ETL Assignment: Data Science Bootcamp

Date: 2/16/2019

**Overview**

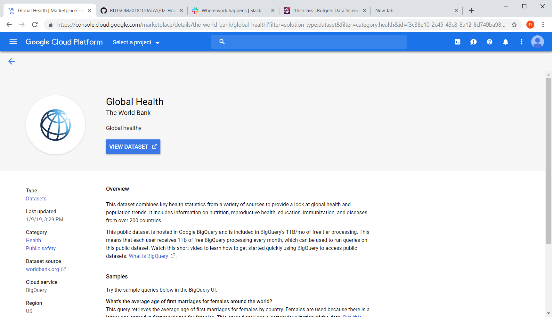
Obesity is becoming a Global trend. It was requested to identify if the countries with the Happiest people in the World also have a higher concentration of obese Adults.

1. Is there a correlation across countries between the percentage of obese Adults and the overall Happiness score.
2. How does the country’s percentage of obese Adults compare with the Healthy Life Expectancy score. Do we feel obesity is accounted for in this score.

<https://www.hsph.harvard.edu/obesity-prevention-source/obesity-trends/>

**EXTRACT : Data Sources**

1. *World Bank Public Data from Big Query (GCP)*



**Data Extraction: World Bank Indicators**

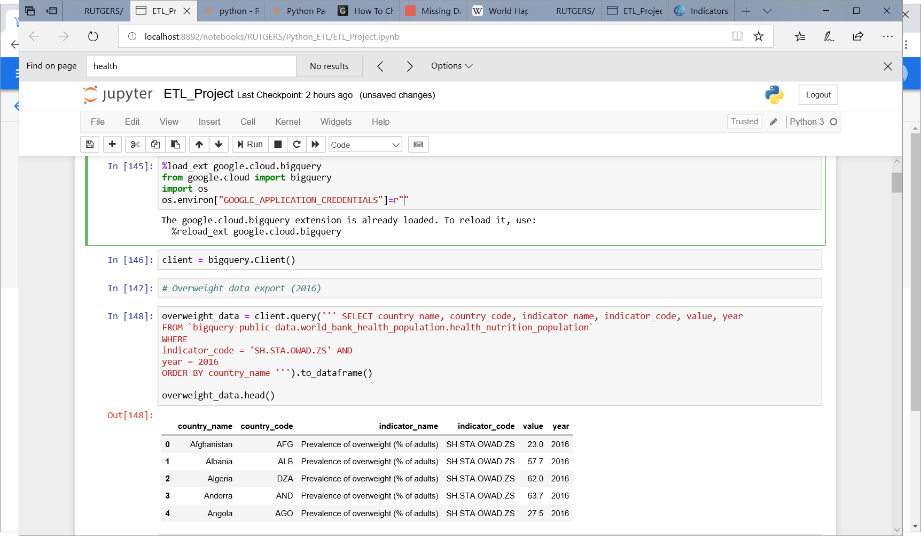
* Prevalence of overweight / obesity (% of adults) : SH.STA.OWAD.ZS (year=2016)
* Domestic general government health expenditure per capita (current US$): SH.XPD.GHED.PP.CD (year=2015)
* <https://data.worldbank.org/indicator> (Indicators)

**File Type / Extraction Method**

* Database Table / SQL Extraction from Big Query to Jupyter Notebook

**Notations**

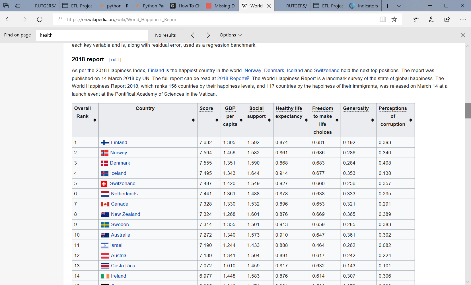
* Overweight adult data was available up to 2016. 2016 was extracted for most current information
* Health expenditure per capita data was available up to 2015. 2015 was extracted for most current information
* Not all countries have complete data
* Not all countries were represented in the data to merge with the Happiest Country list



Code Extraction

World Bank

1. *Wikipedia: Happiest Countries in the World*

**

**Data Extracted: Happiness Score, and associated Sub-Scores**

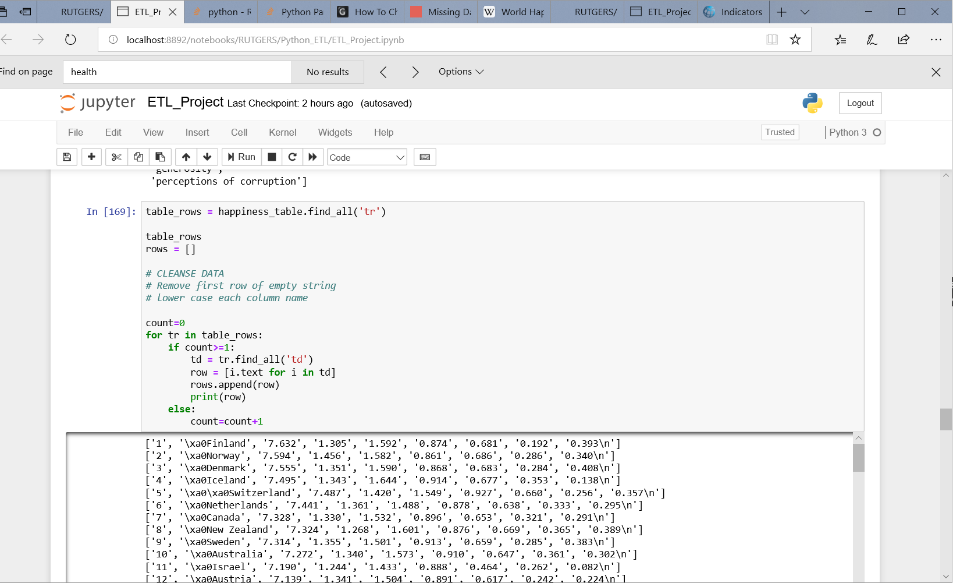
https://en.wikipedia.org/wiki/World\_Happiness\_Report

**File Type / Extraction Method**

* HTML / Beautiful Soup & Pandas

**Notations**

* Happiest Countries in the World Ranking are from 2018

****Code Extraction

Happiest Countries

**TRANSFORM : Process**

World Bank Transform Process

* Identified spaces within World Bank data file were prohibiting the merge of country\_name with Happiest Country file. Therefore, spaces required removal within World Bank country\_name

\*\*\* df\_worldbank.country\_name = df\_worldbank.country\_name.str.strip()

* Column names required \n to be removed
* 6 countries were manually identified that required to be renamed from World Bank to join with the Happiest Country file.

\*\*\*\* df\_worldbank.rename(index={'Yemen, Rep.' :'Yemen',

'Venezuela, RB' :'Venezuela',

'Russian Federation' : 'Russia',

'Trinidad and Tobago' : 'Trinidad & Tobago',

'Syrian Arab Republic': 'Syria',

'Korea, Rep.' : 'South Korea’

* All column names were adjusted to remove spaces and replace with underlines

Happiest Countries in the World Process

* Column Fields were parsed out and were assigned to lower case
* Columns Fields required the removal of \n from each field name (extraneous characters)
* All values within the last column required the removal of \n (extraneous characters)
* All numeric fields were converted from STRING to FLOAT, except Overall Rank which was made INTEGER

NOTE: Numeric value scores from World Bank can be NULL. Missing Value handling will be determined by the Analysts.

**LOAD :**

* **Github: https://github.com/nj08831/Python\_ETL**
* **Final File: World\_Bank\_and\_Wiki\_Happiest\_Countries.xls**

**NOTE:**

* The final dataset of 156 records (countries) includes only the Happiest Countries in the World list.
* 21 of the 156 countries within the Happiest Country list have missing values from the World Bank healthcare expenditure field.
* 16 of the 156 countries within the Happiest Country list have missing values from the World Bank obesity/overweight adult field.

**FINAL DATA FIELDS**

Data columns (total 11 columns):

country\_name 156 non-null object

overall rank 156 non-null int32

score 156 non-null float64

gdp\_per\_capita 156 non-null float64

social\_support 156 non-null float64

healthy\_life\_expectancy 156 non-null float64

freedom\_to\_make\_life\_choices 156 non-null float64

generosity 156 non-null float64

perceptions\_of\_corruption 156 non-null float64

healthcare\_govt\_expenditure 135 non-null float64

per\_overweight\_adults 140 non-null float64