Please read this entire assignment carefully.

Important Due Dates		
Item Due	Submission Method	Due Date
Partner Assignments	Add to GoogleSheet	Tues 2/20 by class
First Draft	Email to reviewers and me	Tues 3/19 by class
Peer Review Comments	Email to authors and me	Fri 3/22 by "class time"
Final Draft *and* Response to Reviewers	Moodle	Wed 3/27 by "class time"

*"class time" is either 10:30AM or 1:30PM depending on your section; it's in quotes because we don't actually have class on these days

Instructions. You will work with a partner on this project. The project is an open-ended, independent project. This means you should synthesize and apply your knowledge base from our class. But if you encounter "unfamiliar" territory, be creative and resourceful: you *may (and should!)* consult textbooks, articles or Internet resources, but you *may not* consult other people besides your partner. You may ask me questions only to clarify *what* the assignment is asking, but not *how* to complete parts of the assignment.

Your task. Write a summary report about *average full market value* of condos in NYC and how it relates to condo characteristics. The only requirements are that your conclusions be supported by the data and be based on a thoughtful data exploration and an appropriate multiple linear regression model. Have fun exploring the data and telling a story (supported by the data) from it!

Data Source Website. The data for this project were downloaded from NYC Open Data[1]. Be sure to read this and familiarize yourself with the data.

Data. The data file (`nyc-condos.csv`) is part of the Midterm .zip file that you downloaded from Moodle. It contains a random sample of 200 condominiums in NYC from the original dataset and the following subset of variables from the original dataset:

- `Boro-Block-Lot`
- `CondoSection`
- `Address`
- `Neighborhood`
- `BldgClassification`
- `TotalUnits`
- `YearBuilt`
- `GrossSqFt`
- `EstGrossIncome`
- `GrossIncomePerSqFt`
- `EstimatedExpense`
- `ExpensePerSqFt`
- `NetOperatingIncome`
- `FullMarketValue`
- `MarketValuePerSqFt`
- `ReportYear`

For variable descriptions, see the Data Source Website (above).

Get to know your data. Before diving into building models (tempting!), be sure to spend time carefully getting to know your data through graphical and numerical exploration. In addition, you should browse the data source website and other relevant sites to become familiar with the variables and problem

[1] Data Source: `https://data.cityofnewyork.us/City-Government/DOF-Condominium-Comparable-Rental-Income-in-NYC/9ck6-2jew`

context.

Be creative and keep it simple. You are free to re-format or transform any variables as you see appropriate. Also, simplicity is okay. I'd rather see a model that is simple, complete, and supported by the data, than one that is complicated and not supported by the data. The data are not perfect and there are likely some variables you'd additionally wish to have, but do what you can with the data we do have. You're welcome, but not expected, to use outside data.

Preparing your report (and `R` code). Use R Markdown to prepare your report, similar to what you've been doing for homework. The .Rmd file should have all the code you used (even for analyses that you may not describe in the report) and be documented with comments, but your compiled report should not have any code in it. Use the optional chunk arguments to customize and control the look of your report.

First Draft: What to submit. Read carefully.

1. **Summary report** (.pdf) that satisfies the following conditions:

    - Written in prose form. Do not use bulleted lists to summarize your methods, results, or discussion points.
    - Approximately 3-5 pages. This includes the Abstract, Introduction, Methods, Results, Discussion, Conclusion, and References (all described below), as well as tables and figures.

      See below (*More on the summary report structure*) for specifics on how each section should look.
    - Figures and tables. Be sure to label any figures/tables with titles, and reference them in your text. Do not include every figure you make; choose figures and tables that will help the reader understand the data and/or model results. For example, you may want to include a "baseline characteristics" table in which you provide summary statistics for the variables in the dataset, or even bivariate analyses (i.e. SLR of the outcome on each predictor variable). Or you may include some graph of the original data to show any potentially interesting observations. Avoid including diagnostic plots (e.g. residual, influence, etc) in the paper.
    - No `R` code should appear in your summary report, but it should all be in your .Rmd file. Use the optional code chunk arguments such as `echo = F, message = FALSE, warning = FALSE` to suppress printing of code, messages and warnings in the compiled report.
        - Helpful resources for Markdown formatting:
            - http://www.stat.cmu.edu/∼cshalizi/rmarkdown
            - https://commonmark.org/help/tutorial/
    - Look at existing journal articles (including our In the Lit article) to serve as templates for paper structure. Also see the section *More on the summary report structure* below.

2. **Source file** (.Rmd) that satisfies the following conditions:

    - Reproducible. I should be able to run the entire script error-free without having to edit any code. To help with this, be sure you create an `R` project and include all files (including data files) needed to run your analysis in the project folder.

- Documented with comments. There should be good documentation describing each step of your analysis and any observations or decisions that you make (but that you may not necessarily include in your final report). Any comments you want to record as reference but not appear in your report should be typed as commented code *inside* the code chunks; otherwise text that is meant to appear in the report should be typed *outside* of code chunks as Markdown / plain text.
- Think of using this file as a way to communicate to your reviewers (who are interested in the details of *how* you arrived at your findings) the details of your analysis that aren't included in your summary report.

3. **Data files** (.csv) - any data files you used, including the original data file I gave you. You are *not* required nor necessarily expected to use **outside data**, but if you do, be sure you include those .csv files as well.

4. **Project file** (.Rproj) - to aid in reproducibility

More on the summary report structure:

1. **Title.** In 80 characters or less, give the essence of your research question and findings (i.e. it's okay if the title is a spoiler; this is typical of academic journal titles)
2. **Abstract.** In 250 words or less, briefly summarize your paper including motivation, methods, key findings, and a concluding remark.
3. **Introduction.** Describe the context and state the research question; motivate the question. For purposes of this assignment, feel free to embellish and be creative within reason, based on any personal experience or knowledge.
4. **Methods.** Describe what you did (e.g. models you used, plots you created, tests you conducted)
    - On the one hand, do not describe every single detail of everything you tried, and on the other, do give an honest picture of your approach. (Your source file will be a place where reviewers can see "everything", step-by-step, that you tried. Feel free to include comments in your source file of things you tried and observations/decisions you made along the way.)
5. **Results.** State your results here. Connect the text in your Results section to the tables and figures you include by references, e.g. "As seen in Table 2, the model tells us..." Save any *discussion* (e.g. interesting findings or implications of your findings) for the Discussion section.

    Assume your reader has basic statistical understanding. This means you do not need to explain common concepts (e.g. p-value, confidence intervals, regression coefficients); you can simply report them. A good way to report coefficients is to describe them qualitatively (using words like "more" or "less", "increase" or "decrease") and then state the point estimate and a CI in parentheses; see our In the Lit or other articles for examples.
6. **Discussion.** Highlight interesting and/or unusual features in your data and findings from your analysis, and provide any possible explanations; state potential implications and limitations of the data, your analysis, and/or your findings
7. **Conclusion.** In approximately one paragraph, summarize your report. Remind the reader of the importance of your work, and the key message of your analysis and how it relates to your original question. End on an appropriately positive and strong note.

The next two sections are to be included as needed:

8. **References.** If you used any references, be sure to list them in a References section
9. **Appendix.** Include any technical details here that aren't essential for understanding the report but that may be of interest to some readers. This might include, for example, explanations of any variables you derived; but would not include explanations of commonly-used or known methods.

A helpful resource on structuring a data analysis report:
http://www.stat.cmu.edu/∼brian/701/notes/paper-structure.pdf

Peer Review / Revision Process. As part of this project, you will have the chance to "peer review" your classmates' work. After the *first draft submission deadline*, you (and your partner) will receive the first draft from another group. You will review the draft and provide comments that help to clarify the writing, suggestions for potential additional analyses to conduct, suggestions for additional discussion items to include in the paper, and general questions about the analysis. See the separate document *Midterm Reviewer Guidelines* on Moodle for more details.

Final Draft: What to submit. After you *receive peer reviewer comments* for your rough draft, you (and your partner) will have a chance to address these comments. In a separate "Response to Reviewers" file, respond to *each comment*, indicating how/where in your revised draft you addressed the comment. You can use an itemized list, with one item per comment and response pair. It should be made clear how you addressed each comment. Upload to Moodle the *final versions* of the following *before class* on the due date:

1. **Response to Reviewers** (.pdf)
2. **Summary report** (.pdf)
3. **Source file** (.Rmd)
4. **Data files** (.csv)
5. **Project file** (.Rproj)

Evaluation. You will be evaluated on your rough and final drafts, and your review comments that you provide to your assigned authors, with most weight being given to the final draft. I will evaluate your final paper with the following aspects in mind:

- Writing Style
  - Clarity
  - Organization
  - Formatting of report (hint: use optional code chunks for the .Rmd file)
- Statistical Analysis
  - Thoroughness (at least all potential "flags" explored, as evidenced in the commented R code, though not necessarily addressed in the paper if found not to be a problem)
  - Appropriateness of final model - have the authors assessed the appropriateness of their final model?
- Conclusion and Discussion
  - Sufficient support for conclusions - are the conclusions supported by the data?
  - Limitations and Strengths - do the authors thoughtfully consider limitations and strengths of their work?
- Figures and Tables
  - Appropriate choice of tables and figures to include in paper
  - Appropriate use of labels - informative and descriptive
  - Formatting of tables and figures (e.g. no copy and paste of R output)