



The Well Being of Women

By: **Betsy Kalkwarf**

Roopa Raghav,

Jessica Martin,

Paola Rivas



Reason Topic was Selected

As women, we care about women's well-being and strive to figure out what aspects of life most impact if women live long and happy lives.



Sources of data

3 Main Sources

1. LivWell (Kaggle)

2. Latitude/Longitude (Kaggle)

3. GDP (world bank data)

1. Livwell (Kaggle)

- a. LivWell is a global longitudinal database
- b. Provides a range of key indicators related to:
 - i. Women's socioeconomic status, health and well-being,
 - ii. Women's access to basic services, and demographic outcomes.
- c. <https://www.kaggle.com/datasets/konradb/wellbeing-of-women-in-52-countries?resource=download>

2. Latitude and Longitude (Kaggle)

- a. Latitude and Longitude for Every Country and State
- b. Provides the GPS coordinates for every world country and every USA state
- c. https://www.kaggle.com/datasets/paultimothymooney/latitude-and-longitude-for-every-country-and-state?select=world_country_and_usa_states_latitude_and_longitude_values.csv

3. GDP (World Bank Data)

- a. World Bank national accounts data, and OECD National Accounts data files
- b. <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD>



Questions the team hopes to answer with data

- Is there a relationship between country demographics and aspects of life indicators (domestic violence rate, marriage age, years of education, and fertility rate) that impact women's overall well-being?
- Does GDP relate to these aspects of life?



Data Exploration and Analysis



Data Exploration

Data Cleaning:

- Pandas will be used to clean the data and perform an exploratory analysis.
- Further analysis will be completed using Python
 - Image below shows the main stats from the LivWell data

```
In [19]: livewell_etl_df = livewell_etl_df.loc[livewell_etl_df['year'] >= 2000]
livewell_etl_df.describe()
```

Out [19]:

	year	DM_age_15.19_p_se	DM_age_20.24_p_se	DM_age_25.29_p_se	DM_age_30.34_p_se	DM_age_35.39_p_se	DM_age_40.44_p_se
count	5967.000000	5967.000000	5967.000000	5967.000000	5967.000000	5967.000000	5967.000000
mean	2007.827216	1.243674	1.23557	1.239853	1.150581	1.063693	0.969010
std	4.755202	0.576403	0.48798	0.463376	0.411243	0.381618	0.374867
min	2000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2004.000000	0.876000	0.920000	0.935167	0.877321	0.808000	0.718000
50%	2008.000000	1.226667	1.182000	1.184000	1.104000	1.020000	0.920000
75%	2012.000000	1.557750	1.467750	1.457321	1.360000	1.262678	1.141833
max	2019.000000	7.924000	6.154000	6.020000	4.656000	4.016000	4.756000

8 rows x 38 columns

Data Exploration

Database Storage:

- AWS RDS is the database setup
- Integration of Postgres sql to display the ETL process for country demographics.
- Created S3 buckets and uploaded data
- Rearranged columns
- Dropped null values into another dataframe
- Uploaded cleaned dataframe into database table in PostgreSQL for further analysis



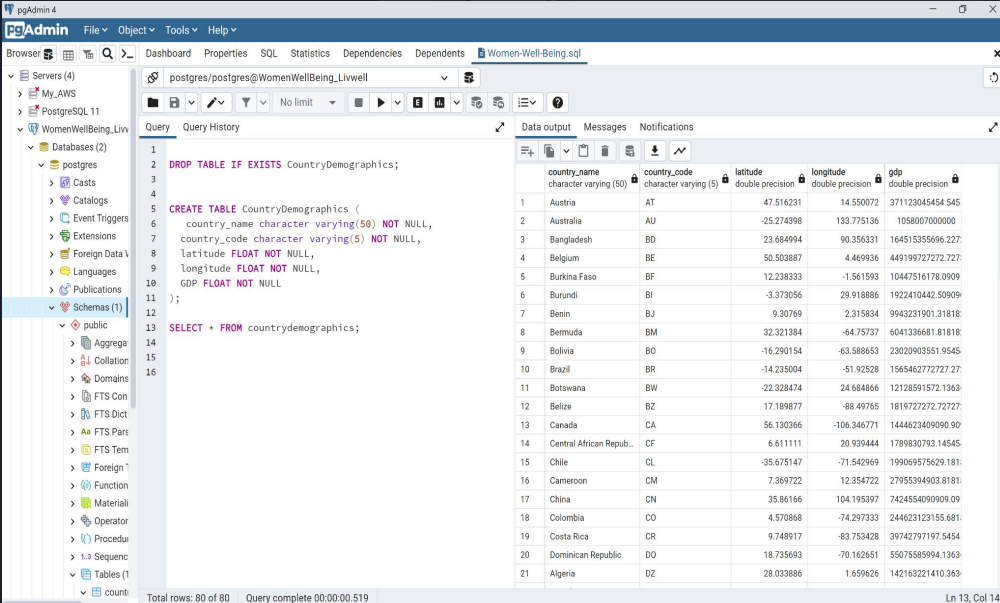
Data Exploration

Database Storage:

- Screenshot of database updations

- Links:

- PostgreSQL
 - <https://github.com/Betsy-Kalkwarf/Women-Well-Being/blob/main/Women-Well-Being.sql>
- CountryETL.ipynb
 - <https://github.com/Betsy-Kalkwarf/Women-Well-Being/blob/main/CountryETL.ipynb>



The screenshot shows the pgAdmin 4 interface. The left sidebar displays the database structure, including Servers, Databases, Schemas, and Tables. The main window shows a SQL query being executed in the 'Query' tab. The query consists of two parts: a table drop statement and a table creation statement. The results are displayed in the 'Data output' tab, showing a table with 5 columns: country_name, country_code, latitude, longitude, and gdp. The table contains 21 rows of data, representing various countries and their corresponding codes, latitudes, longitudes, and GDP values.

	country_name	country_code	latitude	longitude	gdp
1	Austria	AT	47.516231	14.550072	371123045454.545
2	Australia	AU	-25.274398	133.775136	1058007000000
3	Bangladesh	BD	23.684994	90.356331	164515355696.227
4	Belgium	BE	50.503887	4.469936	449199727272.727
5	Burkina Faso	BF	12.238333	-1.561593	10447516178.0909
6	Burundi	BI	-3.373056	29.918886	192241042.50909
7	Benin	BJ	9.30769	2.315834	9943231901.31818
8	Bermuda	BM	32.321384	-64.75737	6041336681.81818
9	Bolivia	BO	-16.290154	-63.588653	23020903551.9545
10	Brazil	BR	-14.235004	-51.92528	156546277272.727
11	Botswana	BW	-22.328474	24.684866	12128591572.1363
12	Belize	BZ	17.189877	-88.49765	1819727272.72727
13	Canada	CA	56.130366	-106.346771	1444623409090.90
14	Central African Repub..	CF	6.611111	20.939444	1789830793.14545
15	Chile	CL	-35.675147	-71.542969	199069575629.181
16	Cameroon	CM	7.369722	12.354722	27955394903.8181
17	China	CN	35.86166	104.195397	742454090909.09
18	Colombia	CO	4.570868	-74.297333	244623123155.681
19	Costa Rica	CR	9.748917	-83.753428	39742797197.5454
20	Dominican Republic	DO	18.735693	-70.162651	58075585994.1363
21	Algeria	DZ	28.033886	1.659626	142163221410.363

Data Analysis

Machine Learning:

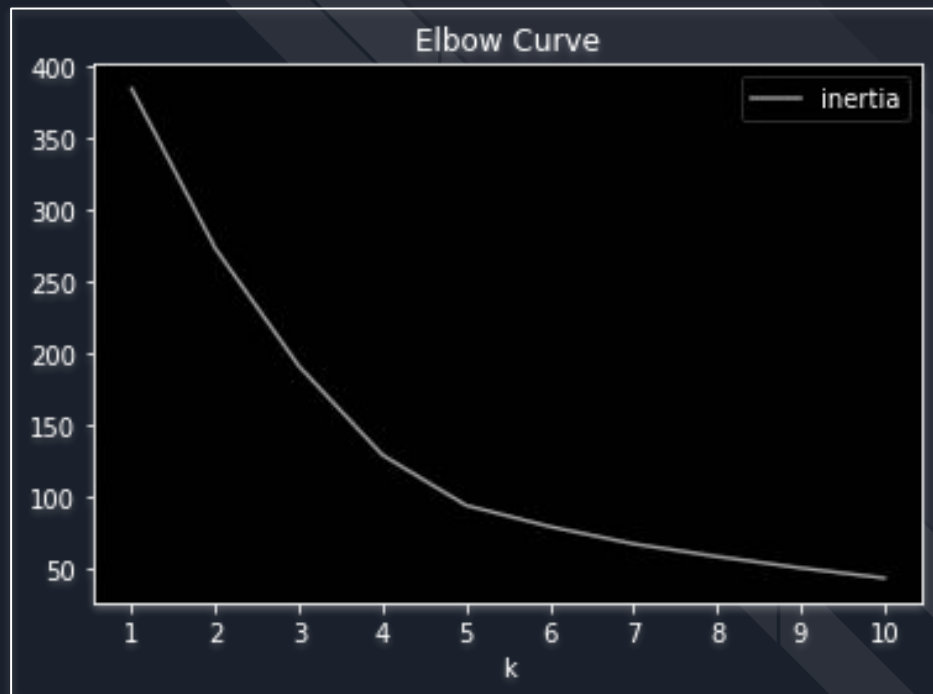
- **Training and Test set up is Unsupervised**
 - Chosen due to source not having any predictions
 - Wanted to cluster indicators chosen based on the country
- **SciKit Learn is the machine learning library we'll be using to create a classifier.**
 - Jupyter notebooks:
 - [ML_indicators.ipynb](#)
 - [CountryETL.ipynb](#)

Data Analysis

Machine Learning:

- Data was retrieved from database
- Set up ML model
- Scaled, fit and transformed the data
- Applied PCA for reduction
- Checked the Elbow Curve to decide the best K-value for clustering

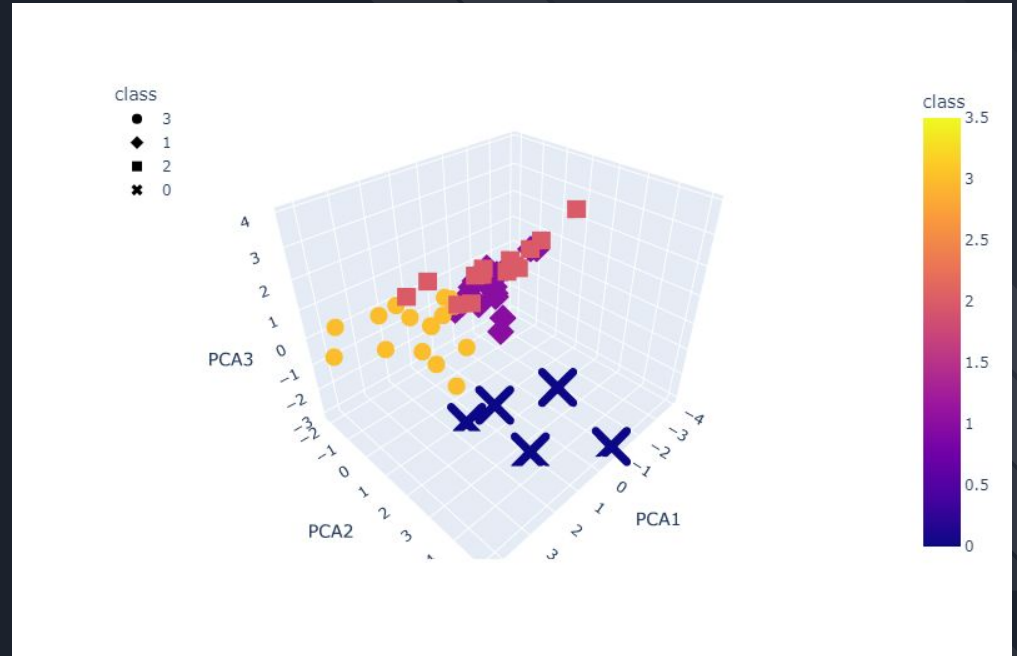
Edit the elbow curve pic to show which data we chose



Data Analysis

Machine Learning:

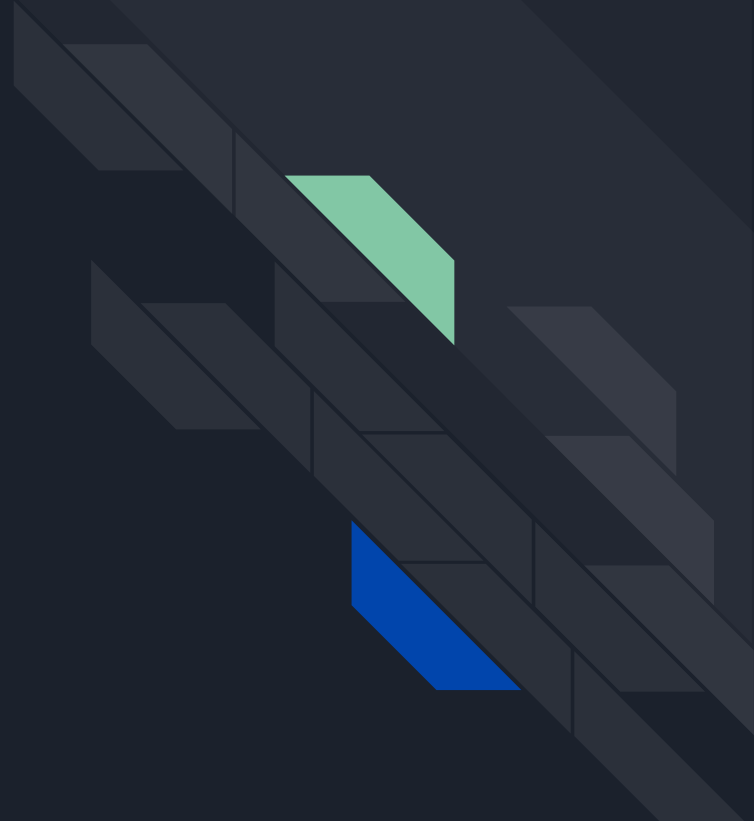
- 3D scatter plot created to check clusters
- Links
 - Code:
 - https://github.com/Betsy-Kalkwarf/Women-Well-Being/blob/main/ML_Indicators.ipynb
 - Picture:
 - <https://github.com/Betsy-Kalkwarf/Women-Well-Being/blob/main/Resources/PCA-Cluster.png>



Data Analysis

Dashboard:

The dashboard is hosted on [Tableau](#).





Thanks!

Are there any
questions?

