Group: Jin, Nate, & Emma
Unit 13 ETL Project

Project Proposal

For our ETL project, we would like to analyze happiness and freedom data. We want to evaluate how a country's perception of political, social, and individual freedoms affect happiness. We would assume that countries with greater freedoms have higher happiness scores.

Finding Data

To find our data sources, we used Kaggle.com as reference and inspiration. As a group, we decided to focus on two csv files:

1. freedom.csv
   Website: https://www.cato.org/human-freedom-index-new
   Kaggle: https://www.kaggle.com/gsutters/the-human-freedom-index#hfi_cc_2018.csv
2. happiness.csv
   Website: https://ourworldindata.org/happiness-and-life-satisfaction

Data Cleanup & Analysis

The freedom.csv focuses on human rights around the world, from various sample years, using a scoring system 0 through 10 with 10 being "more free". The data looks at different categories of "freedoms" relating to violence, female treatment, judicial systems, personal and social rights, military intervention, etc. adding up to about 115 columns worth of indicators of "freedom". From these columns we determined 9 that we wanted to further analyze.

1. #pf_expression, Freedom of expression
2. #ef_legal_protection, Protection of property rights
3. #ef_legal_integrity, Integrity of the legal system
4. #pf_identity_sex, Same-sex ralitionships
5. #pf_religion, Religious freedom
6. #pf_association_assembly, Freedom of assembly
7. #ef_trade_movement_visit, Freedom of foreigners to visit
8. #ef_legal_military, Military interference in rule of law and politics
9. #ef_money_inflation, Inflation: most recent year

After loading the data into Jupyter Notebook, to clean, and determining which columns to keep, we renamed the columns using Pandas for simplicity.

With the variables listed above, we want to compare happiness scores across countries. We would assume that countries with greater freedom report higher happiness scores and therefore are generally happier than countries with lower perceived freedoms.

The happiness.csv shows survey data that measures life satisfaction and happiness throughout the globe. The values range from 0 to 10 with 10 being “happier”. This file is much smaller with only 4 columns: region, country code, year, and happiness score.

Some limitations to our data sets:

1. Happiness and freedom are relative and difficult to accurately measure as human perception of happiness and freedom can be defined differently from person to person.
2. The data between the two csv files are not perfectly matched. There are blanks and null values. The years do not always correlate per country and not each country participated each time, each year so the data may be skewed.

Type of Transformation:

In order to clean the data, we decided to use Pandas to filter and join tables. Utilizing Pandas was the first step in our data cleanup process and using pgAdmin 4 with the help of SQLalchemy was the second step.

We created two tables to be imported into a relational database:

Values Table

code	year	happiness_score	religion	assembly	expression	same_sex	legal_protection	military_interference	legal_system	inflation	foreigners_visit
AGO	2011	5.589001	7	2.5	6.7	0	2.9	3.3	4.2	7.3	0
ALB	2009	5.485470	9.8	-	7.7	10	3.9	8.3	4.2	9.5	0
ARE	2009	6.866063	6	2.5	6.6	0	6.7	8.3	6.7	9.7	3.4
ARG	2008	5.961034	8.9	10	8.6	10	3.2	7.5	4.2	6.9	4.6
ARM	2008	4.651972	7.4	-	5.7	10	5.4	5.8	5	8.2	1.1
...	...	...	...	...	...	...	...	...	...	...	...
VNM	2008	5.480425	6.4	2.5	4.6	10	5.7	5	6.7	5.4	0.6
YEM	2009	4.809259	-	-	-	-	-	-	-	-	-
ZAF	2008	5.346307	8	10	9.1	10	8.1	8.3	4.2	7.7	8
ZMB	2008	4.730263	7.7	5	7.2	5	5.8	8.3	6.7	7.5	3.7
ZWE	2008	3.174264	8.2	5	6	5	1.7	3.3	5	9.5	3.2

code	countries	region
AGO	Angola	Sub-Saharan Africa
ALB	Albania	Eastern Europe
ARE	United Arab Emirates	Middle East & North Africa
ARG	Argentina	Latin America & the Caribbean
ARM	Armenia	Caucasus & Central Asia
...	...	...
VNM	Vietnam	South Asia
YEM	Yemen	Middle East & North Africa
ZAF	South Africa	Sub-Saharan Africa
ZMB	Zambia	Sub-Saharan Africa
ZWE	Zimbabwe	Sub-Saharan Africa

Countries Table

These tables were also created in order to normalize our data.

For our countries table and values table, uploaded into pgAdmin4, we determined country code as the primary key for both. Our tables reflect a one-to-many relationship because one record, country code, relates to the country code in the other table.

Project Report

Extract: Our data was extracted into CSV files and we used Jupyter Notebook and pgAdmin 4 to edit, filter, and execute our data analysis.

Transform: In terms of transforming the data, we first uploaded our two CSVs into Jupyter Notebook. We wanted to merge our freedom.csv and happiness.csv together. We first converted the individual CSVs to tables and then combined the tables using a left join on country code.

Load: After cleaning and merging the data, we converted our tables into SQL files in Jupyter Notebook. And from there, our data was uploaded to pgAdmin4 where we created tables to analyze our data.

Analysis

	countries text	happiness_score double precision
1	Indonesia	4.815309525
2	Bangladesh	4.670460760625
3	Venezuela	4.041114807
4	Cameroon	4.449212074125
5	Uganda	4.2459688662
6	Montenegro	4.8010602
7	Jordan	4.8690702915
8	Dominican Re...	4.7768793103333325
9	Cambodia	4.1540832758
10	Macedonia	4.624328231800001
11	Sri Lanka	4.288902150222222

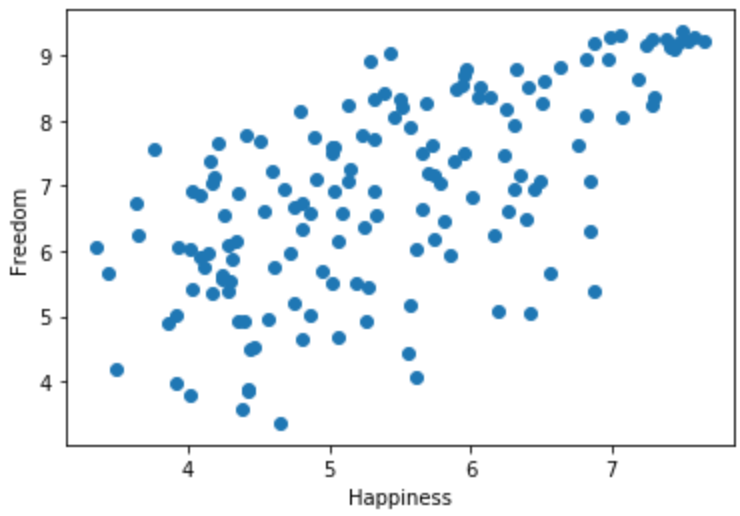
pgAdmin Table #1 - Average Happiness Score

This table displays the average happiness score per country. The results are as expected. The listed countries here are known to have political strife and social-economic issues that would probably generate lower life satisfaction and lesser happiness. More analysis needs to be conducted on this table as we'd like to explore the outliers (if this project was longer and more robust).

pgAdmin Table #2 - A look at one specific year (2008)

	code text	countries text	year bigint	happiness_score double precision	expression text	legal_protection text
1	ARG	Argentina	2008	5.9610342979999995	8.6	3.2
2	AUS	Australia	2008	7.253757477000001	9.4	8.6
3	AUT	Austria	2008	7.180953979	9.3	9
4	BGD	Bangladesh	2008	5.052278519	8.1	4.1
5	BLR	Belarus	2008	5.463332176000001	NA	NA
6	BEL	Belgium	2008	7.116590977	9.7	7.9
7	BOL	Bolivia	2008	5.297872543	8.3	2
8	BWA	Botswana	2008	5.451147079	8	7.2
9	BRA	Brazil	2008	6.691424847	8.7	5.6
10	CAN	Canada	2008	7.485603809	9.5	8.7
11	CHL	Chile	2008	5.789438725	9.1	7.5

With this table, we look at how 2008's economic downturn affected a country's happiness score, as well as some of their freedom scores. If we had more time, we would analyze year-over-year trends of happiness and freedom and compare the data to this specific point in time. With only this table, it is difficult to determine if the economic crash had any effect on happiness. Given our assumptions about happiness and prosperity, we would assume that The Great Recession did play at least some part in happiness declines. For example, if Austria typically reported in happiness scores of 9 or 10 prior to 2008, we could assume that the stock market crash may have been a contributor to this lowered score. Given more time and scope, we would work to investigate further.



Pandas Table - Does Happiness and Freedom correlate? YES

Using a scatter plot, we determined that happiness and freedom have a relatively strong positive correlation with the correlation coefficient between happiness and freedom being 0.63. In order to ready the data for plotting, we needed to find the averages for all of our variables. This graph displays the average freedom and happiness scores of all the countries in our data set.

Please see the files in our project github for the codes and to observe greater detail into how we constructed our tables and cleaned our data. Thank you.