GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 3 - Regression
Monday 4th April

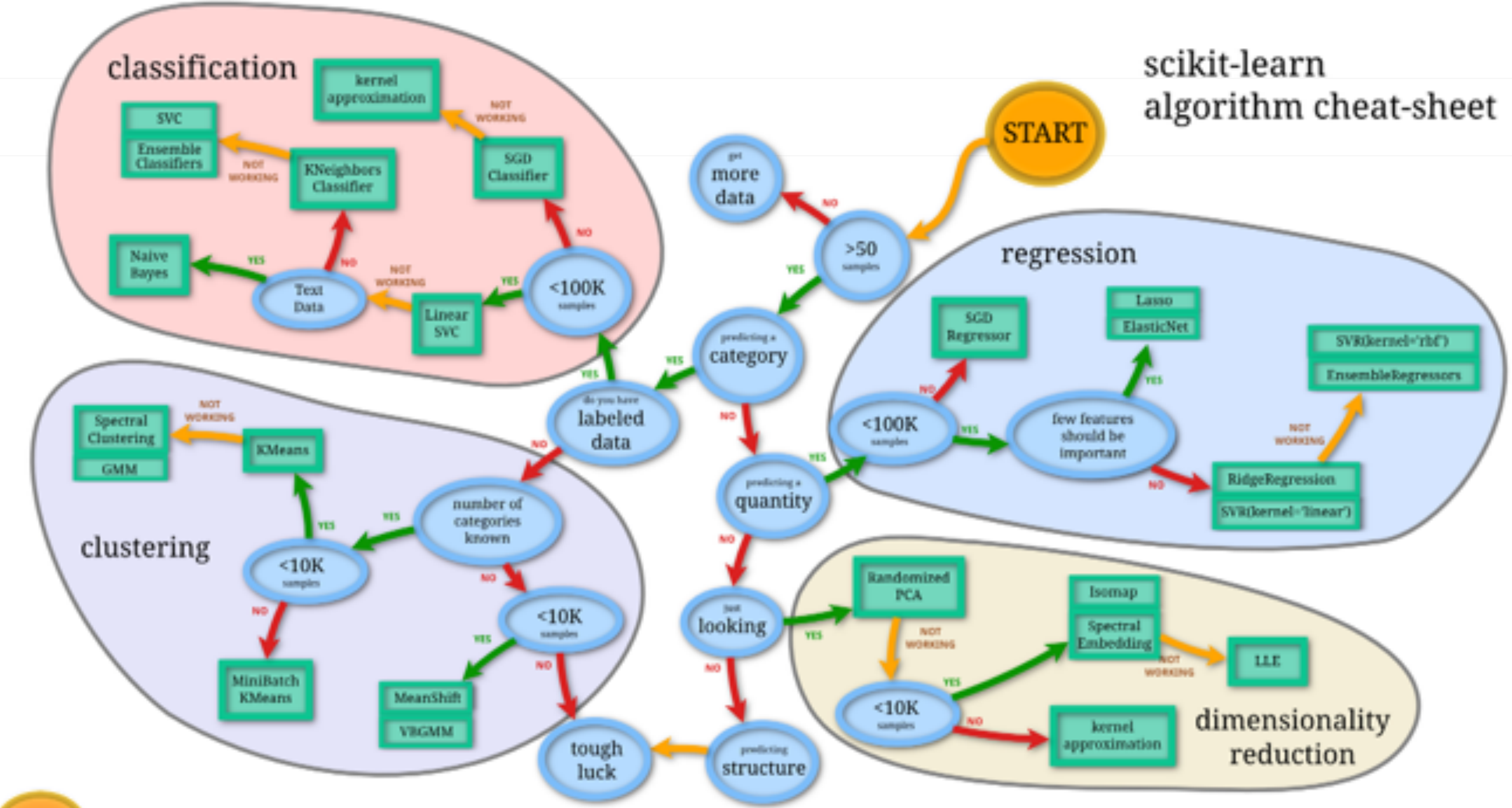
AGENDA

1. Motivation
2. Supervised Vs Unsupervised learning
3. What is Linear Regression?
4. How do Run a Linear Regression Model?
5. Lab
6. Discussion / Review / Homework

DATA SCIENCE PART TIME COURSE

WHAT ARE THE GOALS OF STATISTICAL LEARNING?

scikit-learn algorithm cheat-sheet



Back

We want to predict some value, let's call it **y**, based on some observed data we have, let's call that **x**.

We will use statistical learning to estimate a function that approximates **y** based on the input, **x**.

y is also called; label, dependent variable, target

x is also called; predictor, independent variable, features

SUPERVISED LEARNING - EXAMPLE

We want to predict the price of a house, let’s call it y, based on some observed data we have about the area, number of bedrooms, size of the house, and if it has a pool or not.

The area, number of bedrooms, size of the house, and if it has a pool or not would be our x variables (sometimes you might see this denoted as X)

What we want is $y = f(X)$, a way to describe the house price based on observed data

SUPERVISED LEARNING - REGRESSION & CLASSIFICATION

If the y variable is numeric then we have a regression problem - we are trying to predict a continuous number

If the y variable is a category (for example trying to predict a type of flower) the we have a classification problem - we are trying to classify what group that y belongs to.

UNSUPERVISED LEARNING

We want to find some underlying structure or patterns in the data but in this case we don’t have any labeled data.

So for example, if we have a large group of customers but would like to separate them into groups (or clusters) to better target them.

WHAT IS LINEAR REGRESSION?

We want to model a linear relationship (think straight line) between our target variable y and our input variable x.

$$y = X\beta + \epsilon$$

$$y = X\beta + \epsilon$$

- y = target variable
- X = input variable
- β = coefficients
- ϵ = error term

Note, one of our input variables can be 1 so we have an intercept parameter

LINEAR REGRESSION - REQUIREMENTS

- Linear relationship in the parameters, β, we can transform the actual values of the inputs if we want
- Variance of the error term, ϵ, is constant. This means there is no systematic pattern in the values of X and the variance of ϵ
- The mean of $\epsilon = 0$
- ϵ has a normal distribution
- No perfect (or near perfect) co-linearity between any of the input variables. Otherwise the fitting procedure will break.

HOW TO RUN LINEAR REGRESSION?

ORDINARY LEAST SQUARES

$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

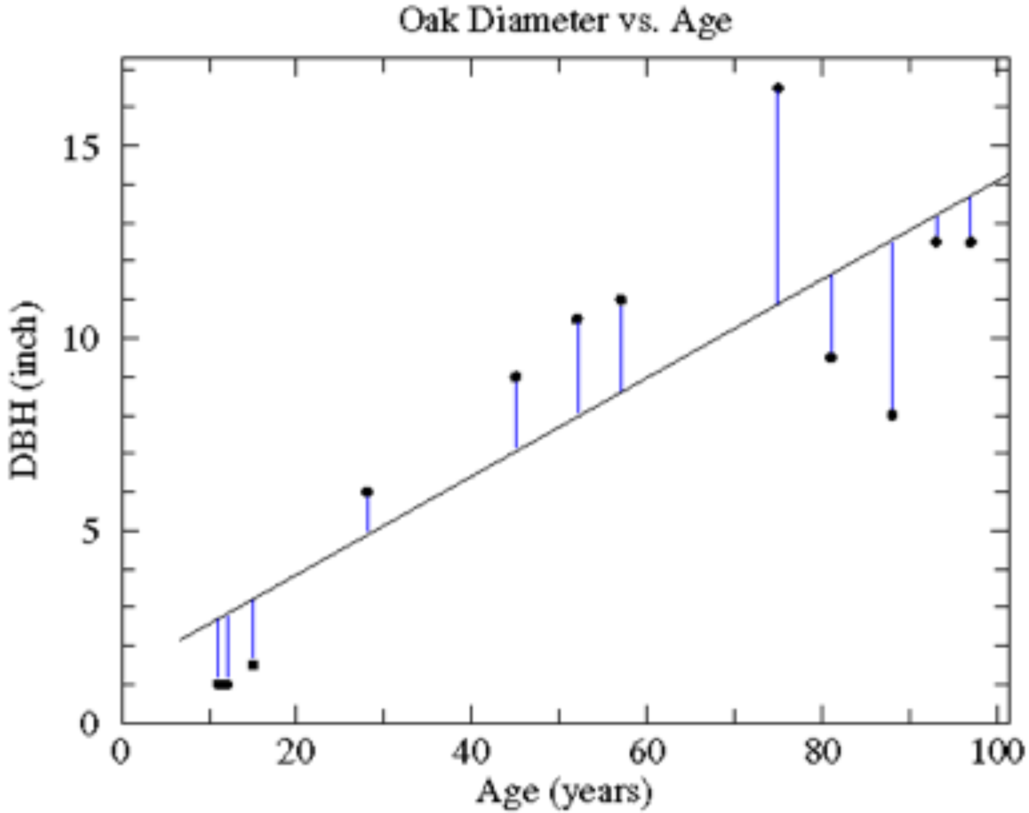
Basically, what we are trying to do is minimise the Residual Sum of Squares. This is the Sum of the squared difference between our observed value and the value from the model

ORDINARY LEAST SQUARES

$$\boxed{SS_{res}} = \boxed{\sum_{i=1}^{n}}\boxed{\left(\boxed{\boxed{y_i} - \boxed{f(x_i)}}\right)^2}$$

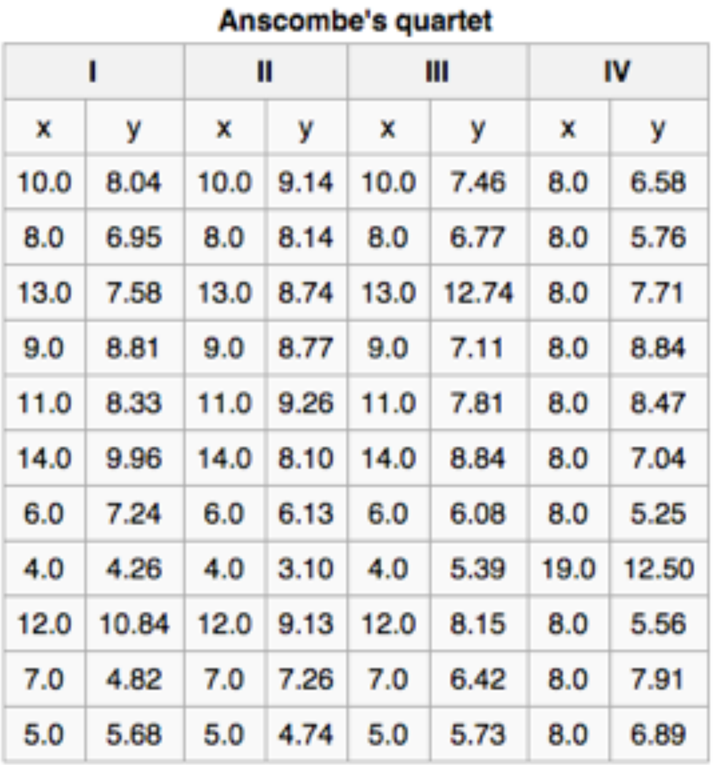
Basically, what we are trying to do is minimise the Residual Sum of Squares. This is the Sum of the squared difference between our observed value and the value from the model

ORDINARY LEAST SQUARES



LINEAR REGRESSION - IMPORTANT STEP

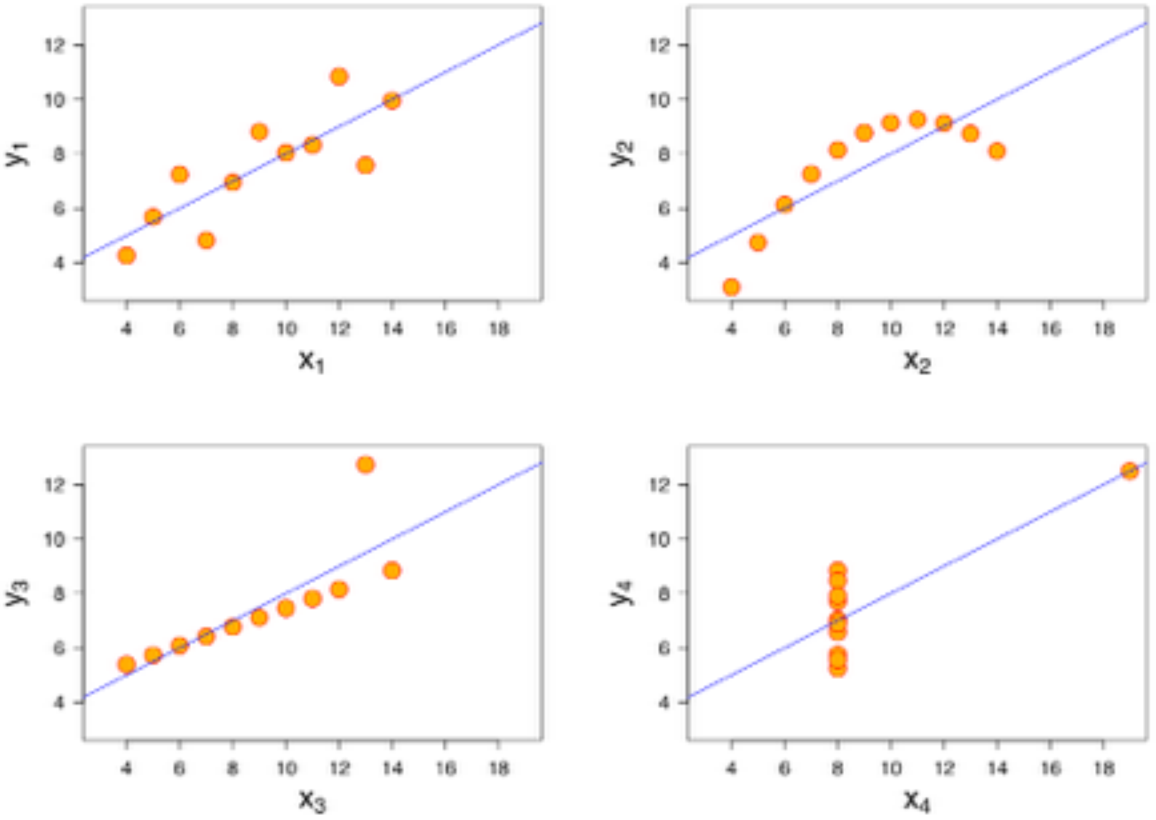
- Make sure you visualise your data and check the actual model fit !!!
- The fitting a model to the four datasets in the table on the right produce the same fit statistics, model coefficients and standard error
- See anything wrong?



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

LINEAR REGRESSION – IMPORTANT STEP



LAB

SYNCHING A FORK

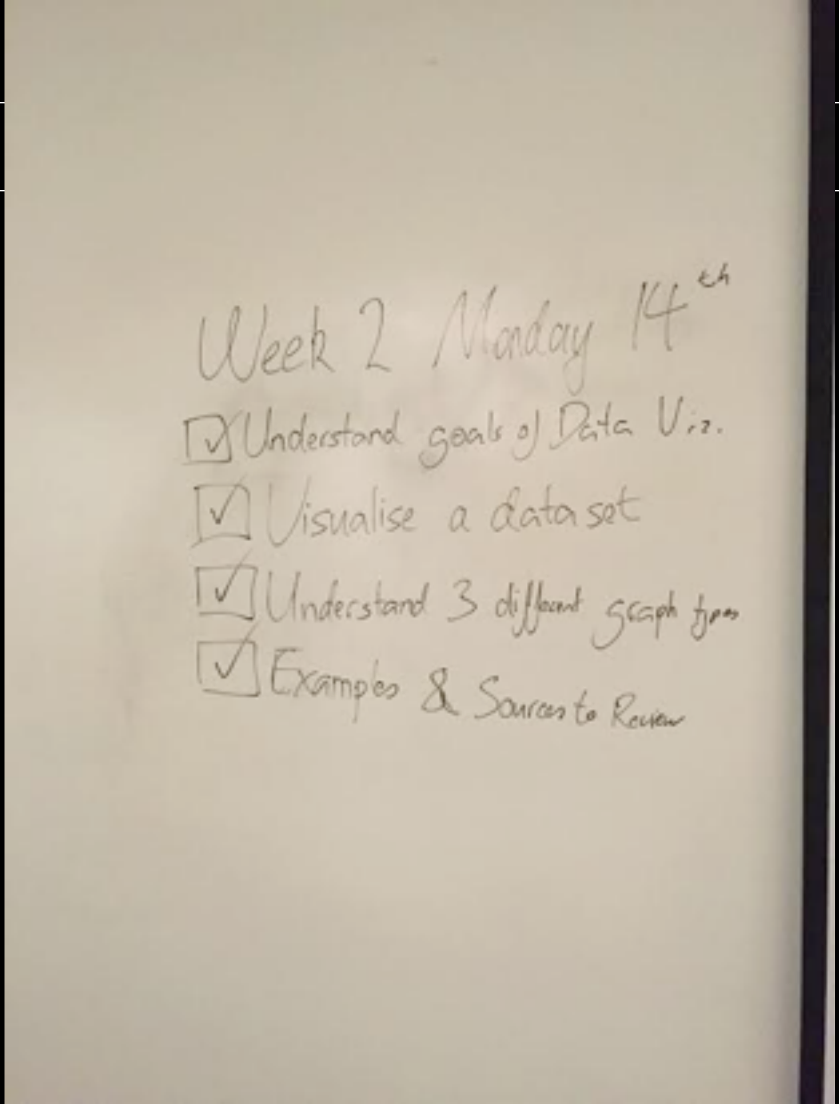
```
git remote -v
git remote add upstream https://github.com/ihansel/SYD_DAT_3.git
git remote -v
git fetch upstream
git checkout master
git merge upstream/master
```

DATA SCIENCE – Week 2 Day 2

DISCUSSION TIME

- Review of last class
- Further Reading for Regression
- Check in with homework/course project

WEEK 2 – Wednesday



Week 2 Monday 14th

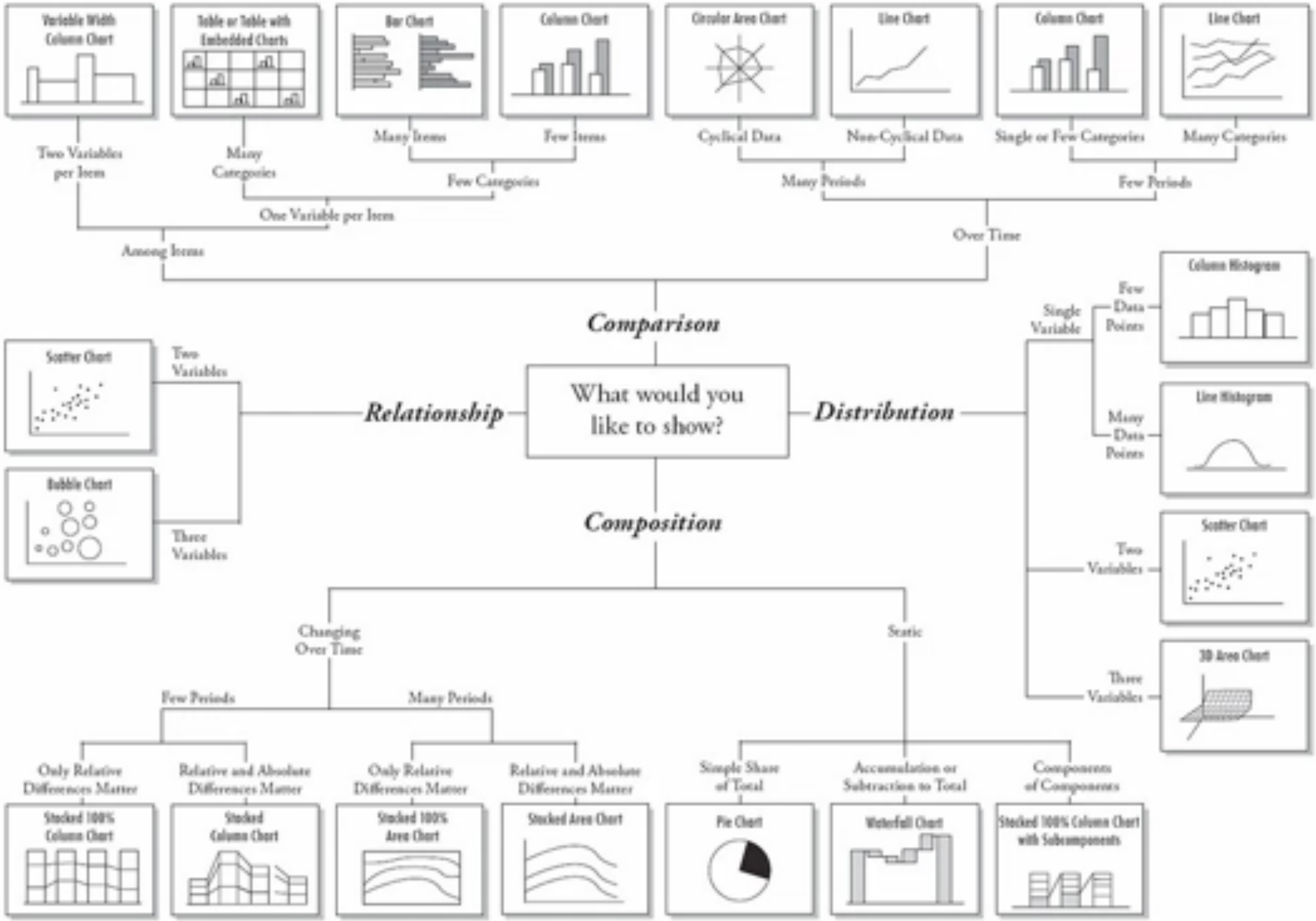
☑ Understand goals of Data Viz.

☑ Visualise a dataset

☑ Understand 3 different graph types

☑ Examples & Sources to Review

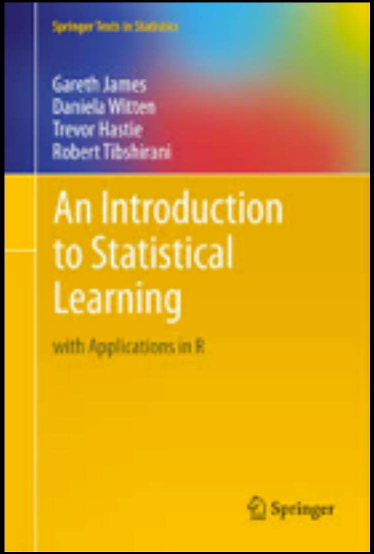
Chart Suggestions—A Thought-Starter



DISCUSSION TIME

An Introduction to Statistical Learning

- Chapter 3 – Linear Regression
- Chapter 6 – Linear Model Selection and Regularization



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

DISCUSSION TIME

Homework/Course Project

- How’s it going ?