Aim of this work

This data analysis work has been made for my application as a PhD student for the ExpBoD.

Its point is to show how I compute and interpret data, while giving a sample of redaction in English.

This doesn’t follow the usual research paper structure. The introduction is minimal, and results and discussion are mixed up. I tried to be as concise as possible while still showing the relevant insight we could extract from the given database.

This work could have been more exhaustive, but I didn’t want it to be too long so that it could be at least partially read during an application selection process, I wanted to finish this work before the 1st of December deadline, and I also didn’t want to be too elusive on the findings that were already at reach without digging for hours.

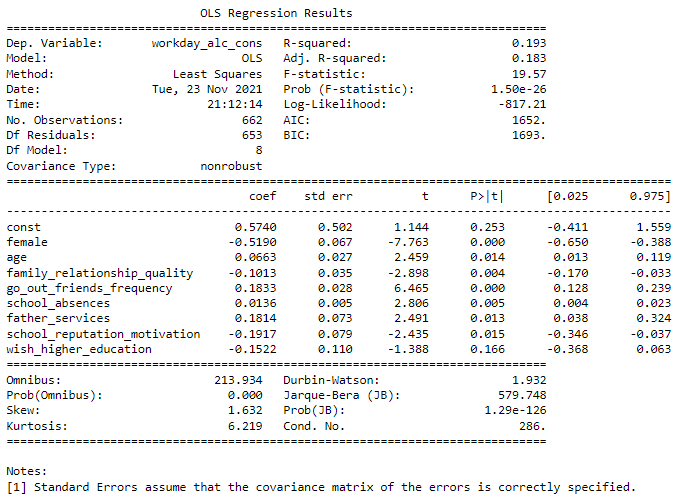
First overview

Alcohol consumption has many negative consequences which can affect the consumer (Birhanu, Bisetegn, & Woldeyohannes, 2014), and more indirectly society (Shimelis & Wosen, 2015). Highschool is one of the most important periods to pay attention to when studying alcohol consumption, as it is associated with both increasing of alcohol usage, and more health impact on short and long-term (Kounnavong, Vonglokham & Okumura, 2021).

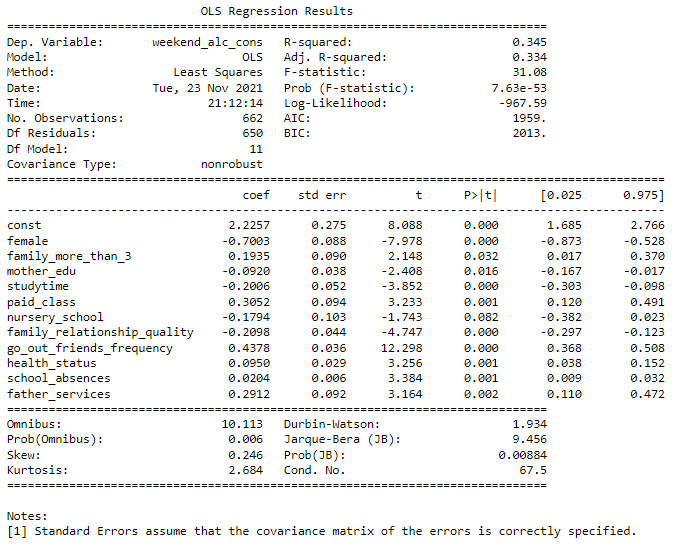
Cortez & Silva (2008) shared a csv-format database that consists in answers a self-reported answers to various questions, and school-related data such as grades or school absences. There are 662 participants for this study, from which 33 measures were gathered including weekend and workday alcohol consumption frequencies (rated with a 5-point likert scale). The results could be biased because of social desirability, especially with alcohol consumption-related questions in a school environment, and authors didn’t share their questionnaire, making difficult some results interpretation. Still, with its various variables and decent number of participants, it can be used to unveil potential links between alcohol consumption and the other featured variables for a small project like this one which doesn’t aim for any publication. It is open source and is available here (https://www.kaggle.com/uciml/student-alcohol-consumption). In fact, it has already been used for alcohol consumption prediction in a previous study (Pagnotta & Amran, 2016), even though they seemed to have had access to a previous and “unclean” version of the data. There are many problems with the method used in this paper. Some choices can be debated (eliminating variables that aren’t directly correlated with the target, which can lead to miss any indirect effect, merging workday and weekend consumption even though they don’t have the same distributions and correlations with many variables, turning continuous into booleans and losing information), but the main problem is the metric used to train their model. They tested several classifiers and optimized them so that they aim for the best accuracy possible. They managed to get 92% accuracy. This number can seem great, but they forgot to include the recall or any confusion matrix. The whole point of the study is to see what can predict alcohol usage behaviour, so it seems more important to have an algorithm that is able to predict problematic alcohol usage even if it gets wrong on non-drinkers (recall-oriented optimization) than the opposite, which would be getting right on non-drinkers, but not being able to predict correctly if someone will be a drinker or not (precision-oriented optimization). I replicated the computing of their target variable (available at the end of my notebook), and the problem (for the data, not the students) is that almost 78 % of the students belong to the “not a drinker” category, and “only” 22% of them belong to the “drinker” category by applying their method. So, if the 8% of students that weren’t categorized well by the algorithm are drinkers, then their recall is 63%, which is far to be as thrilling as the announced 92% of good predictions. Therefore, the whole result and feature importance interpretation would be flawed. But maybe their recall was almost 100% and their 8% of error was on the precision side, which would be less of a problem for this precise subject, but we can’t know for sure. For all these reasons, I didn’t take this paper into account when working on my analysis. Therefore, I used a different method for the data that is all explained and illustrated in my notebook.

Every step, from the pre-processing to the list of correlations and the testing of different algorithms are detailed in the notebook, so you can check it if you want information about the whole process. To make it short, I tried to train linear regression algorithms for both weekend and workday alcohol consumption frequencies, which have decent performance considering the number of data and variables available. Then, I tried to train several classifiers (logistic regression, decision trees and KNN), to predict if a student could be considered as an heavy drinker or not (more than 2/5 in the workday alcohol consumption frequency scale, and more than 3/5 for the weekend), but none of them could beat a 61% recall for the weekend and 36% recall for the workday alcohol consumption, which I considered as not efficient enough to take their result into account.

Therefore, the two kept model are as follow:



Workday alcohol consumption frequency linear regression results

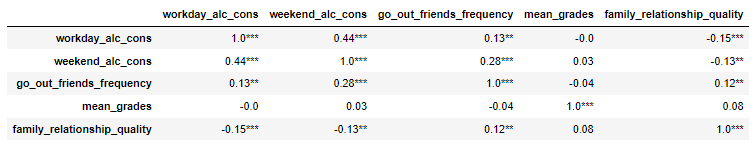


Weekend alcohol consumption frequency linear regression results

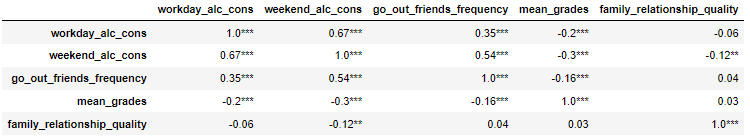
There are many correlations between continuous variables and interesting variance difference for categorical variables for the two kind of alcohol consumption. It would take pages to get into every correlation, but I will outline the most interesting ones while discussing the results of the two linear regressions. But again, if you want the whole detail of every correlation and variance difference, you can check the notebook.

There is a .62 significant correlation between workday and weekend alcohol consumption, so the two are definitely linked. They don’t appear in the regressions as I chose to do the analysis from a “naïve” perspective where the algorithm has no idea of the other alcohol consumption frequency. The algorithm scores way better when we include the other alcohol consumption of course, but it doesn’t give more information than the correlation, as it can be summarized as “people who drink during the week also drink the weekend”, which doesn’t really give any useful insight. Still, the opposite is not true, and students in general drink more during weekend than workday. It is obviously a good thing, but it made workday alcohol consumption frequency harder to predict as the target is less balanced than weekend alcohol consumption frequency, even though it is already unbalanced (there are more students that are not drinking or very little than heavily drinking students). Aside from the prediction quality, coefficients are lower for the workday regression than the weekend. Therefore, while the comparison of the two regressions, it’s important to keep these limitations in mind.

The first thing to notice is that while there are common variables between the two kind of alcohol consumption frequencies, there are also some differences. Moreover, even similar variables don’t always act similarly. For example, going out with friends frequency has a .18 coefficient in the workday alcohol consumption frequency linear regression model, which is important, but not as much as the .44 from the weekend analysis. This effect has already been studied by Lau-Barraco, Braitman, Linden-Carmichael & Stamates (2016), also studying alcohol consumption frequency on “emerging adults”, even though it was on a nonstudent population: “Specifically, with regard to weekday consumption, stronger tension reduction beliefs were associated with more drinks consumed. Social expectancies also predicted weekday drinking but at slightly lower odds. For weekend consumption, drinking was higher than weekday drinking. Interestingly, in contrast to weekday drinking, social expectancies alone predicted stronger weekend increases while tension reduction expectancies did not”. It is difficult to outline tension reduction beliefs from the available data, but we can see social pressure played a greater role during weekend than workdays on alcohol consumption frequency. Another example is gender. Female students tend to drink less often than their male counterpart in general, which is a well-known effect (Gohari, Cook, Dubin & Leatherdale, 2019). But surprisingly, this effect is stronger during weekend, which can have two interpretations. First, it could be female students simply drinking less during weekend, which is the opposite of Braitman, Linden-Carmichael & Stamates (2016) findings, or it could be male students that would be more affected by other variables such as pair pressure, which goes with Schulte, Ramo & Brown (2009) findings. They outlined in their article the “greater social sanctions against drinking” female students could experience compared to male students. A simple correlation analysis comparison shows that both male and female experience a correlation twice as strong between going out with friends and weekend alcohol consumption compared to workday alcohol consumption. Still, the correlation is almost three times stronger with men during workdays, and two times stronger during weekend compared to female students. Another indirect effect is about alcohol use effect on grades, which differs based upon gender. Balsa, Giuliano and French (2010) already documented this effect : “We find that increases in alcohol consumption result in small yet statistically significant reductions in GPA for male students and in statistically non-significant changes for females". However, this effect seems to be even stronger with our dataset, as correlations between workday and weekend alcohol consumption are significant at -.2 and -.3 among male students, which is rather high, and approaches 0 among female students. Gender-specific analysis must be done in order to understand how it affects alcohol use. For example, along with previous studies’ results (Rukundo, Ayebare & David, 2020 ; Choo & Shek, 2013), results show that family relationship quality’s effect on alcohol use and the frequency of going out with friends were gender-dependant. Even though family relationship quality was difficult to interpret as explained later, there seem to be very interesting findings to unravel with gender-specific variability, but I sadly didn’t have the time to deepen the subject while making it on time for the 1st of December deadline.



Correlations of alcohol consumption frequency going out with friends frequency, grades and family relationship quality among female students



Correlations of alcohol consumption frequency, going out with friends frequency, grades and family relationship quality among male students

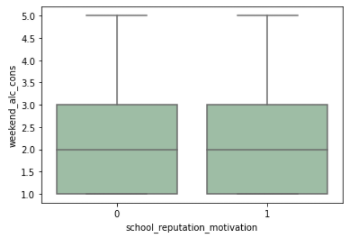
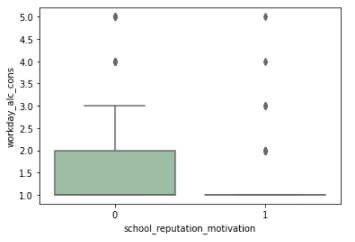
The quality of the relationship with the family is affecting alcohol consumption negatively, and even more on the week-end than during workdays. Still, the difference is rather small and could be only due to the difference between the two model qualities. The effect of family on alcohol use among adolescents is well-studied. Still, it is highly dependent of the family structure, alcohol use in the family, the type of parenthood and the culture (Kounnavong, Vonglokham, Moji & Okumura, 2020 ; Choo & Shek, 2013 ; Kelly et al., 2011). Therefore, it is difficult to interpret further on this matter with a simple 5-point likert scale on the vague subject of “family relationship quality”, which could indistinctly refer to trust, availability, discipline, violence and so on with any member of the family. In the same way, it is difficult to interpret why having a father working in services increases both alcohol use frequencies without any more precision about the job in itself (title, status, salary and so on), especially as it doesn’t seem to be correlated with variables directly linked with both alcohol consumption frequencies (see annexes at the end of the notebook).

Age is known to be positively correlated with alcohol consumptions among adolescents (Behrad & Jalilvand, 2019). Still, in this precise dataset, the strength of the correlation is higher during workday (.13) than weekend (.07), and had a lower p (respectively p>.01 and p>.1), which is probably partly why it doesn’t appear in the weekend alcohol consumption’s linear regression. Another explanation (that isn’t contradictory) is that the “age” variable is an indirect measure of academic difficulties through its negative correlations with parents’ education and mean grades, and positive correlations with previous failures at finals and school absences. It would explain why we don’t find mean grades as an important variable for the linear regression, as age would mathematically be a more “reliable” variable of academic success/failure. If this interpretation is correct, it would also mean that academic success would be more related with workday alcohol consumption than weekend alcohol consumption. However, as far as I know, there are studies comparing weekend and workday alcohol consumption differencies (Lau-Barraco, Braitman, Linden-Carmichael & Stamates, 2016), but none specifically including grades comparison. Therefore, other studies should be found or made on this subject in order to confirm or infirm this statement, especially as it is incoherent with the fact that school absences are considered important for both workday and weekend alcohol use (which has already been studied by Wesley (2012).



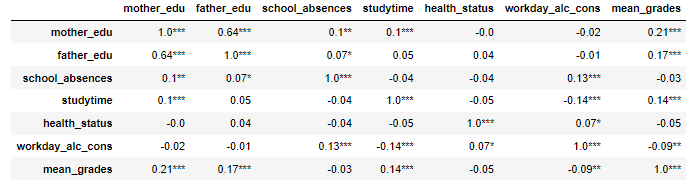
Main correlations with age

In the same way, choosing the school for its reputation decreases workday alcohol consumption in this dataset, but doesn’t seem to affect weekend alcohol consumption.



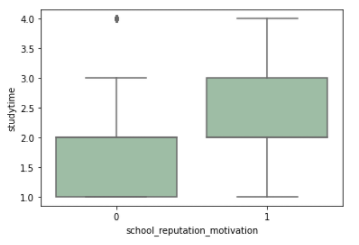
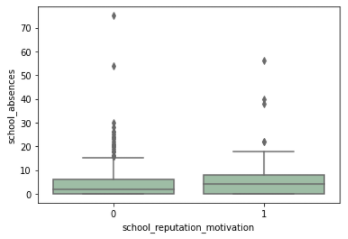
School reputation effect on both alcohol consumptions

It can be an indirect sign of parental academic success pressure and monitoring, but there is no other data to corroborate this. Still, it seems to be linked to better parents’ education level, longer study time, better health status, and better grades (see annexes at the end of the notebook). It’s also associated with slightly more school absences, which can be considered strange as it is itself correlated with more workday alcohol consumption, but looking at a correlation table with all of the previously mentioned variables show that there is a balance between positive and negative coefficient among them.



Correlations between variables affected by school reputation-based choice of establishment

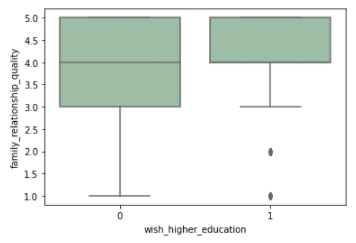
Moreover, the variability between dimensions doesn’t have the same strength among these variables. Variables positively correlated with workday alcohol consumption are less affected by school reputation-based choice than negatively correlated ones.

Impact of reputation-based choice of establishment on study time and school absences

Moreover, half of them is positively correlated with mean grades, two are not correlated, and workday alcohol consumption is negatively correlated with it. Therefore, based on the available data, the choice of school based on reputation can be considered as an indirect sign of academic success.

The wish for higher education doesn’t seem to be linked with workday alcohol consumption, but seems to be with weekend alcohol consumption (see annexes at the end of the notebook), which can be considered odd as it is only considered important in the workday alcohol consumption linear regression. Still, in the same way as the previous examples, it is positively linked with indirect signs of academic success (both parents education level, mean grades), and negatively with indirect signs of academic issues (age, previous failures, going out with friends’ frequency). It is also interesting to see that it seems strongly connected with family relationship quality, which is negatively correlated with both alcohol consumptions.

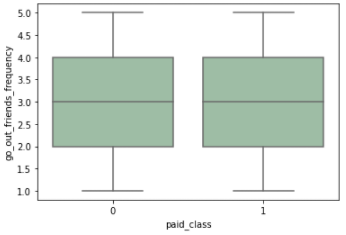
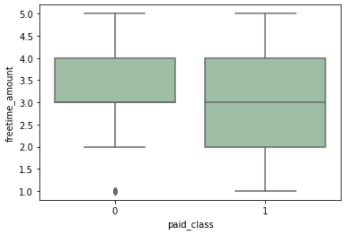


Interaction between family relationship quality and wish for higher education

It may be because higher education often means being dependant of their family longer (need of guarantors for student loan or renting, resume living within the family home).

From these three last findings, it seems that academic success in negatively correlated with workday alcohol consumption, but isn’t correlated with weekend alcohol consumption. There are several possible interpretations that are not exclusive. The most obvious is that courses are to attend during the week, and alcohol use effects (during the day, or even the night before if the student is in hangover) could directly affect the student ability to focus on a course or to perform at a test. But more indirectly, drinking during workdays is an indirect sign of heavy drinking during the weekend. Only 6 of the 80 students that are heavy drinkers during workdays rarely or don’t use alcohol during the weekend, and their consumption mean during workdays is really high (4 on a 5 point likert scale). The remaining 74 have a mean consumption of 3.6 during workdays, and 4.3 during the weekend, which is more than students that are only heavily drinking during the weekend (3.5 for weekend consumption frequency for these 186 students). Therefore, workday alcohol consumption is also an indirect sign of heavy alcohol consumption in general, which seems to have a stronger effect on academic performance than “only” drinking during the weekend and staying sober during the week.

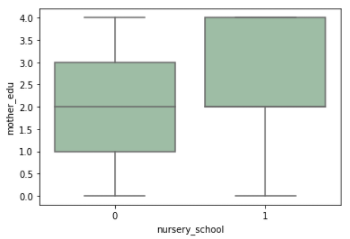
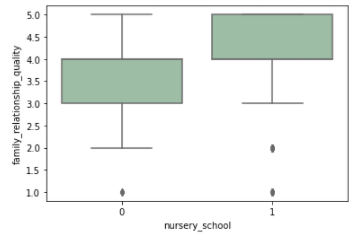
Variables kept only for weekend alcohol consumption frequency linear regression are more numerous, but most of them are whether difficult to interpret because of the lack of information surrounding these variables, or doesn’t give much relevant information. For example, having a family with more than 3 has a positive coefficient in the prediction. Sibling role on adolescent’s alcohol use has already been studied and seem significant, but can have opposite effect depending on their relationship (Low, Shortt, & Snyder, 2012). It also doesn’t seem to affect other continuous variables. Therefore, it’s difficult to interpret the reason why this variable has this effect without more precision about the type of relationship these students share with their siblings. The study time variable negative coefficient can simply mean study time mainly takes place during the week-end, and as alcohol in the week-end is highly linked with going out with friends (even more than workday alcohol consumption), it could simply lead to less time invested in studying (-.08 significant correlation between study time and going out with friends). The paid class variable is also difficult to interpret, because there is no given information on when these courses take place, their cost, their length and so on. If they take place during the weekend, along with study time, it would just mean taking these courses gives less time for free time (as shown is the figure below) and possibly drink. Still, it doesn’t seem to be linked with going out with friends’ frequency, which can be contradictory.



Link between paid class, going out with friends frequency and freetime

Health status is the less explicit variable, and is only correlated with weekend alcohol consumption and family relationship status. Still, Shields and Shooshtari (2001) showed people drinking weekly were less likely to self-report low health status, and women specifically tend to self-report a good health status when doing so, while other factors such as direct physical and psychological health issues had a greater impact on self-reporting low health status. Still, the population wasn’t the same (45-74 years old participants), and their data came from a dataset that is between 22 and 27 years old. Further literature review and/or studies are needed to know if this can apply to high school students nowadays as our dataset result seem to confirm.

Mother’s education level is also considered important for predicting weekend alcohol consumption. Father education is not included in this regression as its effect was not significant, but whereas mother education has a negative and mildly strong coefficient of -.12, father’s education has a low positive coefficient of .06. Chalfin, A. and Monica D. (*2018*) studied parents’ education’s effect on their children alcohol consumption and found it was negatively correlated, but didn’t find a difference between father and mother influence on this precise effect. According to this paper from 2018, there were very few studies on this subject, and it could be interesting to deepen the knowledge of this effect, and especially its link with culture and parental roles. In the same way, attending nursery school have a negative coefficient, but it may be because its distribution seems linked with family relationship quality and mother’s education. There are papers studying long-term effect on attending nursery school (Goodman é Sianesi, 2005), but I couldn’t find any that had relevant insight that could explain the direct impact of this variable, even indirectly.

Austin, W. A. (2012). The Effects of Alcohol Use on High School Absenteeism. The American Economist, 57(2), 238–252. <https://doi.org/10.1177/056943451205700208>

Balsa, A. I., Giuliano, L. M., & French, M. T. (2011). The effects of alcohol use on academic achievement in high school. *Economics of education review*, *30*(1), 1–15. <https://doi.org/10.1016/j.econedurev.2010.06.015>

Behrad, P., & Jalilvand, M. (2019). Prevalence of alcohol consumption and related factors among students of higher education centers in one of the northeastern cities of Iran. AIMS Public Health. 6. 523-533. 10.3934/publichealth.2019.4.523.

Birhanu AM, Bisetegn TA, Woldeyohannes SM. (2014) High prevalence of substance use and associated factors among high school adolescents in Woreta Town, Northwest Ethiopia: multi-domain factor analysis. *BMC Public Health*. doi: 10.1186/1471-2458-14-1186. PMID: 25410657; PMCID: PMC4289242.

Chalfin, A. & Monica D. (2018). "The Effect of Parental Education on Children's Drug and Alcohol Use." *AEA Papers and Proceedings*, 108: 373-78. DOI: 10.1257/pandp.20181105

Choo, H., & Shek, D. (2013). Quality of Parent–Child Relationship, Family Conflict, Peer Pressure, and Drinking Behaviors of Adolescents in an Asian Context: The Case of Singapore. *Social Indicators Research*, *110*(3), 1141–1157. http://www.jstor.org/stable/24719096

Gohari, M., Cook, R., Dubin, J. & Leatherdale, S. (2019). Identifying patterns of alcohol use among secondary school students in Canada: A multilevel latent class analysis. Addictive Behaviors. 100. 106120. 10.1016/j.addbeh.2019.106120.

Goodman, A., & Sianesi, B. (2005). Early Education and Children’s Outcomes: How Long Do the Impacts Last? *Fiscal Studies*, *26*(4), 513–548. http://www.jstor.org/stable/24440002

Kounnavong, T., Vonglokham, M., Moji, K. & Okumura J. (2021). Factors affecting alcohol drinking behaviour among secondary school students in Vientiane Province, Lao People's Democratic Republic: a cross-sectional study. Int Health. doi: 10.1093/inthealth/ihab047. Epub ahead of print. PMID: 34403469.

Kelly, A. Toumbourou, J. O'Flaherty, M., Patton, G., Homel, R., Connor, J. & Williams, J. (2011). Family Relationship Quality and Early Alcohol Use: Evidence for Gender-Specific Risk Processes. Journal of studies on alcohol and drugs. 72. 399-407. 10.15288/jsad.2011.72.399.

Kounnavong T, Vonglokham M, Moji K, Okumura J. (2021). Factors affecting alcohol drinking behaviour among secondary school students in Vientiane Province, Lao People's Democratic Republic: a cross-sectional study*. Int Health*. doi: 10.1093/inthealth/ihab047. Epub ahead of print. PMID: 34403469.

Lau-Barraco, C., Braitman, A. L., Linden-Carmichael, A. N., & Stamates, A. L. (2016). Differences in weekday versus weekend drinking among nonstudent emerging adults. *Experimental and clinical psychopharmacology*, *24*(2), 100–109. <https://doi.org/10.1037/pha0000068>

Low, S., Shortt, J. W., & Snyder, J. (2012). Sibling influences on adolescent substance use: the role of modeling, collusion, and conflict. *Development and psychopathology*, *24*(1), 287–300. https://doi.org/10.1017/S0954579411000836

Pagnotta, F. & Amran, H. (2016). Using data mining to predict secondary school student alcohol consumption. 10.13140/RG.2.1.1465.8328.

Rukundo, A., Ayebare, D., Grace, M., & Steffens, K. (2020). Family Factors Associated with Consumption of Spirits: A Comparative Gender-Based Study of Ugandan Students in Public Secondary Schools. Education Research International. 2020. 1-6. 10.1155/2020/5432545.

Schulte, M. T., Ramo, D., & Brown, S. A. (2009). Gender differences in factors influencing alcohol use and drinking progression among adolescents. *Clinical psychology review*, *29*(6), 535–547. https://doi.org/10.1016/j.cpr.2009.06.003

Shields, M., & Shooshtari, S. (2001). Determinants of self-perceived health. *Health reports*, *13*(1), 35–52.

Shimelis KT and Wosen K. (2015) Assessment of Causes, Prevalence and Consequences of Alcohol and Drug Abuse among Mekelle University, CSSL 2nd Year Students. *American Journal of Applied Psychology*. 3(3):47-56. doi: 10.12691/ajap-3-3-1