kong SAR, China	1842.0	England	Asia	0.0	14.90	2.701361	10.857952	0.246111	8.136364	...	0.0	0.0	
125	Honduras	1524.0	Spain	South America	1.0	78.10	4.357990	8.353066	0.166667	5.318182	...	0.0	0.0	
133	Indonesia	1512.0	Portugal	Asia	0.0	170.00	5.135798	8.508344	0.055600	7.590909	...	0.0	29700.0	
135	India	1765.0	England	Asia	0.0	48.63	3.884241	8.262562	0.222200	8.272727	...	7.0	6750.0	
149	Jamaica	1494.0	Spain	South America	1.0	130.00	4.867535	NaN	0.201667	7.090909	...	0.0	0.0	
157	Kenya	1885.0	Germany	Africa	0.0	145.00	4.976734	7.476214	0.011111	6.045455	...	0.0	0.0	
171	Sri Lanka	1505.0	Portugal	Asia	0.0	69.80	4.245634	8.739891	0.077778	6.045455	...	0.0	0.0	
181	Morocco	1884.0	Spain	Africa	0.0	78.20	4.359270	8.554794	0.355556	7.090909	...	0.0	0.0	
185	Madagascar	1897.0	France	Africa	0.0	536.04	6.284209	6.885859	0.222222	4.454545	...	0.0	0.0	
187	Mexico	1521.0	Spain	South America	1.0	71.00	4.262680	9.725023	0.255556	7.500000	...	4.0	570000.0	
191	Mali	1880.0	France	Africa	0.0	2940.00	7.986165	7.102072	0.188889	4.000000	...	0.0	0.0	
193	Malta	1814.0	England	Other	0.0	16.30	2.791165	10.275515	0.394444	7.227273	...	0.0	0.0	
207	Malaysia	1511.0	Portugal	Asia	0.0	17.70	2.873565	9.749323	0.025556	7.954545	...	0.0	192000.0	
211	Niger	1899.0	France	Africa	0.0	400.00	5.991465	6.499560	0.177778	5.000000	...	0.0	0.0	

	name	yr.col	colonizer	continent	s.america	mortality	logem4	logpgp12	lat_abst	avexpr	...	zinc	oilres	y
213	Nigeria	1900.0	England	Africa	0.0	280.00	5.634789	7.692504	0.111111	5.545455	...	0.0	16800.0	
215	Nicaragua	1502.0	Spain	South America	1.0	163.30	5.095589	8.311885	0.144445	5.227273	...	0.0	0.0	
223	New Zealand	1840.0	England	Other	0.0	8.55	2.145931	10.380326	0.455556	9.727273	...	0.0	48300.0	
227	Pakistan	1857.0	England	Asia	0.0	36.99	3.610648	7.969205	0.333333	6.045455	...	0.0	3220.0	
229	Panama	1538.0	Spain	South America	1.0	163.30	5.095589	9.718080	0.100000	5.909091	...	0.0	0.0	
231	Peru	1529.0	Spain	South America	1.0	71.00	4.262680	9.299445	0.111111	5.772727	...	5.0	16600.0	
243	Paraguay	1537.0	Spain	South America	1.0	78.10	4.357990	8.722185	0.255556	6.954545	...	0.0	0.0	
255	Sudan	1898.0	England	Africa	0.0	88.20	4.479607	7.694067	0.166667	4.000000	...	0.0	10900.0	
257	Senegal	1677.0	France	Africa	0.0	164.66	5.103883	7.572484	0.155556	6.000000	...	0.0	0.0	
259	Singapore	1819.0	England	Asia	0.0	17.70	2.873565	11.031711	0.013556	9.318182	...	0.0	0.0	
261	Sierra Leone	1462.0	Portugal	Africa	0.0	483.00	6.180017	7.214545	0.092222	5.818182	...	0.0	0.0	
283	Togo	1884.0	Germany	Africa	0.0	668.00	6.504288	6.957467	0.088889	6.909091	...	0.0	0.0	
291	Trinidad and Tobago	1498.0	Spain	South America	1.0	85.00	4.442651	10.190422	0.122222	7.454545	...	0.0	442000.0	
293	Tunisia	1881.0	France	Africa	0.0	63.00	4.143135	9.189585	0.377778	6.454545	...	0.0	198000.0	
299	Tanzania	1885.0	Germany	Africa	0.0	145.00	5.634789	7.378226	0.066667	6.636364	...	0.0	0.0	
301	Uganda	1800.0	England	Africa	0.0	280.00	5.634789	7.209137	0.011111	4.454545	...	0.0	0.0	
305	Uruguay	1519.0	Spain	South America	1.0	71.00	4.262680	9.682635	0.366667	7.000000	...	0.0	0.0	
307	United States	1513.0	Spain	Other	0.0	15.00	2.708050	10.819083	0.422222	10.000000	...	13.0	95700.0	
311	Venezuela, RB	1522.0	Spain	South America	1.0	78.10	4.357990	9.509368	0.088889	7.136364	...	0.0	3040000.0	
313	Vietnam	1887.0	France	Asia	0.0	140.00	4.941642	8.198421	0.177778	6.409091	...	0.0	7140.0	
319	South Africa	1652.0	Netherlands	Africa	0.0	NaN	2.740840	9.344904	0.322222	6.863636	...	2.0	0.0	

60 rows × 38 columns

In [34]:

```
df['name'].nunique()
```

Out[34]:

```
60
```

Hence we had 60 individual entries of countries that had all the information we were going to use (compared to the whole uncleaned data set which had a list of 150 countries, but which were missing a lot of the variable data).

Creating subsets of data based on continent

For our analyses, we wanted to look at the characteristics of the data set first, specifically to see what the differences were between colonies from different continents. Hence, we made subsets of the data based on continent - these consisted of colonies from Asia, Africa and South America:

In [38]:

```
Asia
```

Out[38]:

	name	yr.col	colonizer	continent	s.america	mortality	logem4	logpgp12	lat_abst	avexpr	...	zinc	oilres	yellow
25	Bangladesh	1537.0	Portugal	Asia	0.0	71.41	4.268438	7.540647	0.266667	5.136364	...	0.0	0.0	1.0
	Hong Kong													

	name	yr.col	colonizer	continent	s.america	mortality	logem4	logpgp12	lat_abst	avexpr	...	zinc	oilres	yellow
123	Hong Kong SAR, China	1842.0	England	Asia	0.0	14.90	2.701361	10.857952	0.246111	8.136364	...	0.0	0.0	0.0
133	Indonesia	1512.0	Portugal	Asia	0.0	170.00	5.135798	8.508344	0.055600	7.590909	...	0.0	29700.0	1.0
135	India	1765.0	England	Asia	0.0	48.63	3.884241	8.262562	0.222200	8.272727	...	7.0	6750.0	0.0
171	Sri Lanka	1505.0	Portugal	Asia	0.0	69.80	4.245634	8.739891	0.077778	6.045455	...	0.0	0.0	1.0
207	Malaysia	1511.0	Portugal	Asia	0.0	17.70	2.873565	9.749323	0.025556	7.954545	...	0.0	192000.0	1.0
227	Pakistan	1857.0	England	Asia	0.0	36.99	3.610648	7.969205	0.333333	6.045455	...	0.0	3220.0	1.0
259	Singapore	1819.0	England	Asia	0.0	17.70	2.873565	11.031711	0.013556	9.318182	...	0.0	0.0	1.0
313	Vietnam	1887.0	France	Asia	0.0	140.00	4.941642	8.198421	0.177778	6.409091	...	0.0	7140.0	1.0

9 rows × 38 columns

In [39]:

```
Africa
```

Out[39]:

	name	yr.col	colonizer	continent	s.america	mortality	logem4	logpgp12	lat_abst	avexpr	...	zinc	oilres	yellc
3	Angola	1483.0	Portugal	Africa	0.0	280.00	5.634789	8.716929	0.136667	5.363636	...	0.0	146000.0	1
23	Burkina Faso	1896.0	France	Africa	0.0	280.00	5.634789	7.321831	0.144445	4.454545	...	0.0	0.0	1
59	Côte d'Ivoire	1843.0	France	Africa	0.0	668.00	6.504288	7.620299	0.088889	7.000000	...	0.0	7460.0	1
61	Cameroon	1884.0	Germany	Africa	0.0	280.00	5.634789	7.758745	0.066667	6.454545	...	0.0	31900.0	1
63	Congo, Dem. Rep.	1885.0	Belgium	Africa	0.0	240.00	5.480639	8.395268	0.011111	4.681818	...	0.0	338000.0	1
83	Algeria	1830.0	France	Africa	0.0	NaN	4.359270	9.049626	0.311111	6.500000	...	0.0	339000.0	0
87	Egypt, Arab Rep.	1798.0	France	Africa	0.0	67.80	4.216562	8.813373	0.300000	6.772727	...	0.0	111000.0	0
95	Ethiopia	1896.0	Italy	Africa	0.0	26.00	3.258096	7.037666	0.088889	5.727273	...	0.0	0.0	1
103	Gabon	1875.0	France	Africa	0.0	280.00	5.634789	9.685675	0.011111	7.818182	...	0.0	579000.0	1
109	Ghana	1874.0	England	Africa	0.0	668.00	6.504288	7.624223	0.088889	6.272727	...	0.0	30.0	1
111	Guinea	1890.0	France	Africa	0.0	483.00	6.180017	6.974252	0.122222	6.545455	...	0.0	0.0	1
113	Gambia, The	1618.0	England	Africa	0.0	1470.00	7.293018	7.574620	0.147556	8.272727	...	0.0	0.0	1
157	Kenya	1885.0	Germany	Africa	0.0	145.00	4.976734	7.476214	0.011111	6.045455	...	0.0	0.0	1
181	Morocco	1884.0	Spain	Africa	0.0	78.20	4.359270	8.554794	0.355556	7.090909	...	0.0	0.0	0
185	Madagascar	1897.0	France	Africa	0.0	536.04	6.284209	6.885859	0.222222	4.454545	...	0.0	0.0	1
191	Mali	1880.0	France	Africa	0.0	2940.00	7.986165	7.102072	0.188889	4.000000	...	0.0	0.0	1
211	Niger	1899.0	France	Africa	0.0	400.00	5.991465	6.499560	0.177778	5.000000	...	0.0	0.0	1
213	Nigeria	1800.0	England	Africa	0.0	2004.00	7.602901	7.886507	0.111111	5.545455	...	0.0	168000.0	1
255	Sudan	1898.0	England	Africa	0.0	88.20	4.479607	7.694067	0.166667	4.000000	...	0.0	10900.0	1
257	Senegal	1677.0	France	Africa	0.0	164.66	5.103883	7.572484	0.155556	6.000000	...	0.0	0.0	1
261	Sierra Leone	1462.0	Portugal	Africa	0.0	483.00	6.180017	7.214545	0.092222	5.818182	...	0.0	0.0	1
283	Togo	1884.0	Germany	Africa	0.0	668.00	6.504288	6.957467	0.088889	6.909091	...	0.0	0.0	1
293	Tunisia	1881.0	France	Africa	0.0	63.00	4.143135	9.189585	0.377778	6.454545	...	0.0	198000.0	0
299	Tanzania	1885.0	Germany	Africa	0.0	145.00	5.634789	7.378226	0.066667	6.636364	...	0.0	0.0	1
301	Uganda	1800.0	England	Africa	0.0	280.00	5.634789	7.209137	0.011111	4.454545	...	0.0	0.0	1
319	South Africa	1652.0	Netherlands	Africa	0.0	NaN	2.740840	9.344904	0.322222	6.863636	...	2.0	0.0	0

26 rows × 38 columns

In [40]:

```
South America
```

Out[40]:

	name	yr.col	colonizer	continent	s.america	mortality	logem4	logpgp12	lat_abst	avexpr	...	zinc	oilres	yellov
7	Argentina	1512.0	Spain	South America	1.0	68.9	4.232656	NaN	0.377778	6.386364	...	0.0	46900.0	0.(
39	Bolivia	1524.0	Spain	South America	1.0	71.0	4.262680	8.571013	0.188889	5.636364	...	0.0	14300.0	1.(
41	Brazil	1500.0	Portugal	South America	1.0	71.0	4.262680	9.385043	0.111111	7.909091	...	1.0	19400.0	1.(
65	Colombia	1499.0	Spain	South America	1.0	71.0	4.262680	9.266961	0.044444	7.318182	...	0.0	57.0	1.(
71	Costa Rica	1502.0	Spain	South America	1.0	NaN	4.357990	9.468531	0.111111	7.045455	...	0.0	0.0	1.(
81	Dominican Republic	1492.0	Spain	South America	1.0	130.0	4.867535	9.230497	0.211111	6.181818	...	0.0	0.0	1.(
85	Ecuador	1563.0	Spain	South America	1.0	71.0	4.262680	9.189732	0.022222	6.545455	...	0.0	141.0	0.(
119	Guatemala	1519.0	Spain	South America	1.0	71.0	4.262680	8.537425	0.170000	5.136364	...	0.0	2700.0	1.(
125	Honduras	1524.0	Spain	South America	1.0	78.1	4.357990	8.353066	0.166667	5.318182	...	0.0	0.0	1.(
149	Jamaica	1494.0	Spain	South America	1.0	130.0	4.867535	NaN	0.201667	7.090909	...	0.0	0.0	1.(
187	Mexico	1521.0	Spain	South America	1.0	71.0	4.262680	9.725023	0.255556	7.500000	...	4.0	570000.0	1.(
215	Nicaragua	1502.0	Spain	South America	1.0	163.3	5.095589	8.311885	0.144445	5.227273	...	0.0	0.0	1.(
229	Panama	1538.0	Spain	South America	1.0	163.3	5.095589	9.718080	0.100000	5.909091	...	0.0	0.0	1.(
231	Peru	1529.0	Spain	South America	1.0	71.0	4.262680	9.299445	0.111111	5.772727	...	5.0	16600.0	0.(
243	Paraguay	1537.0	Spain	South America	1.0	78.1	4.357990	8.722185	0.255556	6.954545	...	0.0	0.0	0.(
291	Trinidad and Tobago	1498.0	Spain	South America	1.0	85.0	4.442651	10.190422	0.122222	7.454545	...	0.0	442000.0	1.(
305	Uruguay	1519.0	Spain	South America	1.0	71.0	4.262680	9.682635	0.366667	7.000000	...	0.0	0.0	0.(
311	Venezuela, RB	1522.0	Spain	South America	1.0	78.1	4.357990	9.509368	0.088889	7.136364	...	0.0	3040000.0	1.(

18 rows × 38 columns

Descriptive Statistics

These are the basic descriptive statistics for our continent subsets to get a general feel for the data. We specifically looked at the log of GDP per capita variable (or country wealth) across these continents.

Average Log GDP By Continent

In [9]:

```
df.groupby('continent')['logpgp95'].describe()
```

Out[9]:

continent	count	mean	std	min	25%	50%	75%	max
Africa	28.0	7.378724	0.799586	6.109248	6.829750	7.335856	7.815874	8.907883
Asia	14.0	8.593545	1.181823	6.877296	7.447314	8.496450	9.759594	10.146430

Other	7.0	9.496216	0.771366	7.904704	9.356625	9.756104	9.942725	10.215742
South America	18.0	8.510876	0.471677	7.544332	8.229472	8.598778	8.829746	9.133459

In [27]:

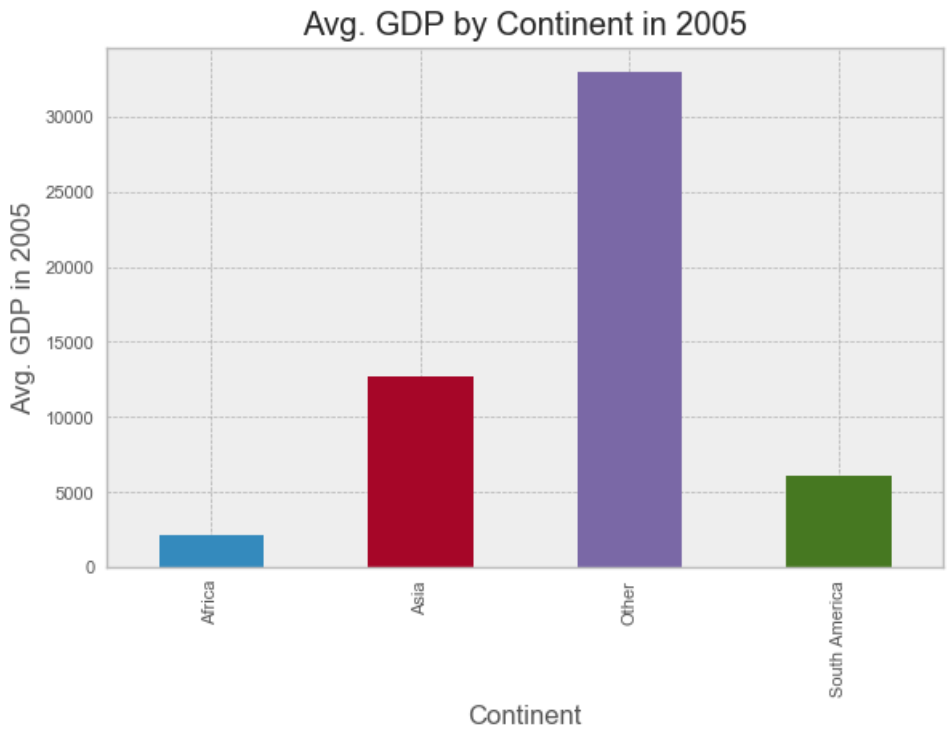
```
Asia = df[df.continent == 'Asia']
Africa = df[df.continent == 'Africa']
South_America = df[df.continent == 'South America']
```

In [99]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
df.groupby('continent')['GDP05'].mean().plot.bar(x = 'continent', y = 'GDP05', ax=ax, figsize = (6,
4))
ax.set_ylabel('Avg. GDP in 2005')
ax.set_xlabel('Continent')
ax.set_title('Avg. GDP by Continent in 2005')
```

Out[99]:

```
Text(0.5,1,'Avg. GDP by Continent in 2005')
```



The continent with the highest average Log GDP is "Other," which includes countries such as the US and Canada. This makes sense given that US and Canada are so called 'Neo-Europes' with very similar institutions as their colonizers like Britain and France. Other countries that are in the Neo-Europe space are Australia and New Zealand.

Similarly, we looked at the GDP distribution within our three continents - Asia, Africa and South America:

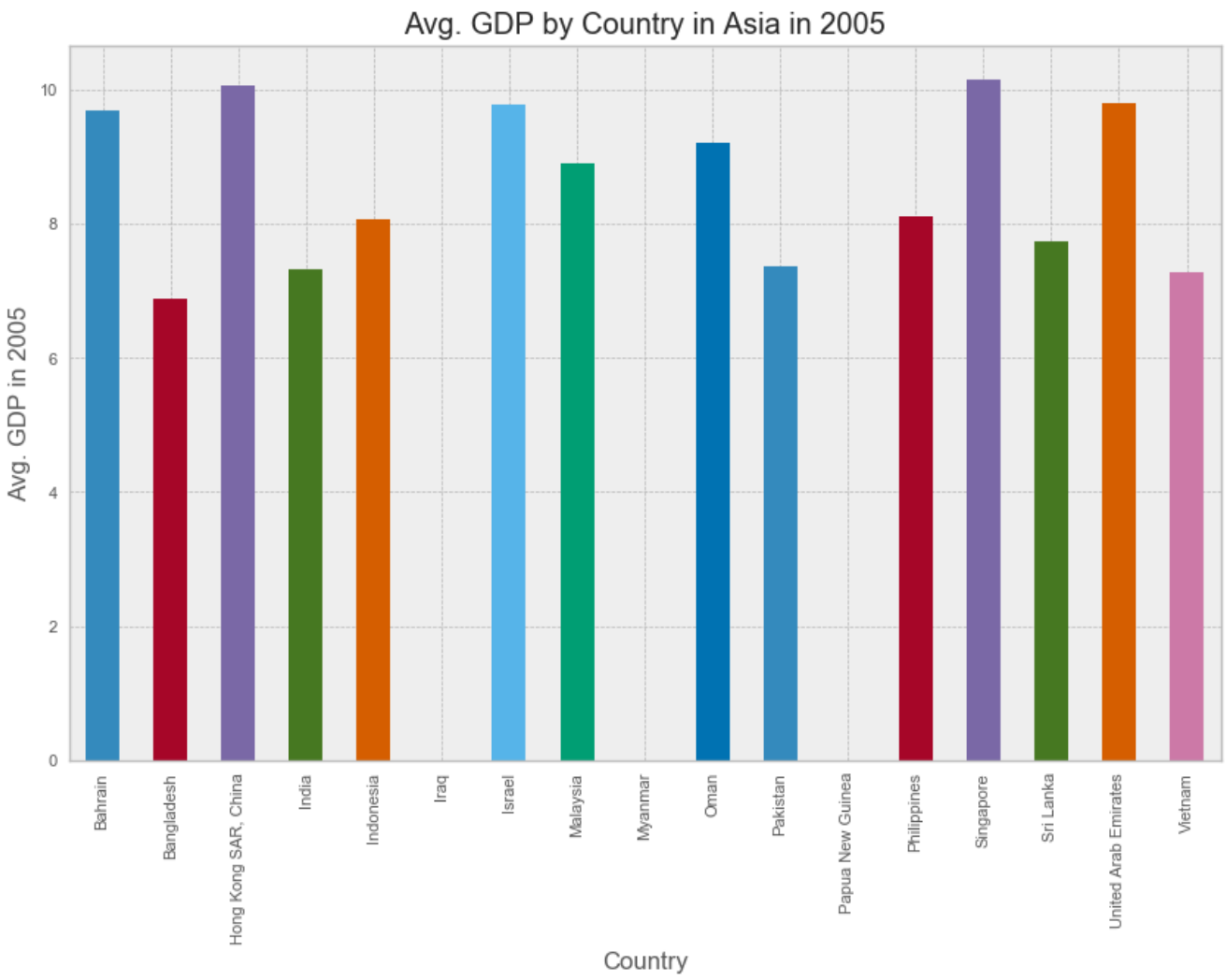
Average Log GDP Within Continent

In [90]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
Asia.groupby('name')['logpgp95'].mean().plot.bar(x = 'name', y = 'GDP05', ax=ax, figsize = (9,6))
ax.set_ylabel('Avg. GDP in 2005')
ax.set_xlabel('Country')
ax.set_title('Avg. GDP by Country in Asia in 2005')
```

Out[90]:

```
Text(0.5,1,'Avg. GDP by Country in Asia in 2005')
```



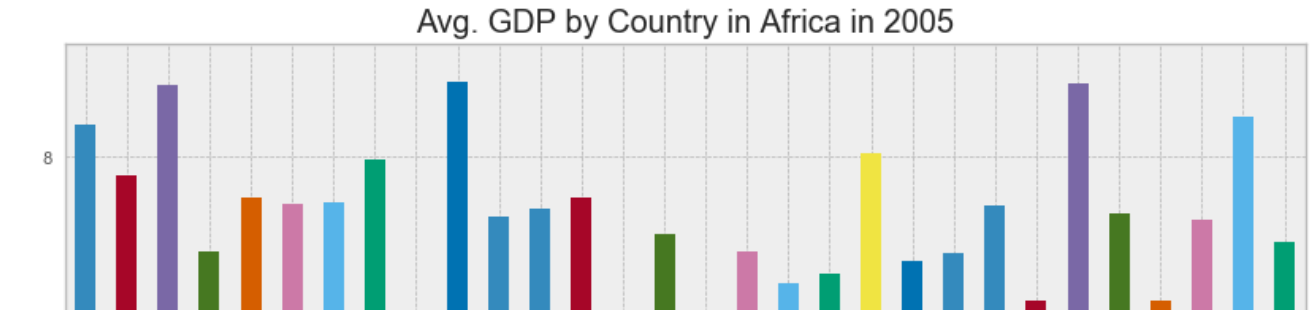
Former colonies in Asia have the second weakthiest among the rest of the three continents, but withing the continent itself there are significant variations in the GDP data. Trading-heavy countries like Singapore and China (Hong Kong) outperform Bangladesh and Pakistan GDP-wise by a significant margin. This makes sense as colonizers would set up trading ports in coastal colonies which would encourage investment in local infrastructure and insitutions (when compared to landlocked countries).

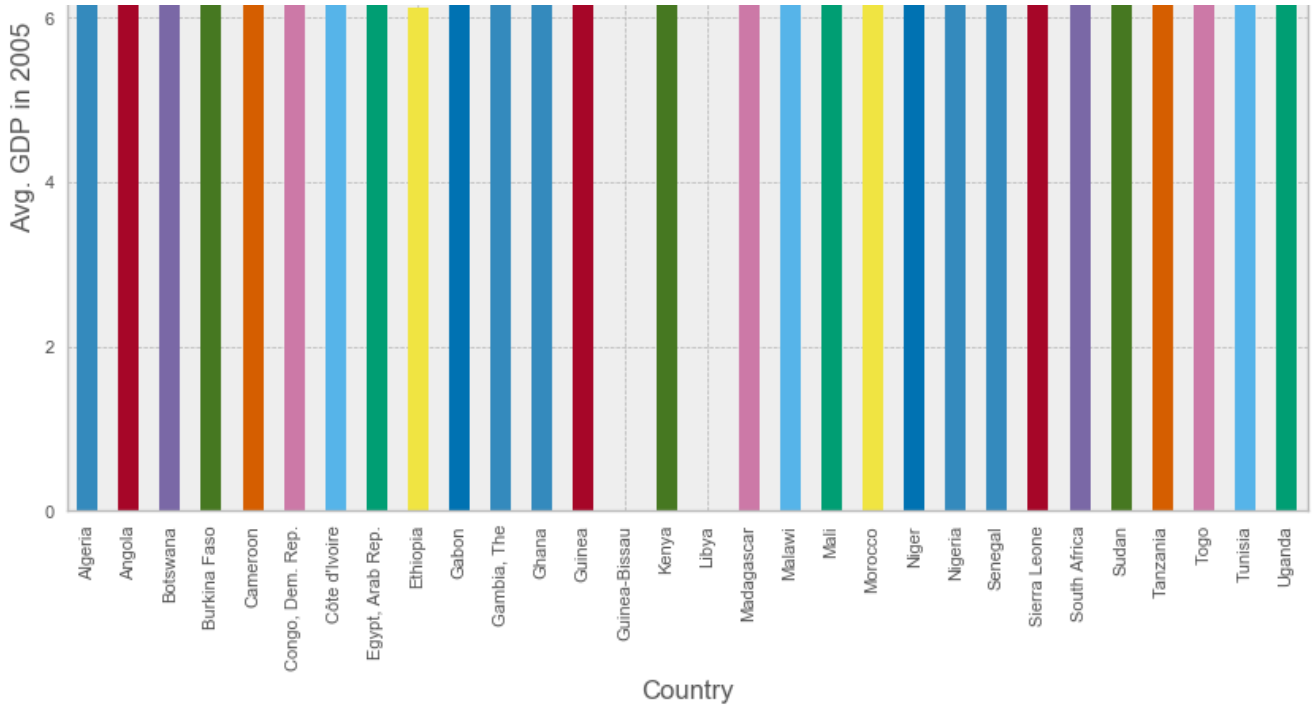
In [89]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
Africa.groupby('name')['logpgp95'].mean().plot.bar(x = 'name', y = 'GDP05', ax=ax, figsize = (9,6))
ax.set_ylabel('Avg. GDP in 2005')
ax.set_xlabel('Country')
ax.set_title('Avg. GDP by Country in Africa in 2005')
```

Out[89]:

```
Text(0.5,1,'Avg. GDP by Country in Africa in 2005')
```





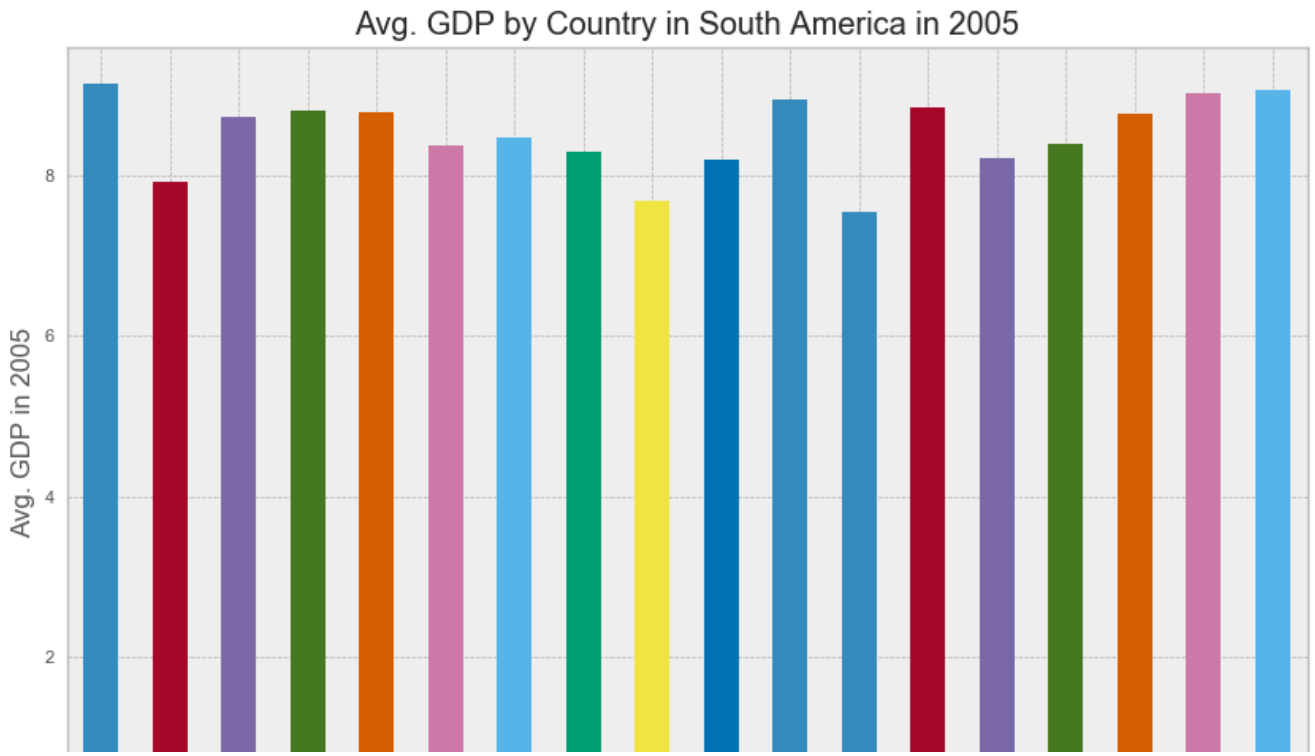
Most of the GDP data in Africa seems to be on the same level - besides some outliers like Gabon, South Africa and Botswana. It would be interesting to see whether these differences are due to the natural resources the country has (i.e. South Africa has a huge diamond manufacturing industry, the presence of oil etc.). It would also be interesting to see whether trading-heavy countries outperformed the non-trading heavy countries, as shown in the Asia analysis.

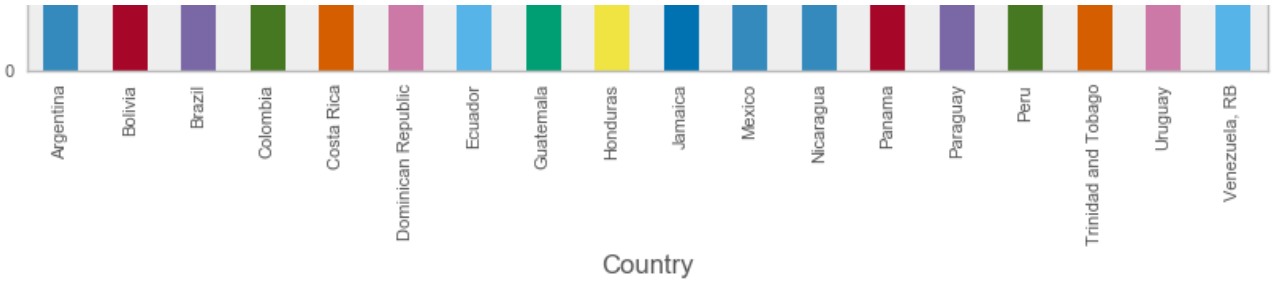
In [86]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
South_America.groupby('name')['logpgp95'].mean().plot.bar(x = 'name', y = 'GDP05', ax=ax, figsize =
(9,6))
ax.set_ylabel('Avg. GDP in 2005')
ax.set_xlabel('Country')
ax.set_title('Avg. GDP by Country in South America in 2005')
```

Out[86]:

```
Text(0.5,1,'Avg. GDP by Country in South America in 2005')
```





Low standard deviations are observed in GDP data in South America - showing these countries have similar wealth.

Ethnographic Factor Analysis

Next we wanted to see how each continent differed by ethnicity, religion and language diversity.

Former African colonies have a significantly higher mean ethnic diversity score than other groups. Former African colonies as a whole also score the highest in language diversity. Institutional quality scores are somewhat similar among former Asian, African and South-American colonies with Asia out perform South-America by less than 0.3 points.

Average Ethnic Diversity By Continent

In [32]:

```
df.groupby('continent')['Ethnic'].describe()
```

Out[32]:

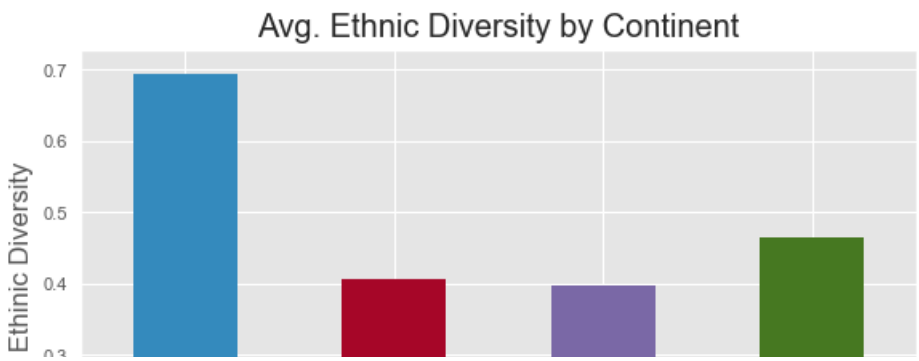
continent	count	mean	std	min	25%	50%	75%	max
Africa	30.0	0.693037	0.206160	0.039400	0.678450	0.738340	0.816371	0.930175
Asia	17.0	0.405358	0.198509	0.045434	0.271800	0.415000	0.506186	0.735134
Other	7.0	0.396580	0.250744	0.041432	0.244876	0.422845	0.554805	0.712420
South America	18.0	0.464937	0.177307	0.168900	0.294473	0.504400	0.589250	0.739625

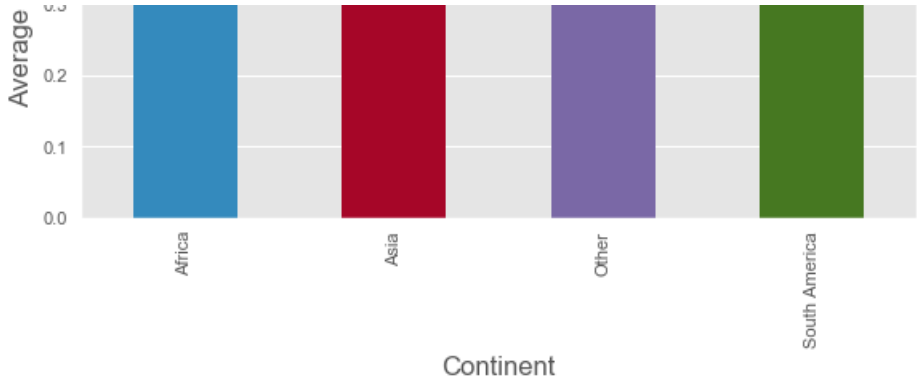
In [82]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
df.groupby('continent')['Ethnic'].mean().plot.bar(x = 'continent', y = 'avexpr', ax=ax, figsize = (
6, 4))
ax.set_ylabel('Average Ethinic Diversity')
ax.set_xlabel('Continent')
ax.set_title('Avg. Ethnic Diversity by Continent')
```

Out[82]:

```
Text(0.5,1,'Avg. Ethnic Diversity by Continent')
```





Average Language Diversity By Continent

In [34]:

```
df.groupby('continent')['Language'].describe()
```

Out[34]:

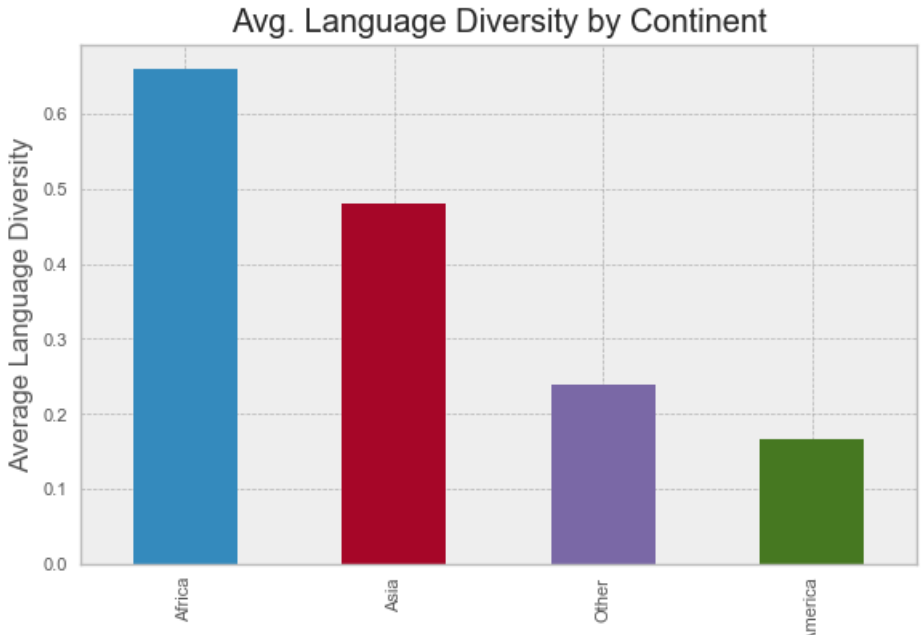
continent	count	mean	std	min	25%	50%	75%	max
Africa	30.0	0.658580	0.282280	0.012422	0.614731	0.777302	0.847449	0.922679
Asia	17.0	0.481059	0.213443	0.092485	0.356743	0.464456	0.596952	0.835952
Other	7.0	0.239164	0.174610	0.068804	0.128172	0.185494	0.293160	0.577184
South America	18.0	0.166068	0.167775	0.019268	0.050506	0.095763	0.205766	0.597501

In [84]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
df.groupby('continent')['Language'].mean().plot.bar(x = 'continent', y = 'avexpr', ax=ax, figsize =
(6,4))
ax.set_ylabel('Average Language Diversity')
ax.set_xlabel('Continent')
ax.set_title('Avg. Language Diversity by Continent')
```

Out[84]:

```
Text(0.5,1,'Avg. Language Diversity by Continent')
```



Average Religion Diversity By Continent

In [36]:

```
df.groupby('continent')['Religion'].describe()
```

Out[36]:

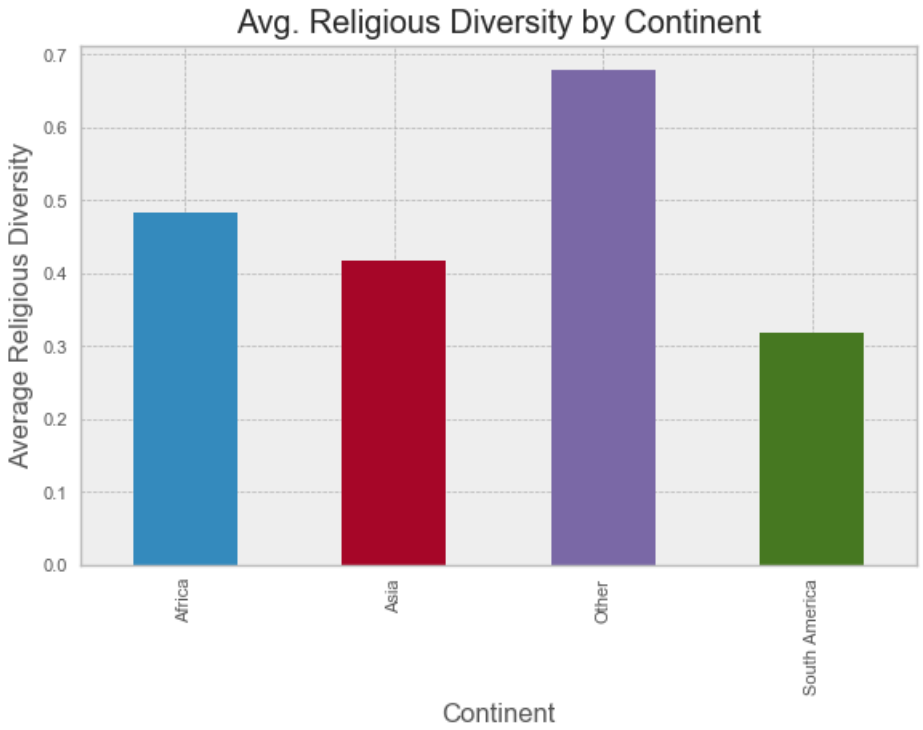
continent	count	mean	std	min	25%	50%	75%	max
Africa	30.0	0.482906	0.282221	0.003463	0.198726	0.605722	0.693466	0.860260
Asia	17.0	0.417099	0.144154	0.197385	0.326023	0.419110	0.508025	0.665688
Other	7.0	0.677616	0.251865	0.122324	0.688614	0.787592	0.816047	0.824078
South America	18.0	0.319095	0.185745	0.135030	0.201203	0.238337	0.370201	0.793610

In [85]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
df.groupby('continent')['Religion'].mean().plot.bar(x = 'continent', y = 'avexpr', ax=ax, figsize =
(6,4))
ax.set_ylabel('Average Religious Diversity')
ax.set_xlabel('Continent')
ax.set_title('Avg. Religious Diversity by Continent')
```

Out[85]:

```
Text(0.5,1,'Avg. Religious Diversity by Continent')
```



Links to Paper - Mortality and Institution Quality Variables

The two sets of graphs below show what looks like an inverse relationship between the mortality rate and the institution quality (like

found in the paper). So for example, Africa looks like it has the highest mortality rate - but it also has the lowest score for institution quality.

Average Mortality By Continent

In [38]:

```
df.groupby('continent')['logem4'].describe()
```

Out[38]:

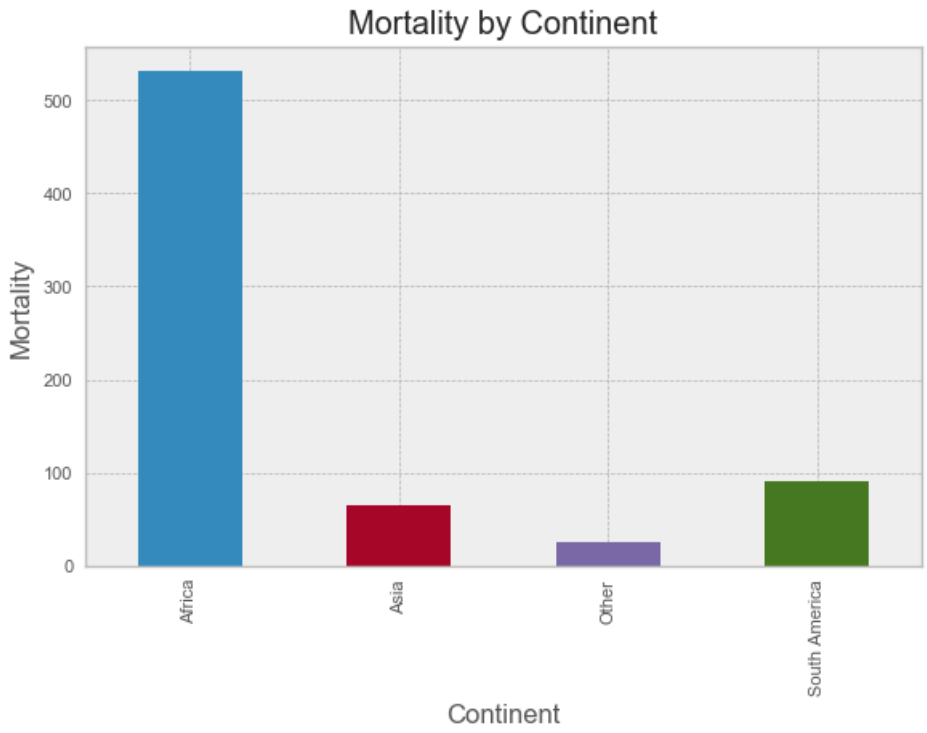
continent	count	mean	std	min	25%	50%	75%	max
Africa	27.0	5.520789	1.226028	2.740840	4.728170	5.634789	6.232113	7.986165
Asia	11.0	3.928595	0.902237	2.701361	3.208710	3.884241	4.605040	5.135798
Other	7.0	2.926270	0.806432	2.145931	2.426991	2.778819	3.131255	4.442651
South America	18.0	4.451942	0.301373	4.232656	4.262680	4.310335	4.421486	5.095589

In [96]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
df.groupby('continent')['mortality'].mean().plot.bar(x = 'continent', y = 'mortality', ax=ax, figsi
ze = (6,4))
ax.set_ylabel('Mortality')
ax.set_xlabel('Continent')
ax.set_title('Mortality by Continent')
```

Out[96]:

```
Text(0.5,1,'Mortality by Continent')
```



Average Institutional Quality By Continent

In [40]:

```
df.groupby('continent')['avexpr'].describe()
```

Out[40]:

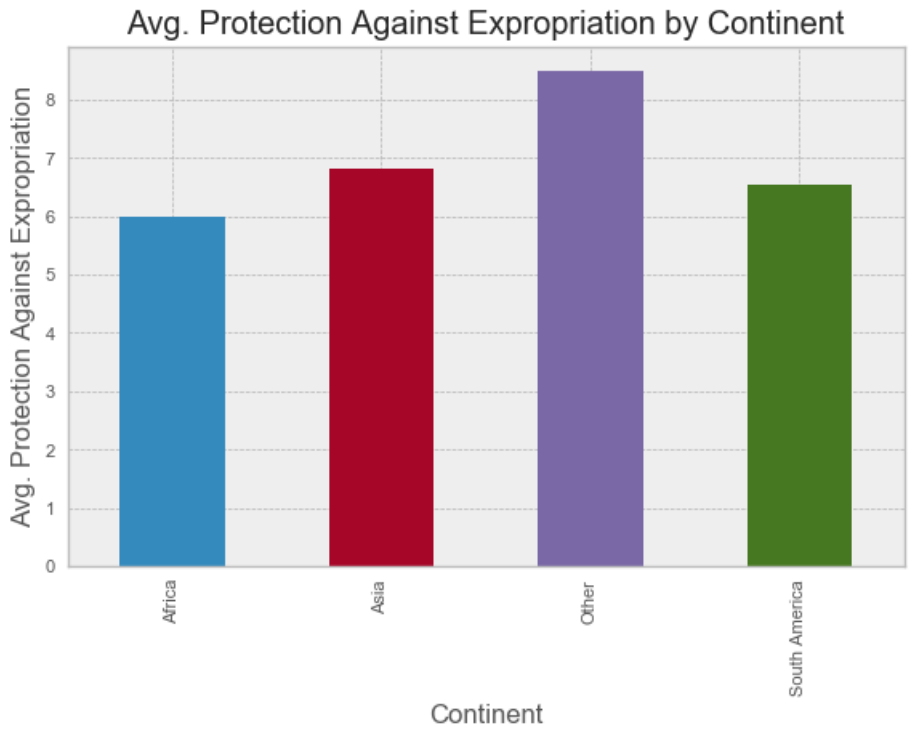
continent	count	mean	std	min	25%	50%	75%	max
Africa	30.0	5.983333	1.162397	4.000000	5.068182	6.159091	6.806818	8.272727
Asia	17.0	6.818182	1.781954	1.636364	6.045455	7.181818	8.000000	9.318182
Other	7.0	8.483766	1.601088	5.886364	7.363637	9.318182	9.727273	10.000000
South America	18.0	6.529040	0.863489	5.136364	5.806818	6.750000	7.125000	7.909091

In [97]:

```
fig,ax = plt.subplots()
plt.style.use('bmh')
df.groupby('continent')['avexpr'].mean().plot.bar(x = 'continent', y = 'avexpr', ax=ax, figsize = (
6,4))
ax.set_ylabel('Avg. Protection Against Expropriation')
ax.set_xlabel('Continent')
ax.set_title('Avg. Protection Against Expropriation by Continent')
```

Out[97]:

```
Text(0.5,1,'Avg. Protection Against Expropriation by Continent')
```



Most popular colonizers by continent

England and Portugal were the two biggest colonizers in Asia, together they owned over 75% of the former colonies in the region. There were more diversity among colonizers in Africa, and France, England and Germany together owned around 60% of the colonies in the Region. Former South American colonies were almost all owned by Spain, Portugal had a minor presence in the region and owned about 10% of the colonies in the region; this explains the former South American colonies' low scores in language and religion diversity.

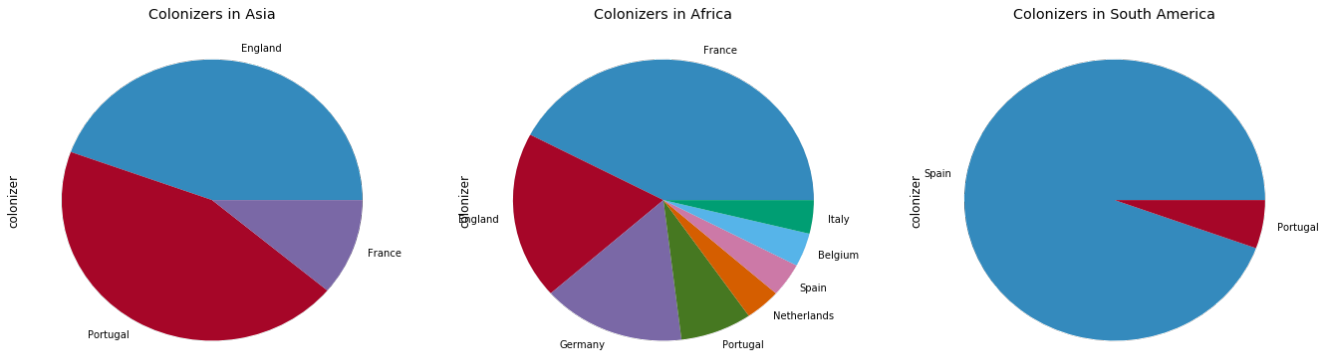
In [49]:

```
fig, ax = plt.subplots(1,3)
plt.style.use('bmh')
Asia['colonizer'].value_counts(normalize = True).plot.pie(ax=ax[0], figsize = (23,6.5))
Africa['colonizer'].value_counts(normalize = True).plot.pie(ax=ax[1], figsize = (23,6.5))
South_America['colonizer'].value_counts(normalize = True).plot.pie(ax=ax[2], figsize = (23,6.5))
```

```
ax[0].set_title('Colonizers in Asia', loc = 'center')
ax[1].set_title('Colonizers in Africa', loc = 'center')
ax[2].set_title('Colonizers in South America', loc = 'center')
```

Out[49]:

```
Text(0.5,1,'Colonizers in South America')
```



Correlation Analysis

Now that we have a basic understanding of what the data set looks like, we set out to look at the relationships between variables that have peaked our interest. We found out before that Africa had the greatest ethnic and language diversity - yet it had the worst institutions and the highest mortality rate. Hence, we first looked at the ethnographic factor correlations with the key variables of interest from the paper - the log GDP per capita and the avg. protection against expropriation risk (institution quality).

Correlation Matrix - Asia

In former Asian colonies, there is a significant positive correlations (0.71) between Log GDP and institutional qualities. Religion diversity has a moderate position correlation with Log GDP (0.41). Log Mortality Rate appears to have a significant negative correlation (-0.71) with Log GDP. The relationship between variables such as language and ethnic diversity do not seem to correlate with Log GDP.

Among dependent variables, language diversity and ethnic diversity appear to be somewhat strongly correlated (0.62); Log mortality rate and institutional quality are inversely correlated to a moderate extent (-0.40).

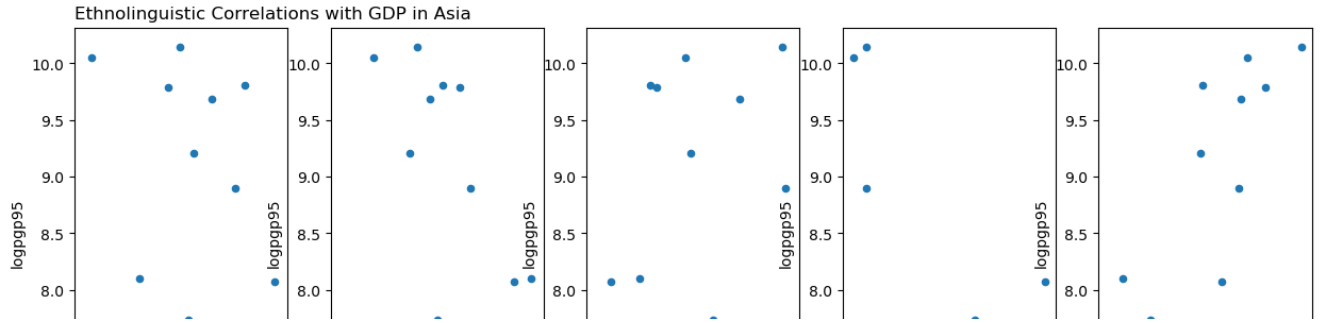
In [121]:

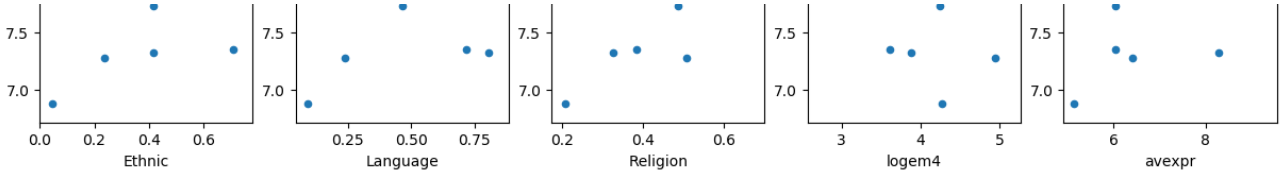
```
# Part (a)
fig, ax = plt.subplots(1,5)
plt.style.use('default')
Asia[['Ethnic', 'logpgp95']].plot.scatter('Ethnic', 'logpgp95', ax=ax[0], figsize = (15,5))
Asia[['Language', 'logpgp95']].plot.scatter('Language', 'logpgp95', ax=ax[1])
Asia[['Religion', 'logpgp95']].plot.scatter('Religion', 'logpgp95', ax=ax[2])
Asia[['logem4', 'logpgp95']].plot.scatter('logem4', 'logpgp95', ax=ax[3])
Asia[['avexpr', 'logpgp95']].plot.scatter('avexpr', 'logpgp95', ax=ax[4])

ax[0].set_title('Ethnolinguistic Correlations with GDP in Asia', loc = 'left')
```

Out[121]:

```
Text(0,1,'Ethnolinguistic Correlations with GDP in Asia')
```





Correlation Matrix - South America

In former South American colonies, Institutional Quality is highly correlated with Log GDP (0.67) as found in the paper; Log Mortality Rate has a moderate negative correlation with Log GDP (-0.34). Other variables do not appear to have meaningful correlations with Log GDP.

Among Variables, Log Mortality rate has a positive correlation with Religion diversity (0.34) and a negative correlation with institutional quality. Other variables do not appear to correlate significantly with each other.

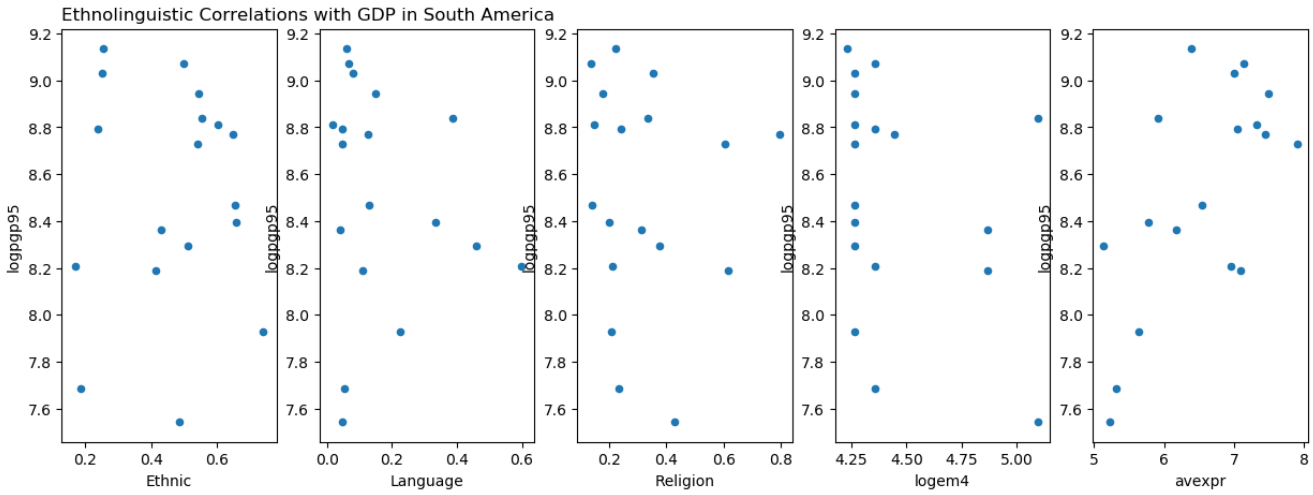
In [122]:

```
fig, ax = plt.subplots(1,5)
plt.style.use('default')
South_America[['Ethnic', 'logpgp95']].plot.scatter('Ethnic', 'logpgp95', ax=ax[0], figsize = (15,5)
)
South_America[['Language', 'logpgp95']].plot.scatter('Language', 'logpgp95', ax=ax[1])
South_America[['Religion', 'logpgp95']].plot.scatter('Religion', 'logpgp95', ax=ax[2])
South_America[['logem4', 'logpgp95']].plot.scatter('logem4', 'logpgp95', ax=ax[3])
South_America[['avexpr', 'logpgp95']].plot.scatter('avexpr', 'logpgp95', ax=ax[4])

ax[0].set_title('Ethnolinguistic Correlations with GDP in South America', loc = 'left')
```

Out[122]:

```
Text(0,1,'Ethnolinguistic Correlations with GDP in South America')
```



Correlation Matrix - Africa

Ethnicity (-0.48), Language (-0.32) and Log Mortality Rate (-0.43) all have moderate negative correlation with Log GDP. Institutional Quality is somewhat significantly correlated with Log GDP (0.49).

There is observed positive positive correlations within Language diversity, Religion and Ethnic diversity.

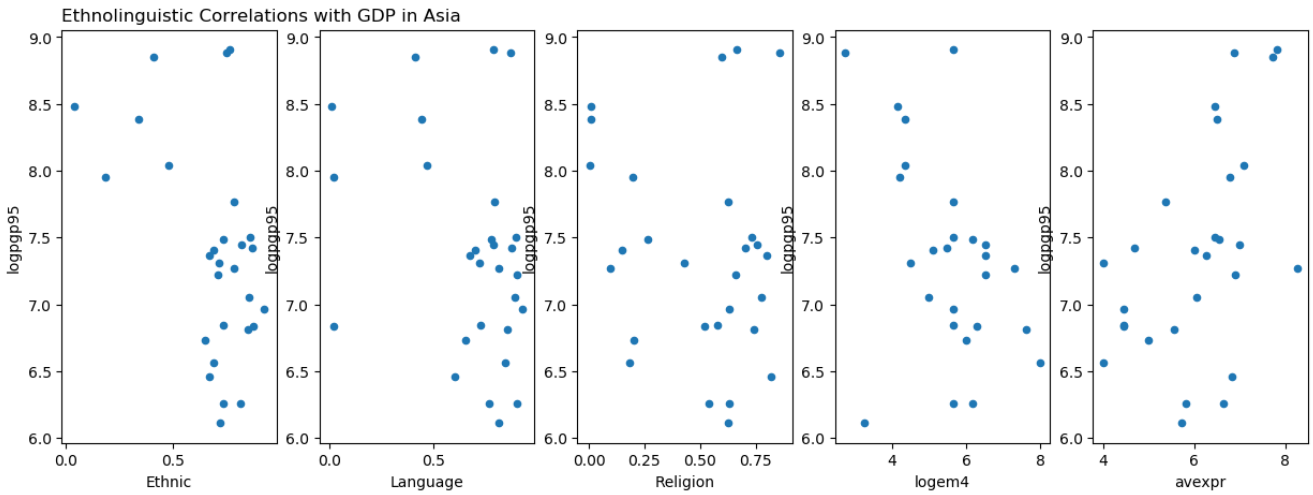
In [123]:

```
fig, ax = plt.subplots(1,5)
plt.style.use('default')
Africa[['Ethnic', 'logpgp95']].plot.scatter('Ethnic', 'logpgp95', ax=ax[0], figsize = (15,5))
Africa[['Language', 'logpgp95']].plot.scatter('Language', 'logpgp95', ax=ax[1])
Africa[['Religion', 'logpgp95']].plot.scatter('Religion', 'logpgp95', ax=ax[2])
Africa[['logem4', 'logpgp95']].plot.scatter('logem4', 'logpgp95', ax=ax[3])
Africa[['avexpr', 'logpgp95']].plot.scatter('avexpr', 'logpgp95', ax=ax[4])
```

```
ax[0].set_title('Ethnolinguistic Correlations with GDP in Asia', loc = 'left')
```

Out[123]:

```
Text(0,1,'Ethnolinguistic Correlations with GDP in Asia')
```



Regression Analysis

Links to Paper - Base Regressions

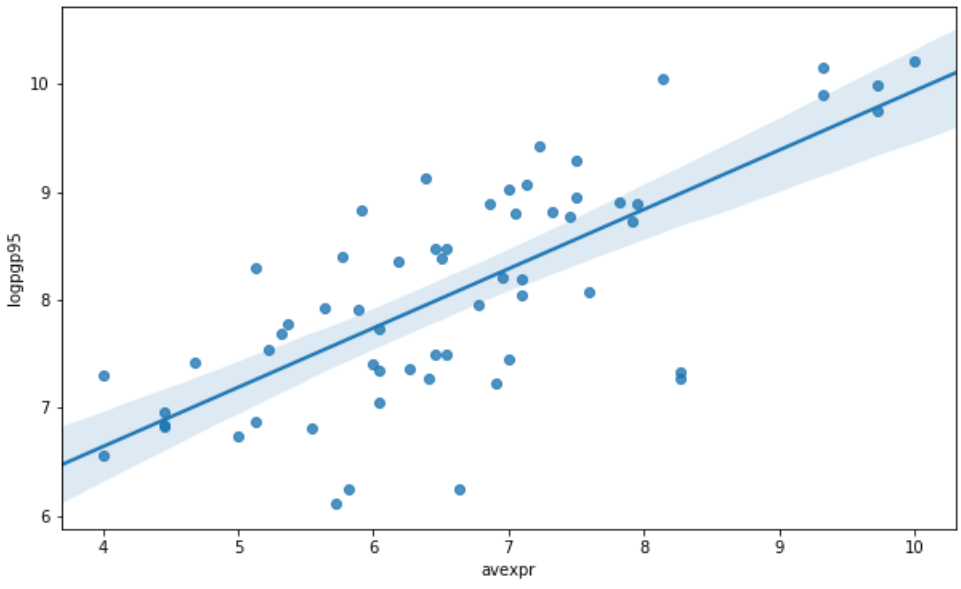
In [16]:

```
import seaborn as sns
fig,ax = plt.subplots(figsize=(10, 6))
sns.regplot(x="avexpr", y="logpgp95", data=df, ax=ax)
```

```
/Users/Sweta/anaconda3/lib/python3.7/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2318fa58>
```



As expected, the paper's main findings are very significant - with p values being almost 0 (this is in the regression output we ran as

As expected, the paper's main findings are very significant - with p-values being almost 0 (this is in the regression output we ran as part of our longer code - to make things simpler we just showed this relationship through the graph above). This shows that settler mortality rate affects the avg. protection against expropriation risk (institution quality) and institution quality affects the log GDP per capita.

Now to test for our hypothesis that:

1. Ethnographic factors also impact the GDP

We expect that with greater ethnolinguistic diversity, there would be more chance for civil unrest which would make running the country difficult - hence institutions would be weaker and thus GDP would be lower

In [128]:

```
print(smf.ols('logpgp95 ~ Ethnic + Religion + Language + avexpr', data = df).fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                logpgp95   R-squared:                       0.640
Model:                            OLS   Adj. R-squared:                  0.617
Method:                 Least Squares   F-statistic:                     27.54
Date:                Wed, 19 Dec 2018   Prob (F-statistic):           3.71e-13
Time:                        00:40:49   Log-Likelihood:                -65.606
No. Observations:                  67   AIC:                             141.2
Df Residuals:                      62   BIC:                             152.2
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.5076      0.528     10.422      0.000       4.451       6.564
Ethnic        -0.4762      0.475     -1.002      0.320      -1.427       0.474
Religion      -0.0327      0.383     -0.085      0.932      -0.798       0.733
Language      -0.9740      0.335     -2.908      0.005      -1.644      -0.304
avexpr         0.5034      0.068      7.449      0.000       0.368       0.638
==============================================================================
Omnibus:                        1.295   Durbin-Watson:                   2.170
Prob(Omnibus):                  0.523   Jarque-Bera (JB):                1.197
Skew:                          -0.318   Prob(JB):                        0.550
Kurtosis:                       2.840   Cond. No.                         54.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Findings:

It looks like ethnicity and religion are not significant regressors of GDP; however, language turned out to be a very significant (p-value <0.01) regressor of GDP. That means countries with more diverse languages tended to do worse than those with fewer languages. Possible explanations could be that with more languages, there is more disconnect between linguistic groups in the country. Maybe this would make it more difficult to implement standardized policies as people speak different languages and it could be hard to unify them to follow the economic policies.

We wanted to see whether this effect held true for each continent and so ran the same regressions based on each subset of continent data - Asia, Africa and South America. What we found, interestingly, was that none of the continent regressions showed the same pattern of having language as a significant regressor. Maybe the trend in the overall data set reflects the trend in Europe and the 'Neo-Europes' like Australia and Canada and these countries largely speak English. To check all the three regression outputs, you can take a look at the code we ran which looks similar to the regression output above.

We also wanted to see in case there was any relationship between the institution quality and these ethnographic factors, even though the correlation analysis before did not render very strong results:

In [67]:

```
print(smf.ols('avexpr ~ Ethnic + Religion + Language + logem4', data = df).fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  avexpr   R-squared:                       0.323
Model:                            OLS   Adj. R-squared:                  0.276
```

```
Method:                 Least Squares   F-statistic:                     6.906
Date:                Tue, 18 Dec 2018   Prob (F-statistic):           0.000129
Time:                        19:47:15   Log-Likelihood:                -97.359
No. Observations:                  63   AIC:                             204.7
Df Residuals:                      58   BIC:                             215.4
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.6777      0.658     13.183      0.000       7.360       9.995
Ethnic        -1.2776      0.918     -1.392      0.169      -3.115       0.560
Religion       1.3238      0.676      1.957      0.055      -0.030       2.677
Language       0.1684      0.637      0.264      0.793      -1.107       1.444
logem4        -0.4445      0.147     -3.016      0.004      -0.740      -0.149
==============================================================================
Omnibus:                        2.131   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.344   Jarque-Bera (JB):                1.417
Skew:                           0.334   Prob(JB):                        0.492
Kurtosis:                       3.305   Cond. No.                         34.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Findings

It doesn't look like any of them affect the institution quality - except the mortality rate and the religion variable (at a 10% sig. level) but we'll still check by continent. After running the same regressions on the three different data sets the results were inconclusive. None of the factors were significant in Africa or South America. Religion was significant in Asia at the 10% level. Possible explanations could be that greater religious diversity could promote tolerance and cooperation when making economic policies, which leads to more stable institutions. A good example of this would be India, which has a great religious diversity apart from being a former British colony that is performing very well economically in the past decade.

Limitations and Further Research:

The main limitation with our analyses is that we did not run 2-stage least squares (2SLS) regressions to test for endogeneity between variables. In the paper, because mortality affected institution quality and institution quality affected GDP, the researchers used 2SLS regressions to show this relationship more precisely. Then they also used instrumental variables regressions to control for endogeneity (when one variable affects both the dependent and the independent variable) by including them in the second stage. However, since we didn't conduct our tests using IV and 2SLS regressions, the results may be a little biased.

So for future consideration, it would be helpful to see whether having 2-stage least square regressions and instumental variables improve our predictions and R^2 values. Maybe by controlling for our Ethnicity, Religion and Language variables, we can get more accurate/unbiased results. For further research, it would also be interesting to see if there are any other variables that affect institution quality - like geography or climate.

Conclusion

Even though our correlation matrices show that mortality and institutional quality are two factors that are significantly correlated with GDP in all three continents, causal effects can not be established based on our first-stage least squares regressions. We hypothesized that institutional quality are affected by ethnographic factors such as language, ethnic and religious diversity, but the regression output shows that none of the variables are significant in our model (except for Religion in Asia).

In the future, we would need to run 2-stage least squares regressions to improve the validity of the our analysis. Introducing more variables to our analysis in the future could also help us capture all factors that have an impact on Log GDP. Multicollinearity among variables might further affect the regression output, therefore it would be helpful to use partial least squares regressions or principal component analysis by combining similar variables.

Link to full code: https://github.com/sg7731/Data-Bootcamp-Final-Project