Deliverable 4: Final Report
Due week 14, one submission per team
Your final report should document your analysis, communicating your findings in a way that is technically
precise, clear, and persuasive.
The maximum length is 4 pages using standard pdf_document output in RStudio, and including all tables, appendices, and references. This limit is strict.
The exact format of your report is flexible (form follows function), but it should include the following elements.

You can have a fifth page for references only. Reference any statement or source which was not developed by the team

1. An Introduction
Your introduction should present a research question and motivate its importance. It should draw the reader's attention to specific X and Y concepts in a way that makes the reader care about them. After reading the introduction, the reader should be prepared to understand why the models are constructed the way that they are. It is not enough to simply say, "We are looking for product features that enhance product success." Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also a good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.

Take a perspective as to who you are who is your client. Why is your client engaging you. What is your research thesis. Take a position as to x drives changes in Y in what way. Involve at least conceptually your treatment X, output Y. Mention covariates. No fishing statements Propose an initial causal path diagram for your research

2. A Brief Description of the Data
You should assume that your reader is not familiar with the data you are using. Provide basic information such as the organization that collected the data, whether it is experimental or observational, and how units of observation were selected.

Mention how you are going to split data into train and confirmation sets

3. A Discussion of How Key Concepts are Operationalized
You should explain which variables are used to represent your X and your Y, and how well they match these concepts. Identify key gaps between the conceptual and operational definitions. If there are alternative variables that you considered, highlight them and explain how you made your decision

Clear operational definitions of concept to operation. So, emphasize the match between concepts and actual features used (table).

4. An Explanation of Key Modeling Decisions

1. How many observations were removed from the data, and for what reasons?

2. What transformations did you apply to your variables and why? Are they supported by scatterplots, statistical tests, or existing theory?

3. Are there covariates that were intentionally left out of your models and why? For example, did they reduce your precision too much, or are they outcome variables?

Accounting table is required. Perform EDA. Emphasis on what supports the use of transformations, if you performed them.

5. A Table or Visualization
You will be graded on your visual design. In particular:
1. Plots should be easy to navigate, with useful titles and axis labels.
2. Do not include raw R output. All output, including variable names, should be formatted to make it easy for an English speaker to read.
3. Plots should have a high information-to-ink ratio. If you are only communicating 2-4 numbers, a table is generally more effective than a plot.
4. Any plot or table you include must be commented on in your narrative. In other words, no output dumps

Emphasis. Plot and tables are important. Beyond all the quality issues, every graph or table must have an accompanying reason for its inclusions and particular placement in the paper. Multiple visuals showing the same observation are not to be used. If comparisons are made make sure the tables or visual are structured to be compared by the reader. No code display is necessary or appropriate.

6. A Well-Formatted Regression Table.
It is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects. You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table. As you select your model specification, your goal is to encircle the space of reasonable modeling choices, and to give an overall understanding of how these choices impact results. You should strive to make your models different from each other. However, each individual model must be defensible.
At a minimum, you need to estimate at least three model specifications. The first model you create should include only the key variables you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute

minimum number of covariates (usually zero or one covariate that is so crucial it would be unreasonable to omit it). The structure of the other models is more flexible. Most often, you will see researchers add a block of covariates from one model to the next. Each model should be defensible, and should continue to tell the story of how product features contribute to product success. This might mean including additional covariates to remove omitted variable bias; or, instead, it might mean estimating a model that operationalizes your X or Y in a different way (be sure the operationalization is substantially different). You may also create a model tailored to investigating a heterogeneous effect.

Up to this point use the train data set. This table is to be run on the confirmation data set. Model 1. treatment and output. Model 2. Model 1 plus covariates you think most important. No more than 5 covariate concepts (dummies not counted as individual concepts). Model 3. Kitchen sink.

7. A Discussion of Results

In your text, comment on both statistical significance and practical significance. You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

As part of practical significance refer back to your purpose for your client and your thesis. How do they compare? Discussion the salient material from the stargazer. What surprised you, what did not? Interpret the coefficient of your treatment variable. Do you think your client could or should use this to manipulate the process?

8. A Discussion of Limitations

8a. Statistical limitations of your model Make sure to evaluate all of the large sample model assumptions (or the CLM if you have a small sample). However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify,
describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies. Note that you may need to change your model specifications in response to violations of the large sample model.

Even if you are claiming large sample model assumptions, I expect you to evaluate the CLM assumptions and highlight in your paper examplar (not all) assumptions that might pose significant problems for your analysis. Convince me that you understand where you might start to improve your model.

8b. Structural limitations of your model What are the most important omitted variables that you were not able to include? For each variable you name, reason about the direction of bias caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting.

Is there a possibility of reverse causality? If so, reason about the direction of bias this causes. Are there any outcome variables on the right hand side? If so, reason about the direction of bias this causes.

Provide two OVB thought experiments (unused covaiates) versus the treatment variable. Assume that the treatment estimated coefficient represents β^r and using the OVB analysis we have discussed to tell me why and in what direction β^r will move in moving towards β^c. Provide an updated causal pathway diagram which you believe represent your findings.

9. A Conclusion
Make sure that you end your report with a discussion that distills key insights from your work, addresses your research question, and leaves the reader with a sense of broader context or impact.

Plus where would you seek learning in a future analysis of the topic

10 Reference Section (A section is optional, references are not).