Executive Summary

Overview: With the rise of the video game industry's popularity among the youth, many in academia have speculated as to the efficacy of using the medium for educational purposes- such as Play Forward developed by the play2PREVENT lab at the Yale School of Medicine. The game, which follows a decision making storyline, requires all players to make virtually the same decisions to reach the "epilogue" and complete the game. In this analysis, we chose to study the correlation and general trends between student demographics and Total Time Played, as those who have less knowledge in the subjects of sexual health and substance abuse may take more time navigating to the correct answers and thus the next event in the storyline progression. This empirical measure could also indicate the subject was more invested in the gameplay, or was distracted.

Analysis: The main focus of our analysis was to look for correlation between demographic variables and Total Time Played. We cut our dataset down from the full log to a set of 166 observations (one for each observed player) with player_id, race, gender, school, Total Time played, and wave variables. After running a linear regression, we recognized major outliers in the dataset which were creating an intensely skewed right distribution. These observations were dropped based on a threshold of 1.5*IQR above and below the first and fourth quartile, thus narrowing our sample size from 166 observations to 128. After rerunning the linear regressions between time and age and time and gender (producing a p-value of 0.6587 and 0.4560 respectively), it was concluded that there was no evidence of a correlation between the variables. Multiple regressions including time and age+gender, time and race, time and age+gender+race produced similarly insignificant results. Furthermore, a one-way ANOVA was performed to test for a significant difference in Total Time Played between ages and a similarly insignificant conclusion was found. Most notably from our analysis, we found the greatest significance to come from the variables Wave and School, though Wave was the only characteristic variable to break the 0.05 alpha value. The p-value for the school's significance sat on the precipice of a 0.1 alpha value with a p-value of 0.1009. Despite the lack of statistical significance found, there were interesting trends found in subsequently produced plots of said regressions, which were further discussed in our oral presentation.

Conclusion: Of all of the variables that we selected to study along with the Total Time, the wave variable turned out to be the most significant. The wave represents the time period in which the student was playing the game. Overall it is our belief that characteristics such as age, gender, or race may not play as large of a role as society thinks in terms of predicting behaviors related to this specific gameplay. Our team hopes that our analysis can help contribute to dismantling the stereotypes typically associated with certain groups and push towards improvement in education concerning sexual health and substance abuse for demogprahics where this education is lacking.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	3	1.791e+08	59698084	0.536	0.6587
gender	1	6.238e+07	62377509	0.560	0.4560
race	2	2.845e+08	142226781	1.277	0.2835
school	11	2.004e+09	182175382	1.636	0.1009
wave	15	3.324e+09	221568575	1.990	0.0238 *
Residuals	95	1.058e+10	111350907		

Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				