Final Project Proposal

Team Members
Minh Vu, 33077769
Jana Chittarath, 87884193

Github Repo
https://github.com/minhVu03/Bayesian-Data-Analysis-Project.git

Datasets

Dataset 1

Data Name: Crude Suicide Rate (Per 100,000 Population)
Source: https://www.who.int/data/gho/data/themes/mental-health/suicide-rates
Description: The raw dataset has notable features like country, age group, sex, and suicide rate (per 100,000 people) that can be extracted.

```
> head(suicide_data)
# A tibble: 6 × 34
  IndicatorCode Indicator  ValueType ParentLocationCode ParentLocation `Location type` SpatialDimValueCode Location
  <chr>         <chr>      <chr>     <chr>              <chr>          <chr>           <chr>                <chr>
1 SDGSUICIDE    Crude sui… text      AMR                Americas       Country         VCT                  Saint V…
2 SDGSUICIDE    Crude sui… text      EMR                Eastern Medit… Country         OMN                  Oman
3 SDGSUICIDE    Crude sui… text      EMR                Eastern Medit… Country         PSE                  occupie…
4 SDGSUICIDE    Crude sui… text      EMR                Eastern Medit… Country         JOR                  Jordan
5 SDGSUICIDE    Crude sui… text      EMR                Eastern Medit… Country         KWT                  Kuwait
6 SDGSUICIDE    Crude sui… text      EMR                Eastern Medit… Country         SYR                  Syrian …
# ℹ 26 more variables: `Period type` <chr>, Period <dbl>, IsLatestYear <lgl>, `Dim1 type` <chr>, Dim1 <chr>,
#   Dim1ValueCode <chr>, `Dim2 type` <chr>, Dim2 <chr>, Dim2ValueCode <chr>, `Dim3 type` <lgl>, Dim3 <lgl>,
#   Dim3ValueCode <lgl>, DataSourceDimValueCode <lgl>, DataSource <lgl>, FactValueNumericPrefix <lgl>,
#   FactValueNumeric <dbl>, FactValueUoM <lgl>, FactValueNumericLowPrefix <lgl>, FactValueNumericLow <dbl>,
#   FactValueNumericHighPrefix <lgl>, FactValueNumericHigh <dbl>, Value <chr>, FactValueTranslationID <lgl>,
#   FactComments <lgl>, Language <chr>, DateModified <dttm>
```

Dataset 2

Data name: Weather Forecast Accuracy
Source: https://github.com/rfordatascience/tidytuesday/tree/main/data/2022/2022-12-20
Description: The dataset includes information related to weather forecasting, but we will focus on variables relating to the observed high and low temperature over a 16-month period for cities across the United States.

```
> head(weather_forecasts)
# A tibble: 6 × 10
  date       city    state high_or_low forecast_hours_before observed_temp forecast_temp observed_precip
  <date>     <chr>   <chr> <chr>                       <dbl>         <dbl>         <dbl>           <dbl>
1 2021-01-30 ABILENE TX    high                           48            70            NA               0
2 2021-01-30 ABILENE TX    high                           36            70            NA               0
3 2021-01-30 ABILENE TX    high                           24            70            NA               0
4 2021-01-30 ABILENE TX    high                           12            70            70               0
5 2021-01-30 ABILENE TX    low                            48            42            NA               0
6 2021-01-30 ABILENE TX    low                            36            42            NA               0
# ℹ 2 more variables: forecast_outlook <chr>, possible_error <chr>
```

Project Theme

This project will focus on using a Bayesian spatial model to analyze a specific variable of interest, depending on the chosen dataset.

(Dataset 1) The variable of interest for this dataset is the mortality rate by suicide in countries around the world. According to the Suicide Prevention Resource Centre, "In 2020, [...] Suicide was the third leading cause of death for ages 15 to 24, the fourth leading cause of death for ages 35 to 44, and the seventh leading cause of death for ages 55 to 64." Hence, it is essential to analyze historical suicide data around the world to estimate the severity of the issue and prevent new attempts. Since on average the suicide rate seems to be higher for younger people, we would like to analyze the data to find the posterior mean of the suicide rate for the age groups 15-24, and hence determine the average suicide rate. Furthermore, the suicide rates in countries in similar geographic areas may be correlated, given their location and cultural influence on each other. Therefore, we can use a spatial structure.

(Dataset 2) The variable of interest for this dataset is the average high temperature during the summer in cities across the United States. With the increasingly pressing issue of global warming, we are interested in investigating whether there is a significant increase in average temperature after a single year. As this dataset is from 2021-2022, it is relatively recent. This analysis may help to inform the rate at which temperatures are rising, and thus how quickly we need to take further action to reduce the effects of global warming.

Summary of Potential Approaches

(Dataset 1) A prior probability/belief needs to be established. We can do some literature review to obtain this belief and use a distribution like Poisson (since the suicide rate cannot be negative) as a prior distribution. To resolve the issue of correlated data between countries near each other, we can apply a spatial model to account for this (we're still doing some research on this topic since it's pretty new to us). Then, we can use the Metropolis-Hastings algorithm (MCMC) to start calculating the posterior distribution of the suicide rate and obtain the mean and its confidence intervals. *If time permits (or if you'd recommend it)*, we can compare this value to one of another age group (which can be obtained the same way as described above) to answer our research question: Which age group has a higher suicide rate?

(Dataset 2) This dataset includes a few irrelevant variables and focuses on the summer season, we can tidy and wrangle it before moving on to the analysis. Similar to the approach for Dataset 1, we will obtain a prior probability based on literature review, and use a spatial model to account for correlated data between cities near each other. We can then use MCMC to calculate the posterior mean of the high temperature for each region and a credible interval for each estimate.

Collaboration Plan

We will work on the project together in person after office hours on Thursdays or Tuesdays after lecture, if needed. We'll also communicate via Discord and can collaborate during virtual meetings. This project needs to be done chronologically, so we will primarily be working together for ideation (coming up with models, literature review, etc.). We can then put together the final written report on an R Notebook file and commit to Github as we work.

- Jana: Intro, Literature Review, Data Cleaning, Problem Formulation, spatial model implementation into prior, Challenges, Conclusion
- Minh: Intro, Literature Review, Dataset description, Problem Formulation, Prior Model, MCMC calculation to obtain posterior mean & CI, Analysis of Mean, Conclusion

If we decide to compare the mean suicide rate of both age groups (Dataset 1) then we would divide the work: Each person does one age group for the prior formulation, spatial model, prior and posterior distribution.

I. Sources

https://sprc.org/about-suicide/scope-of-the-problem/suicide-by-age/
https://www.paulamoraga.com/book-spatial/bayesian-spatial-models.html
https://becarioprecario.bitbucket.io/inla-gitbook/ch-spatial.html
https://cran.r-project.org/web/packages/geostan/vignettes/custom-spatial-models.html
https://rpubs.com/chrisbrunsdon/503833