World Happiness and Development Indicators Project

Participants: David Fried, Carol Schiro, Jose Diaz, Eric Donnelly, Bryan Evans, Jill Amirakul

Goal: Download the “World Happiness report up to 2020” dataset and the “World Development Indicators” dataset from Kaggle. Extract the data from the CSV files, Transform and Clean the data. Then upload it to a final database. Data could be used to see what kind of relationship there is between happiness and world development data.

Extract Process: The “World Happiness Report Up To 2020” included 5 CSV files. Each file represented one year starting with 2015. We downloaded the 2015 CSV file from Kaggle and uploaded it into Pandas.

With this data we can see if there is any correlation between Happiness and Family, Generosity, Freedom, Economy (GDP) and other factors within the file itself. To see examples of this analysis please see the appendix A.

To make the fictitious analysis more interesting, we decided to combine this data with a separate dataset to determine if happiness had any correlation with different factors not available in the “World Happiness Report Up To 2020”.

We thought the indicators CSV file from the “World Development Indicators” would provide some interesting analysis points including:

- Does a country's development have an impact on how happy they are?
- Can we predict if a country will be happy based on how developed they are?
- How or does happiness have a positive impact on the environment particularly birds, fish, and mammals?

The “World Development Indicators” was found on Kaggle as well and downloaded to a CSV file and uploaded into pandas.

Transforming Data: The “World Happiness Report Up To 2020” was exceptionally clean data, with no missing values or duplicate data.

The “World Development Indicators” was not formatted in the same way that the “World Happiness Report Up To 2020” was. The variables in the “World Development Indicator file were not listed in multiple columns. This report had one variable column where sets of different variables were repeated for each country. Since this was the case, we were not able to merge the files together without pivoting the data so that each of the variables had their own column.

After pivoting the “World Development Indicators” data and merging it with the “World Happiness Report Up To 2020” we noticed that several columns from the “World Development Indicators” file had a multitude of missing values. Due to the amount of missing data in these columns, we decided to only keep the variable that had less than 100 missing values for the 139 countries that were shared between the “World Happiness Report Up To 2020” and the “World Development Indicators.” To see what data was kept and available in the SQL database please see Appendix B.

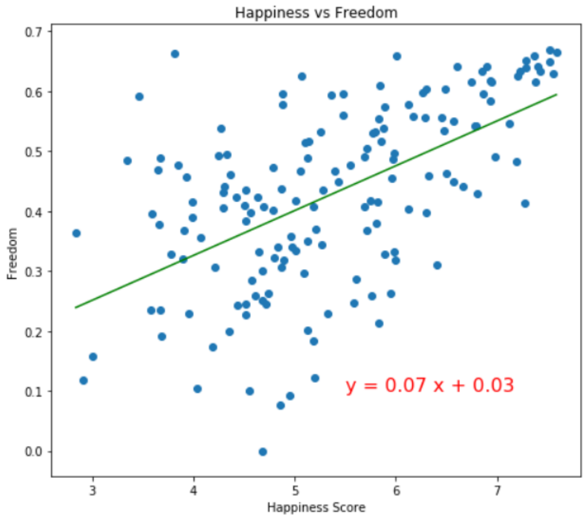
Last, we cut the merged file into 3 separate CSV files. The file was cut into three so any queries ran on the database could be done quicker. Each file contained the country name and would serve as the primary/foreign keys for the SQL file.

Loading the Data: The choice was made to upload the final 3 data sets into SQLAlchemy. This allows Python code to map from the database schema to the applications' Python objects. This allows for complex queries and mappings in Python.

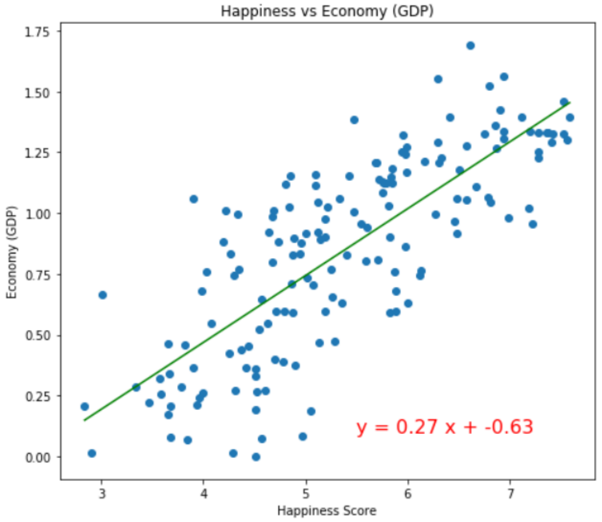
Appendix A: Correlation Examples

Below are examples showing the relationships between the data in the SQLAlchemy database which contains happiness and world development information.

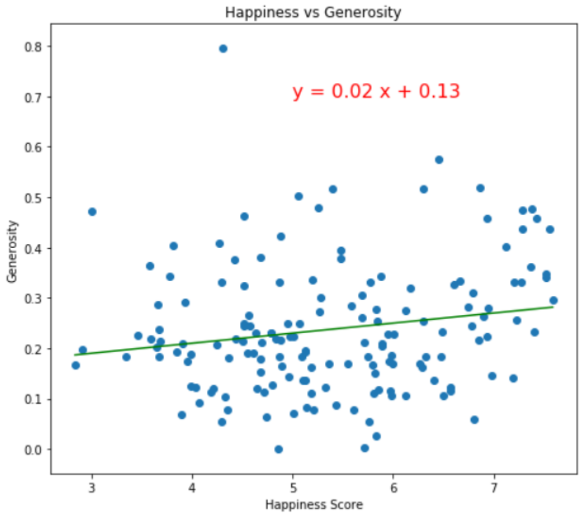
```
Coef. of Corr. Happiness Score vs Freedom = 0.57
```



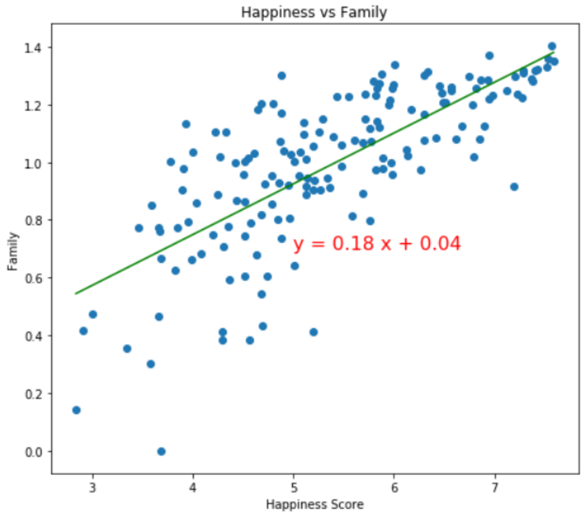
```
Coef. of Corr. Happiness Score vs Economy (GDP) = 0.78
There is a strong relationship between happiness and the performance of the economy
```



```
Coef. of Corr. Happiness Score vs Generosity = 0.18
There is not relationship between happiness and Generosity
```



```
Coef. of Corr. Happiness Score vs Family = 0.74
There is a strong relationship between happiness and Family
```



Appendix B:

Data Included in final SQL database.

Variable Name	Variable Definitions
Country	
Region	
Happiness Rank	
Happiness Score	The national average response to the question of life evaluations
Standard Error	
Economy (GDP per Capita)	GDP per capita in purchasing power parity
Family	
Health (Life Expectancy)	Healthy life expectancy at birth
Freedom	Responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
Trust (Government Corruption)	The measure is the national average of the survey responses to the questions in the Gallup World Poll: "Is corruption widespread throughout the government"
Generosity	Generosity is the residual of regressing national average in response to the question "Have you donated money to a charity in the past month?" on GDP per capita.
Dystopia Residual	The score of a hypothetical country that has a lower rank than the lowest ranking country on the report, plus the residual value of each country (a number that is left over from the normalization of the variables which cannot be explained)
Bird species, threatened	Birds are listed for countries included within their breeding or wintering ranges. Threatened species are the number of species classified by the IUCN as endangered, vulnerable, rare, indeterminate, out of danger, or insufficiently known.

Business extent of disclosure index	Disclosure index measures the extent to which investors are protected through disclosure of ownership and financial information. The index ranges from 0 to 10, with higher values indicating more disclosure.
Cost of business start-up procedures	Cost to register a business is normalized by presenting it as a percentage of gross national income (GNI) per capita.
Depth of credit information index	Depth of credit information index measures rules affecting the scope, accessibility, and quality of credit information available through public or private credit registries. The index ranges from 0 to 8, with higher values indicating the availability of more credit information, from either a public registry or a private bureau, to facilitate lending decisions.
Distance to frontier score	Distance to frontier score illustrates the distance of an economy to the "frontier," which represents the best performance observed on each Doing Business topic across all economies and years included since 2005. An economy's distance to frontier is indicated on a scale from 0 to 100, where 0 represents the lowest performance and 100 the frontier. For example, a score of 75 in 2012 means an economy was 25 percentage points away from the frontier constructed from the best performances across all economies and across time. A score of 80 in 2013 would indicate the economy is improving.
Ease of doing business index	Ease of doing business ranks economies from 1 to 189, with first place being the best. A high ranking (a low numerical rank) means that the regulatory environment is conducive to business operation. The index averages the country's percentile rankings on 10 topics covered in the World Bank's Doing Business. The ranking on each topic is the simple average of the percentile rankings on its component indicators.
Fish species, threatened	Fish species are based on Froese, R. and Pauly, D. (eds). 2008. Threatened species are the number of species classified by the IUCN as endangered, vulnerable, rare, indeterminate, out of danger, or insufficiently known.
Improved sanitation facilities	Access to improved sanitation facilities refers to the percentage of the population using improved sanitation facilities. Improved sanitation facilities are likely to ensure hygienic separation of human excreta from human contact. They include flush/pour flush (to piped sewer system, septic tank, pit latrine), ventilated improved pit (VIP) latrine, pit latrine with slab, and composting toilet.

Improved sanitation facilities, rural	Refers to the percentage of the population using improved sanitation facilities. Improved sanitation facilities are likely to ensure hygienic separation of human excreta from human contact. They include flush/pour flush (to piped sewer system, septic tank, pit latrine), ventilated improved pit (VIP) latrine, pit latrine with slab, and composting toilet.
Improved sanitation facilities, urban	Access to improved sanitation facilities refers to the percentage of the population using improved sanitation facilities. Improved sanitation facilities are likely to ensure hygienic separation of human excreta from human contact. They include flush/pour flush (to piped sewer system, septic tank, pit latrine), ventilated improved pit (VIP) latrine, pit latrine with slab, and composting toilet.
Improved water source (% of population with access)	Access to an improved water source refers to the percentage of the population using an improved drinking water source. The improved drinking water source includes piped water on premises (piped household water connection located inside the userÂ's dwelling, plot or yard), and other improved drinking water sources (public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs, and rainwater collection).
Improved water source, rural	Access to an improved water source refers to the percentage of the population using an improved drinking water source. The improved drinking water source includes piped water on premises (piped household water connection located inside the userÂ's dwelling, plot or yard), and other improved drinking water sources (public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs, and rainwater collection).
Improved water source, urban	Access to an improved water source refers to the percentage of the population using an improved drinking water source. The improved drinking water source includes piped water on premises (piped household water connection located inside the userÂ's dwelling, plot or yard), and other improved drinking water sources (public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs, and rainwater collection).
Lifetime risk of maternal death (%)	Life time risk of maternal death is the probability that a 15-year-old female will die eventually from a maternal cause assuming that current levels of fertility and mortality (including maternal mortality) do not change in the future, taking into account competing causes of death.

Lifetime risk of maternal death	Life time risk of maternal death is the probability that a 15-year-old female will die eventually from a maternal cause assuming that current levels of fertility and mortality (including maternal mortality) do not change in the future, taking into account competing causes of death.
Mammal species, threatened	Mammal species are mammals excluding whales and porpoises. Threatened species are the number of species classified by the IUCN as endangered, vulnerable, rare, indeterminate, out of danger, or insufficiently known.
Maternal mortality ratio	Maternal mortality ratio is the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births. The data are estimated with a regression model using information on the proportion of maternal deaths among non-AIDS deaths in women ages 15-49, fertility, birth attendants, and GDP.
Mortality rate, infant (per 1,000 live births)	Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year.
Mortality rate, infant, female (per 1,000 live births)	Infant mortality rate, female is the number of female infants dying before reaching one year of age, per 1,000 female live births in a given year.
Mortality rate, infant, male (per 1,000 live births)	Infant mortality rate, male is the number of male infants dying before reaching one year of age, per 1,000 male live births in a given year.
Mortality rate, neonatal (per 1,000 live births)	Neonatal mortality rate is the number of neonates dying before reaching 28 days of age, per 1,000 live births in a given year.
Mortality rate, under-5 (per 1,000)	Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.
Mortality rate, under-5, female	Under-five mortality rate, female is the probability per 1,000 that a newborn female baby will die before reaching age five, if subject to female age-specific mortality rates of the specified year.
Mortality rate, under-5, male	Under-five mortality rate, male is the probability per 1,000 that a newborn male baby will die before reaching age five, if subject to male age-specific mortality rates of the specified year.
Number of maternal deaths	Maternal deaths is the number of women who die during pregnancy

	and childbirth.
Plant species (higher), threatened	Higher plants are native vascular plant species. Threatened species are the number of species classified by the IUCN as endangered, vulnerable, rare, indeterminate, out of danger, or insufficiently known.
Private credit bureau coverage (% of adults)	Private credit bureau coverage reports the number of individuals or firms listed by a private credit bureau with current information on repayment history, unpaid debts, or credit outstanding. The number is expressed as a percentage of the adult population.
Procedures to build a warehouse (number)	Number of procedures to build a warehouse is the number of interactions of a company's employees or managers with external parties, including government agency staff, public inspectors, notaries, land registry and cadastre staff, and technical experts apart from architects and engineers.
Procedures to register property (number)	Number of procedures to register property is the number of procedures required for a business to secure rights to property.
Proportion of seats held by women in national ...	Women in parliaments are the percentage of parliamentary seats in a single or lower chamber held by women.
Public credit registry coverage (% of adults)	Public credit registry coverage reports the number of individuals and firms listed in a public credit registry with current information on repayment history, unpaid debts, or credit outstanding. The number is expressed as a percentage of the adult population.
Start-up procedures to register a business	Start-up procedures are those required to start a business, including interactions to obtain necessary permits and licenses and to complete all inscriptions, verifications, and notifications to start operations. Data are for businesses with specific characteristics of ownership, size, and type of production.
Strength of legal rights index	Strength of legal rights index measures the degree to which collateral and bankruptcy laws protect the rights of borrowers and lenders and thus facilitate lending. The index ranges from 0 to 12, with higher scores indicating that these laws are better designed to expand access to credit.
Tax payments (number)	Tax payments by businesses are the total number of taxes paid by businesses, including electronic filing. The tax is counted as paid once a year even if payments are more frequent.

Time required to build a warehouse (days)	Time required to build a warehouse is the number of calendar days needed to complete the required procedures for building a warehouse. If a procedure can be speeded up at additional cost, the fastest procedure, independent of cost, is chosen.
Time required to enforce a contract (days)	Time required to enforce a contract is the number of calendar days from the filing of the lawsuit in court until the final determination and, in appropriate cases, payment.
Time required to get electricity (days)	Time required to get electricity is the number of days to obtain a permanent electricity connection. The measure captures the median duration that the electricity utility and experts indicate is necessary in practice, rather than required by law, to complete a procedure.
Time required to register property (days)	Time required to register property is the number of calendar days needed for businesses to secure rights to property.
Time required to start a business (days)	Time required to start a business is the number of calendar days needed to complete the procedures to legally operate a business. If a procedure can be speeded up at additional cost, the fastest procedure, independent of cost, is chosen.
Time to prepare and pay taxes (hours)	Time to prepare and pay taxes is the time, in hours per year, it takes to prepare, file, and pay (or withhold) three major types of taxes: the corporate income tax, the value added or sales tax, and labor taxes, including payroll taxes and social security contributions.
Time to resolve insolvency (years)	Time to resolve insolvency is the number of years from the filing for insolvency in court until the resolution of distressed assets.
Total tax rate (% of commercial profits)	Total tax rate measures the amount of taxes and mandatory contributions payable by businesses after accounting for allowable deductions and exemptions as a share of commercial profits. Taxes withheld (such as personal income tax) or collected and remitted to tax authorities (such as value added taxes, sales taxes or goods and service taxes) are excluded.
Number of infant deaths	Number of infants dying before reaching one year of age.
Number of neonatal deaths	Number of neonates dying before reaching 28 days of age.
Number of under-five deaths	Number of children dying before reaching age five.

Note: Definitions were taken from the “World Development Indicators” and documentation from the World Happiness Report.