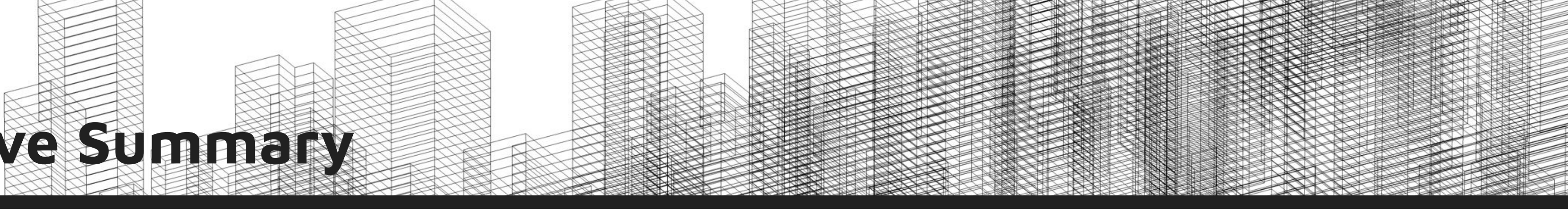
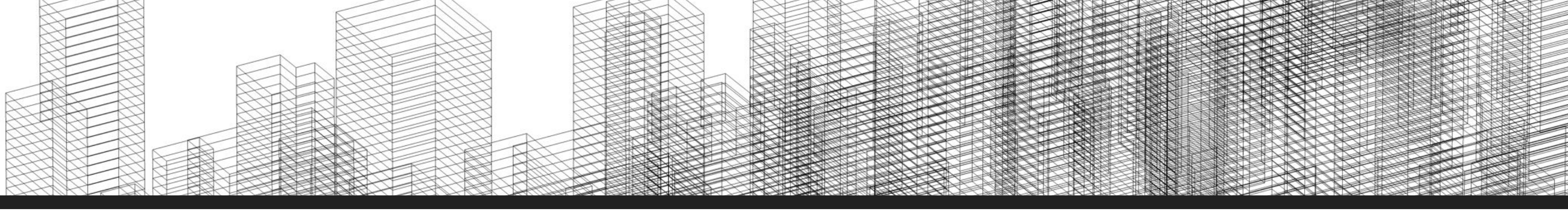
What Determines Happiness?

Final Project

Executive Summary



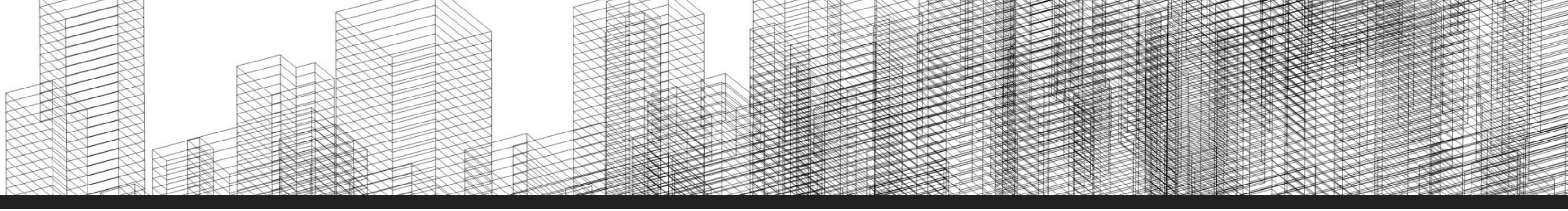


Team Members / Roles

Marc Corti - Data Engineer

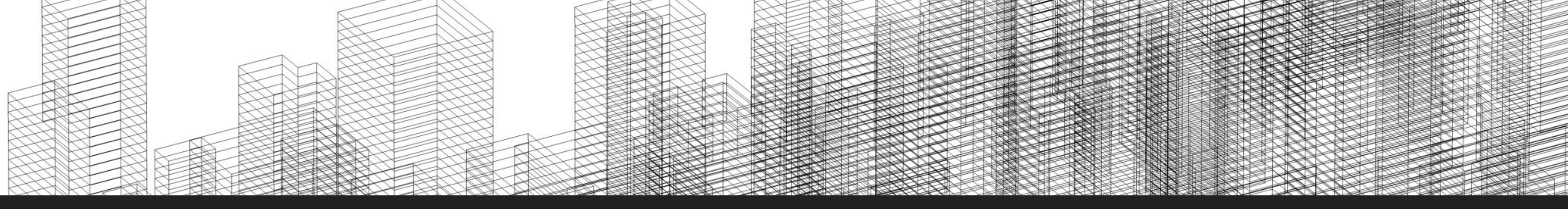
Zohair Khan - Machine Learning

Nancy Condon - Dashboard



Happiness

“Happiness is an emotional state characterized by feelings of joy, satisfaction, contentment, and fulfillment. While happiness has many different definitions, it is often described as involving positive emotions and life satisfaction.” - verywellmind.com

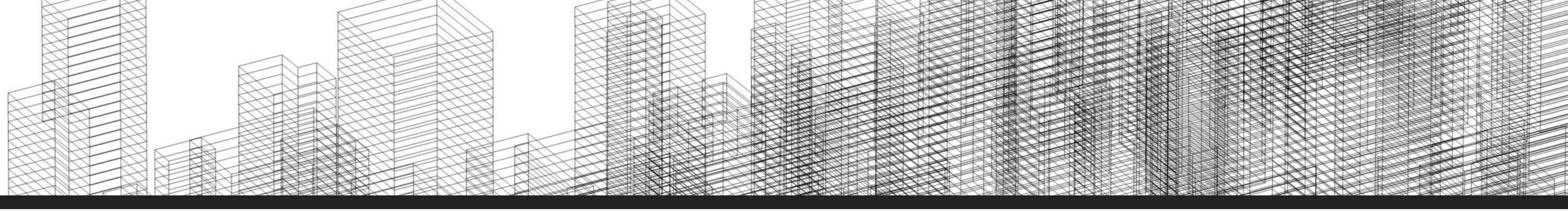


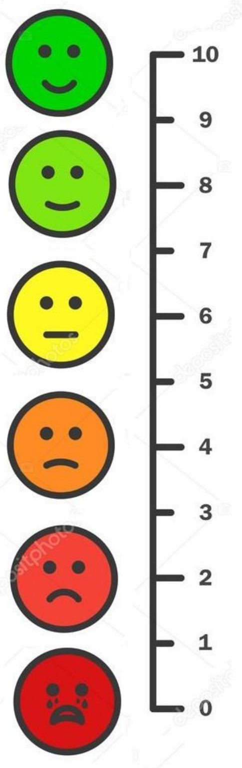
World Happiness Report 2019

The World Happiness Report is a publication of the Sustainable Development Solutions Network.

The 2019 report uses data from the Gallup World Poll surveys.

World rankings are based on answers to the main life evaluation question asked in the poll. Based on the Cantril Self-Anchoring Striving Scale (Cantril, 1965), respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. (This is known as the Life Ladder)



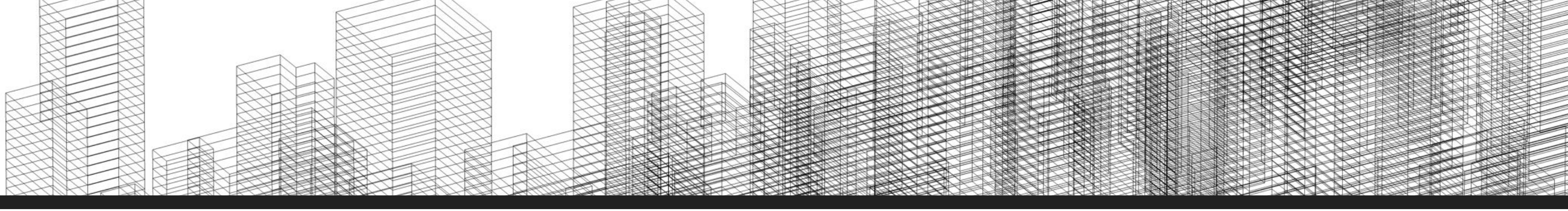


2019 World Happiness Rankings

1. Finland (7.78)
2. Switzerland (7.694)
3. Denmark (7.693)

....

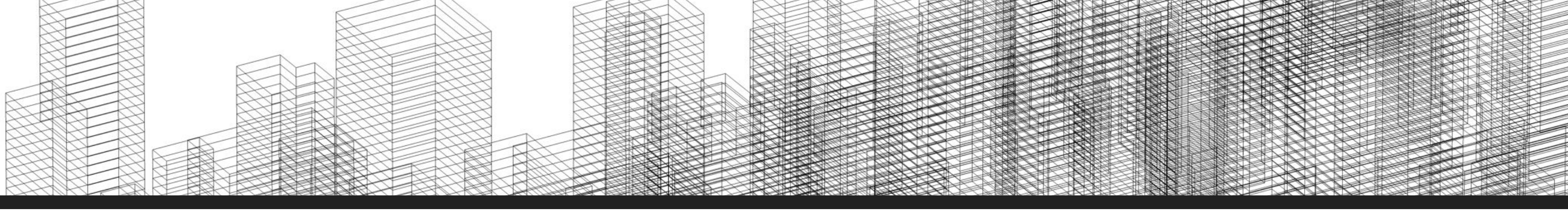
138. India (3.249)
139. Zimbabwe (2.694)
140. Afghanistan (2.375)



Purpose of Analysis

Each country’s happiness score is an average of each respondents ***subjective*** evaluation to the Life Ladder question in the Gallup World Survey.

Can country specific socio economic and political factors help explain the scores for each country? And can these results help draft policies that could lead to increased happiness for each country’s population?



Questions to be Considered

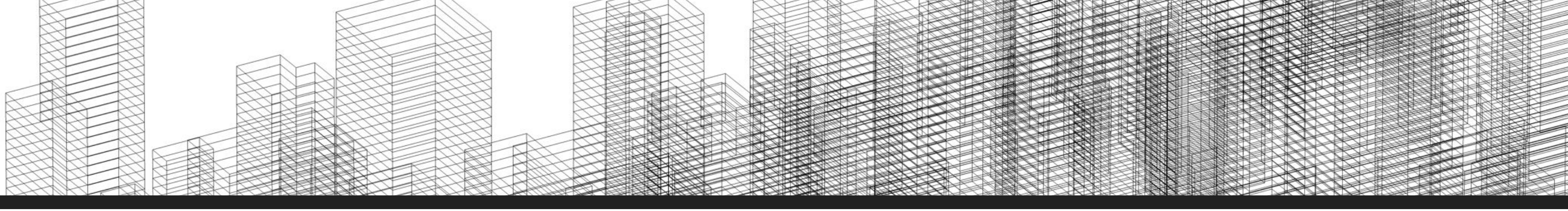
Which features within the dataset will have a significant effect on happiness?

WHR
World Happiness Report

Is happiness related to increase in a nation’s wealth?

GROSS DOMESTIC PRODUCT
GDP

Is there an opportunity to use public policy to improve happiness?



Methodology

Due to the onset of COVID-19 in 2020 and the inconsistency of available 2020 & 2021 datasets from the United Nations (UN), the data used in the following analysis comes from 2019 datasets.

Using modern data analysis techniques (our “Tool Box”), the team analyzed a multitude of available countrywide datasets.

Data Sources:

- World Happiness Report
- The Economist Intelligence Unit - Democracy Index
- United Nations Data

Tool Box

amazon S3

Algorithms

Linear Regression

Random Forest Regressor

Gradient Boosting

tableau

amazon RDS

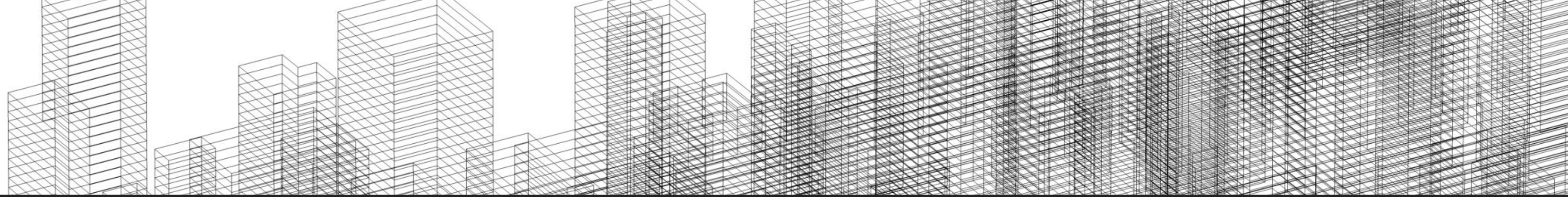
PostgreSQL

PySpark

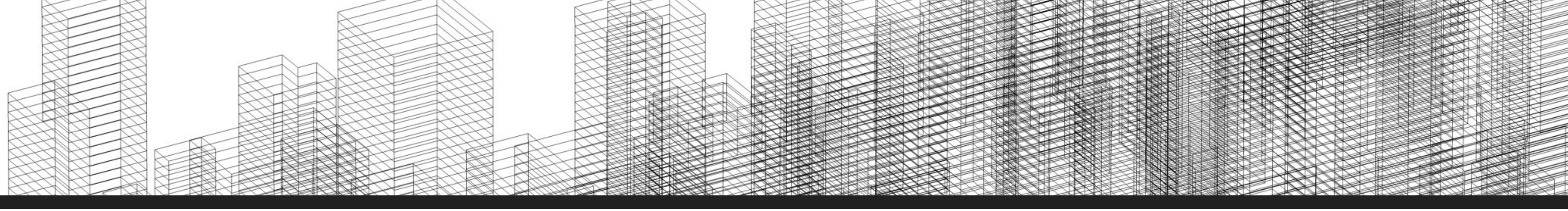
python™

jupyter

Data



Name	Description
Clean Water	Safely managed drinking water sources, total (Proportion of population with access).
Consumer Price Index	Indexed cost of a basket of similar goods by country.
Democracy Index	The Economist Intelligence Unit’s Democracy Index
GDP Per Capita	A country's GDP divided by population.
Gender Ratio	Sex ratio (males per 100 females)
Happiness ("Life Ladder") Country Scores	2019 Life Ladder scores for all available countries.
Infant Mortality	Infant mortality for both sexes (per 1,000 live births).
Life Expectancy	Life expectancy at birth for both sexes (years).
Parliament Seats held by Women	Seats held by women (as a % of total) in parliament.
Population Density	Population per square kilometer.
Unemployment Rate	% of total population that is unemployed.



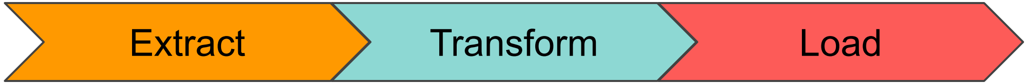
Data Pipeline

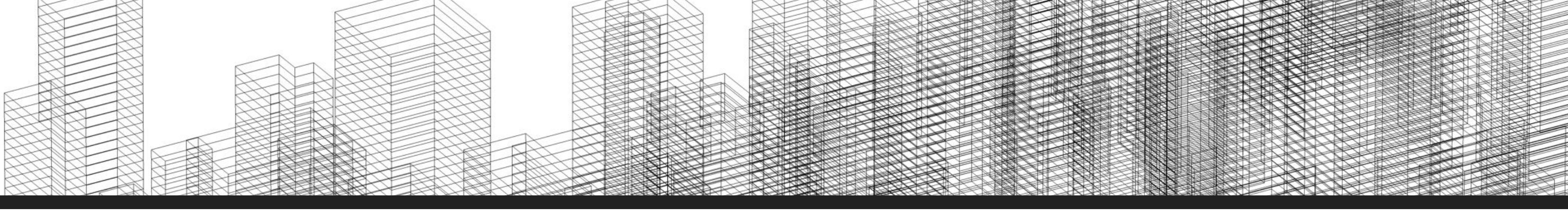
All original source data placed in an S3 bucket for all team members to access.

Using Google Collaboratory and the PySpark library, data went through an ETL process in accordance with project ERD.

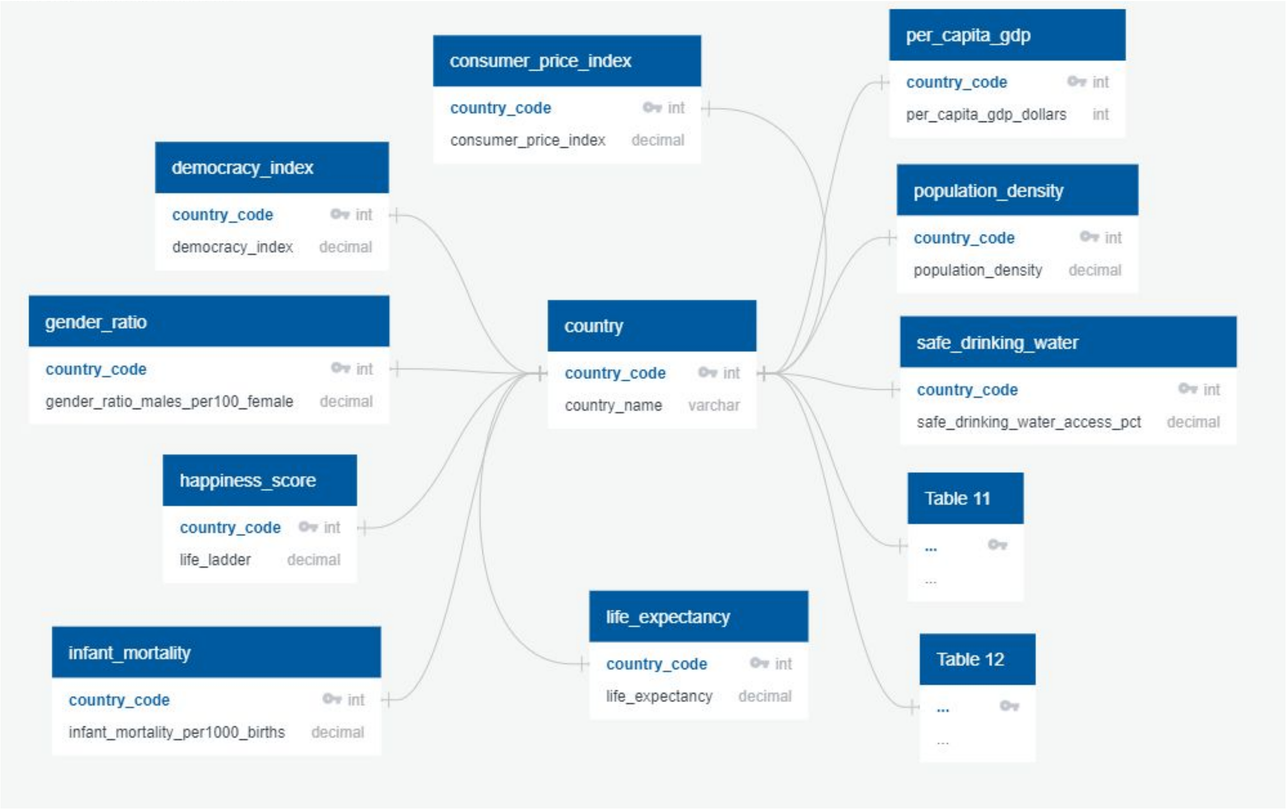
RDS PostgreSQL instance housed each cleaned dataset in their own respective tables.

PostgreSQL used to query data, to create final dataset, to be used in the analysis.





ERD



consumer_price_index		
country_code	🔑	int
consumer_price_index		decimal

per_capita_gdp		
country_code	🔑	int
per_capita_gdp_dollars		int

democracy_index		
country_code	🔑	int
democracy_index		decimal

population_density		
country_code	🔑	int
population_density		decimal

gender_ratio		
country_code	🔑	int
gender_ratio_males_per100_female		decimal

country		
country_code	🔑	int
country_name		varchar

safe_drinking_water		
country_code	🔑	int
safe_drinking_water_access_pct		decimal

happiness_score		
country_code	🔑	int
life_ladder		decimal

Table 11		
...	🔑	
...		

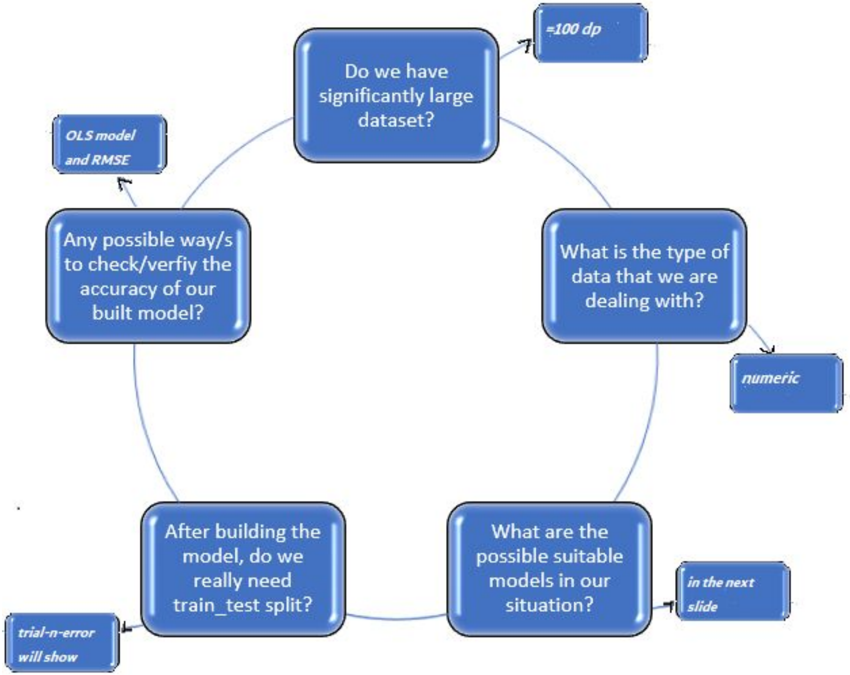
life_expectancy		
country_code	🔑	int
life_expectancy		decimal

infant_mortality		
country_code	🔑	int
infant_mortality_per1000_births		decimal

Table 12		
...	🔑	
...		

Machine Learning:

A never-ending cycle of non-directional questions



Do we have significantly large dataset?

=100 dp

What is the type of data that we are dealing with?

numeric

What are the possible suitable models in our situation?

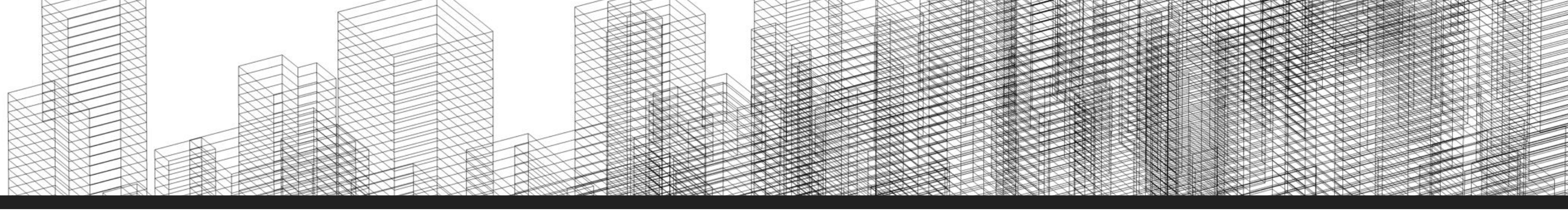
in the next slide

After building the model, do we really need train_test split?

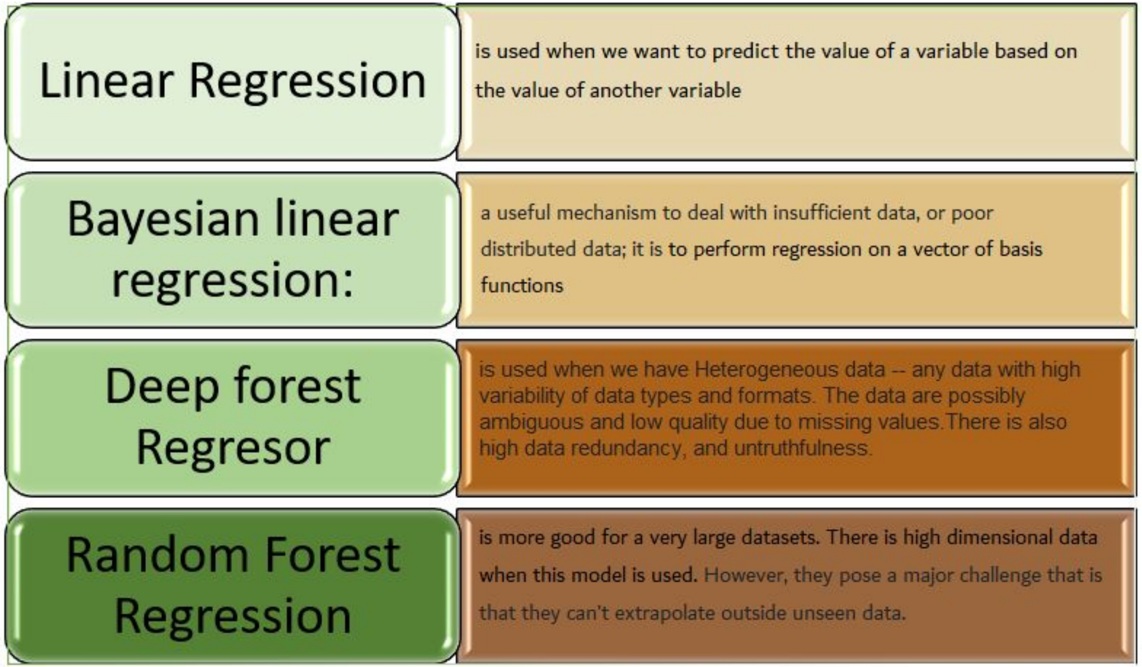
trial-n-error will show

Any possible way/s to check/verfiy the accuracy of our built model?

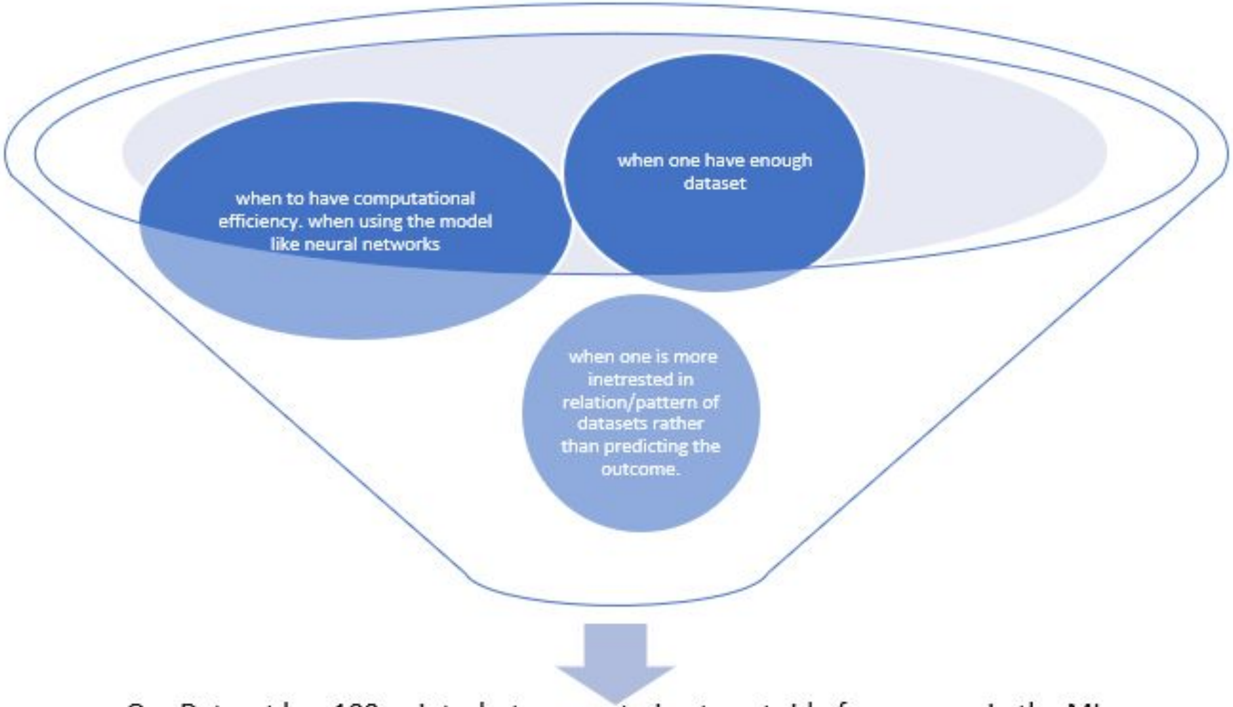
OLS model and RMSE



What are the possible suitable Models for our ML?



☑ Linear Regression	is used when we want to predict the value of a variable based on the value of another variable
Bayesian linear regression:	a useful mechanism to deal with insufficient data, or poor distributed data; it is to perform regression on a vector of basis functions
Deep forest Regresor	is used when we have Heterogeneous data -- any data with high variability of data types and formats. The data are possibly ambiguous and low quality due to missing values.There is also high data redundancy, and untruthfulness.
Random Forest Regression	is more good for a very large datasets. There is high dimensional data when this model is used. However, they pose a major challenge that is that they can't extrapolate outside unseen data.



when to have computational efficiency. when using the model like neural networks

when one have enough dataset

when one is more inetrested in relation/pattern of datasets rather than predicting the outcome.

Our Dataset has 100 points, but we are trying to get rid of any errors in the ML model by verifying its accuracy. Also, We are more interested in finding the linear pattern/relationship on the x-variables and the Happiness.

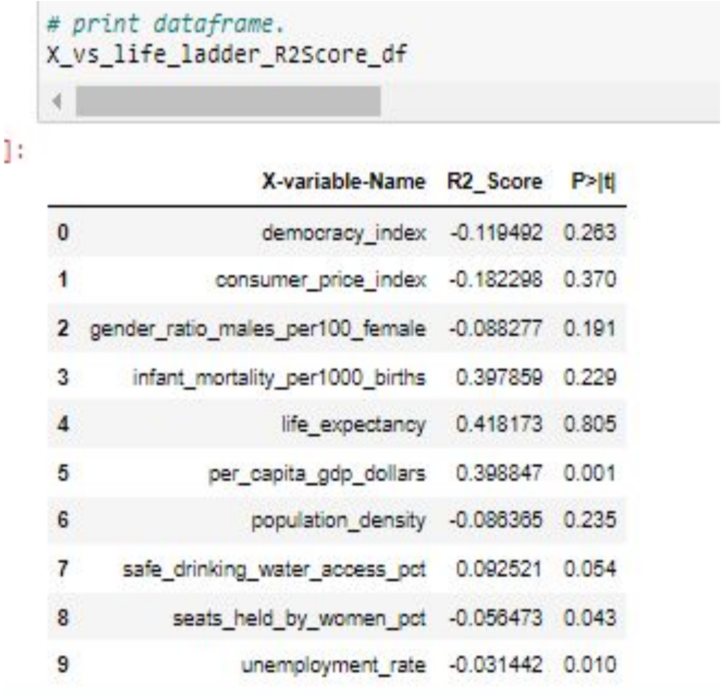
For Our Model, is it necessary to put “Train_Test_Splits”?

Decision was narrowed down into:

to use the train-test splits as on trial_N_error basis

Steps to the best Fit and accurate Linear-Regression Model’s results:

1. Find the P_values and the R2 scores of the each variable with the constant y-axis (life_ladder).

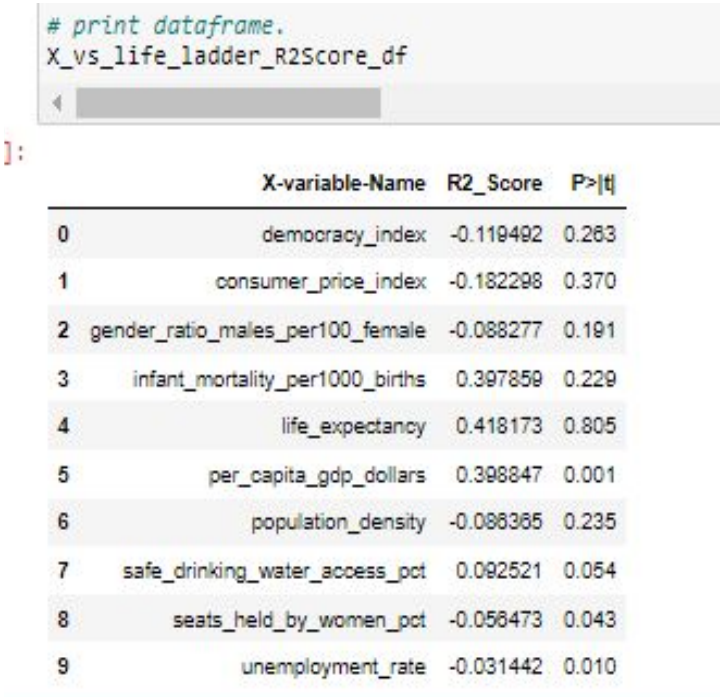


```
# print dataframe.
X_vs_life_ladder_R2Score_df
```

]:

	X-variable-Name	R2_Score	P>|t|
0	democracy_index	-0.119492	0.263
1	consumer_price_index	-0.182298	0.370
2	gender_ratio_males_per100_female	-0.088277	0.191
3	infant_mortality_per1000_births	0.397859	0.229
4	life_expectancy	0.418173	0.805
5	per_capita_gdp_dollars	0.398847	0.001
6	population_density	-0.088365	0.235
7	safe_drinking_water_access_pct	0.092521	0.054
8	seats_held_by_women_pct	-0.058473	0.043
9	unemployment_rate	-0.031442	0.010

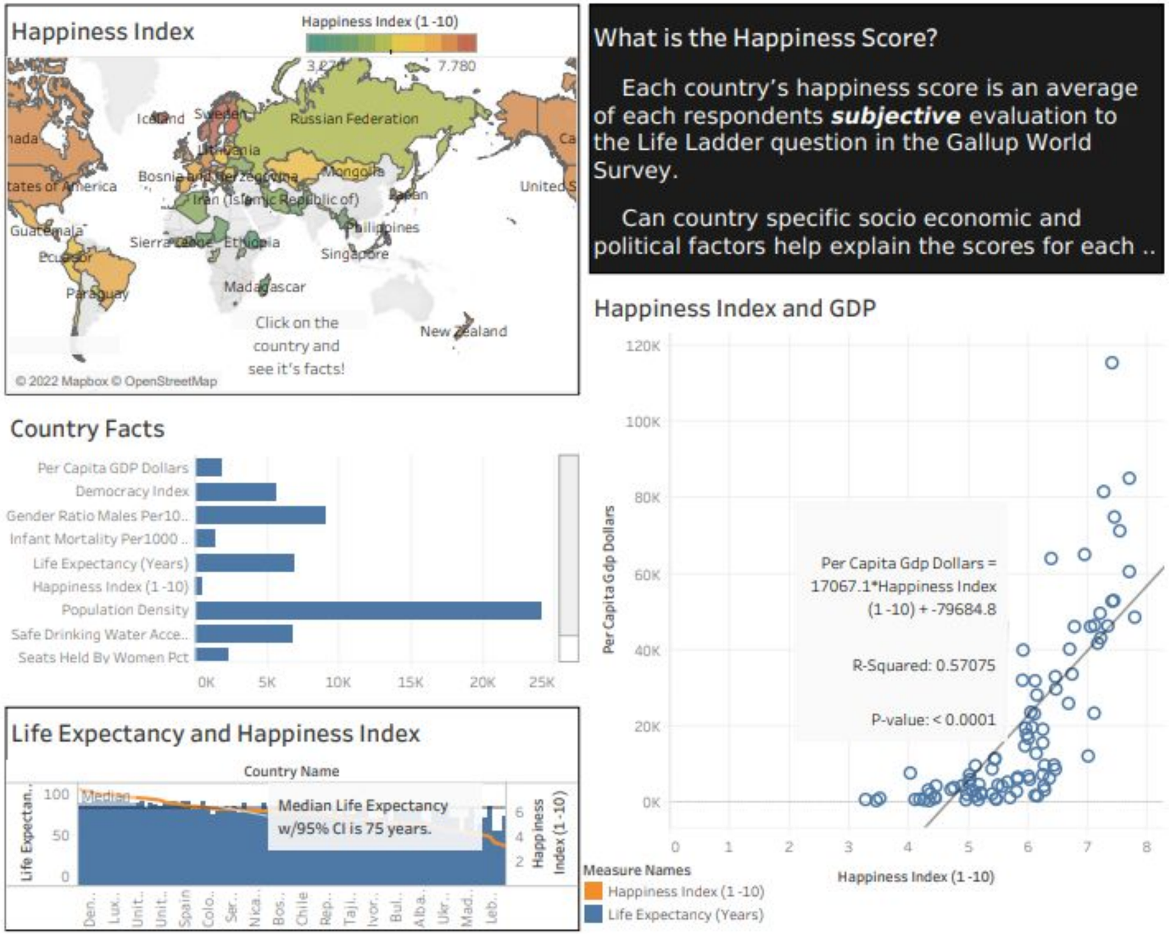
Steps to the best Fit and accurate Linear-Regression Model’s results:



```
# print dataframe.
X_vs_life_ladder_R2Score_df
```

	X-variable-Name	R2_Score	P>|t|
0	democracy_index	-0.119492	0.263
1	consumer_price_index	-0.182298	0.370
2	gender_ratio_males_per100_female	-0.088277	0.191
3	infant_mortality_per1000_births	0.397859	0.229
4	life_expectancy	0.418173	0.805
5	per_capita_gdp_dollars	0.398847	0.001
6	population_density	-0.086365	0.235
7	safe_drinking_water_access_pct	0.092521	0.054
8	seats_held_by_women_pct	-0.056473	0.043
9	unemployment_rate	-0.031442	0.010

1. Determined the P_values and the R2 scores of the each variable with the constant y-axis (life_ladder).
2. Eliminated the highest P-value scores’ x-dependent variables and rerun the multiple linear regression model one-by-one.



What is the Happiness Score?

Each country's happiness score is an average of each respondents ***subjective*** evaluation to the Life Ladder question in the Gallup World Survey.

Can country specific socio economic and political factors help explain the scores for each ..

Per Capita Gdp Dollars = 17067.1*Happiness Index (1-10) + -79684.8

R-Squared: 0.57075

P-value: < 0.0001

Median Life Expectancy w/95% CI is 75 years.

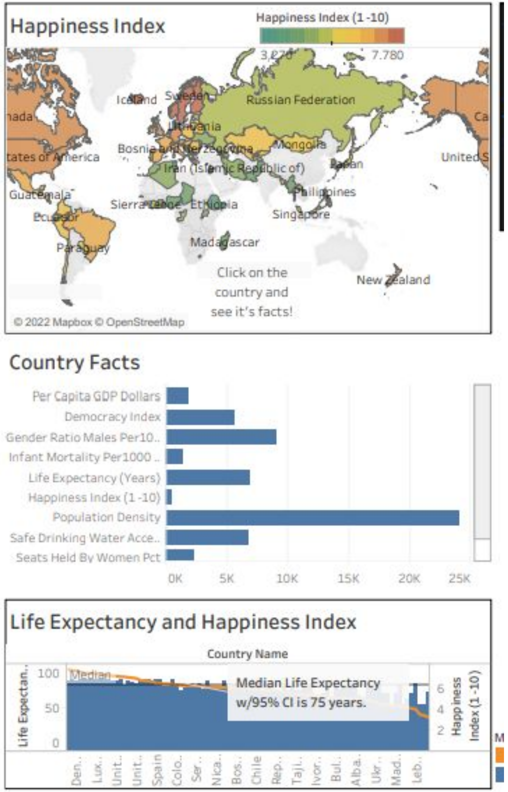
Click on the country and see it's facts!

Dashboard

Tableau Public was used to make a dashboard to allow the customer to:

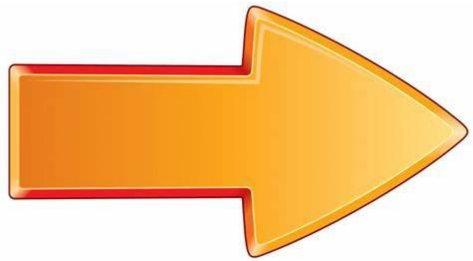
1. Interact with the 12 data tables
2. Understand the ease and capability of Tableau.

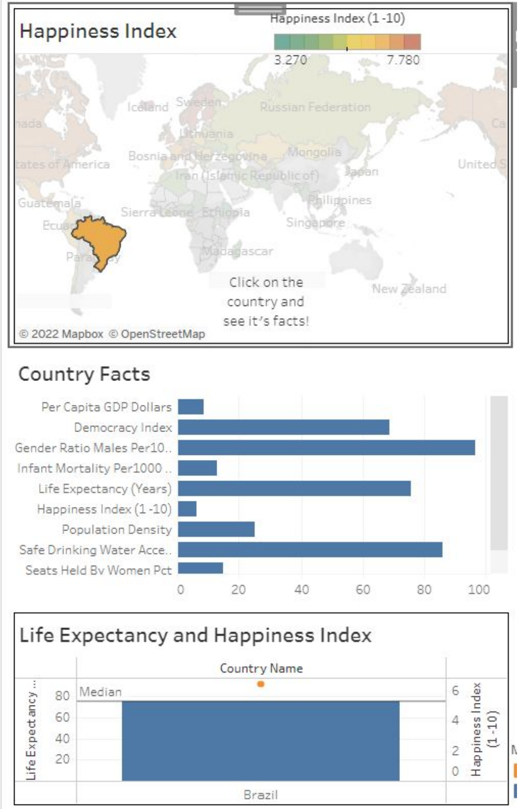
Dashboard Interactive Features



"Click" on a country:

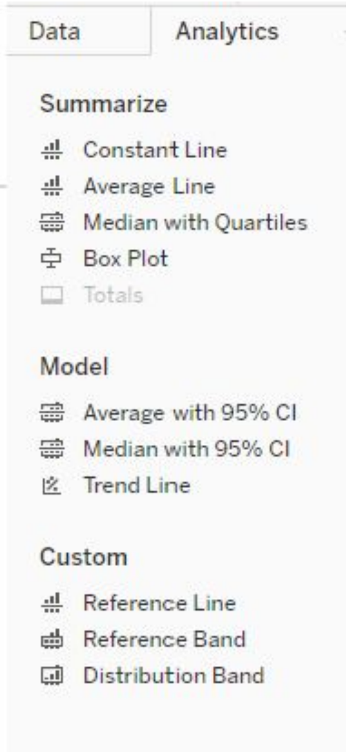
The country facts change and Life Expectancy and Happiness graph update to that country's information.





Other Dashboard Features:

- Hover over the country and it shows the index and country name.
- A URL is listed for the user to click on to read the World Happiness Report 2021.
- Utilizing the analytics in Tableau:
  - The Life expectancy and Happiness Index is a 2 axis graph that calculated the median life expectancy with a 95% confidence interval.
  - A static scatter graph was developed for the Happiness Index and GDP. A trend line was added and a R squared calculated with a p-value.


Data
Analytics
Summarize
Constant Line
Average Line
Median with Quartiles
Box Plot
Totals
Model
Average with 95% CI
Median with 95% CI
Trend Line
Custom
Reference Line
Reference Band
Distribution Band