*I will not follow this script or memorize it.
Intro: Name, Title
Background:
My project revolves around COVID-19, as many of you know the disease that has caused the pandemic for the past couple years. Many companies have since created vaccines to minimize the spread and symptoms of COVID-19, many of which succeeding. To make these vaccines, companies must go through a long and exhaustive process of testing the effectiveness. In this project, we will recreate 29148 observations for data representing the outcome for the Moderna results. We will be focused on the time until an event, where in this case "time" represents days and event represents a symptomatic COVID case. The number of days after receiving a shot for someone to have symptomatic COVID. As with many tests there will be two testing groups, the placebo group who doesn't receive the vaccine and an experimental group who receive the real vaccine. One problem with this data is censoring, as this represents those who participate in the experiment and receive the first shot, experiment or placebo, but leave the experiment before they experience any event. This is due to various reasons, as they may have decided to no longer wish to participate in the study. (Explain vis 3, X is an event. C, E censored. B at least 12 weeks). The survival function can help answer this question, as the survival function gives the probability a person does "survives"/does not get the event longer than some time. This is always decreasing. (Possibly explain why)

Hazard Function:
One function that is very important and tricky to grasp is the hazard function. Unlike the survival function, which gives a probability, the hazard function gives the potential per unit time for an event to happen, given an individual has survived up to some time (little t). This does not represent a probability and can be a number greater than 1, it is rather the potential within some time interval for an event to happen. One example could be that some person experiences an event 0.6 times a day, or 4.2 times per week, depending on the time measurement. The hazard function is synonymous with the conditional failure rate, for those who may be familiar.

This function is very different from the survival function. As mentioned before the hazard function is NOT a probability, rather a potential in some time interval. The survival function also focuses on not failing, creating the graph based on who survives past some time t. However, the hazard function focuses on how many failures / event experiences one may have within some time interval.

Kap + Cox:
We will focus on the Kaplan Meier Curves and Cox PH Model. These both help produce our conclusion. First, the Kaplan-Meier curve is a non-parametric model as it uses estimations from incomplete observations (censoring). However the Cox PH Model is semiparametric since it has parametric and nonparametric components, baseline hazard unspecified. Parametric model is one whose functional form is specified. Next, the Kaplan-Meier curves represent the survival function, while the Cox PH represents the hazard function. Kaplan-Meier uses the random censoring, which basically means that each subject has a censoring time should be representative of all study participants who were still in the study at time t. Subjects who have

been censored are assumed to have equal failure rate for those who remained in the study and not censored. The Cox PH Model also has this assumption, but also one more which is the PH (Proportional Hazards) Assumption. This assumption basically means that the hazard for one individual is proportional to the hazard of another, must be a constant. One way to check this is to use an R function, which we will use, which tests the proportional hazards assumption.

Finally, both discover different, but similar conclusions. The Kaplan-Meier outputs the graphs for estimated survival curves. Using these curves we can see how different one or another may be, and also observe measurements and trends in the graphs such as drawing a line at the probability 0.5 mark to see when 50% of individuals experience an event. The Cox PH Model outputs the hazard ratio and can be used to test vaccine efficacy, the reduction in cases for one group.

Kaplan-Meier:
We can now use the Kaplan-Meier method to visualize the survival curves separated by placebo and treatment group. We can observe a significant difference between the two, which is the curve for the treatment group represented by the dashed line and placebo by the solid. The treatment group's curve is significantly greater than the placebo groups, as we see an obvious difference between the two. The survival probability is nearly 1 for the treatment group, meaning almost nobody experienced symptomatic Covid-19 during the duration of the study. However, there were more people who experienced the event in the placebo group.

We can use a test known as the Log-Rank test to determine if the curves are statistically different. We will use a null hypothesis that the curves are the same, alternatives the curves are different. We receive a very small p-value of <2.2e-16, so we reject the null hypothesis and conclude the two curves are statistically different.

Now we would like to get a better conclusion and maybe some numeric values for how different these may be, so we transition to the Cox PH model.

Cox PH Model:
We observe the output for the Cox PH Model, which as a reminder helps to calculate the hazard ratio which in turns calculates the vaccine efficacy, using the equation 1 - HR. (Discuss output on screen, hazard ratio 95% CI). We can then use this equation to calculate the vaccine efficacy rate, which turns out to be around 95%. We see a 95% reduction in symptomatic Covid-19 cases for those in the treatment group.

We can see below the test for the model assumption. A null hypothesis is that the hazards are proportional, alternative is they're not proportional. P-value of 0.15, fail to reject null hypothesis.

(Discuss takeaways with remaining time, length depends on time left)

Perhaps it could be helpful to tell people that you tried to recreate the results from the paper of the Moderna vaccine results, and that your mentors simulated the data. And that they did not tell you how the data was simulated so you have a chance to decide how to analyze the data.