DATA 2010 Group Project

Lucas Berzuk, Ethan Robson, Hugo Ng, Rodell Salonga

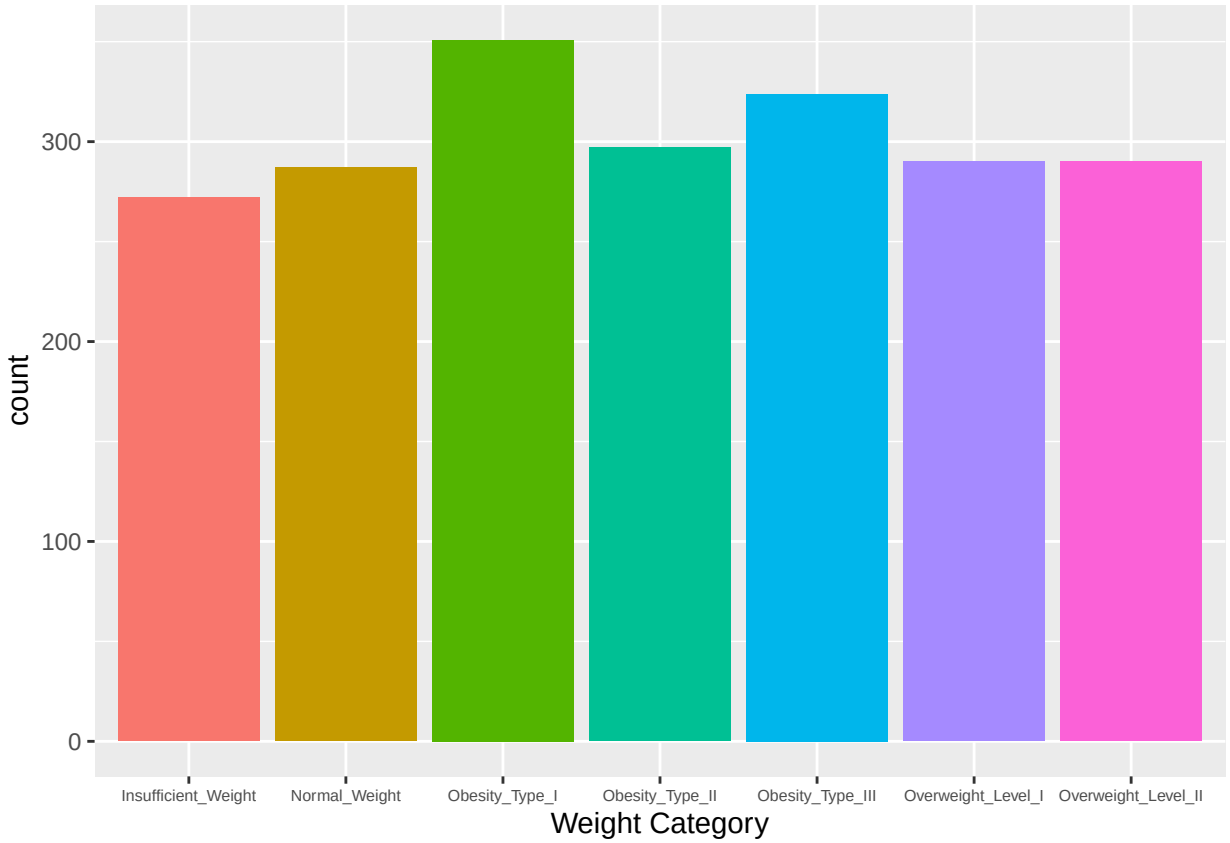
Due on 015/03/2024

```
library(ggplot2)
```

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

```
ObesityDataset = read.csv("ObesityDataSet_raw_and_data_sinthetic.csv")

# A Plot of how many people analyzed were in each weight category
ObesityDataset |>
    ggplot(aes(x = NObeyesdad, fill = NObeyesdad)) + geom_bar() + theme(axis.text.x = element_text(size
    legend.position = "none") + xlab("Weight Category")
```



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 people's eating habits and their physical condition have an effect on their level of obesity. It is important to note that up to 77% of this data has been synthetically generated because of greatly unbalanced number of samples between the different weight categories in the sample data. Therefore the data set is not completely real data. This was an online survey that was 16 questions with a variety of responses. It was accessible for 30 days for users to complete. The researchers use the equation Mass Body Index = Weight / (Height^2) to come up with the mass body index of every individual and then they compared it to data that was provided by the WHO to come up with these weight classes: ***Underweight Less than 18.5, Normal 18.5 to 24.9, Overweight 25.0 to 29.9, Obesity I 30.0 to 34.9, Obesity II 35.0 to 39.9 and Obesity III Higher than 40****.*

At this stage in our data analysis, we are observing how many people fall under each weight category and what the possible reasons are. The count of samples in each weight category is shown to be relatively even based on the plot shown above, so we know that there are many data points in each weight category that we can analyze to figure out possible reasons that people are overweight (or at least what factors correlate to being overweight). What we have been able to find out thus far (based on another plot that has not been included here) is that the amount of people that 'sometimes' drink alcohol are always the majority in every weight category, with the people that answered 'no' always being the second highest answer. Minimal people have answered 'frequently' and 'always', but we believe that this data may be skewed as a result of the data being collected voluntarily by members of the public who may not feel comfortable disclosing an alcohol problem which may be seen as a negative trait. We also noticed that in the 'Overweight_Level_I', 'Obesity_Type_II' and 'Obesity_Type_III' Category, the number of people that answer 'sometimes' for alcohol is significantly higher than the number of people that answer 'no', compared to the rest of the categories.

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