SAMPLING PLAN

1. Sampling Strategy

**Objectives and Reliability Requirements**

Gender inequality is a type discrimination against a sex, resulting in a particular individual being treated disadvantageously because of their sex. It is something that has deeply afflicted women in work environments since time immemorial and still persists in spite of all the leaps and bounds various women’s rights movements have made. The main objective of this research is to decipher whether or not there is inequality in the labor force, if a gender gap exists and if so is it stronger in certain industries than in others. Because the data provided contains over 500 participants, I will need to sample the data thus the need for creating a sampling plan.

**Target Population**

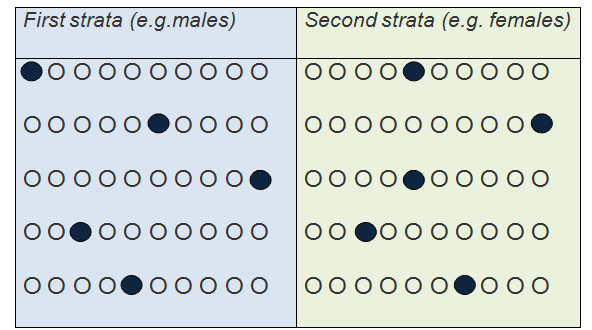
This dataset was retrieved from the Bureau of Labor Statistics and downloaded from Kaggle. It highlights the weekly median incomes of 535 different occupations in the United States of America therefore encompassing information for all working American citizens as of January 2015. The different incomes are broken down into male and female statistics, preceded by the total median income when including both genders.

The features of the target population are as seen below:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Occupation | Job title as given by the Bureau. Industry summaries are given in ALL CAPS. |
| All\_workers | Number of workers male and female, in thousands. |
| All\_weekly | Median weekly income including male and female workers, in USD. |
| M\_workers | Number of male workers, in thousands. |
| M\_weekly | Median weekly income for male workers, in USD. |
| F\_workers | Number of female workers, in thousands. |
| F\_weekly | Median weekly income for female workers, in USD. |

**Sampling Method**

The sample approach to be carried out will be the ***stratified sampling method*** where the entire heterogeneous population is divided into multiple,non overlapping,homogenous groups/strata and then proceed to randomly choose samples from the various strata. Because gender is the deciding factor in this research, the sampling will be done in accordance to gender



**Sampling Size**

The total population of the data stands at 538, and because it isn’t too large there won’t be any wastage of resources throughout the analysis. To obtain the sample size we will take 10% of the total population, thus giving us 54 observations to work with.

**Sampling Frame**

This is where I go ahead and clarify all units of the population of interest by creating a list of said units.

|  |
| --- |
| Occupation type |
| All workers |
| Median weekly income for both male and female |
| Male workers |
| Median weekly income for male workers |
| Female workers |
| Median weekly income for female workers |

**Sample Representativeness**

Once we have obtained our strata, there will be a random selection of the sample where the male and female ratio are equal thus ensuring good representation of the population.

1. Data

**Field Measurements**

The variables to be measured are seen below:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Occupation | Job title as given by the Bureau. Industry summaries are given in ALL CAPS. |
| All\_workers | Number of workers male and female, in thousands. |
| All\_weekly | Median weekly income including male and female workers, in USD. |
| M\_workers | Number of male workers, in thousands. |
| M\_weekly | Median weekly income for male workers, in USD. |
| F\_workers | Number of female workers, in thousands. |
| F\_weekly | Median weekly income for female workers, in USD. |

**Quality Assurance / Quality Control**

For me to set the seal on the high quality of the data to be used, I was sure to use data that was collected in a professional manner by the Bureau of Labor Statistics where all the essential features are properly captured by the data. Any missing data, anomalies and outliers will be handled accordingly before any analysis is performed on the data.

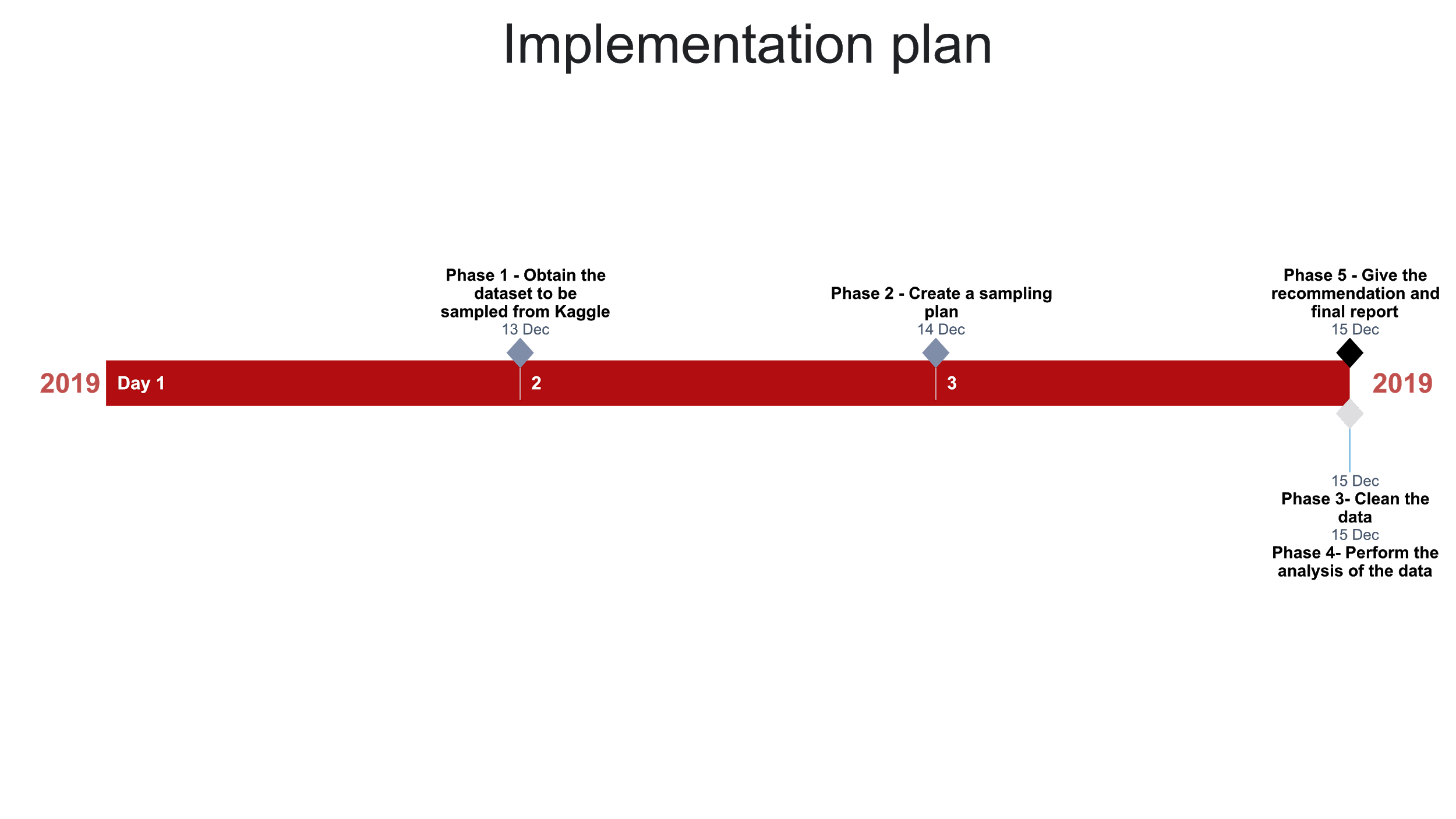
**Analysis**

Once the missing data, anomalies and outliers have been dealt with, analysis will be the next step so as to determine whether gender inequality is still prevalent in the labor force. Methods of analysis to be used are the bivariate and multivariate analysis.

1. Implementation

**Implementation Plan**

I will give an overview of the timeline for the research



Below is a kaggle link from where I sourced the dataset.

**Source:** [**https://www.kaggle.com/jonavery/incomes-by-career-and-gender**](https://www.kaggle.com/jonavery/incomes-by-career-and-gender)