Extraction, Transformation & Load Technical Report

The Impact of Non-Traditional Variables on Happiness Metrics

The Happiness Group:

Edgar Mendez

Daniah Al-Bayati

Ralph Baroud

Table of Contents:

I. Introduction:

1.1 Summary

The objective of this project is to provide enough data to try and understand through if certain non-traditional factors or metrics could play a significant role in determining the happiness of a country's citizens. While the world happiness report does incorporate significant factors to calculate a final score, there are important variables we believe to be relevant in determining happiness that were not present. These non-traditional metrics include the number of days of sunshine per year and the fertility/birth rate, among others. Our findings range from the years 2015 to 2019 inclusive.

1.2 Scope

The scope of this project included gathering all the relevant data from the appropriate sources for our chosen timeframe. This entailed the extraction of information from supranational and governmental organizations that compile the data on a yearly basis. Following extraction, the data was cleaned and further manipulated to deliver it in a concise and straightforward manner.

1.3 Technologies

The technologies and/or software used in this project include the following:

- Microsoft Office Suite –Microsoft Excel for data and Microsoft Word for the writing of this report
- Python – Pandas Library
- SQL – PostgreSQL

1.4 Definitions & Sources

- **World Happiness Report**: the world happiness report is an annual publication by the United Nations Sustainable Development Solutions Network. The report contains rankings of citizens' happiness within a country based on respondent ratings while correlating with several life factors.
- **Fertility/Birth Rate**: The fertility rate (births per woman) is compiled by the World Bank on an annual basis.
- **Crime Index**: the crime index is a representation of crime statistics by country. The report is compiled primarily from surveys of the population and statistics provided by the relevant law enforcement authorities.
- **Global Innovation Index**: The Global Innovation Index is "an annual ranking of countries by their capacity for, and success in, innovation". This index was used as a proxy for technological advancement as technology is – in economic terms – a broad measure of efficiency which in turn is impossible without innovation. The index is published by Cornell University, INSEAD, and the World Intellectual Property Organization.
- **Days of Sunshine per Country**: these statistics were produced by the World Meteorological Organization. Its parent organization is the United Nations Economic and Social Council.

II. ETL Details

2.1 Data Import/Extract Method and Transformation/Manipulation

The bulk of our data was download directly from the above-mentioned sources in Excel or Comma-separated values files for the 2015-2019 timeframe inclusive. After the necessary cleaning using both Pandas and Microsoft Excel, the data was subsequently loaded in PostgreSQL where a relational database with primary and secondary keys in each table was created. The primary and secondary keys are country and year, respectively.

Redundant columns were eliminated using either Pandas or Microsoft Excel. Pandas was also used to obtain a definitive list of countries that is consistent across our database. The data was then sorted in an ascending fashion by criteria value with the corresponding year beginning from 2015.

Averaging was carefully made for data related to weather since it was reported for specific stations by country and not on a country-wide basis. We averaged the monthly hours of sunshine for each station in each country to reach an annual average which we then divided by 24 (hours) so that we could convert the metric to days of sunshine per year as opposed to hours of sunshine per year.

2.2 Data Integrity

We believe that the data's integrity was not comprised during any step. Measures were taken to ensure that the data downloaded is relevant, that the sources are reliable, and that it fits our end-goal. Some conservative calculations were made to fill missing data points as the data is not available for countries with lax and/or no adequate reporting during certain years.

The conservative calculations entailed the calculation of geometric mean based on the present beginning and ending values. The geometric mean enabled us to obtain a certain growth rate (in percentage) which was used to estimate some missing values.

It is possible to subscribe and receive notification regarding the release and/or update of data with some of the public sources listed such as the World Bank, but others require the manual checking and retrieval of the data. We would continue to regularly check and send updates to the client and/or user. However, due to the nature of the data and different reporting standards among the supranational and governmental agencies, it would be difficult to accurately predict when and where the data will be released hence manual checking is required.

2.3 Data Refresh Frequency

The data in question would be refreshed annually upon its publication/release by the relevant organizations except for data related to weather since it is significantly difficult to obtain. The World Meteorological Organization does not report on a continuous basis.

2.4 Data Loading & Availability

The cleaned data can be readily loaded into any database software post-cleaning while the raw data is publicly available from the previously mentioned sources. Our final data can be accessed using the provided comprehensive Excel file, individual comma-separated values files, or the relational database that was created using PostgreSQL.

2.5 Data Security

The data extracted has no anonymity or security requirements that must be satisfied since it is considered in the public domain and is available online for free from the above-mentioned organizations.

III. Data Quality

3.1 Success Criteria

The success criteria for this project would be the accurate compilation, cleaning, and manipulation of our data sets with the objective of creating a clear and concise database that would offer additional insights once statistical calculations are made. We believe that we have met our success criteria as can be seen from our relational database or the comprehensive Microsoft Excel or comma separated values files.

3.2 User Acceptance Testing

Due to the mostly descriptive nature of our data, there is no real process of verifying that the data we have provided works for the user apart from making sure that the data is complete across the selected criteria. We believe this step was taken when we created our relational database, implemented keys, and subsequently made sure it would return desired queries when prompted. Furthermore, this would entail that any desired statistical analysis can be performed by the client and/or end user.

3.3 Assumptions

We assume that the extracted/downloaded data to be accurate since it is compiled by reliable supranational and governmental sources.

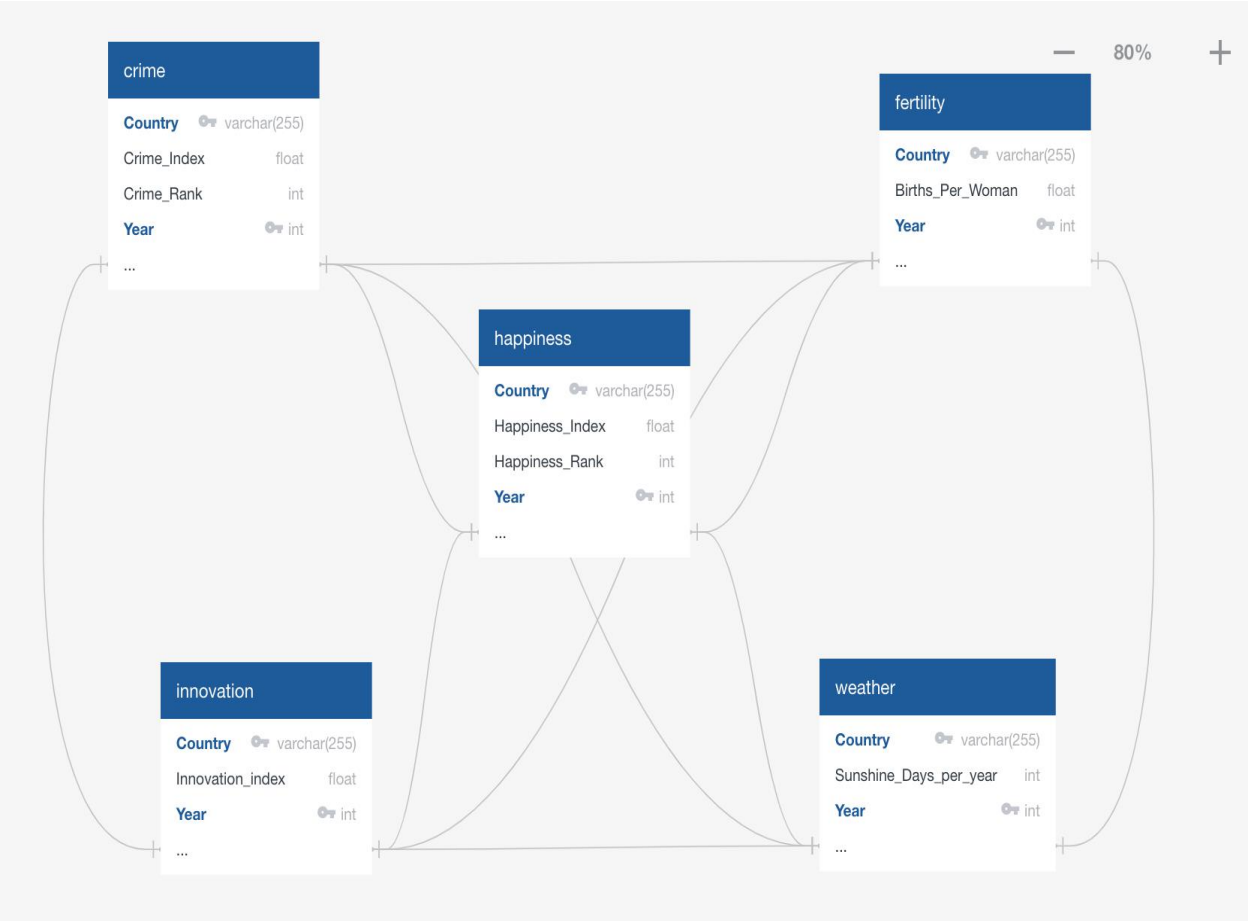
With respect to the averaging of the weather data provided for periods ranging from the year 1960 up to the year 2000 inclusive, we assume that weather patterns have remained mostly unchanged throughout the decades up until the present day. This insinuates that while some countries have been warmer or colder due to recent climate change, the number of hours and/or days of sunshine a country receives has not significantly changed.

We believe the growth rate calculated using a geometric average to be an accurate measure of the underlying trend since we have eliminated countries with 3 data points or less for all our selected criteria and across the years. This was done as to not infringe on the integrity of the data as it would cause the geometric mean – and consequently the estimated growth rate – to be inaccurate due to the little or no inputs.

IV. Database Information

PostgreSQL was our database of choice due to the hierarchical similarity between the tables and the potential expandability of the underlying database once new data is released by the relevant supranational or governmental agencies.

ENTITY RELATIONSHIP DIAGRAM



PGADMIN4. Happiness_db data samples:

select * from happiness limit 5;

happiness_db/postgres@PostgreSQL 12

Data Output

	country [PK] character varying (255)	happiness_index double precision	happiness_rank integer	year [PK] integer
1	Switzerland	7.587	1	2015
2	Iceland	7.561	2	2015
3	Denmark	7.527	3	2015
4	Norway	7.522	4	2015
5	Canada	7.427	5	2015

select * from crime limit 5;

happiness_db/postgres@PostgreSQL 12

Data Output

	country [PK] character varying (255)	crime_index double precision	crime_rank integer	year [PK] integer
1	South Sudan	85.32	1	2015
2	Venezuela	84.07	2	2015
3	Guatemala	79.34	3	2015
4	South Africa	78.44	4	2015
5	Afghanistan	77.34	5	2015

select * from innovation limit 5;

happiness_db/postgres@PostgreSQL 12

Data Output

	country [PK] character varying (255)	innovation_index double precision	year [PK] integer
1	Switzerland	68.3	2015
2	Sweden	62.4	2015
3	United Kingdom	62.4	2015
4	Netherlands	61.6	2015
5	United States	60.1	2015

select * from weather limit 5;

happiness_db/postgres@PostgreSQL 12

Data Output

	country [PK] character varying (255)	sunshine_days_per_year integer	year [PK] integer
1	Afghanistan	124	2015
2	Argentina	55	2015
3	Australia	114	2015
4	Bahamas	119	2015
5	Bahrain	139	2015

select * from fertility limit 5;

happiness_db/postgres@PostgreSQL 12

Data Output

	country [PK] character varying (255)	births_per_woman double precision	year [PK] integer
1	Aruba	1.85	2015
2	Afghanistan	4.98	2015
3	Angola	5.77	2015
4	Albania	1.68	2015
5	United Arab Emirates	1.54	2015

```
SELECT a.country, a.happiness_index, a.happiness_rank, a.year, b.births_per_woman
 FROM happiness a, fertility b
 WHERE a.country = b.country and a.year = b.year
 order by b.births_per_woman desc;
```

happiness_db/postgres@PostgreSQL 12

Data Output

	country character varying (255)	happiness_index double precision	happiness_rank integer	year integer	births_per_woman double precision
1	Niger	3.845	138	2015	7.17
2	Niger	3.856	137	2016	7.09
3	Niger	4.028	133	2017	7
4	Niger	4.166	131	2018	6.91
5	Nigeria	5.265	84	2019	6.85
6	Mali	3.995	133	2015	6.15
7	Mali	4.073	131	2016	6.06
8	Chad	3.667	143	2015	6.05
9	Serbia	5.603	70	2019	6
10	Mali	4.19	125	2017	5.97
11	Chad	3.763	138	2016	5.95
12	Mali	4.447	116	2018	5.88
13	Chad	3.936	135	2017	5.85
14	Malta	6.726	22	2019	5.81
15	Angola	4.033	132	2015	5.77
16	Chad	4.301	128	2018	5.75
17	Burundi	2.905	151	2015	5.7

Linking all 5 tables:

```
SELECT h.country, h.happiness_index, h.happiness_rank, h.year, f.births_per_woman, c.crime_index, c.crime_rank,
w.sunshine_days_per_year, i.innovation_index

 FROM happiness h, fertility f, crime c, weather w, innovation i

 WHERE h.country = f.country and f.country = c.country and c.country = w.country

    and w.country = i.country and h.year = f.year and f.year = c.year and c.year = w.year

    and w.year = i.year

    and h.year = 2015;
```

happiness_db/postgres@PostgreSQL 12

Data Output

	country character varying (255)	happiness_index double precision	happiness_rank integer	year integer	births_per_woman double precision	crime_index double precision	crime_rank integer	sunshine_days_per_year integer	innovation_index double precision
1	United Arab Emirates	6.901	20	2015	1.54	22.22	118	75	40.1
2	Argentina	6.574	29	2015	2.3	62.4	27	55	34.3
3	Australia	7.284	10	2015	1.81	42.16	73	114	55.2
4	Belgium	6.937	19	2015	1.7	42.04	75	62	50.9
5	Bulgaria	4.218	129	2015	1.53	43.45	69	88	42.2
6	Bahrain	5.96	47	2015	2.06	39.97	80	139	37.7
7	Belarus	5.813	57	2015	1.72	30.5	104	74	38.2
8	Brazil	6.983	16	2015	1.75	68.95	18	94	34.9
9	Switzerland	7.587	1	2015	1.54	26.77	114	97	68.3
10	China	5.14	82	2015	1.67	41.75	78	100	47.5
11	Colombia	6.477	32	2015	1.86	56.88	34	82	36.4
12	Cyprus	5.689	65	2015	1.35	31.56	102	138	43.5
13	Czech Republic	6.505	30	2015	1.57	32.89	101	66	51.3

V. Sources

5.1 World Happiness Index:

https://www.kaggle.com/unsdsn/world-happiness

https://countryeconomy.com/demography/world-happiness-index

5.2 Fertility/Birth Rate

https://data.worldbank.org/indicator/SP.DYN.TFRT.IN

5.3 Crime Index

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2015&displayColumn=0

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2016&displayColumn=0

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2017&displayColumn=0

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2018&displayColumn=0

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2019&displayColumn=0

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2020&displayColumn=0

5.4 Global Innovation Index

https://www.globalinnovationindex.org/analysis-indicator

5.5 Days of Sunshine per Country

http://data.un.org/Data.aspx?q=sunshine&d=CLINO&f=ElementCode%3a15