Last week, you did a full-fledged analysis after some Data Camp Exercises. Congrats!!! As time moves forward, your ability to do that (with more varied data) will increase. You should be proud of yourself. As time moves forward, you will begin to figure out how to use scripts (canned text) to rerun and recreate an analysis so you can keep yourself moving forward.

This week, you will first do a brief Datacamp exercise and then work with data from the American Community Survey. Each of you will be visualizing data, visual data tells stories and as a Data Scientist, this is important. You will produce Histograms, Probability Plots, and Descriptive Statistics. These will help you answer a series of questions.

Last week, you looked at data from the American Community Survey and produced Histograms, Probability Plots, and Descriptive Statistics. This allowed you to answer a series of specific questions based on the variables assigned. Nice work! We are taking steps into an increasingly complex world.

Don’t forget to give yourself a breather now and then.

The most important goal from my perspective last week was to remember from last week was that you should start to understand what a normal distribution is and how to use common visualizations to get a snapshot or sense of the “normality” of the data.

This week, we’ll continue building on normal distributions and focus specifically on the resources provided with the describe() function in R to further explore the distribution of variables. Having a firm understanding of the statistics provide from this function will be important in later sessions (e.g. Regression).

Remember that both textbooks for this course include many examples that are similar to your weekly assignments. Practice using the textbook examples to build your confidence and proficiency is recommended. While I can force you to do things, you doing stuff on your own will be way more valuable.

This book will provide you with a lot of help if you give it some time:

Xie, Yihui; Allaire, J. J.; & Grolemund, Garrett. (2019). [R Markdown: The Definitive Guide](https://bookdown.org/yihui/rmarkdown/).

Also, don’t be afraid to poach things from:

RStudio. [RStudion. R Markdown](https://rmarkdown.rstudio.com/lesson-1.html" \t "_blank).

Last Week, we built on normal distributions by focusing on the resources provided with the describe() function. We also used R Markdown as part of our Data Science workflow for report generation. Markdown is a weird language, honestly, but it translates to so many other aspects of what the computer-folks do that it is necessary to learn.

This week, we will analyze how we can express the relationships between variables statistically by looking at two measures: Covariance and the correlation coefficient. We will also discover and practice how to carry out and interpret correlations in R. Recall that both textbooks for this course include many examples that are similar to your weekly assignments. Continue to practice using the textbook examples to build your confidence and proficiency.

For those of you wondering about correlation, there are a lot of examples and ideas around the web. I find this one particularly useful:

Magnusson, Kristoffer. [Interpreting Correlations](https://rpsychologist.com/d3/correlation/).

Similarly, this one is also useful:

Wagih, Omar. [Guess the Correlation](http://guessthecorrelation.com/).

Finally, this item from week 1 has some discussion about correlation:

Kunin, Daniel. [Seeing Theory](https://seeing-theory.brown.edu/).

Last week, you analyzed the relationship between variables by looking at two measures: covariance and the correlation coefficient. You also discovered and practiced how to carry out and interpret correlations in R, including how to control and make interpretations about a third variable. Correlation and regression are the bread and butter of frequentist statistics.

Congratulations of making your first food! You can now start to live on your own. This week, we figure out how to garnish a meal. We will take the methods from last week and move a step forward. We will begin to predict outcomes using what is known as regression. We will look at how influential cases and outliers can affect the accuracy of a regression model and methods used to identify those factors. You will perform regression using R, interpret the output, create regression models and test those models’ reliability and generalizability.

Last Week, we expanded on correlation by introducing prediction with regression. Influential cases and outliers can influence the accuracy of a regression model. In that light, methods that are commonly used to identify those factors were highlighted. You did an analysis using R, interpreted the output, created regression models and tested those models reliability and generalizability.

This week, we extend our statistical methods even further. If we’re sticking with the food metaphor. In week 5, we found some bread. In week 6, we made a sandwich, maybe a cheese sandwich? In week 7, we make some grilled cheese and tomato soup. To be less metaphorical, this week we will expand linear regression to include a set of predictors along with a response variable.

Some Statistics Tips!

An important reminder before starting this week’s assignments: One thing not to do is select hundreds of random predictors, bring them all into a regression analysis and hope for the best.

Also remember that R is a tool and it will perform calculations on the data you ‘feed’ it; garbage in garbage out may result if you are not careful with the quality of the data inputs.

In addition to the problem of selecting predictors, there are several ways in which variables can be entered into a model. When predictors are all completely uncorrelated, the order of variable entry has very little effect on the parameters calculated; however, we rarely have uncorrelated predictors and so the method of predictor selection is crucial. Keep all of this in mind as you work through this week’s activities.

Alright, no more jokes about food. I’m sorry, I kept writing when I was hungry. This week is about logistic regression. This form of regression is appropriate when the dependent variable is a binary. Like all regression, this binary-based is predictive as well. Logistic regression describes the relationship between one binary variable (we call it the dependent variable) and one or more variables of any kind (these are independent).

This form of regression is important for a lot of things. In fact, you may not know it but it is a form of machine learning. More on this next week.

Here is a breakdown of your assignments and tasks this week:

Read:

Chapter 8 from Discovering Statistics Using R.

[Understanding the Bias-Variance Tradeoff](http://scott.fortmann-roe.com/docs/BiasVariance.html), Scott Fortmann-Roe, 2012

[Calculating UAC: The Area Under a ROC Curve](https://www.r-bloggers.com/calculating-auc-the-area-under-a-roc-curve/), Bob Horton, R-bloggers, 2016

8.1 Assignment: Logistic Regression

8.2 Assignment: Fit Logistic Regression Model

8.3 Discussion: Tech Support Fort

<https://www.youtube.com/watch?v=XycruVLySDg&feature=emb_logo>

This week, we begin to disconnect from the DataCamp world. While we’ve gotten pretty good at making lessons like these, they often provide you with a false sense of “knowing what you’re doing.” This isn’t to say that you don’t, I’m sure you do; however, this week you’re going to be finding your own data, asking your own questions, and doing your own analysis. That, in and of itself, is an insane thing.

So, we’re adding some stuff to this week as well. This is the big first step also: applying what you’ve learned with another aspect of, “doing it yourself.” Did you know that much of what you’ve done is machine learning? Yes, that machine learning.

The rest, keep.

Here is a breakdown of your assignments and tasks this week:

Read the following:

Machine Learning Fundamentals

Bernard Marr. (2016). [Supervised V Unsupervised Machine Learning – What’s The Difference?](https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats-the-difference/#4d2a4cf2485d)

Bernard Marr. (2016). [What Is The Difference Between Artificial Intelligence And Machine Learning?](https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/)

Bernard Marr. (2016). [What Is The Difference Between Deep Learning, Machine Learning and AI?](https://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/)

K-Means Clustering

Sejal Jaiswal. (2018). [K-Means Clustering in R Tutorial](https://www.datacamp.com/community/tutorials/k-means-clustering-r)

Andrea Trevino. [Introduction to K-means Clustering](https://www.datascience.com/blog/k-means-clustering)

Nearest Neighbors Classification

Kevin Zakka. (2016). [A Complete Guide to K-Nearest-Neighbors with Applications in Python and R](https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/)

Scikit Learn. [Nearest Neighbors Classification](http://scikit-learn.org/stable/modules/neighbors.html#classification)

9.1 Discussion: Final Project Step 1: Getting Started

9.2 Assignment: Introduction to Machine Learning

9.3 Assignment: Clustering