**Proposal**

**Section 1 - Introduction: The introduction should introduce your general research question and your data (where it came from, how it was collected, what are the cases, what are the variables, etc.).**

**Research question:**

Among non-elderly adults living in the US in the year 2020, what are the distribution and predictors of high risk of psychological distress across the US regions?

We are considering an assessment of the risk for serious mental health distress in the US using the Kaiser Psychological distress scale (K6). K6 is a six-item self-reported measure of psychological distress. We will show how selected sociodemographic factors influence the observed risk.

**Data paragraph:**

The Urban Institute introduced the Well-Being and Basic Needs Survey (WBNS) in 2017 [(Zuckerman, 2020)](https://www.zotero.org/google-docs/?lJ3lQV). The survey has the purpose of tracking the well-being of non-elderly adults and their families across the country. It is designed to collect data covering topics such as health, material hardship, and safety net programs. The results of this survey aim to inform policymakers in the reform of federal safety net programs serving low-income families. The data were collected using a stratified random sample of about 7500 individuals between 18 and 64 years old living in the United States in the year 2020.

The dataset that is going to be used for this project comes from the 2020 WBNS survey [(Zuckerman, 2022)](https://www.zotero.org/google-docs/?ffa0Zn).  It is important to consider that this survey has a greater representation of individuals with low household income (under 150 percent of the federal poverty level). This decision was made with the purpose of refining the estimates for this group. The data about the individuals included in the sample was gathered from KnowledgePanel. This internet panel is probability-based and represents the whole nation. The survey could be completed in English or Spanish. Internet access was provided for individuals who did not have web connectivity.

The data is distributed by the inter-university consortium for political and social research (ICPSR).

Our data consists of a major outcome variable; risk of severe mental disorder as measured by K6 and the predictor variables age, education, race/ethnicity, gender, household head,

household income, MSA area, region, LGBT, marital status, number of children age 0-18 in household,  any own child age 0-18 in household,  and having moved in the last 12 months.

**Section 2 - Data: Place your data in the `/data` folder and use the dataReporter package to create a codebook. Do not include unnecessary columns in your dataset.**

Report can be found in the zip file for the proposal.

* Codebook file names: proposal\_codebook\_Mar7.Rmd, proposal\_codebook\_Mar7.html
* Files for data visualizations and data exploration: Proposal\_final\_project.Rmd, Proposal\_final\_project.html

**Section 3 - Data analysis plan:**

**Describe the key variables to answer your question:**

The predictor variables we will explore are the following:

* Age: participant age
* Education: less than high school; high school; some college; bachelor’s degree or higher
* Race/Ethnicity: white, non-hispanic ; black, non-hispanic ; other, non-hispanic, hispanic
* Gender: male; female
* Household head: yes; no
* Household income: between less than $5000 to more than $250000
* MSA area: metropolitan area; non-metropolitan area
* Region: northeast, midwest, south, west
* LGBT: yes; no
* Married: yes; no
* Number of children age 0-18 in household: 0 - 6 or more
* Any own child age 0-18 in household: yes; no
* Survey question: have you moved in the last 12 months? Yes; no

Age is a numeric variable while all other are categorical

The outcome variable (risk of severe mental disorder) will be calculated using the following variables (categorical) defined by the K6 psychological distress scale. The survey includes the six questions that compose this scale:

* During the past 30 days, about how often did you feel: nervous?
* During the past 30 days, about how often did you feel: hopeless?
* During the past 30 days, about how often did you feel: restless or fidgety?
* During the past 30 days, about how often did you feel: so sad that nothing could cheer you up?
* During the past 30 days, about how often did you feel: that everything was an effort?
* During the past 30 days, about how often did you feel: worthless?

The answer to all six questions can be any of the following options: “0 none of the time”, “1 a little of the time”, “2 some of the time”, “3  most of the time”, “4 all of the time”. The sum of the sores lies between 0 and 24. A score of 0 to 7 implies “low distress”, 8 to 12 “moderate distress”, and 13 to 24 “high risk of psychological distress” [(Donker et al., 2010; Hilton & Whiteford, 2010; Hozawa et al., 2009)](https://www.zotero.org/google-docs/?DzHqFU). In our study, this variable will be treated as a binary variable with levels 1 to 12 “low to moderate distress” and 13 to 24 “high risk of psychological distress”.

**Preliminary exploratory data analysis**

**Chart, bar chart

Description automatically generated**

**Figure 1: Distribution of the variables that will be used to generate our outcome variable**

**How this can help us learn more about our data:**

Based on these visualizations, we can see that there is variability in the responses for each one of the K6 questions. Our goal would be to determine who are the individuals that are showing the highest score in the K-6 scale and the sociodemographic factors they have in common.

Our next steps will include some data wrangling to tidy the answer responses. We will remove white space, recode the responses to the numerical scale according to the K6 indicator scores, and categorize them as low to moderate distress or high risk of psychological distress.

**Graphical user interface, chart

Description automatically generated**

**Figure 2: Distribution of our predictor variables**

**How this can help us learn more about our data:**

Having plots for the predictor/exposure variables allows us to see the distribution patterns in the population. For age, we can see that the participants ranged from age 18 to 64. For race/ethnicity we can see that the majority of the participants belong to the white/non-hispanic group. We intend to assess if being part of any minority group could be a predictor for psychological distress. With the education variable we can see that the majority of the participants had at least a high school degree.  We have about an even distribution in the gender variable. The distribution of the income variable shows that the income of the participants ranged from below $5000 to above $250000. There is a fair representation from the four regions in the US.

Overall, we can see that our predictor variables have a fair representation in each category. This will allow us to identify if there are any specific predictors playing a role in determining the risk for psychological distress in the year 2020. With these visualizations we can have a general idea of the composition of our study group and could also help us narrow down the predictors we could use for our model.

**Useful methods in answering your question(s).**

We will compute the proportion of individuals with high risk of psychological distress across the US and by region. Then we will map out the proportion by region for the various sociodemographic characteristics. We will then write a model that can be used to predict high risk of psychological distress for the general US population and across the different regions. With this data, we will set up a creative dashboard to share our results.

**What results from these specific statistical methods are needed to support your hypothesized answer?**

We would need the proportion of individuals with high risk of psychological distress for the US as well as in the four regions to test the null hypothesis that our identified sociodemographic factors are not associated with high risk of psychological distress. We can then evaluate our results in the overall US sample and stratified by region.

H0: There is no association between sociodemographic factors and risk of psychological distress among non-elderly adults living in the United States in 2020

HA: There is an association between sociodemographic factors and risk of psychological distress among non-elderly adults living in the United States in 2020

**References**

[Donker, T., Comijs, H., Cuijpers, P., Terluin, B., Nolen, W., Zitman, F., & Penninx, B. (2010). The validity of the Dutch K10 and extended K10 screening scales for depressive and anxiety disorders. *Psychiatry Research*, *176*(1), 45–50. https://doi.org/10.1016/j.psychres.2009.01.012](https://www.zotero.org/google-docs/?iaVppx)

[Hilton, M. F., & Whiteford, H. A. (2010). Interacting with the public as a risk factor for employee psychological distress. *BMC Public Health*, *10*(1), 435. https://doi.org/10.1186/1471-2458-10-435](https://www.zotero.org/google-docs/?iaVppx)

[Hozawa, A., Kuriyama, S., Nakaya, N., Ohmori-Matsuda, K., Kakizaki, M., Sone, T., Nagai, M., Sugawara, Y., Nitta, A., Tomata, Y., Niu, K., & Tsuji, I. (2009). Green tea consumption is associated with lower psychological distress in a general population: The Ohsaki Cohort 2006 Study. *The American Journal of Clinical Nutrition*, *90*(5), 1390–1396. https://doi.org/10.3945/ajcn.2009.28214](https://www.zotero.org/google-docs/?iaVppx)

[Zuckerman, S. (2020). *Well-Being and Basic Needs Survey, United States, 2017: Archival Version* (Version v0) [Data set]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR37513](https://www.zotero.org/google-docs/?iaVppx)

[Zuckerman, S. (2022). *Well-Being and Basic Needs Survey, United States, 2020: Version 1* (Version v1) [Data set]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR38388.V1](https://www.zotero.org/google-docs/?iaVppx)