# Analyzing Happiness on a Global Scale

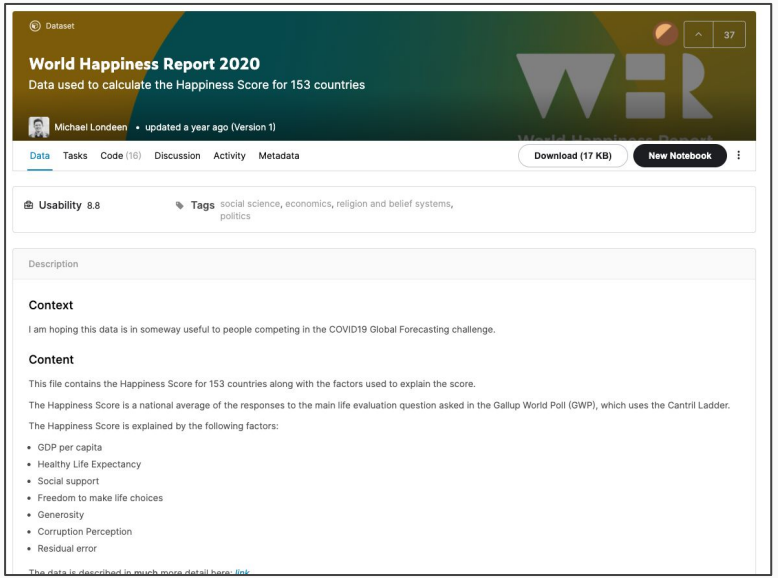
By Valerie Andrade

# Motivation & Summary

- Interested in anything social science-related
- Found an interesting dataset that was based on a survey study taken on samples of the population in 153 countries that calculates an average Happiness Score per country.
- Decided to analyze Happiness Scores per country with variables such as Fertility Rates, Median Age, Population Growth Rate and Urbanization Rate

# Dataset #1 - World Happiness Report

- Gallup is a global analytics and research firm that since 2005 has produced a “World Happiness Report” using data from a survey called the Gallup World Poll
- This dataset contains a “Happiness Score” for 153 countries and uses data collected from 2017-2019. It was released Feb. 2020
- The Happiness Score is a national average of the responses to the main life evaluation question asked in the Gallup World Poll (GWP), which uses the Cantril Ladder Scale. Sample question on next slide.

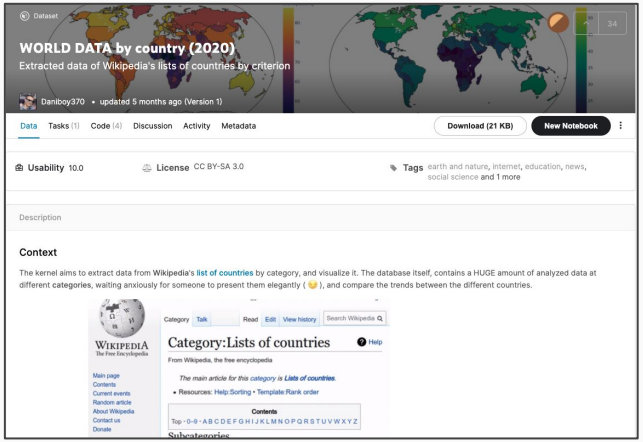


# Dataset #1 - World Happiness Report


- Their survey included questions for life evaluations including the Cantril Ladder. Sample question:
  - **“Are you satisfied or dissatisfied with your freedom to choose what you do with your life?** Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"
- The Happiness Score is calculated by the following factors:
  - GDP per capita
  - Healthy Life Expectancy
  - Social support
  - Freedom to make life choices
  - Generosity
  - Corruption Perception
  - Residual error

# Dataset #2 - World Data by Country

- This dataset includes extracted global data from Wikipedia by 9 different categories. For this project I used the following:
  - **Fertility Rate in 2018 for 202 countries (via World Bank)**
    - “The fertility rate is the expected number of children born per woman in her child-bearing years.”
  - **Median Age in 2018 for 224 countries (via CIA World Factbook)**
    - “Median age is the age that divides a population into two numerically equally sized groups - that is, half the people are younger than this age and half are older.”
  - **Urbanization population rate in 2019 for 213 countries (via World Bank)**
    - “Urban population describes the percentage of the total population living in urban areas, as defined by the country.”
  - **Population Growth Rate from 2015 - 2020 for 208 countries (via UN)**



# Questions

- Question 1:
  - Are any of the following 4 variables (Fertility Rate, Median Age, Urbanization Rate and Population Growth Rate) correlated to a country's Happiness Score (from per Gallup World Poll)?

- Question 1a:
  - For each variable, how strong, moderate or weak is the correlation?

# Data Cleanup & Exploration

- Merged 5 datasets into 1 main dataframe
- Removed unrelated columns from Happiness Score dataset
- Resolved duplicate country issue that created multiple rows
- Had to go through the CSVs to resolve country names that were inconsistent with another and reimported
- Renamed columns
- Confirmed no missing values
- Grouped countries by region indicators (groupby)
- Noticed about 79 countries from World Data dataset were dropped as Happiness Score dataset only measured 153 countries

# Data Cleanup & Exploration - Examples

## Merged 5 datasets into 1 main dataframe

```
In [2]: world_happiness_report = "Resources/rawdata/WHR20_DataForFigure2.1.csv"
        fertility_world_data = "Resources/rawdata/World Data/Fertility.csv"
        median_age_world_data = "Resources/rawdata/World Data/Median age.csv"
        population_growth_world_data = "Resources/rawdata/World Data/Population growth.csv"
        urbanization_rate_world_data = "Resources/rawdata/World Data/Urbanization rate.csv"

        whr_df = pd.read_csv(world_happiness_report)
        fertility_df = pd.read_csv(fertility_world_data)
        median_age_df = pd.read_csv(median_age_world_data)
        population_growth_df = pd.read_csv(population_growth_world_data)
        urbanization_rate_df = pd.read_csv(urbanization_rate_world_data)

        fertility_merge = pd.merge(whr_df, fertility_df, on="Country")
        fertility_merge
```

Out[2]:

| | Country | Regional indicator | Ladder score | Standard error of ladder score | upperwhisker | lowerwhisker | Logged GDP per capita | Social support | Healthy life expectancy | Freedom to make life choices | ... | Ladder score in Dystopia | Explained by: Log GDP per capita | Explained by: Social support |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | South Asia | 2.5669 | 0.031311 | 2.628270 | 2.505530 | 7.462861 | 0.470367 | 52.590000 | 0.396573 | ... | 1.972317 | 0.300706 | 0.396434 |
| 1 | Albania | Central and Eastern Europe | 4.8827 | 0.056116 | 4.992687 | 4.772713 | 9.417931 | 0.671070 | 68.708138 | 0.781994 | ... | 1.972317 | 0.906553 | 0.830484 |
| 2 | Algeria | Middle East and North Africa | 5.0051 | 0.044236 | 5.091802 | 4.918397 | 9.537965 | 0.803385 | 65.905174 | 0.466611 | ... | 1.972317 | 0.943856 | 1.143004 |
| 3 | Argentina | Latin America and Caribbean | 5.9747 | 0.053442 | 6.079446 | 5.869954 | 9.810955 | 0.900568 | 68.803802 | 0.831132 | ... | 1.972317 | 1.028466 | 1.372544 |
| 4 | Armenia | Commonwealth of Independent States | 4.6768 | 0.058596 | 4.791646 | 4.561953 | 9.100476 | 0.757479 | 66.750656 | 0.712018 | ... | 1.972317 | 0.808262 | 1.034577 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 143 | Venezuela | Latin America and Caribbean | 5.0532 | 0.064281 | 5.179190 | 4.927210 | 8.977794 | 0.890408 | 66.505341 | 0.623278 | ... | 1.972317 | 0.770239 | 1.348547 |
| 144 | Vietnam | Southeast Asia | 5.3535 | 0.033801 | 5.419749 | 5.287251 | 8.809546 | 0.849987 | 67.952736 | 0.939593 | ... | 1.972317 | 0.718092 | 1.253075 |

## Made new DF containing relevant columns

```
In [21]: main_df = df2[['Country', 'Regional indicator', 'Ladder score',
                       'Fertility', 'Median age', 'Population growth',
                       'Urbanization rate', 'ISO-code_y']].reset_index()
         main_df = main_df.drop(columns={"index"})
         main_df
```

Out[21]:

| | Country | Regional indicator | Ladder score | Fertility | Median age | Population growth | Urbanization rate | ISO-code_y |
|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | South Asia | 2.5669 | 4.5 | 27.4 | 2.41 | 25.754 | AFG |
| 1 | Albania | Central and Eastern Europe | 4.8827 | 1.6 | 32.9 | 0.26 | 61.229 | ALB |
| 2 | Algeria | Middle East and North Africa | 5.0051 | 3.0 | 28.1 | 1.89 | 73.189 | DZA |
| 3 | Argentina | Latin America and Caribbean | 5.9747 | 2.3 | 31.7 | 0.88 | 91.991 | ARG |
| 4 | Armenia | Commonwealth of Independent States | 4.6768 | 1.8 | 35.1 | 0.17 | 63.219 | ARM |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 139 | Venezuela | Latin America and Caribbean | 5.0532 | 2.3 | 28.3 | 1.53 | 88.240 | VEN |
| 140 | Vietnam | Southeast Asia | 5.3535 | 2.0 | 30.5 | 1.06 | 36.628 | VNM |
| 141 | Yemen | Middle East and North Africa | 3.5274 | 3.8 | 19.5 | 2.33 | 37.273 | YEM |
| 142 | Zambia | Sub-Saharan Africa | 3.7594 | 4.6 | 16.8 | 3.19 | 44.072 | ZMB |
| 143 | Zimbabwe | Sub-Saharan Africa | 3.2992 | 3.6 | 20.0 | 2.70 | 32.210 | ZWE |

144 rows × 8 columns

## Renamed column names and reordered them

```
In [22]: main_df = main_df.rename(columns = {
             'Regional indicator': 'Region',
             'Ladder score': 'Happiness Score (0 - 10)',
             'Fertility': 'Fertility Rate in 2018 (births/woman)',
             'Median age': 'Median Age in 2018 ',
             'Population growth': 'Population Growth: 2015-2020 (%)',
             'Urbanization rate': 'Urbanization Rate in 2019 (%)',
             'ISO-code_y': 'Country Code'
         })

         main_df
```

Out[22]:

| | Country | Region | Happiness Score (0 - 10) | Fertility Rate in 2018 (births/woman) | Median Age in 2018 | Population Growth: 2015-2020 (%) | Urbanization Rate in 2019 (%) | Country Code |
|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | South Asia | 2.5669 | 4.5 | 27.4 | 2.41 | 25.754 | AFG |
| 1 | Albania | Central and Eastern Europe | 4.8827 | 1.6 | 32.9 | 0.26 | 61.229 | ALB |
| 2 | Algeria | Middle East and North Africa | 5.0051 | 3.0 | 28.1 | 1.89 | 73.189 | DZA |
| 3 | Argentina | Latin America and Caribbean | 5.9747 | 2.3 | 31.7 | 0.88 | 91.991 | ARG |
| 4 | Armenia | Commonwealth of Independent States | 4.6768 | 1.8 | 35.1 | 0.17 | 63.219 | ARM |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 139 | Venezuela | Latin America and Caribbean | 5.0532 | 2.3 | 28.3 | 1.53 | 88.240 | VEN |
| 140 | Vietnam | Southeast Asia | 5.3535 | 2.0 | 30.5 | 1.06 | 36.628 | VNM |

# Data Cleanup & Exploration - Examples

## 8 duplications of Guinea found

```
In [11]: data_merge['Country'].value_counts()

Out[11]: Guinea          8
         Russia          2
         Uzbekistan      1
         Austria         1
         South Korea     1
                        ..
         Mali            1
         France          1
         Ghana           1
         Argentina       1
         Luxembourg      1
         Name: Country, Length: 145, dtype: int64
```

```
In [12]: data_merge.loc[data_merge['Country'] == 'Guinea']
```

Out[12]:

| | Country | Regional indicator | Ladder score | Standard error of ladder score | upperwhisker | lowerwhisker | Logged GDP per capita | Social support | Healthy life expectancy | Freedom to make life choices | ... | Explained by: Healthy life expectancy | Explained by: Freedom to make life choices | Explained by: Generosity | Exp Perce corr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | Guinea | Sub-Saharan Africa | 4.9493 | 0.073042 | 5.092463 | 4.806137 | 7.75099 | 0.637573 | 54.4678 | 0.706847 | ... | 0.333655 | 0.371878 | 0.249491 | 0. |
| 50 | Guinea | Sub-Saharan Africa | 4.9493 | 0.073042 | 5.092463 | 4.806137 | 7.75099 | 0.637573 | 54.4678 | 0.706847 | ... | 0.333655 | 0.371878 | 0.249491 | 0. |
| 51 | Guinea | Sub-Saharan Africa | 4.9493 | 0.073042 | 5.092463 | 4.806137 | 7.75099 | 0.637573 | 54.4678 | 0.706847 | ... | 0.333655 | 0.371878 | 0.249491 | 0. |
| 52 | Guinea | Sub-Saharan Africa | 4.9493 | 0.073042 | 5.092463 | 4.806137 | 7.75099 | 0.637573 | 54.4678 | 0.706847 | ... | 0.333655 | 0.371878 | 0.249491 | 0. |
| 53 | Guinea | Sub-Saharan | 4.9493 | 0.073042 | 5.092463 | 4.806137 | 7.75099 | 0.637573 | 54.4678 | 0.706847 | ... | 0.333655 | 0.371878 | 0.249491 | 0. |

## Resolved by making new DF & dropping Guinea

```
In [13]: df1 = data_merge.loc[data_merge['Country'] != 'Guinea']
         df1
```

Out[13]:

| | Country | Regional indicator | Ladder score | Standard error of ladder score | upperwhisker | lowerwhisker | Logged GDP per capita | Social support | Healthy life expectancy | Freedom to make life choices | ... | Explained by: Healthy life expectancy | Explained by: Freedom to make life choices | Explai Genero |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | South Asia | 2.5669 | 0.031311 | 2.628270 | 2.505530 | 7.462861 | 0.470367 | 52.590000 | 0.396573 | ... | 0.266052 | 0.000000 | 0.135 |
| 1 | Albania | Central and Eastern Europe | 4.8827 | 0.056116 | 4.992687 | 4.772713 | 9.417931 | 0.671070 | 68.708138 | 0.781994 | ... | 0.846330 | 0.461946 | 0.171 |
| 2 | Algeria | Middle East and North Africa | 5.0051 | 0.044236 | 5.091802 | 4.918397 | 9.537965 | 0.803385 | 65.905174 | 0.466611 | ... | 0.745419 | 0.083944 | 0.118 |
| 3 | Argentina | Latin America and Caribbean | 5.9747 | 0.053442 | 6.079446 | 5.869954 | 9.810955 | 0.900568 | 68.803802 | 0.831132 | ... | 0.849774 | 0.520840 | 0.070 |
| 4 | Armenia | Commonwealth of Independent States | 4.6768 | 0.058595 | 4.791646 | 4.561953 | 9.100476 | 0.757479 | 66.750656 | 0.712018 | ... | 0.775857 | 0.378076 | 0.107 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 148 | Venezuela | Latin America and Caribbean | 5.0532 | 0.064281 | 5.179190 | 4.927210 | 8.977794 | 0.890408 | 66.505341 | 0.623278 | ... | 0.767026 | 0.271717 | 0.087 |
| 149 | Vietnam | Southeast Asia | 5.3535 | 0.033801 | 5.419749 | 5.287251 | 8.809546 | 0.849987 | 67.952736 | 0.939593 | ... | 0.819134 | 0.650836 | 0.136 |
| 150 | Yemen | Middle East and North Africa | 3.5274 | 0.054158 | 3.633550 | 3.421250 | 7.759683 | 0.817981 | 56.727283 | 0.599920 | ... | 0.415000 | 0.243721 | 0.094 |
| 151 | Zambia | Sub-Saharan Africa | 3.7594 | 0.060677 | 3.878326 | 3.640474 | 8.224720 | 0.698824 | 55.299377 | 0.806500 | ... | 0.363593 | 0.491318 | 0.250 |
| 152 | Zimbabwe | Sub-Saharan Africa | 3.2992 | 0.058674 | 3.414202 | 3.184198 | 7.865712 | 0.763093 | 55.617260 | 0.711458 | ... | 0.375038 | 0.377405 | 0.151 |

145 rows × 25 columns

# Data Cleanup & Exploration - Examples

**Found 2 rows of Russia with different values for Fertility Rates**

```
In [16]: df1.loc[data_merge['Country'] == 'Russia']
```

Out[16]:

| | Country | Regional indicator | Ladder score | Standard error of ladder score | upperwhisker | lowerwhisker | Logged GDP per capita | Social support | Healthy life expectancy | Freedom to make life choices | ... | Explained by: Healthy life expectancy | Explained by: Freedom to make life choices | Explaine b Generosi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 116 | Russia | Commonwealth of Independent States | 5.546 | 0.03961 | 5.623635 | 5.468365 | 10.128872 | 0.903151 | 64.100456 | 0.729893 | ... | 0.680446 | 0.3995 | 0.0990 |
| 117 | Russia | Commonwealth of Independent States | 5.546 | 0.03961 | 5.623635 | 5.468365 | 10.128872 | 0.903151 | 64.100456 | 0.729893 | ... | 0.680446 | 0.3995 | 0.0990 |

2 rows × 25 columns

**Resolved by dropping the lower value. Used the code below:**

- duplicate_row = df1.loc[(df1['Country'] == "Russia") & (df1['Fertility'] == 1.60)].index
- df2 = df1.drop(duplicate_row)
- df2

# Data Cleanup & Exploration - Examples

## Made new DF to groupby Region & Country

```
In [27]: region_groupby = main_df.groupby('Region', as_index = True)['Country']
         region_counts = pd.DataFrame(region_groupby.value_counts())
         region_counts
```

Out[27]:

| Region | Country | Country |
|---|---|---|
| Central and Eastern Europe | Albania | 1 |
| | Bosnia and Herzegovina | 1 |
| | Bulgaria | 1 |
| | Croatia | 1 |
| | Czech Republic | 1 |
| | Estonia | 1 |
| | Hungary | 1 |
| | Latvia | 1 |
| | Lithuania | 1 |
| | Montenegro | 1 |

## Made another new DF to groupby Region & get the .mean() for each column

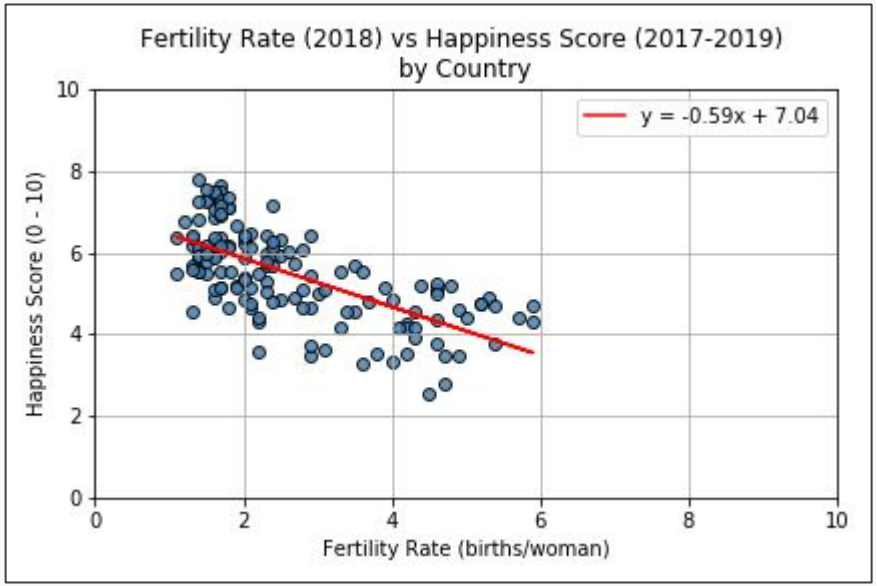
```
In [28]: region_group = main_df.groupby("Region")

         region_df = pd.DataFrame(region_group.mean())
         region_df2 = region_df.sort_values(by='Happiness Score (0 - 10)', ascending = False)
         region_df2
```

Out[28]:

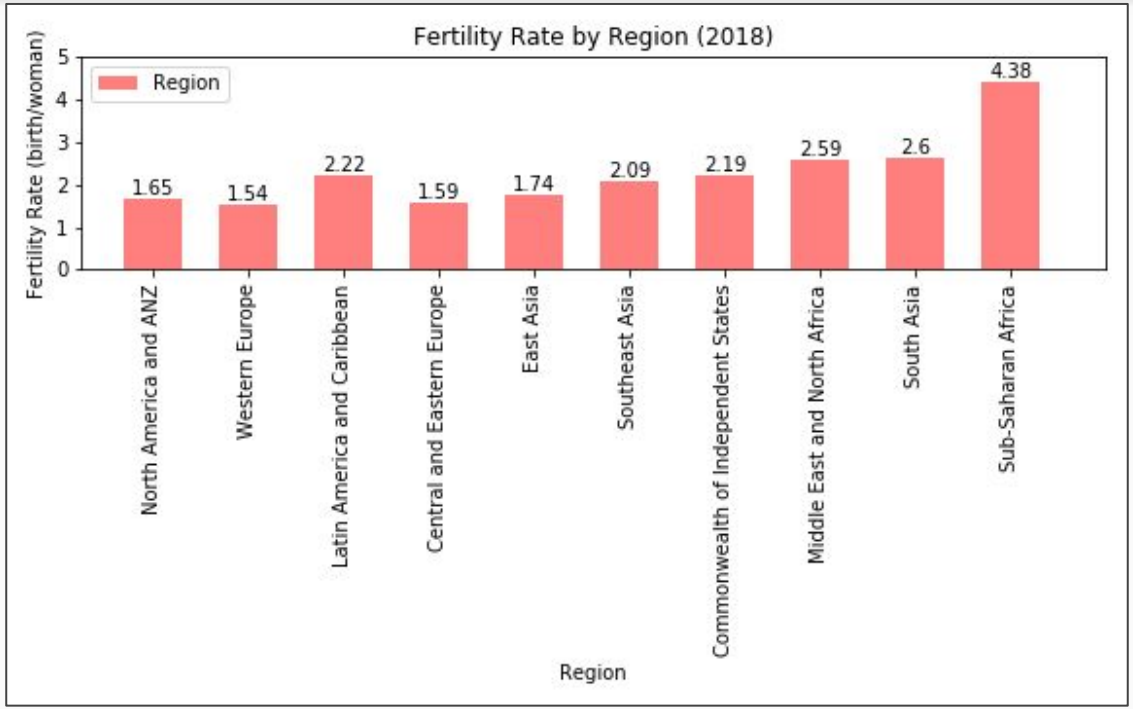
| Region | Happiness Score (0 - 10) | Fertility Rate in 2018 (births/woman) | Median Age in 2018 | Urbanization Rate in 2019 (%) | Population Growth: 2015-2020 (%) |
|---|---|---|---|---|---|
| North America and ANZ | 7.173525 | 1.650000 | 39.225000 | 84.170000 | 1.027500 |
| Western Europe | 6.967405 | 1.535000 | 41.530000 | 80.710050 | 0.580500 |
| Latin America and Caribbean | 5.981786 | 2.219048 | 28.571429 | 72.448143 | 1.249524 |
| Central and Eastern Europe | 5.875664 | 1.592857 | 41.764286 | 63.249286 | -0.315714 |
| East Asia | 5.566740 | 1.740000 | 39.840000 | 80.395800 | 0.686000 |
| Southeast Asia | 5.517788 | 2.087500 | 29.162500 | 53.313875 | 1.512500 |
| Commonwealth of Independent States | 5.358342 | 2.193333 | 33.375000 | 55.670000 | 0.835000 |
| Middle East and North | | | | | |

# Data Analysis - Fertility Rate vs Happiness Score
## *Scatter Plot & Linear Regression*

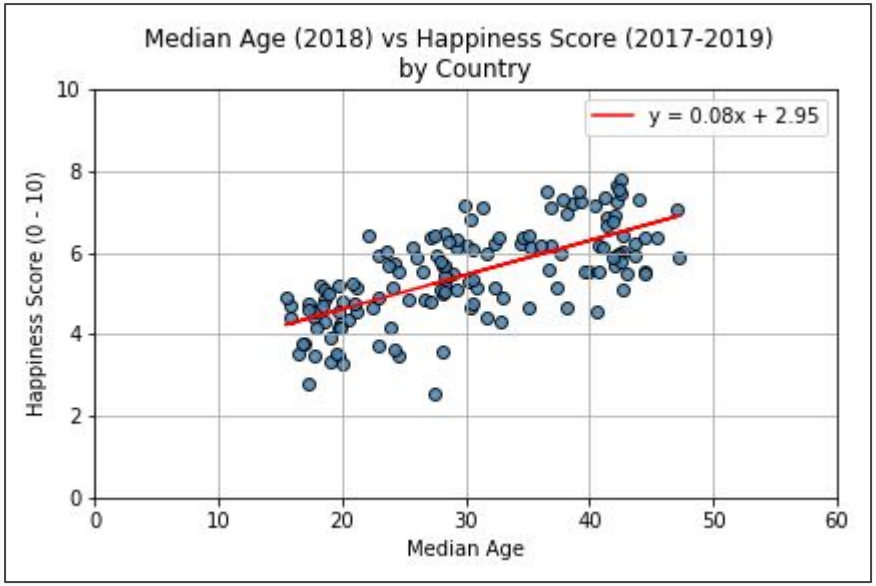



- The correlation coefficient for these two variables is R=-0.6618.
- The scatter plot indicates a moderate negative linear association between a country's Happiness Score and its average Fertility Rates.
- There appears to be a moderate relationship between the two variables.
- It's interesting to see the data show us that for many countries the Happiness Score decreases as the Fertility Rate increases.

# Data Analysis - Average Fertility Rate by Region
*Bar Chart*

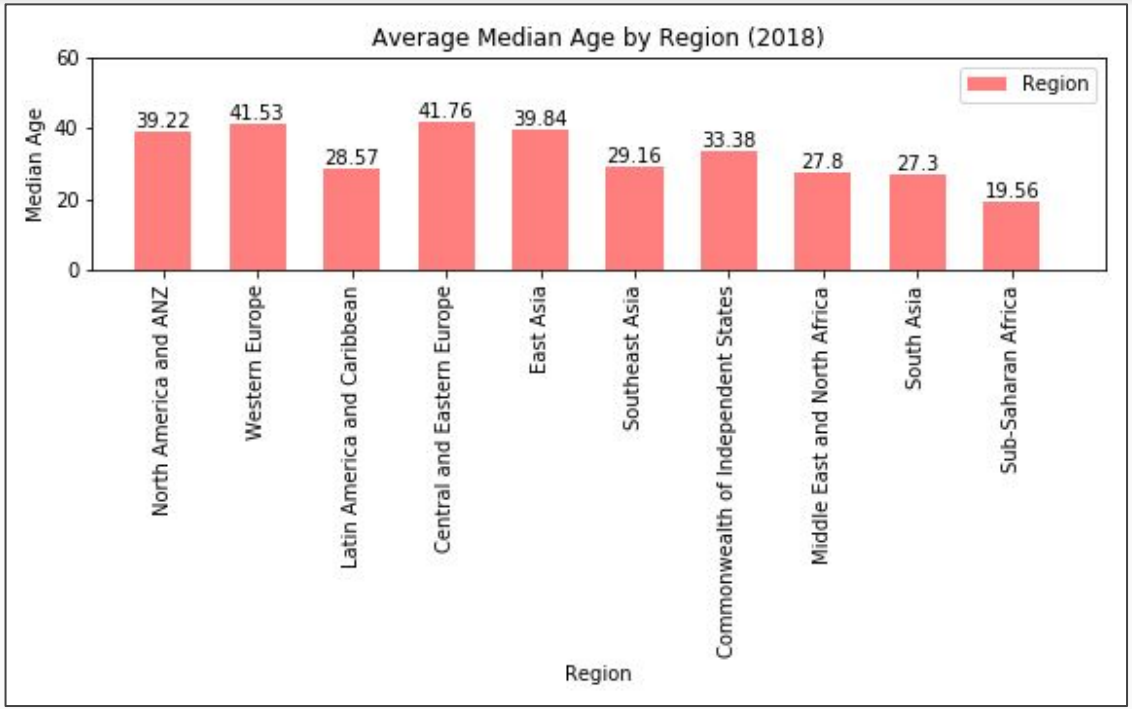


# Data Analysis - Median Age vs Happiness Score
## *Scatter Plot & Linear Regression*

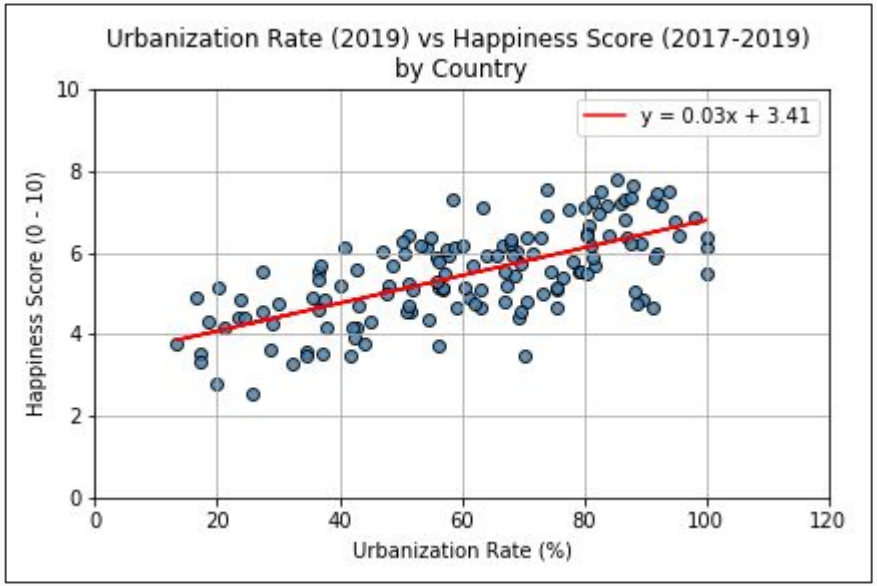



- The correlation coefficient for these two variables is R=0.6753.
- The scatter plot indicates a moderate positive linear association between a country's Happiness Score and its Median Age.
- There appears to be a moderate relationship between the two variables.
- It's interesting to see the data show us that for many countries, the Happiness Score increases as the Median Age increases.

# Data Analysis - Average Median Age by Region
*Bar Chart*

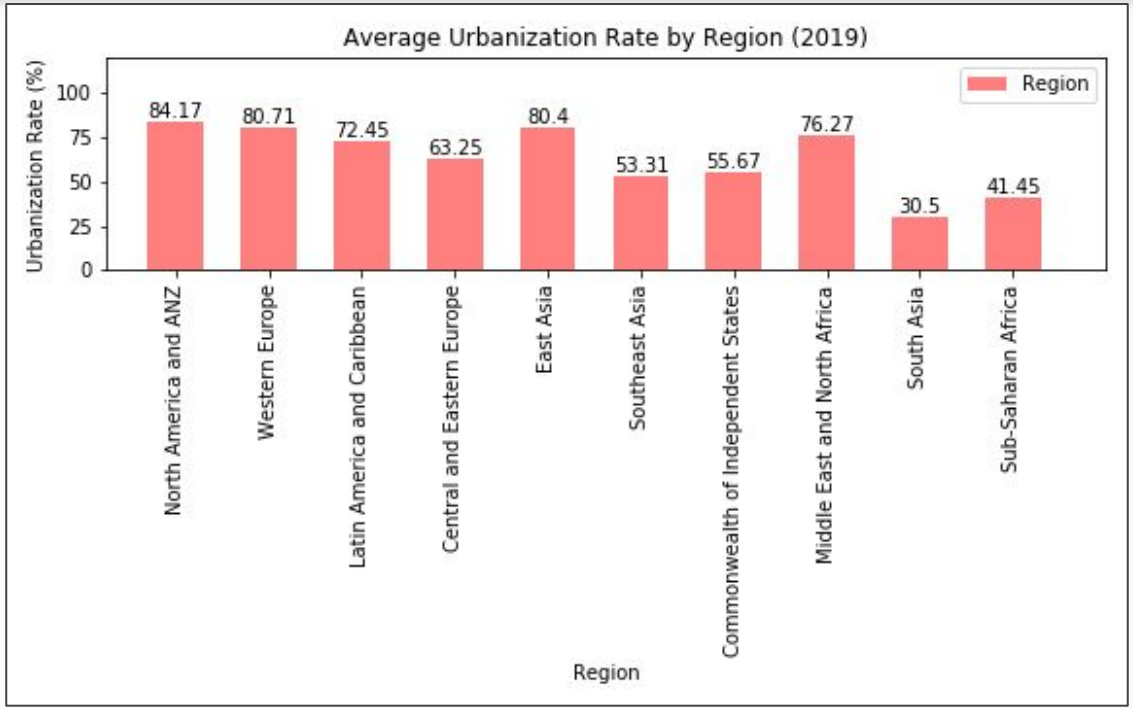


# Data Analysis - Urbanization vs Happiness Score
## *Scatter Plot & Linear Regression*

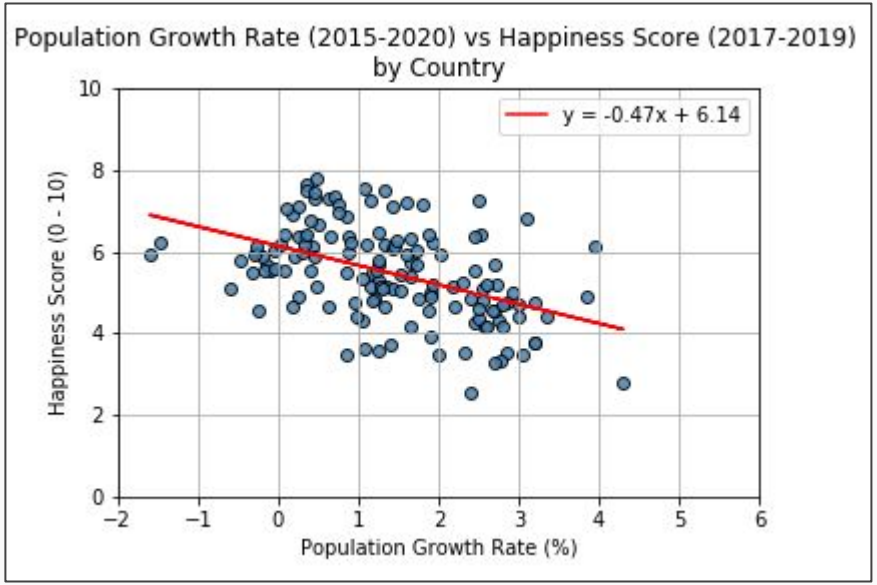



- The correlation coefficient for these two variables is R=0.6676.
- The scatter plot indicates a moderate positive linear association between a country's Happiness Score and Urbanization Population in 2019.
- There appears to be a moderate relationship between the two variables.
- It's interesting to see the data show us that for many countries the Happiness Score is higher when more of the population is urbanized.

# Data Analysis - Avg. Urbanization Rate by Region
*Bar Chart*

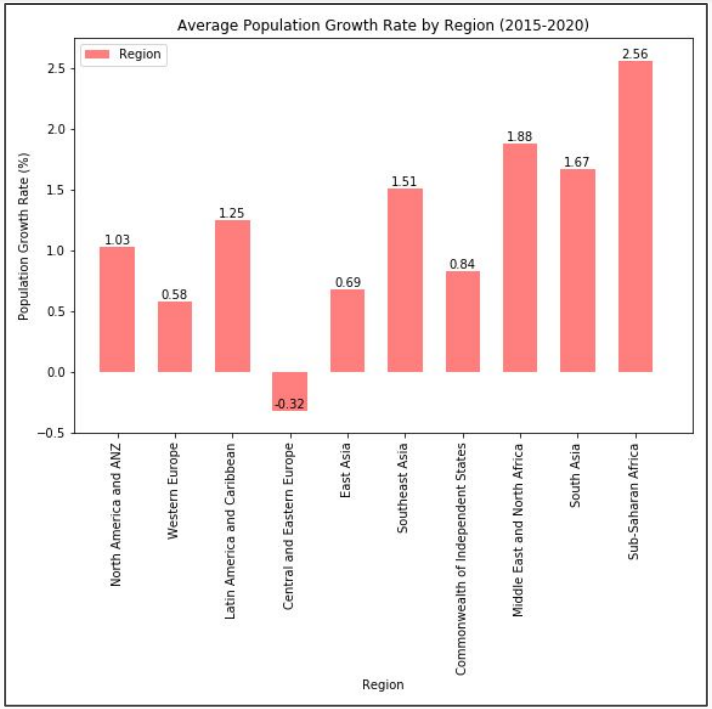


# Data Analysis - Pop. Growth vs Happiness Score
## *Scatter Plot & Linear Regression*




- The correlation coefficient for these two variables is R=-0.4712.
- The scatter plot indicates a semi-moderate negative linear association between a country's Happiness Score and Population Growth Rate (estimated 2015-2020).
- There appears to be a semi-moderate relationship between the two variables.
- It's interesting to see the data show us that for some countries the Happiness Score decreases as the Population Growth Rate increases.

# Data Analysis - Avg. Pop. Growth Rate by Region
*Bar Chart*

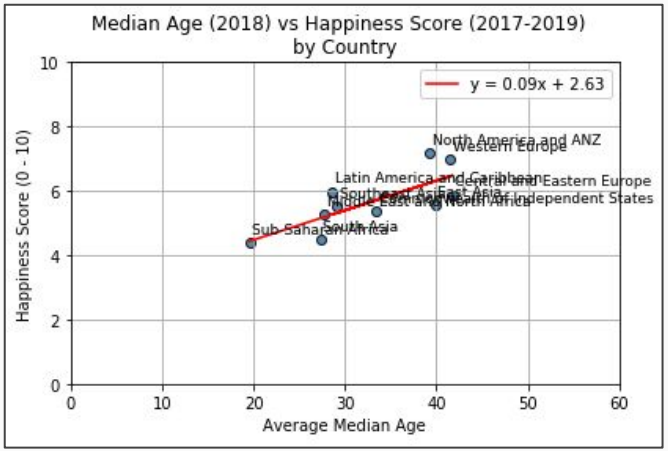


# Findings & Conclusion

- I expected to find stronger correlations than what I did.
- I believe there is a moderate correlation between the World Happiness Score and 3 out of the 4 variables, which were: Fertility Rate, Median Age, and Urbanization Population Rate.
- The correlation between World Happiness Score and Population Growth Rate was the weakest out of the 4 variables, though still semi-moderate.

# Final Thoughts

- Difficulties that arose mainly came during the plotting of the bar charts (as the region names were very long and overlapped each other) and finding the best way to display the data in a way that makes sense to others.
- I also tried very hard to make scatter plots by region with the names of the region next to each plot, but the labels overlapped too much and wasn't readable (as shown to the right).
- Overall, this was an interesting topic to analyze. It made me curious about the endless possibility of variables I could compare against.




**Example of difficulties that arose when plotting**

# THANK YOU