**World Happiness Final Project**

**Kristen Goh and Jonathan Goh**

**Woz U & Bethel School of Technology**

**Variables**

World Happiness 2020 Data Set

* **Country Name**
* **Regional Indicator**
* **Happiness Score**

World Health 2020 Data Set

* **cleanFuelAndTech**.csv -> Proportion of population with primary reliance on clean fuels and technologies (%)
* **basicDrinkingWaterServices**.csv ->Population using at least basic drinking water services (%)
* Sanitation and Hygiene
* **atLeastBasicSanitizationServices**.csv - > Population using at least basic sanitation services (%)
* roadTrafficDeaths.csv -> Estimated road traffic death rate per 100,000 population
* **AlcoholSubstanceAbuse**.csv -> Total (recorded + unrecorded) alcohol per capita (15 +) consumption’s.
* **crudeSuicideRates**.csv -> Crude suicide rates per 100,000 population
* **infantMortalityRate**.csv-> Probability of dying between birth and age 1 per 1000 live births.
* **lifeExpectancyAtBirth**.csv -> Life expectancy at birth, country wise mentioned in age (years).

**Evaluation Questions**

1. Identify how happiness differs by country.
   * **Variables**: Happiness Score (continuous DV), Country (categorical IV)
   * **Statistical Test**: ANOVA
   * *Possible to add population as co-variate.*
   * *Possible to run analysis by country type – developing, developed…*
2. Identify health metrics that influence happiness.
   * **Variables**: Happiness Score (continuous - DV), infant mortality rate (continuous - IV), life expectancy at birth (continuous - IV), suicide rates (continuous - IV), road traffic deaths (continuous - IV)
   * **Statistical Test**: Multiple linear regression
   * *At least 20 rows per independent variable – can add a few more variables*
3. Identify the environmental and lifestyle indicators that influence happiness.
   * **Variables**: Happiness Score (continuous - DV), clean fuel and tech (continuous), basic sanitation services (continuous), basic drinking water services (continuous), alcohol substance abuse (continuous)
   * **Statistical Test**: Multiple linear regression

**Additional descriptive statistics**

* Graph happiness scores by region from 2015-2020
* Graph average, range, and standard deviation of total happiness scores in 2020

*Probably better for R Studio at the moment*

*Data wrangling in Python and then move to R-studio*

*Add a machine learning in Python*

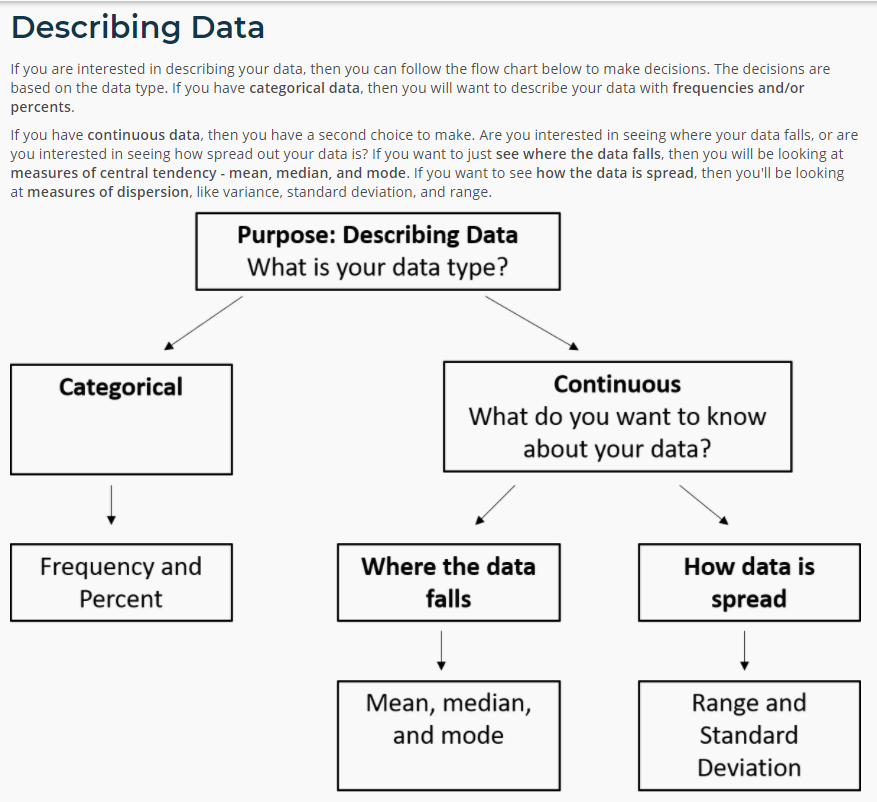
*Python –*

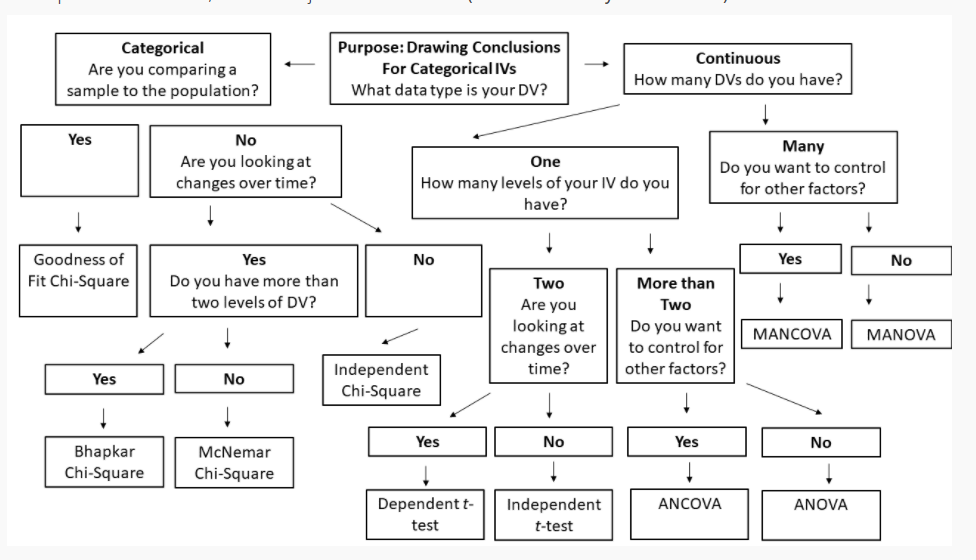
*Melt code – transformation with specific variables – change orientations with this held and aggregated*

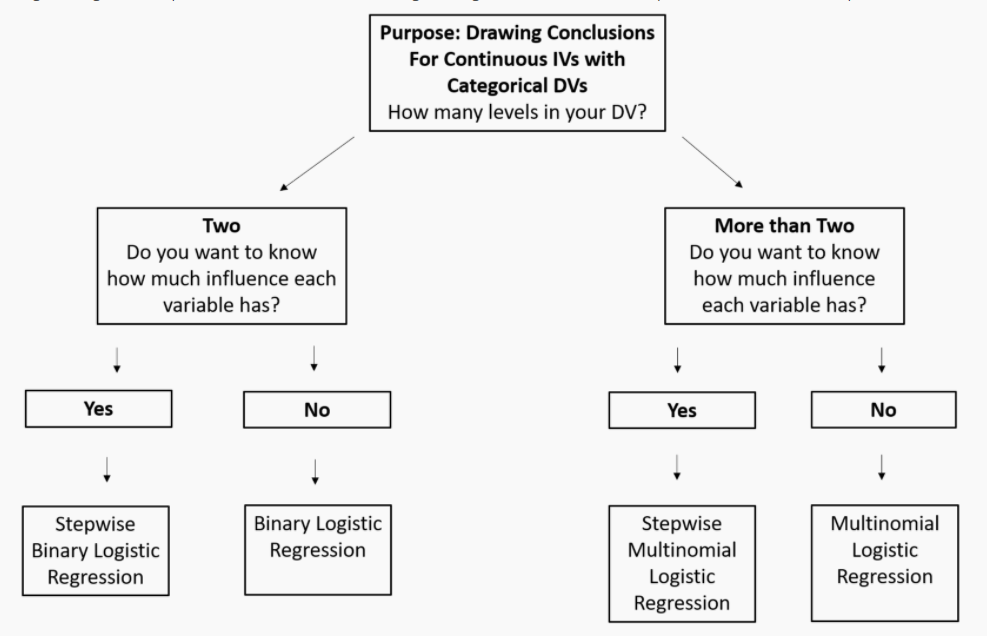
**1/13 – Questions**

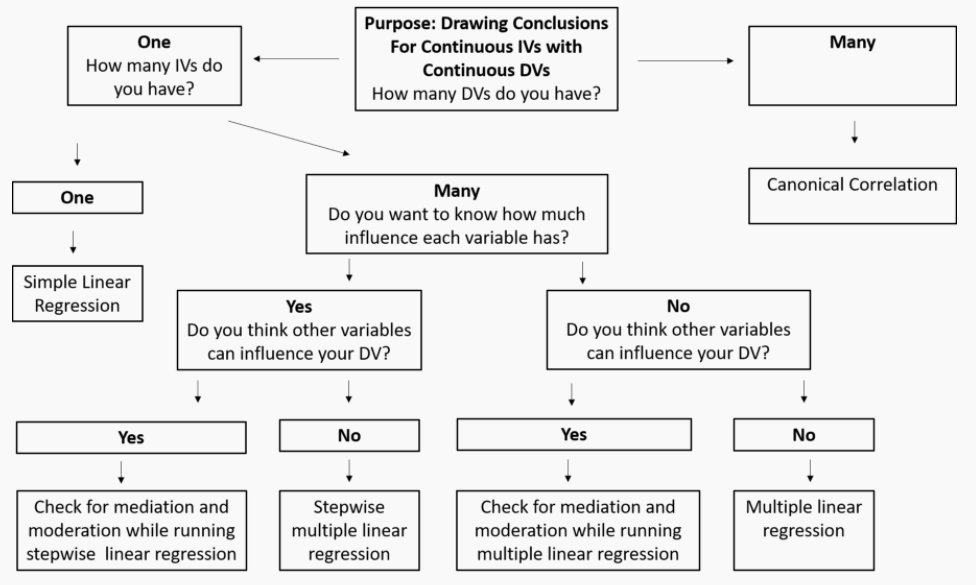
* Are variables that are measured as probability continuous?
* Are variables that are measured as rates (proportion/percentage) continuous?
* Keep the country denominator based on the world happiness data set?
* Can we use machine learning to predict happiness score? Would it be applicable?

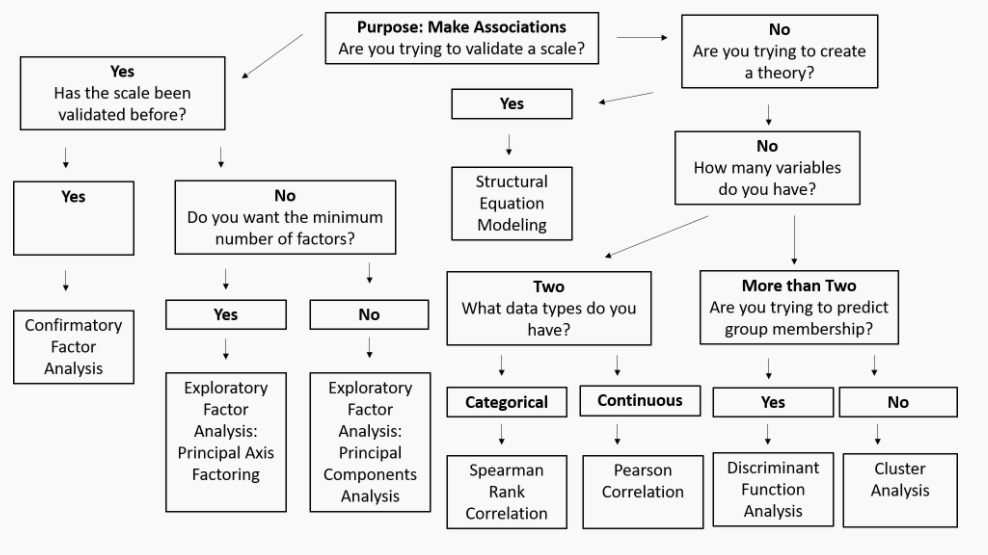
Random forests if happiness scores are categorical.











|  |  |
| --- | --- |
| **Keyword** | **Description** |
| **Inferential Statistics** | Branch of statistics when you are drawing conclusions from your data. |
| **Frequency and Percent** | A statistic for when you want to describe your data and your data is categorical. |
| **Measures of Central Tendency** | Mean, median, and mode, for when you want to describe your data, your data is continuous, and you want to know where the data falls. |
| **Measures of Distribution** | Range and standard deviation, for when you want to describe your data, the data is continuous, and you want to know how the data is spread. |
| **Goodness of Fit Chi-Square** | For when you are drawing conclusions about your data, have a categorical IV and a categorical DV, and are comparing a sample to a population. |
| **Bhapkar Chi-Square** | For when you are drawing conclusions about your data, have a categorical IV and a categorical DV, are looking at changes over time, and have more than two levels of your dependent variable. |
| **McNemar Chi-Square** | For when you are drawing conclusions about your data, have a categorical IV and a categorical DV, are looking at changes over time, and have only two levels of your dependent variable. |
| **Independent Chi-Square** | For when you are drawing conclusions about your data, have a categorical IV and a categorical DV, and are not looking at changes over time. |
| **Dependent t-test** | For when you are drawing conclusions about your data, have a categorical IV with two levels and a continuous DV, and are looking at changes over time. |
| **Independent t-test** | For when you are drawing conclusions about your data, have a categorical IV with two levels and a continuous DV, and are not looking at changes over time. |
| **Analysis of Covariance (ANCOVA)** | For when you are drawing conclusions about your data, have a categorical IV with two or more levels and a continuous DV, and want to control for other factors. |
| **Analysis of Variance (ANOVA)** | For when you are drawing conclusions about your data, have a categorical IV with two or more levels and a continuous DV, and do not want to control for other factors. |
| **Multivariate Analysis of Covariance (MANCOVA)** | For when you are drawing conclusions about your data, have a categorical IV with two levels or more and multiple continuous DVs, and want to control for other factors. |
| **Multivariate Analysis of Variance (MANOVA)** | For when you are drawing conclusions about your data, have a categorical IV with two levels or more and multiple continuous DVs, and do not want to control for other factors. |
| **Stepwise Binary Logistic Regression** | For when you are drawing conclusions about your data, have a continuous IV with a categorical DV with two levels, and want to see how much influence each individual variable has. |
| **Binary Logistic Regression** | For when you are drawing conclusions about your data, have a continuous IV with a categorical DV with two levels, and do not want to see how much influence each individual variable has. |
| **Stepwise Multinomial Logistic Regression** | For when you are drawing conclusions about your data, have a continuous IV with a categorical DV with more than two levels, and want to see how much influence each individual variable has. |
| **Multinomial Logistic Regression** | For when you are drawing conclusions about your data, have a continuous IV with a categorical DV with more than two levels, and do not want to see how much influence each individual variable has. |
| **Simple Linear Regression** | For when you are drawing conclusions about your data, and have a continuous IV with a continuous DV. |
| **Mediation and Moderation while running Stepwise Linear Regression** | For when you are drawing conclusions about your data, have a continuous IV with a continuous DV, want to know how much influence each individual variable has, and think other variables might influence your DV. |
| **Stepwise Linear Regression** | For when you are drawing conclusions about your data, have a continuous IV with a continuous DV, want to know how much influence each individual variable has, and do not think other variables might influence your DV. |
| **Mediation and Moderation while running Multiple Linear Regression** | For when you are drawing conclusions about your data, have a continuous IV with a continuous DV, do not want to know how much influence each individual variable has, and think other variables might influence your DV. |
| **Multiple Linear Regression** | For when you are drawing conclusions about your data, have a continuous IV with a continuous DV, do not want to know how much influence each individual variable has, and do not think other variables might influence your DV. |
| **Confirmatory Factor Analysis** | For when you are making associations about your data, are validating a scale, and it has been validated before. |
| **Exploratory Factor Analysis - Principal Axis Factoring** | For when you are making associations about your data, are validating a scale, you are validating that scale for the first time, and you want the minimum number of factors. |
| **Exploratory Factor Analysis - Principal Components Analysis** | For when you are making associations about your data, are validating a scale, you are validating that scale for the first time, and you don't care if you have the minimum number of factors. |
| **Structural Equation Modeling** | For when you are making associations about your data, and are trying to create a theory. |
| **Spearman Rank Correlation** | For when you are making associations about your data and have two categorical variables. |
| **Pearson Correlation** | For when you are making associations about your data and have two continuos variables. |
| **Discriminant Function Analysis** | For when you are making associations about your data, have more than two variables, and are trying to predict group membership. |
| **Cluster Analysis** | For when you are making associations about your data, have more than two variables, and are not trying to predict group membership. |