For your end of semester project, you will select a dataset and carry out a linear regression analysis. There are multiple sections to this project, so you can start certain parts of the project now even we've only briefly covered linear regression.

Note: The project will be officially due on the last day of the module, December 13. Please note that unlike other deadlines, this is set by the college–I cannot accept anything after noon on December 13!

Dataset Selection (10 points)

You may pick any dataset for your final project, provided that the dataset has:

- At least fifty cases/rows;
- A numerical response–if discrete, there should be a large range of values (20 or more).
- At least one numerical explanatory variable (you can have more if you like).
- At least one categorical explanatory variable (you can have more if you like).

I recommend looking at the sites in the "Project Datasets" handout included in the Final Project folder on Moodle–there are many datasets that are suitable for this class! You are welcome to find a dataset from a different site, or even use your own data. I recommend that you check with me before moving forward with the rest of the project so that you don't have to start over if something turns out wrong–this will be one of the questions in our regular progress checks.

Exploratory Analysis and Data Summary (15 points)

Analysis of a dataset often begins with summarizing each variable in a dataset. A discussion of graphs and summary statistics for each variable is a good way to introduce your audience to the data. Your project should include:

1. Graphical displays for each variable. Points will be taken off if you use an inappropriate graph to summarize the distribution of a variable (e.g., a bar chart for a numerical variable). Graphs should have titles and axes labeled, plus legends if you include different colors.
2. Summary statistics for each variable. You may choose which summary statistics to report, but you should have a table of the summary statistics for the numerical variables and a table of the summary statistics for the categorical variables.
3. One written paragraph per variable, referencing the summary statistics and commenting on the graphs.

To submit Homework 14, please turn in a draft of these tables and graphs by Monday, December 6. The grade will be completion-based, but you have to submit something!!

Note: If you have one and only one categorical variable in your dataset, you may choose to include it in various graphical displays rather than creating a separate graphical display. For example, you may choose to color code your graphs for numerical variables by male versus female observations rather than create a bar chart specifically displaying the distribution of sex. If you have questions, please ask.

Simple Linear Regression (50 points)

To continue your data analysis, select one numerical variable to be the response variable. The selection will depend on the context of your dataset. If you have questions, please ask me during one of our project work days.

Next, select one numerical variable to be your explanatory variable. This selection is not as important–you may have multiple explanatory variables, so you can change the variable if you want–but the relationship you are investigating should make sense.

Carry out a full simple linear regression analysis of these two variables. Your analysis should include:

1. A plot of the two variables. The relationship displayed in this plot should be discussed, including the form, direction, and strength of the relationship as well as any outliers that may be present. (5 pts)
2. Correlation, R, between the two variables. The correlation can be used to summarize the strength of the relationship in part (a) (2 pts).
3. Interpretation of R^2 for these two variables. (3 pts)
4. Calculations of the slope and intercept for the linear model of these two variables. Both the slope and intercept should be interpreted. When interpreting the intercept, clearly if whether you are extrapolating. (5 pts)
5. Confidence intervals for the slope and intercept parameters. (5 pts)
6. Plots for checking the assumptions of a linear model (a scatterplot comparing the explanatory variable to the residuals, and a histogram of the residuals). These plots should be fully discussed (form, direction, strength and outliers for a scatterplot, center, spread, shape, and outliers for the histogram). (10 pts)
7. A discussion of the influential points in the dataset. If there are none, clearly state that there are none. If there are influential points, discuss their influence–how would the relationship change if you removed the point? (5 pts)
8. A discussion of whether the linear regression model is appropriate for the dataset. Do you need to use a more complicated method for this dataset? (5 pts)
9. A discussion of the relationship at hand–is it a strong relationship? A weak relationship? Is the explanatory variable a significant predictor of the response? This section must be supported with statistical evidence. (10 pts)

Adding an Indicator Variable (15 points)

Next, add a categorical variable to your linear relationship. Be sure to appropriately add the variable if it has more than two levels!

Carry out an analysis of the categorical variable. Your analysis should include:

(a) A new scatterplot between the original, numerical predictor and the response variable. Color the points according to your categorical variable, and comment on the new relationship between the three variables. Does knowing the value of the categorical variable add any information to your prediction? (5 pts)

(b) Interpretation of the adjusted R^2 value. Discuss the change in adjusted R^2—does it confirm your theory from part (a)? (2 pts)

(c) Calculation and interpretation of the coefficient on the categorical variable. (3 pts)

(d) A confidence interval for the coefficient on the categorical variable. (5 pts)

Report (10 points)

Finally, write up all of your analyses in a data analysis report. The report should include all elements mentioned previously. It should be formatted neatly with few grammatical and spelling errors. It should not include unnecessary tables or graphs—if you do not feel that a table or graph is important enough to discuss, it is not important enough to include in the document! Finally, the report should not include your code—I do not need to see how you came up with the answers. Essentially, you should be writing this report to an intelligent audience who understands the context of the dataset and generally how to use linear regression, but who does not understand the code or work underneath the hood.

As always, if you have any questions, please ask me during one of our in-class work days, or make an appointment to discuss your project. You may also email me before the work days if you are getting a jump start on the project, but it will be easiest for me to help or debug when we are together.