Capstone Project 1- Milestone Report

Introduction

This project investigates the alcohol consumption of college students. The potential client for this project is college level administrators and working groups aimed at identifying which students are at risk of alcohol abuse in college so that they can provide interventions to help these students. Further, by identifying and providing these students interventions the college may be able to stop alcohol related deaths on their campus.

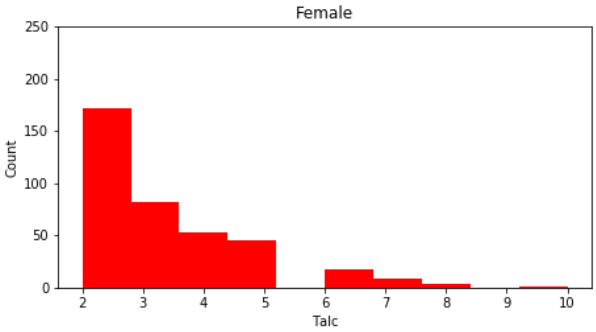
Dataset

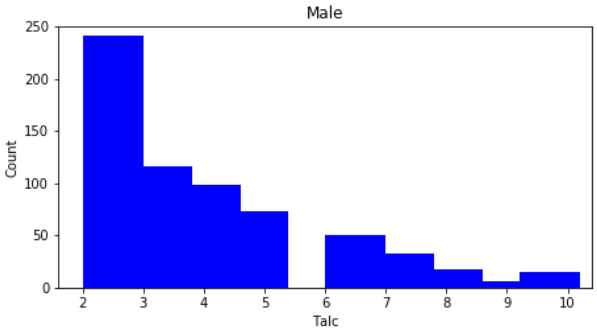
The dataset used in this project was the UCI Machine Learning Student Alcohol Consumption dataset located on Kaggle (https://www.kaggle.com/uciml/student-alcohol-consumption/data). This dataset contains 33 variables related to 649 students who were registered in a math Portuguese language course at one college. The variables contained in this dataset relate to the student's sex and age as well as information about their family (e.g., family size, parent's education level, parent's job), school related variables (e.g, travel time to school, studytime, and past course failures), student's social life (e.g., if they are in a romantic relationships, time going out with friends, and free time), alcohol consumption (both workday and weekday), and academic information (e.g., absences from school and grades from first period, 2nd period and final grade in the Portuguese class). The workday and weekend alcohol consumption variable are numeric ratings from 1(very low) to 5 (very high). This dataset did not have any missing values for any of the variables. I created a new composite variable of total alcohol consumption by adding the value of the workday and weekend variables together. Additionally I created dummy variables for several of the categorical variables in the dataset such as sex, address (urban or rural), family size, mother's education, and father's job. No other data wrangling or data cleaning was needed.

Potential other datasets could be internal student information systems/databases a particular college has on their own students.

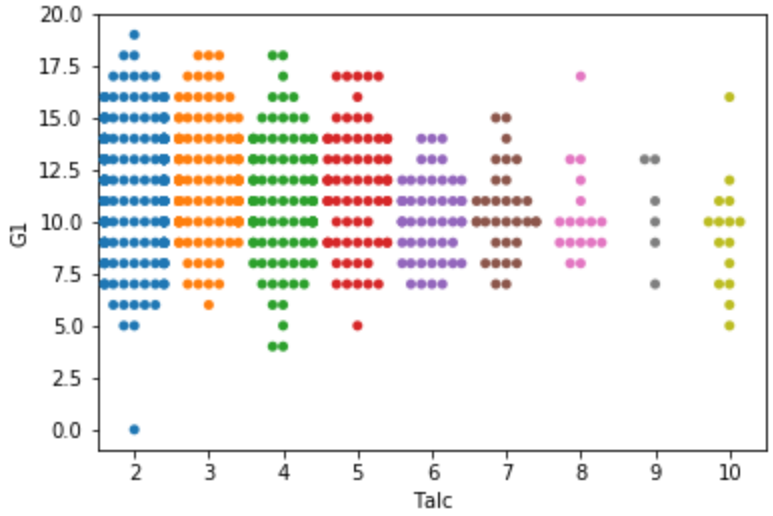
Initial Findings

Based on descriptive statistics, exploratory data analysis, and correlations male students have a higher mean total alcohol consumption that female students.





The average total alcohol consumption was 3.78 for all students regardless of sex. The average age of students in the dataset was 16.74. There was a moderate and positive correlation between workday and weekend alcohol consumption (pearson r= .62) which was statistically significant. The amount of time students go out with friends was weakly and positively correlated with the total alcohol consumption of students (person r=.36). Previous class failures had a weak and negative correlation to grades in the course (pearson r's ranged from -.38 to -.39). Workday and weekend alcohol consumption had strong and positive correlations to total alcohol consumption (.86 and .93 respectively).



According to my swarmplot there are many more students who have a low total alcohol consumption than those who have higher levels of total alcohol consumption. Finally, grades from first period, second period, and final grades were all strongly and positively correlated with pearson r values ranging from .82 to .92.