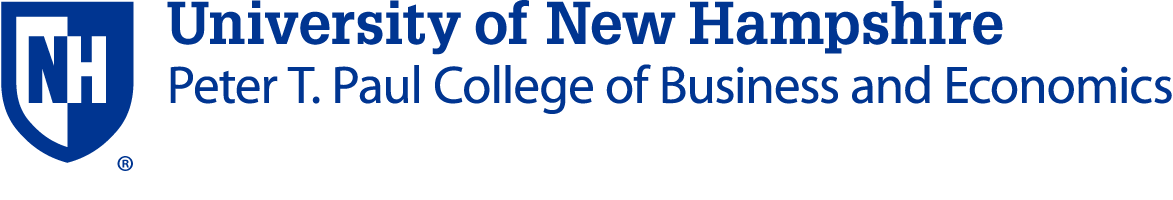
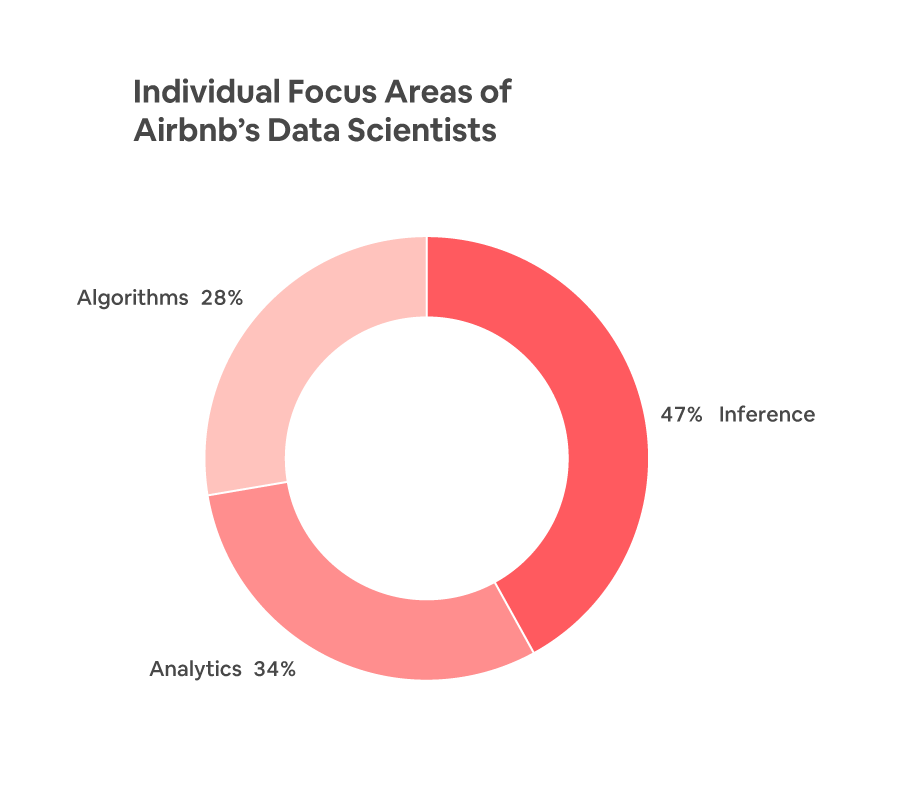
****

**LAB III - CAUSAL DISCOVERY & PANEL DATA ANALYSIS (2 in 1)**

Airbnb’s former Head of Data Science defines most of today’s data science as causal inference:**[[1]](#footnote-0)**





This will be a two-part lab. The two parts are completely independent of each other (so, why are they together? You need to read my announcement to get an answer to that question). You will also use two different lab files on Posit Cloud to answer the two parts below. Feel free to upload two code PDFs, but please keep your answers to the whole lab in the same document. As always, the necessary data file is uploaded to the data folder of both labs on Posit Cloud.

### Part A

In this part, you will use causal discovery algorithms to explain the relationships in the Airbnb data. If possible, use the full data (i.e., the 10k/150k sample in */data*). The more data you feed into a causal discovery algorithm, the more reliable your conditional independence tests will be.

1. In the Airbnb data, develop causal discovery graphical models using the PC algorithm and another algorithm of your choice that relaxes the causal sufficiency assumption. As you develop the causal graphs, make a use case for why you are developing the graphs.

**Hint 1:** Any categorical variables must be dummy coded. Instead of dummy coding for all levels, though, I would recommend you create variables of interest so that the output is meaningful. For example, if you want to add “cancellation policy” to your DAG, create a binary that shows whether the cancellation policy is flexible (or strict, for that matter).

**Hint 2:** Start simple with a few numeric variables and extend it from there. Remember, the more variables you use, the more relationships the algorithm will be able to reveal. Why?

**HInt 3:** I did *not* use high\_booking because that’s a variable created through feature generation. I started with a list of numeric variables that made sense. I removed those that did not provide an opportunity for an intervention (*Bedrooms* was an example but you don’t have to remove it - Just didn’t fit the use case I had in mind for the graph). The causal graph I created using PC ended up explaining the review scores rating quite well.

1. Report the *CPDAG* the PC algorithm creates. Does it make sense? Why or why not?
2. Report the *PAG* from the algorithm you chose. Compare and explain the differences.
3. In a single paragraph, discuss your findings and list your key takeaways and lessons learned from the analysis. How do you think you can use causal discovery in the future?

### 

### Part B

In this part, you will develop panel data models. Use the fixest library and use the feols() function. Again, there is no reason why you shouldn’t use the full data here. Do *not* split it.

1. Rebuild the model from Lab II using the minimum required variables\* *and add* price. Provide a numerical interpretation of the effect size of the treatment in plain English.

*\* Except for state and city. Do not add state and city to your model (yet).*

1. Is the effect size consistent with what the causal forest reported in Lab II?\* Why or why not do you think it is? Which one do you think is a more reliable estimate of CATE?

*\* The effect size from the causal forest was “95%CI for the ATE: 0.071 +/- 0.047”*

1. Add *fixed effects* to the model at Airbnb level (using IDs). Has the effect size changed? If so, how do you explain the change in the effect size after adding fixed effects? Now, is this a better estimate than what the causal forest reported earlier? Why or why not?
2. In the model with fixed effects, fixest clusters the standard errors by group by default. What does this clustering mean *in this domain for this problem* (so, don’t ask ChatGPT)?
3. Between the models *with* and *without* fixed effects, we observe some dramatic changes in the statistical significance levels (and sometimes magnitudes) of other coefficient estimations such as *Accommodates*. How do you explain such dramatic differences? Please provide a reasoning.
4. Add state and city *as fixed effects* to your model (in addition to the ID fixed effects). Has the effect size for treatment changed? Why or why not? Please provide an explanation.
5. Remove all fixed effects and develop a model explaining *price* instead of *high\_booking*.\* What is the effect of the treatment on price? Do you think this is a good model to make such an inference? Why or why not? Now, add the ID fixed effects. How about now?

*\* Just flip the locations of high\_booking and price in the model formulation (DV vs. IV).*

1. During your analysis until Q7, was there a suspicious coefficient? If so, which one is it? What would be a quick way to check and resolve your suspicion about its estimation?

### Deliverables:

You will submit **two separate PDF documents on Canvas**:

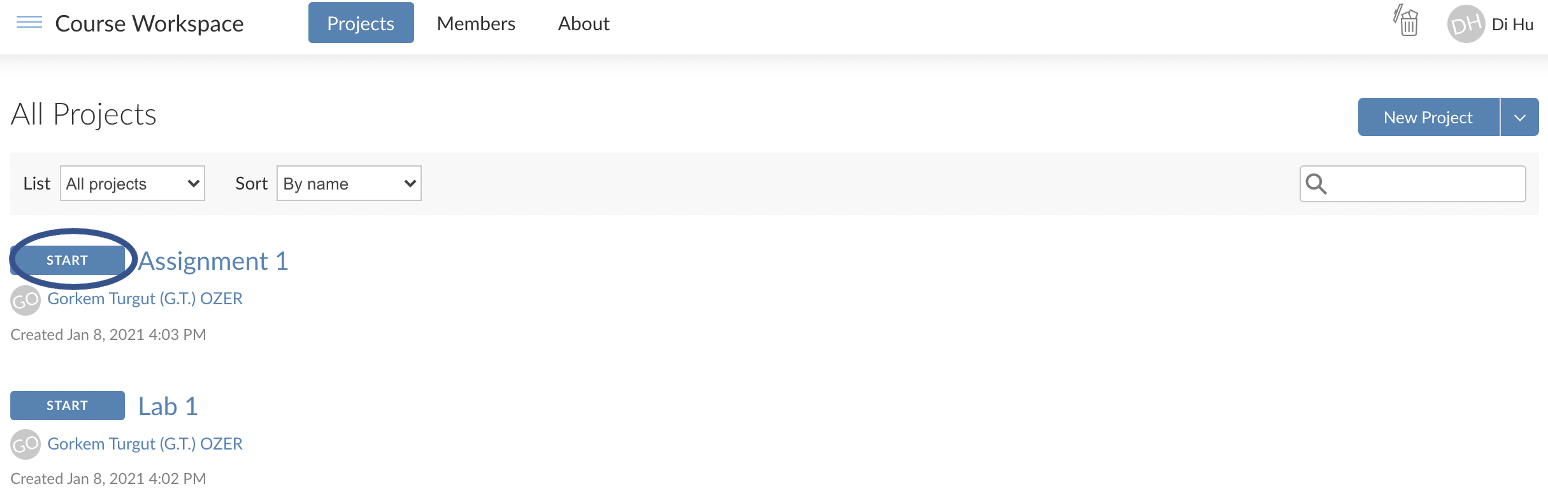
1. The first PDF will include your answers to the lab questions. You may want to create your own copy of the Google Doc, or download it as a MS Word document. Do whichever you like. When you are done answering the questions, please save it using the following naming convention: **LastName-FirstName-Answers.PDF** *-This is the main submission file*
2. The second PDF will be the code file you will generate in Posit Cloud as a PDF file when you are done working on the code. Please see the guidelines starting from the next page. Name it as as **LastName-FirstName-Code.PDF** *-This is the supplemental file*

**Step 0:**

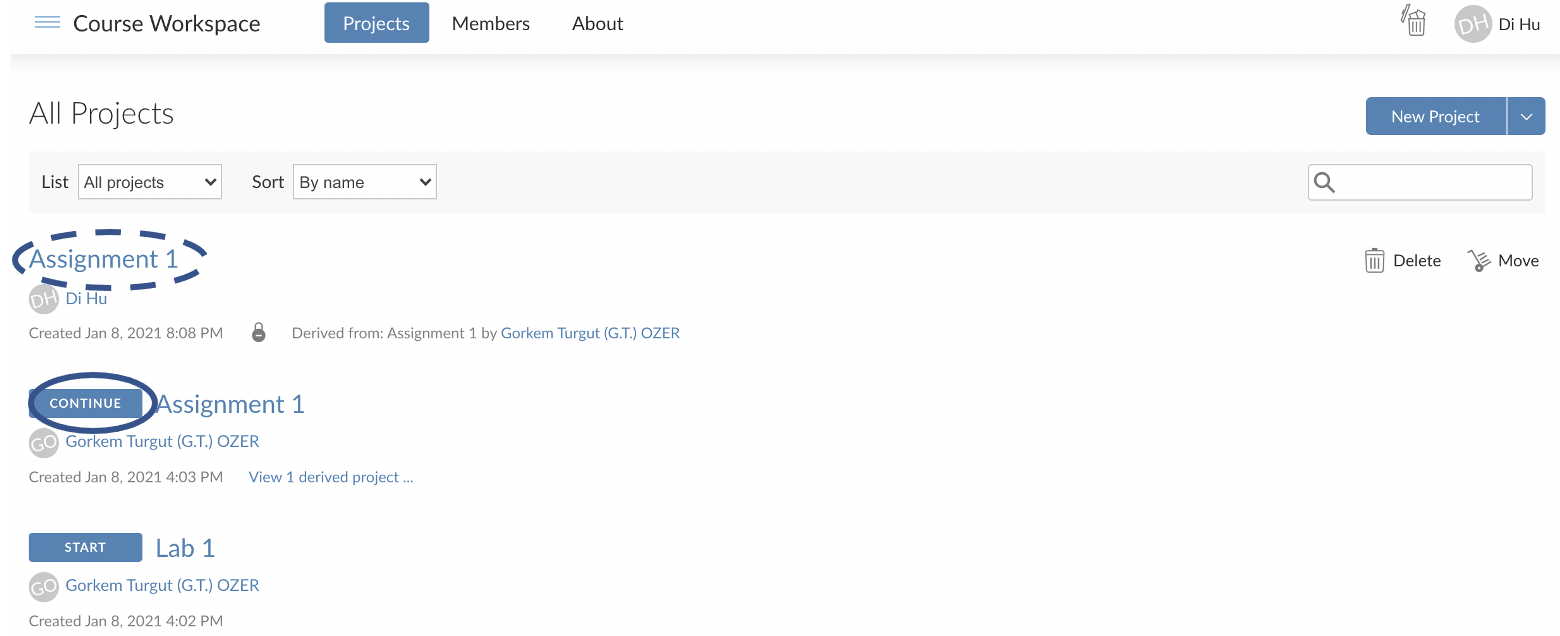
First, read [**How to use Posit Cloud**](https://docs.google.com/document/d/1CcAbFVQYwOJb5AiYusQ68_BX1WIPnz6-7N1_dWYs44M/edit?usp=sharing) in its entirety.

**Step 1:** In Posit Cloud, either start with the lab (where 10k Airbnb data is uploaded) or use your project. Start with clicking the ‘Start’ button to make your own derived copy (or to continue your work on your copy by clicking the ‘Continue’ Button). The assignment will open in an IDE.

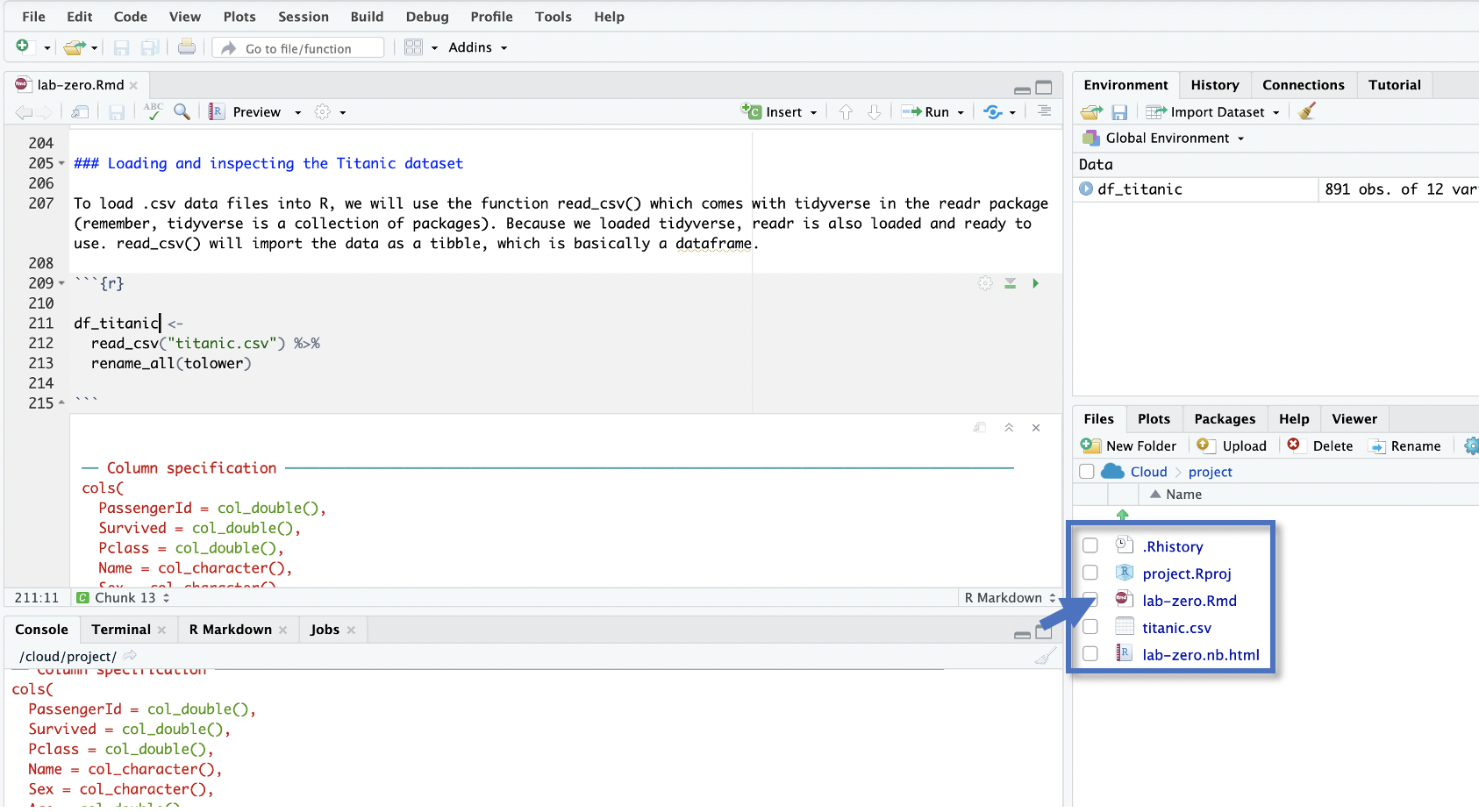
Start working on a lab:



If you already started once, continue working on a lab:



**Step 2:** When the lab opens, look to the right bottom of the window, and you will see several files and folders there. The name of the file will correspond to the name of the lab.



**Step 3:** As you know, you’ll enter your R commands between special markers called “[chunks](https://rmarkdown.rstudio.com/lesson-3.html).” Labs may come with some comments and an empty chunk to start with. A good practice is to use a different chunk for each question (or subquestion in a long question).

```{r}

*Your R commands here*

```

**Step 4:** When you are done answering the questions, style your code using **StyleR**. Install styler and style your code file following the example [here](https://www.tidyverse.org/blog/2017/12/styler-1.0.0/).

**Step 5:** After you fix any styling issues you may have, follow the instructions on the next page to create a **PDF** of your R Notebook with the naming convention **LastName-FirstName-Code.PDF**

**Step X:** Submit the two PDFs you now have through Canvas under the respective assignment:

1. LastName\_FirstName-**Answers**.PDF (will include your answers to the questions)
2. LastName\_FirstName-**Code**.PDF (will include your report including your code)

**How to create a PDF of your code in an R Notebook**

~~0. First, run the following two lines of code in your R Console to install the necessary components:~~

~~tinytex::install\_tinytex() => This should not be needed. Try it if you have trouble.~~

1. To create a PDF (or Word) version of your work from an R Notebook, you need to have the following lines of code at the top of your R notebook. You can copy and paste from below and change the title. When you are using a provided lab/assignment template, these lines should already be in the file.

---

title: "Your Title"

output:

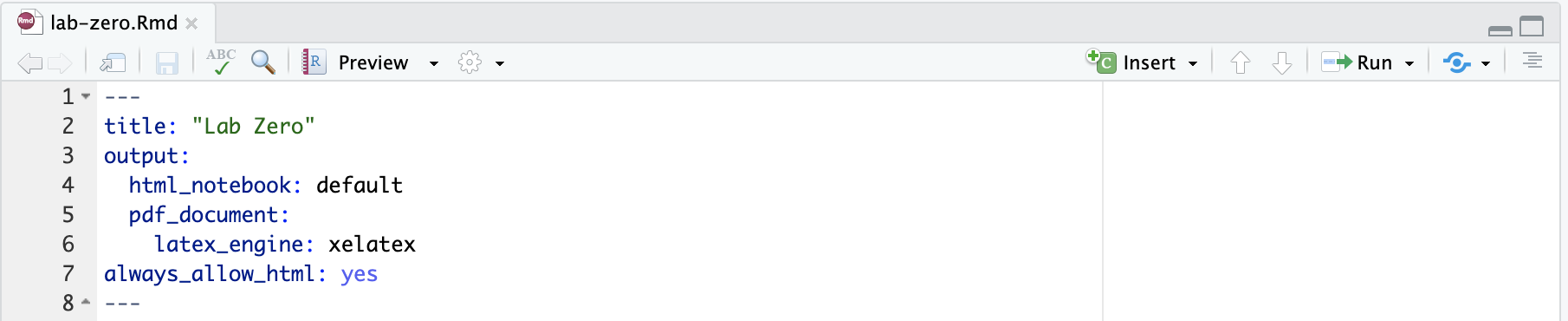
html\_notebook: default

pdf\_document:

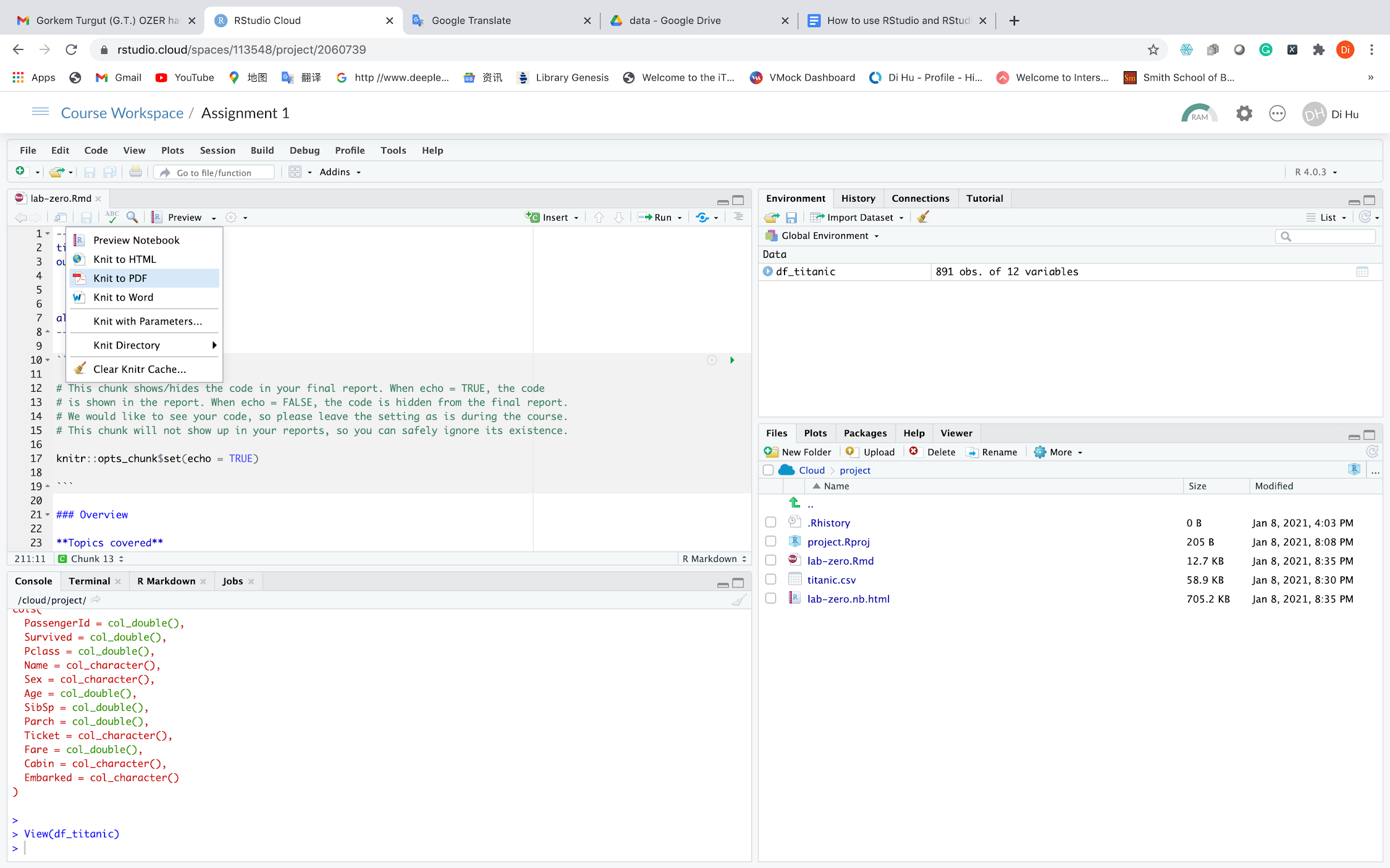
latex\_engine: xelatex

always\_allow\_html: yes

---



2. Save your file after adding these lines. Then, click the down arrow next to the “Preview.” This will show the options as shown below. Choose the PDF (or PDF) option to create a report of your work in seconds:

****

3. Here you go:

- If you choose the PDF option, your PDF report will open in a new window automatically for you to save.

- If you choose the Word option (not needed for the course), you will see a prompt to download the file.

1. **Source:** Article “One Data Science Job Doesn’t Fit All” by the former Head of Data Science at Airbnb: <https://www.linkedin.com/pulse/one-data-science-job-doesnt-fit-all-elena-grewal> [↑](#footnote-ref-0)