Explain the dataset and summarize the analysis
by Indri Himawan

With this dataset, my goal is to find the most effective cause of suicide by comparing likelihood of suicide with the following:

A. Income

Method of analysis	Summary of results
Using stacked barplot	The larger the income the less suicide probability. Doesn't tell much more than that. Let's find out if it is statistically significant?
Using Logical Regression	Finding that the logical regression values fluctuates, which suggests that this relationship is probably not statistically significant.
Using permutation test	Finding out that the p-value is too large to be statistically significant. The p-value, in fact, suggests that we do not reject the null hypothesis that income and probability of suicide is independent.

Conclusion: It looks like there can be slight correlation between income and probability of suicide, makes sense too because the more money you have, the more opportunity you would have to access things like hospitals and therapy. However, the logical regression and permutation test suggests that they may be independent.

B. Number of friends

Method of analysis	Summary of results
Using histogram	Seeing that the more friends the less likely a suicide attempt. In fact, the histogram looks like an exponential distribution
Comparing to exponential distribution, using deciles, chi squared test	There is not enough evidence against the null hypothesis that our data did come form an exponential distribution. This can mean that having friends/loved ones can be especially helpful for those in risk of suicide.
Using permutation test while keeping skewness between -1 and 1	Found that even when adjusting the skewness of the data, the relationship between friends is still statistically significant
Using Logical Regression while keeping skewness between -1 and 1	Found that even though subset is statistically significant, the logical regression values can still fluctuate.
CLT	Illustrates Central Limit Theorem

Conclusion: Having friends can exponentially decrease suicide attempts.

C. Whether or not they experience depression

Method of analysis	Summary of results
Using Contingency Table	Found a p-value close to zero, which means reject the null hypothesis that depression and suicide attempt are independent.

Conclusion: Decreasing depression decreases likelihood of suicide. Efforts to reduce suicide also means efforts to reduce depression.

D. Age and Sexuality

Method of analysis	Summary of results
Using ggplot	Finding that the hotspot of people who attempted suicide and that those who are part of the LGBT has higher risk of suicide.

With this project, I think that I deserve these additional points as listed in the project description:

#	Description	Why I should get a point
1	**A data set with lots of columns, allowing comparison of many different variables**	With this dataset, I was able to evaluate the relationship of the probability of suicide with other variables such as income level, age, number of friends, sexuality, whether or not they experience depression. This is helpful in finding most likely cause of suicide so those who experience suicidal thoughts can try help themselves on what to work on.
2	**Appropriate use of R function for a probability distribution other than binomial, normal, or chi-square**	I compared the relationship of probability of suicide with number of friends and I was using R function for exponential distribution. And it turns out the relationship has a high enough chance for it to come from exponential distribution.
3	**Defining and using your own functions**	I defined the function **findPValueWithTrimmedFriends** to figure out the p-value based on a subset of dataset. The subset of the dataset is the set of data whose number of friends is trimmed by certain percentage (because the data was very skewed). This is to figure out if given a smaller skewness, would this data still be statistically significant? **getLogicalRegressionValues** will try to figure out the if given a certain threshold, what does the logical regression results look like. **trimFriendData** gives you the subset of the dataset where a percentage of the data is cut out.
4	**Nicely labeled graphic using ggplot, with good use of color, line styles, etc that tell a convincing story**	I used ggplot to illustrate how being part of the LGBTQ community affects the likelihood of suicide. Honestly I'm not quite sure what "with good use of color, line styles, etc" mean because ggplot already provides that without having to do any modifications on my side.
5	**Appropriate use of novel statistics (trimmed mean, maximum or minimum, skewness, ratios)**	Used trimmed mean, maximum, minimum to determine that there was a skewness and used ratios to find the threshold in which the largest values needed to be cut out from the data to find a p-value that tells us that 2 variables in the subset is statistically significant. (see point #3)

#	Description	Why I should get a point
6	**Calculation and display of logistic regression curve**	Logistic regression calculation and curve is used for the relationship between likelihood of suicide and number of friends as well as likelihood of suicide and number and income.
7	**Appropriate use of covariance or correlation**	cov() is used to find slope for logical regression line to the plot of the data.
8	**Use of theoretical knowledge of chi-square, gamma, or beta distribution**	Usage of chi-square to demonstrate CLT and usage of Pearson's Chi-squared test to find p-value.
9	**Calculation of confidence interval**	This is used when finding the likelihood of suicide based on age.
10	**Use theoretical knowledge of sampling distribution**	Illustrated Central Limit Theorem and used permutation tests.
11	**Team consists of exactly two members (1 or 3 is a possibility)**	I am the only who worked in this project.