DATA 2010 Group Project

Lucas Berzuk, Ethan Robson, Hugo Ng, Rodell Salonga

Due on 015/03/2024

```
# No one touch this IMPORTANT Before being able to run this code you have to click on the
# .csv In the repository, click import dataset, and then import
ObesityDataset = read.csv("ObesityDataSet_raw_and_data_sinthetic.csv")
```

This data comes from a study performed in Mexico, Peru and Colombia. It has 17 attributes and 2111 data points. It is the study of how peoples eating habits and their physical condition has an effect on their level of obesity. It is important to not that up to 77% of this data has been synthetically generated because of a balancing issue so this is not all 100% real data. There are many variables that have unclear descriptions as headers so here is a list for clarity. Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS) This is where we are currently at with analyzing the data.

```
library(ggplot2)
```

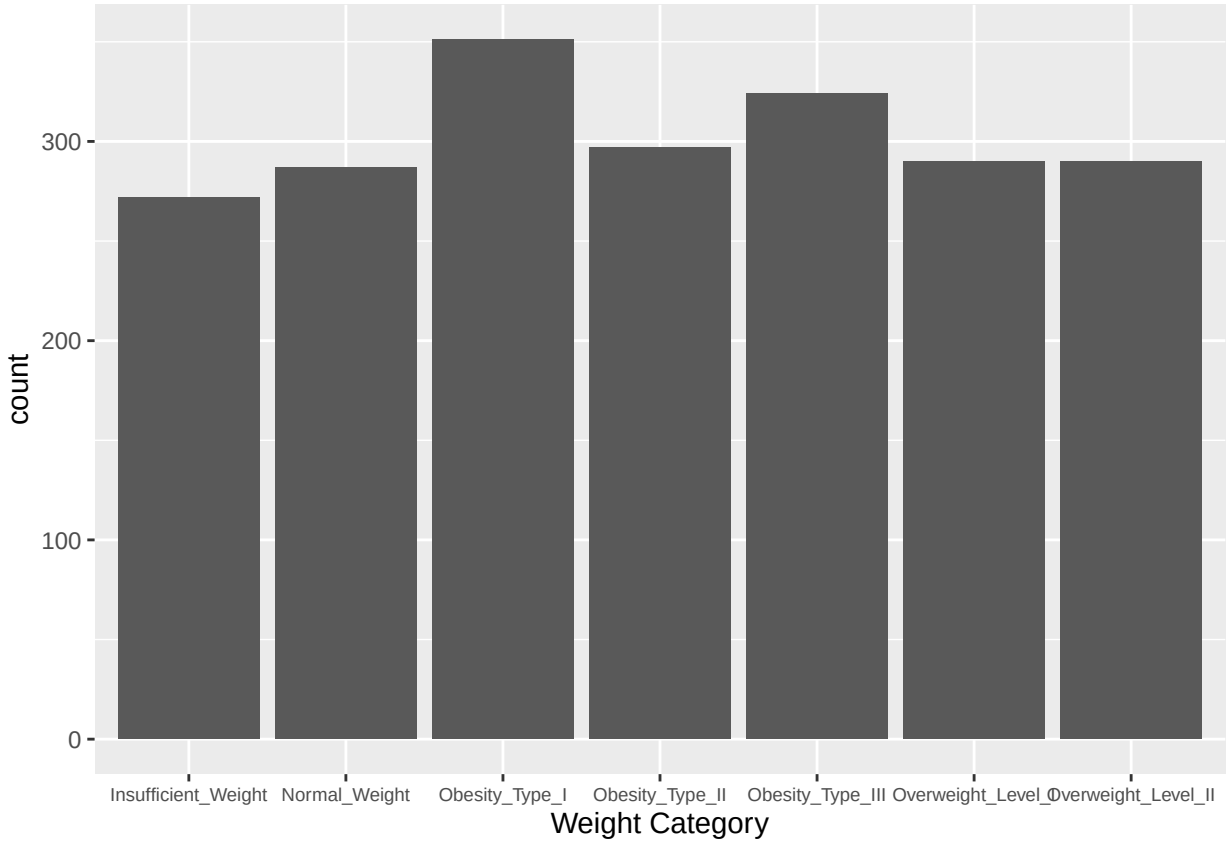
```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

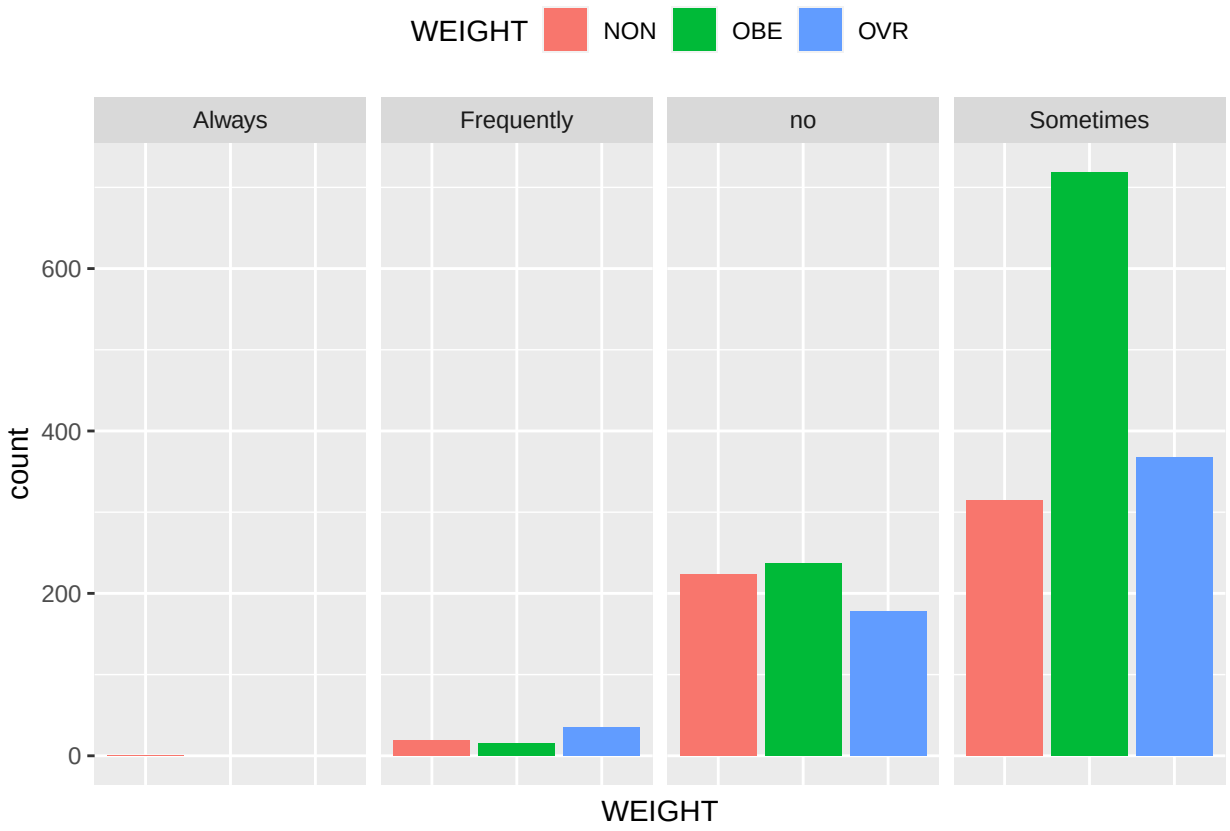
```r
# A Plot of how many people analyzed were in each weight category
ObesityDataset |>
    ggplot(aes(x = NObeyesdad)) + geom_bar() + theme(axis.text.x = element_text(size = 7)) +
    xlab("Weight Category")
```




```r
# A more generalized plot combining all obesity levels into one and all overweight levels
# into one.
mapping = function(x) {
    new_vars = c()
    for (i in 1:length(x)) {
        if (x[i] == "Insufficient_Weight" || x[i] == "Normal_Weight") {
            new_vars = c(new_vars, "NON")
        } else if (x[i] == "Overweight_Level_I" || x[i] == "Overweight_Level_II") {
            new_vars = c(new_vars, "OVR")
        } else {
            new_vars = c(new_vars, "OBE")
        }
    }
    return(new_vars)
}
```

```r
NewObesityDataset = ObesityDataset |>
    mutate(NObeyesdad = factor(NObeyesdad, levels = c("Insufficient_Weight", "Normal_Weight",
        "Overweight_Level_I", "Overweight_Level_II", "Obesity_Type_I", "Obesity_Type_II", "Obesity_Type_
        CALC = factor(CALC), WEIGHT = mapping(NObeyesdad))

NewObesityDataset |>
    ggplot(aes(x = WEIGHT, fill = WEIGHT)) + geom_bar() + theme(legend.position = "top", axis.text.x =
    axis.ticks.x = element_blank()) + facet_grid(~CALC)
```



The direction that we want to go with this is... (we need to come to a conclusion here