Data for our graphs was acquired through responses to our survey, which was hosted on Qualtrics. Publicized through our Facebook page, our website, and our print edition, the survey was initially released in late July and responses were collected over the course of August. The survey was closed before classes began, in order to record the unique characteristics of the Class of 2021 before they settled in to campus life.

There were a total of 1141 responses in the initial dataset collected. Of these responses, 551 were dropped if they were not freshmen and had their survey deemed incomplete by the Qualtrics application. This was done to avoid missing values in the resulting data as much as possible, since it cannot be determined whether a missing value is a conscious refusal to give answers, or a result of one terminating their session before completion. Moreover, large amounts of missing values could lead to potentially misleading statistical conclusions. While imputation methods - filling in missing values with mean, or median values - could be an alternative to dropping these incomplete responses, considering the additional unknown bias this could introduce, we deemed it the best course of action to limit the responses used for imputation. While it is also possible that this selection method is biased towards a certain group, namely those with the willingness and patience to answer several dozen questions, our analysis indicated that the resulting 590 responses followed similar distributions in most if not all of the fields. Following similar logic, if a question's response count was too low, we chose not to include it as the sparse results could be misleading and statistically incorrect.

Many responses were categorical in nature, such as degree of religiosity, race, college, social media accounts, and income group. If an inherent order could be determined (such as religiosity and income group), then numeric integer encoding was used - 0 for lowest and 6 for highest for a field with 6 possible choices, for example. For categorical fields with no such inherent order, a simple one-hot encoding was used - if answer for race was Asian, value 1 was assigned to a new Asian column, while 0 for all other race columns. Numeric encoding allowed ordinal categorical values to be used in correlation analysis, regression analysis, and frequency comparisons.

In terms of analysis and choice of visualizations, we focused on simple linear relationships between variables - correlation and counts - to establish which plots would be displayed, and to determine which questions would be used in the visualizations. To facilitate fair comparisons, each field was scaled so that it would be distributed between -1 and 1 and have a mean at 0. This made sure that fields with inherently large numbers or range and fields that are relatively smaller in range and size be considered on equal scale.

For any questions on the specifics of our data analysis methods, please contact cornelldatascience@gmail.com. For any specific questions about the project at large, please contact editor@cornellsun.com.