Rosie Martinez
Capstone 1 Proposal

Background:
Human infertility is a complex disorder that is becoming more prevalent. In 2010, an estimated 48.5 million couples worldwide were unable to have a child after five years of trying to conceive. Worldwide 1 in 6 couples have troubles getting pregnant or sustaining their pregnancy and approximately 6.1 million American women struggle with issues of infertility. According to the National Center for Health Statistics, in the United States 12.1% of women aged 15-44 have impaired fecundity (the ability to have kids) and 6.7% of married women in the same age range are infertile.

Infertility is influenced by a broad range of physical, anatomical, hormonal, genetic and environmental stressors. About 1/3 of infertility is attributed to female issues, 1/3 is attributed to male issues, and 1/3 is attributed to unknown factors. The burden of infertility high and remains an ongoing global reproductive health issue.

The clinical diagnosis of infertility is defined as the failure to conceive within 12 months and affects 7% to 8% of reproductive-aged American women. The American Society for Reproductive Medicine recommends that a woman should consult her physician if she is under 35 years of age and has been trying to conceive for more than 12 months or over 35 years of age and has been trying for 6 or more months.

Question: Can I predict infertility among women based on self-reported risk factor data?

Description of the data:
While we infertility is not only a "female" issue, for the purposes of this capstone, I will be focusing on infertility among women. To do this, I will be using the NHANES data, the National Health and Nutrition Examination Survey from 2015-2016 (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015)
In this data, there are many different datasets based on questionnaires given, but I will be focusing my project on those questions dealing with the known risk factors of infertility (**Shape: 9971, 35**):

- Age (Question ID (QID): RIDAGEYR)
- Race/Ethnicity (QID: RIDRETH3)
- BMI (body mass index) (QIDs: WHD010, WHD020) – will calculate BMI from height and weight
- Pre-existing hormonal (QID: MCQ160m) or other diseases (some STIs (QIDs: SXQ260, SXQ265, SXQ270, SXQ272, SXQ753), Pelvic Inflammatory Disease (QID: RHQ078)) – might make an indicator of STIs, but unsure
- Physical activity (QID: PAQ605, PAQ610, PAD615, PAQ620, PAQ625, PAD630, PAQ635, PAD645, PAD660, PAD675, PAQ706) – I will be making a physical activity indicator based on the CDC's recommendation of daily exercise
- Alcohol consumption (QID: ALQ130, ALQ141U) – I will be making an indicator of alcohol consumption
- Smoking (QIDs: SMQ020, SMQ040) – will make an indicator of smoking
- Irregular periods (QIDS: RHQ031)

I will have to exclude women who have had a hysterectomy (QID: RHD280) or had their ovaries removed (QID: RHQ305) as this will inherently make them infertile. I will also be excluding those women who have had uterine (MCQ240cc), cervical (MCQ240f), or ovarian cancer (MCQ240s), as these cancers can lead to a hysterectomy or ovary removal. (**Shape: 9387, 35**). After I dropped those observations, I dropped these columns to get a shape of (**Shape: 9387, 30**).

How do I define infertility in the NHANES data: In the NHANES dataset, there is a question asked "Have you/spouse ever attempted to become pregnant over a period of at least a year without becoming pregnant?" (QID: RHQ074) Additionally they have another question "Have you or your spouse ever been to a doctor or other medical provider because you or she has been unable to become pregnant?" (QID: RHQ076)
I will define those participants that said "Yes" to one of those two questions as infertile and those that said "No" as fertile. The final shape after inputting this new outcome variable and removing the missing from these two variables the final shape is (**Shape: 1581, 31**).

Minimum Viable Product (MVP):

- **MVP:** To run cross-validation logistic regression on the NHANES 2015-2016 dataset.
- **MVP+:** To run other classification regularization techniques to see if another model can be used to better predict my outcome
- **MVP++:** To merge data from prior NHANES dataset years to increase sample size, where my predictors are the same (could include these datasets: 2013-2014, 2011-2012, 2009, 2010, 2007-2008, 2005-2006, 2003-2004, 2001-2002, 1999-2000). Examine same analyses (CV-logistic regression) on a larger sample size to see if my prediction model gets better.
- **MVP+++:** To use other classification regularization techniques to see if the model can be better predicted on the merged full dataset.