Requirements:

* 5 variables that are not identifiers
  + At least 2 should be categorical (or we should have some plan on how to treat 2 quantitative variables as categorical)
* Each group member must design at least 2 initial questions - general but not trivial
  + In total we need 8 questions

Proposal must include:

<https://www.cdc.gov/brfss/annual_data/annual_2019.html>

<https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete>

Tips from Sam:

* Try to make sure that we are not all answering the same question - pick something that genuinely interests you
* Be well prepared for the presentation (Creator) - make sure to time yourself so that you can convey all the information within the given time. Make sure that you can field questions about someone else’s work/analysis.

Roles:

* Creator: Kevin Lu
* Interpreter: Aditi Chiney
* Orator : Jessica Barta
* Deliverer: Julia Mengxuan Yu

JY: World health

1. What’s the relationship between life expectancy and alcohol substance abuse in groups of males and females? Is there a significant difference between these two groups?
2. How can the variable medical doctor influence life expectancy and alcohol substance abuse? What is the result of the first question after accounting for the influence of medical doctor?

KL: World Health

1. What is the relationship between universal health coverage and crude suicide rates?
2. What is the relationship between alcohol substance abuse and road traffic deaths?

JB: World Health

1. Looking at our most recent measurements of health outcomes, are there any countries that are above average in most health outcomes, but do very poorly in others? Similarly, do some countries exhibit generally poor health outcomes except in just a few areas? Think about why this might be the case.
2. **The datasets contain measurements for the same countries over several different years. Which countries are improving the most overall? (This could be in all health outcomes or just a subset.) Which countries are getting worse in terms of health outcomes?**
3. **Which health outcomes show the most, and the least, variation among countries? Why might this be?**
4. Do neighboring countries tend to show similar health outcomes?
5. Which health outcomes are most representative of a country’s overall health? Are there variables that we can combine to produce an overall health index?

AC: World Health

1. Does increased UHC coverage predict better health outcomes for women across the following variables: (1) average reproductive age for women who have access to family planning services, (2) teenage pregnancy birth rates, (3) women who have faced sexual violence, and (4) healthy life expectancy at birth for women?
2. How does the probability of dying between ages 30-70 from a chronic disease (CVD, cancer, diabetes, chronic respiratory illness) vary with UHC coverage? If we stratify the chronic disease data by country using a categorical air pollution classification (we’ll create bins/ranges using the variable airPollutionDeathRate and sort into Low Risk, Moderate Risk, High Risk), is there a pattern among countries with high probabilities of dying from chronic disease and their UHC coverage?

**Variables of interest & Description (5 minimum) - list type (at least 2 categorical)**

* Numerical variables
  + HALElifeExpectancyAtBirth - Healthy life expectancy (HALE) at birth, country wise mentioned in age (years)
  + AlcoholSubstanceAbuse - Total (recorded + unrecorded) alcohol per capita (15 +) consumption’s
  + medicalDoctors - Medical doctors per 10,000 population
  + uhcCoverage - UHC index of service coverage (SCI)
  + crudeSuicideRates - Crude suicide rates per 100,000 population
  + roadTrafficDeaths - Estimated road traffic death rate per 100,000 population
  + reproductiveAgeWomen - Married or in-union women of reproductive age who have their need for family planning satisfied with modern methods (%)
  + adolescentBirthRate - Adolescent birth rate per 1000 women aged 15-19 years
  + eliminateViolenceAgainstWomen - Proportion of ever-partnered women and girls aged 15-49 years subjected to physical and/or sexual violence by a current or former intimate partner in previous 12 months
  + 3070cancerChdEtc - Probability of dying between the age of 30 and exact age of 70 from any of the cardiovascular disease, cancer, diabetes, or chronic respiratory disease
* Categorical variables
  + Gender - Male, female, both sexes
  + airPollutionDeathRate - can be transformed into categorical by sorting into 3 “bins”/ranges which will be renamed as Low Risk, Moderate Risk, and High Risk (of dying due to ambient and household air pollution-related causes)
    - Ambient and household air pollution attributable death rate per 100,00 population and the same data with age-standardized

Second Paragraph:

The first follow up question explored the relationship between universal health coverage and crude suicide rates while considering gender as a factor. The new figure showed that the positive correlation observed in the initial question is almost entirely due to a strong positive correlation for males, while there was little to no change for females based on universal health coverage. The figure also provides insight that female suicide rates are much lower than male suicide rates overall. The second follow up question explored the distribution of changes in the healthcare index. Most countries didn't see much change in their coverage between 2015 and 2017, based on the low standard deviation, about 1.7, and boxplot showing little variation. There were only a couple outliers noted above. The lack of variation brings up questions about how healthcare coverage was measured. Should the WHO have used a more sensitive measurement that would pick up more variation? It seems unlikely that nearly all the countries surveyed changed very little, or the same amount, over time.

Does UHC coverage account for the trend in risk of mortality across both genders? Is there a significant difference if we stratify by gender?

----

Final report:

Editing - Julia, everyone else take a look

Introduction - Jessica

* 2 questions clearly defined
* Rationale

Data - Aditi

* Describe source and variables
* 1 table
* 1 descriptive figure

Results - Everyone

* Multiple models used
* Explanation of findings & methods
* At least 4 figures/tables

Conclusion - Julia

* Summary of findings
* Re-connect findings to rationale
* Future analyses proposed

Potential Questions & Figures:

* Q1: Model to predict which regressors are most significant in predicting female HALE (ex. Healthcare coverage index, access to family planning services, teen pregnancy rates, medical doctor availability, midwife/doula availability, etc.) - test at least 5-6 combinations of regressors (i.e. 5-6 models) and look at MSE, MAE, adjusted R-squared values, p-values
  + Possible plots: bar plots of MAE/MSE, residual plots for all models
* Q2: WHO Healthcare coverage index over time (gganimation) - organize countries into world regions and animate HCI over periods of time
  + Instead of animation, could also plot for each period of time and then stack plots
  + Also include heatmap showing decrease/increase in coverage over time
* Research which health outcomes are most often used to gague health progress in different countries. Use R shiny to produce a graph that can be adjusted by country.
* Could comment on the data itself. How does the data compare to other datasets that also are reporting on countries’ health outcomes? Is this data inadequate for some reason? (We noticed that the years tend to not match up.) If so, we could recommend different data collection methods for the WHO to use.

<https://odum.unc.edu/statistics-help-desk/>