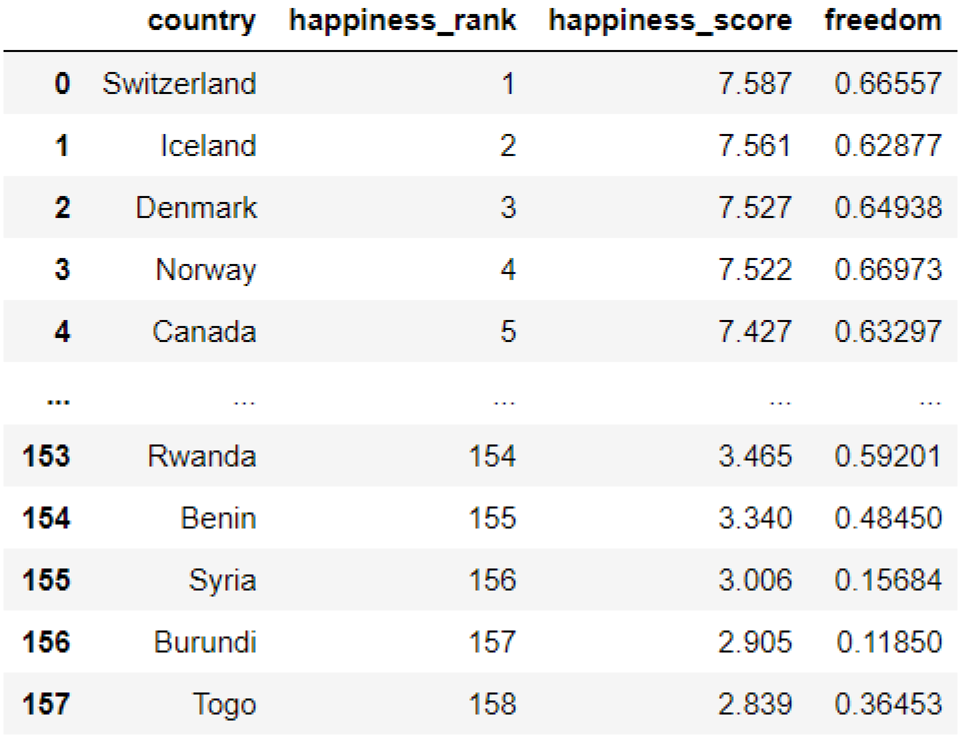
ETL Project
By: Catie Lutz, Chris Segretto, Summer Baptiste

Extraction

We used two Kaggle datasets: One is derived from happiness data collected by various countries in 2015 and the second is compiled of global suicide data between the years 1985 to 2015.

Transformation

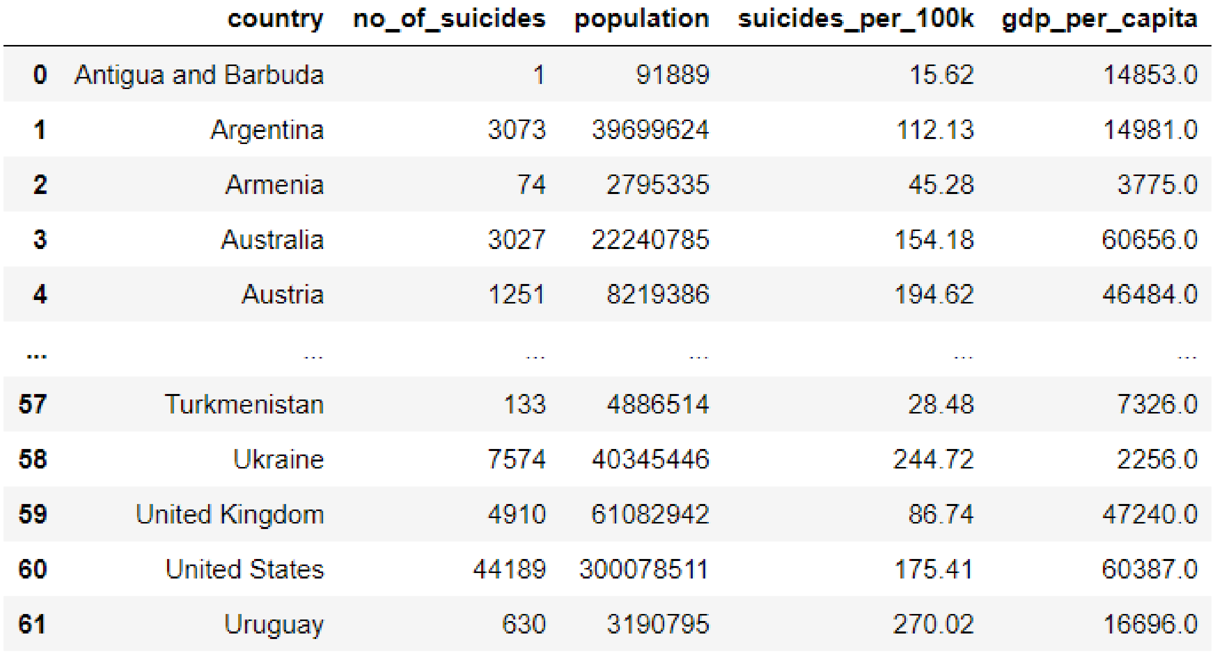
The first steps we took in cleaning the data was selecting all the variables that were relevant and deleting all the variables that were not relevant. Referring to the figure below, we selected the variables country, happiness rank, happiness score and freedom. The dataset that we pulled this from was based on the happiness collected from various countries (happy15.csv).



	country	happiness_rank	happiness_score	freedom
0	Switzerland	1	7.587	0.66557
1	Iceland	2	7.561	0.62877
2	Denmark	3	7.527	0.64938
3	Norway	4	7.522	0.66973
4	Canada	5	7.427	0.63297
...	...	...	...	...
153	Rwanda	154	3.465	0.59201
154	Benin	155	3.340	0.48450
155	Syria	156	3.006	0.15684
156	Burundi	157	2.905	0.11850
157	Togo	158	2.839	0.36453

We did the same with the second dataset by selecting all the variables that were relevant and deleting all the variables that were not relevant. Also, we did a groupby function for this particular dataset to filter out by country and run calculations to avoid duplicate data. Referring to the figure below, we selected the variables country, number of suicides, population, suicides per 100k and GDP per capita.

In this particular dataset (sad.csv) we extracted all the data only from 2015 because the dataset (happy15.csv) only had data from that particular year. We did this because we wanted the data to be pulled from each dataset only from year 2015. The dataset that we pulled this from was based on the number of suicides within the population of each country referring to the sadness (sad.csv).



	country	no_of_suicides	population	suicides_per_100k	gdp_per_capita
0	Antigua and Barbuda	1	91889	15.62	14853.0
1	Argentina	3073	39699624	112.13	14981.0
2	Armenia	74	2795335	45.28	3775.0
3	Australia	3027	22240785	154.18	60656.0
4	Austria	1251	8219386	194.62	46484.0
...	...	...	...	...	...
57	Turkmenistan	133	4886514	28.48	7326.0
58	Ukraine	7574	40345446	244.72	2256.0
59	United Kingdom	4910	61082942	86.74	47240.0
60	United States	44189	300078511	175.41	60387.0
61	Uruguay	630	3190795	270.02	16696.0

Load

Lastly, we were tasked with loading the final database and tables/collections. In doing so, we created a database and utilized Python Pandas to turn our cleaned *happy* and *sad* data frames into tables. Then, we used SQLAlchemy to connect and load our Postgres database.