MSDS 660 — Week 1 Intro: ISLR

Sean O'Malley

3/19/2017

- Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised.
- Statistical learning involves building a statistical model for predicting, or estimating an output based on one or more inputs.

Wage Data

- We wish to understand the association between an employee's **age** and **education**, as well as the calendar **year** on his/her **wage**.

First, lets import and view the data we have to work with

We see that Wage data involves predicting a continuous or quantitative output value, often referred to as a regression problem

```
data(Wage)

glimpse(Wage)
```

```
## Observations: 3,000
## Variables: 12
## $ year       <int> 2006, 2004, 2003, 2003, 2005, 2008, 2009, 2008, 200...
## $ age        <int> 18, 24, 45, 43, 50, 54, 44, 30, 41, 52, 45, 34, 35,...
## $ sex        <fctr> 1. Male, 1. Male, 1. Male, 1. Male, 1. Male, 1. Ma...
## $ maritl     <fctr> 1. Never Married, 1. Never Married, 2. Married, 2....
## $ race       <fctr> 1. White, 1. White, 1. White, 3. Asian, 1. White, ...
## $ education  <fctr> 1. < HS Grad, 4. College Grad, 3. Some College, 4....
## $ region     <fctr> 2. Middle Atlantic, 2. Middle Atlantic, 2. Middle ...
## $ jobclass   <fctr> 1. Industrial, 2. Information, 1. Industrial, 2. I...
## $ health     <fctr> 1. <=Good, 2. >=Very Good, 1. <=Good, 2. >=Very Go...
## $ health_ins <fctr> 2. No, 2. No, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Y...
## $ logwage    <dbl> 4.318063, 4.255273, 4.875061, 5.041393, 4.318063, 4...
## $ wage       <dbl> 75.04315, 70.47602, 130.98218, 154.68529, 75.04315,...
```

Given the age and wage data in a plot

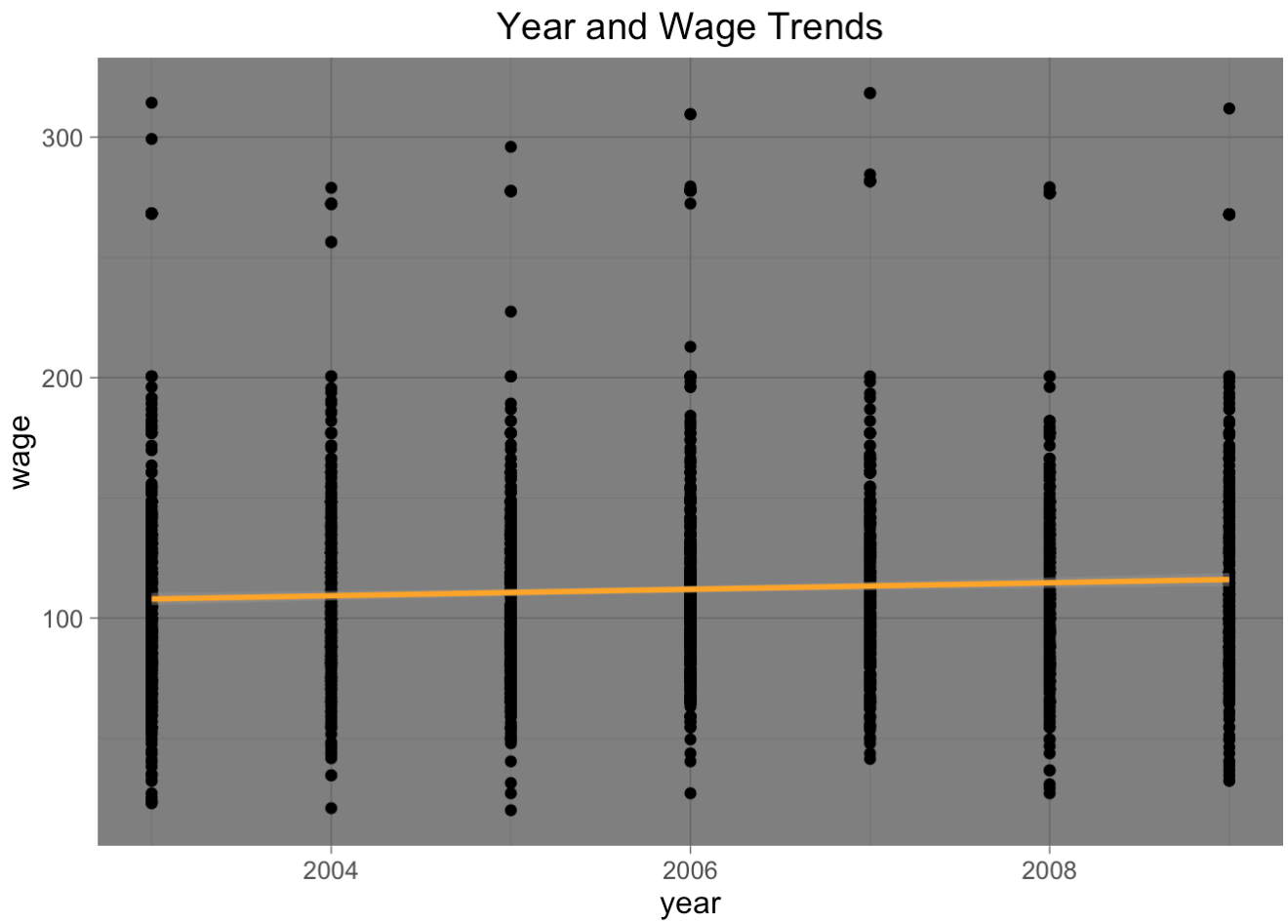
We see there is a significant amount of variability associated with this average value, and so age alone is unlikely to provide an accurate prediction of a particular man's age

```
ggplot(data = Wage,
        aes( x = age, y = wage)) +
        geom_point() +
        geom_smooth(method = "auto", color = "#40E0D0") +
        ggtitle("Age's Affect On Wage") +
        theme_dark()
```

How does **wage** data trend over the **year** variable in a plot?

Our visualization below shows us that wages increase by approximately $10k linearly between 2003 and 2009

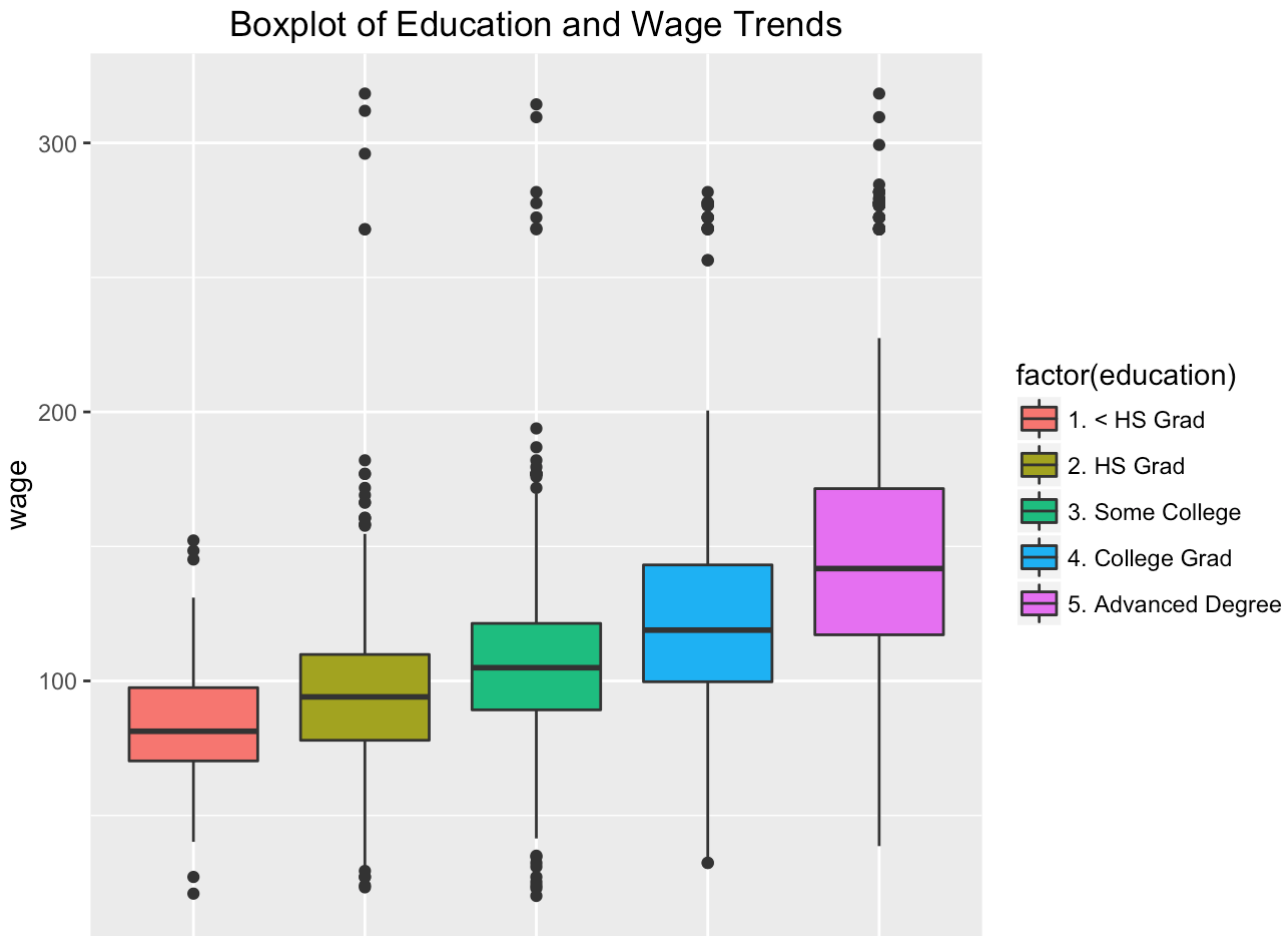
```
ggplot(data = Wage,
        aes( x = year, y = wage)) +
        geom_point() +
        geom_smooth(method = "lm", color = "orange") +
        ggtitle("Year and Wage Trends") +
        theme_dark()
```



How does **wage** data trend over the **year** var

This boxplot plainly illistrates that wages are typically greater for those with higher education values

```
ggplot(data = Wage,
        aes( x = factor(education), y = wage)) +
        geom_boxplot(aes(fill = factor(education))) +
        ggtitle("Boxplot of Education and Wage Trends") +
        theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=el
ement_blank())
```



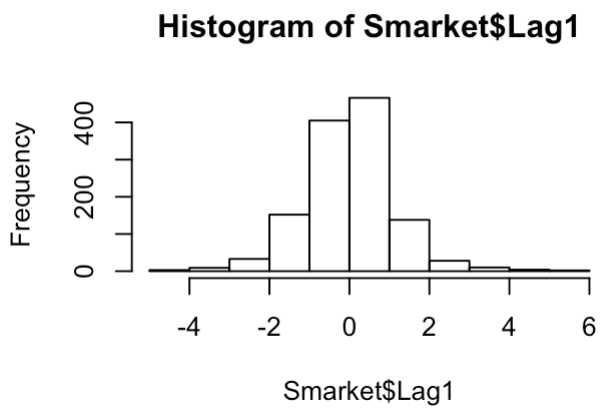
Stock Market Data

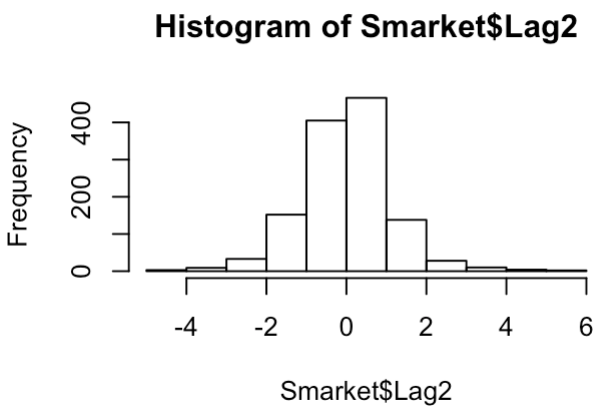
- We examine a stock market data that only contains daily movements in the S&P 500 index over a 5 year period from 2001-2005
- The goal is to predict whether the index will increase or decrease given the past 5 days' percentage changes in the index
- This involves predicting whether the stock value will fall in the **up** value or **down** value, this is classification

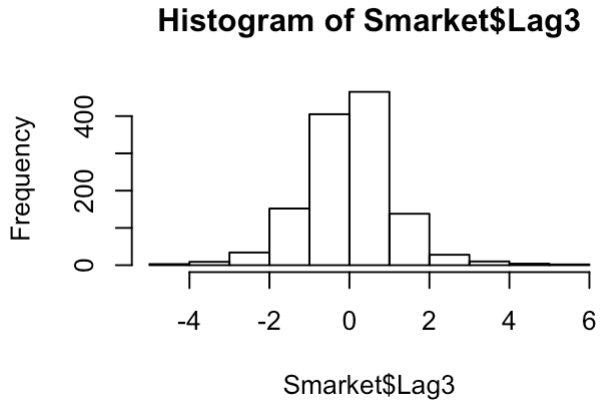
```
data(Smarket)

glimpse(Smarket)
```

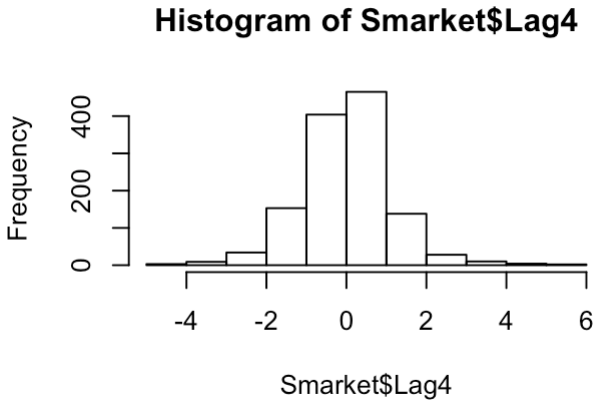
```
## Observations: 1,250
## Variables: 9
## $ Year      <dbl> 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001...
## $ Lag1      <dbl> 0.381, 0.959, 1.032, -0.623, 0.614, 0.213, 1.392, -0...
## $ Lag2      <dbl> -0.192, 0.381, 0.959, 1.032, -0.623, 0.614, 0.213, 1...
## $ Lag3      <dbl> -2.624, -0.192, 0.381, 0.959, 1.032, -0.623, 0.614, ...
## $ Lag4      <dbl> -1.055, -2.624, -0.192, 0.381, 0.959, 1.032, -0.623,...
## $ Lag5      <dbl> 5.010, -1.055, -2.624, -0.192, 0.381, 0.959, 1.032, ...
## $ Volume    <dbl> 1.1913, 1.2965, 1.4112, 1.2760, 1.2057, 1.3491, 1.44...
## $ Today     <dbl> 0.959, 1.032, -0.623, 0.614, 0.213, 1.392, -0.403, 0...
## $ Direction <fctr> Up, Up, Down, Up, Up, Up, Down, Up, Up, Up, Down, D...
```

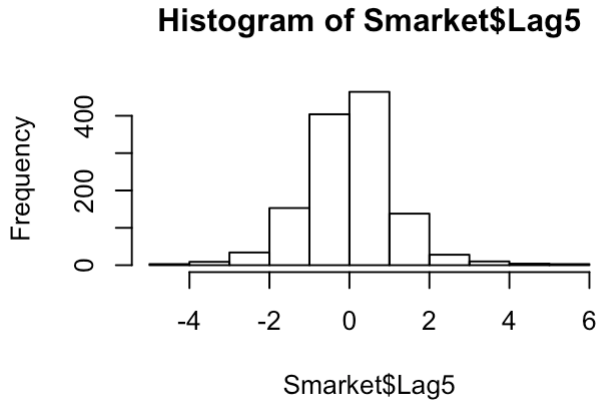
Now lets use some histograms of the data to get a better picture of what's going on with our numerical data

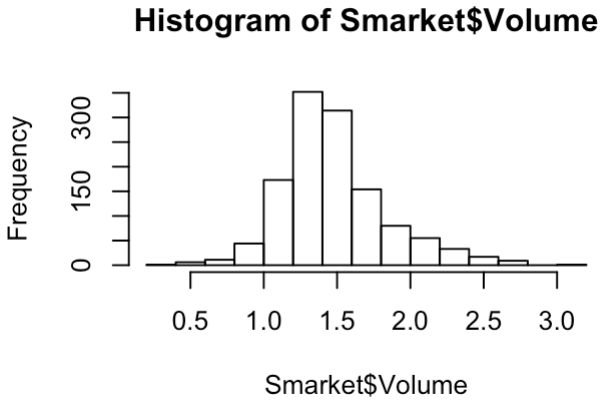

Histogram of Smarket$Lag1
Frequency
0
200
400
-4
-2
0
2
4
6
Smarket$Lag1

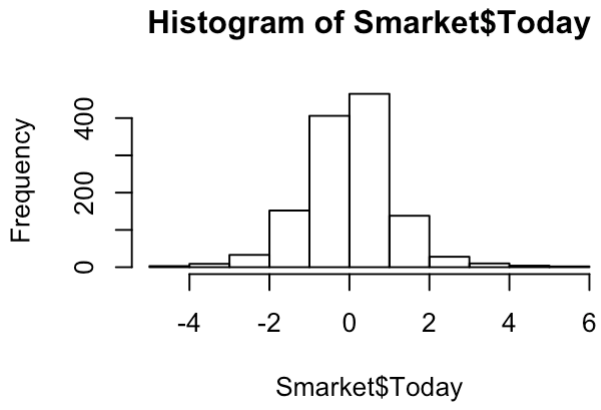

Histogram of Smarket$Lag2
Frequency
0
200
400
-4
-2
0
2
4
6
Smarket$Lag2


Histogram of Smarket$Lag3
Frequency
0
200
400
-4
-2
0
2
4
6
Smarket$Lag3


Histogram of Smarket$Lag4
Frequency
0
200
400
-4
-2
0
2
4
6
Smarket$Lag4


Histogram of Smarket$Lag5
Frequency
0
200
400
-4
-2
0
2
4
6
Smarket$Lag5

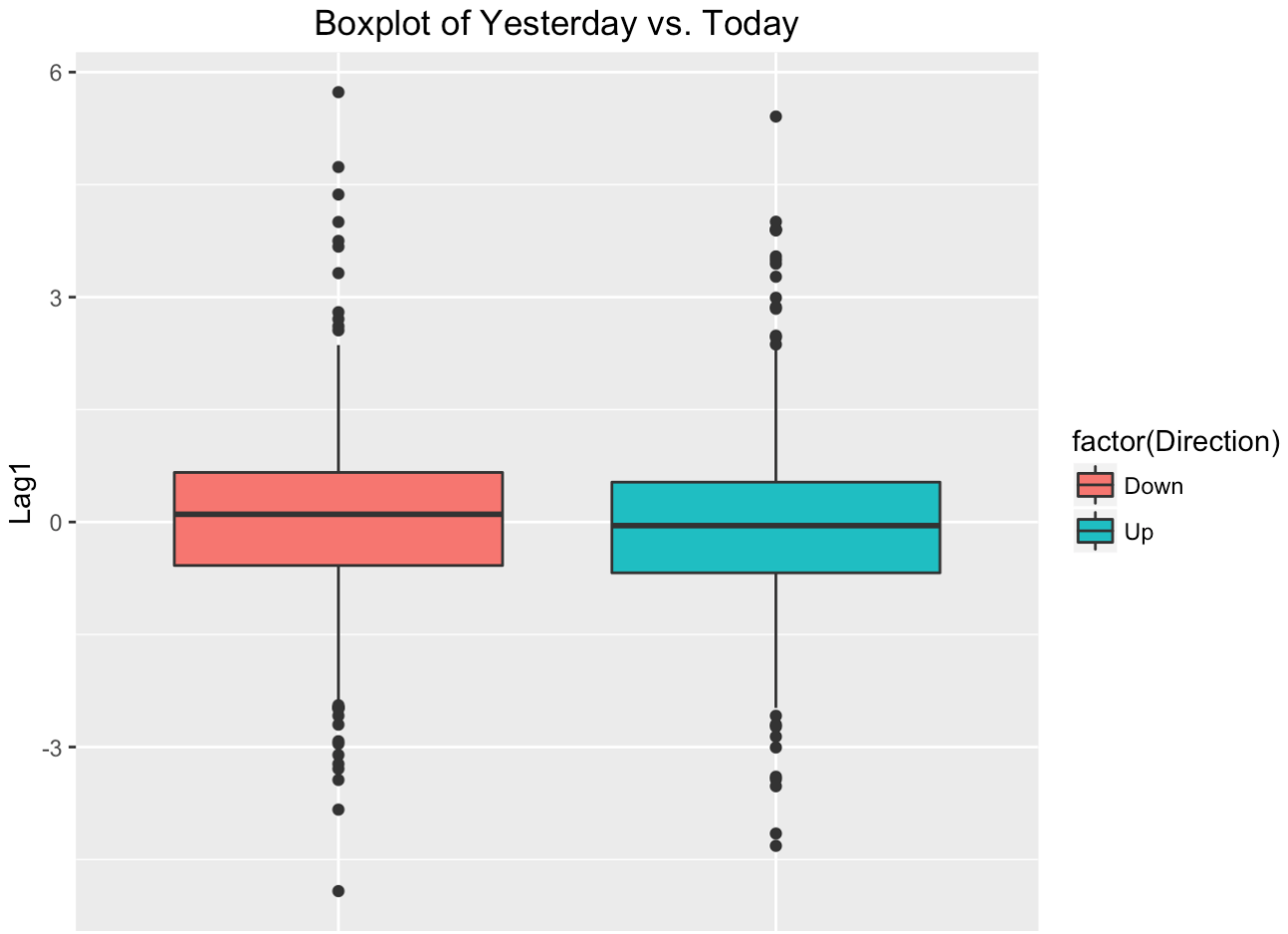

Histogram of Smarket$Volume
Frequency
0
150
300
0.5
1.0
1.5
2.0
2.5
3.0
Smarket$Volume


Histogram of Smarket$Today
Frequency
0
200
400
-4
-2
0
2
4
6
Smarket$Today

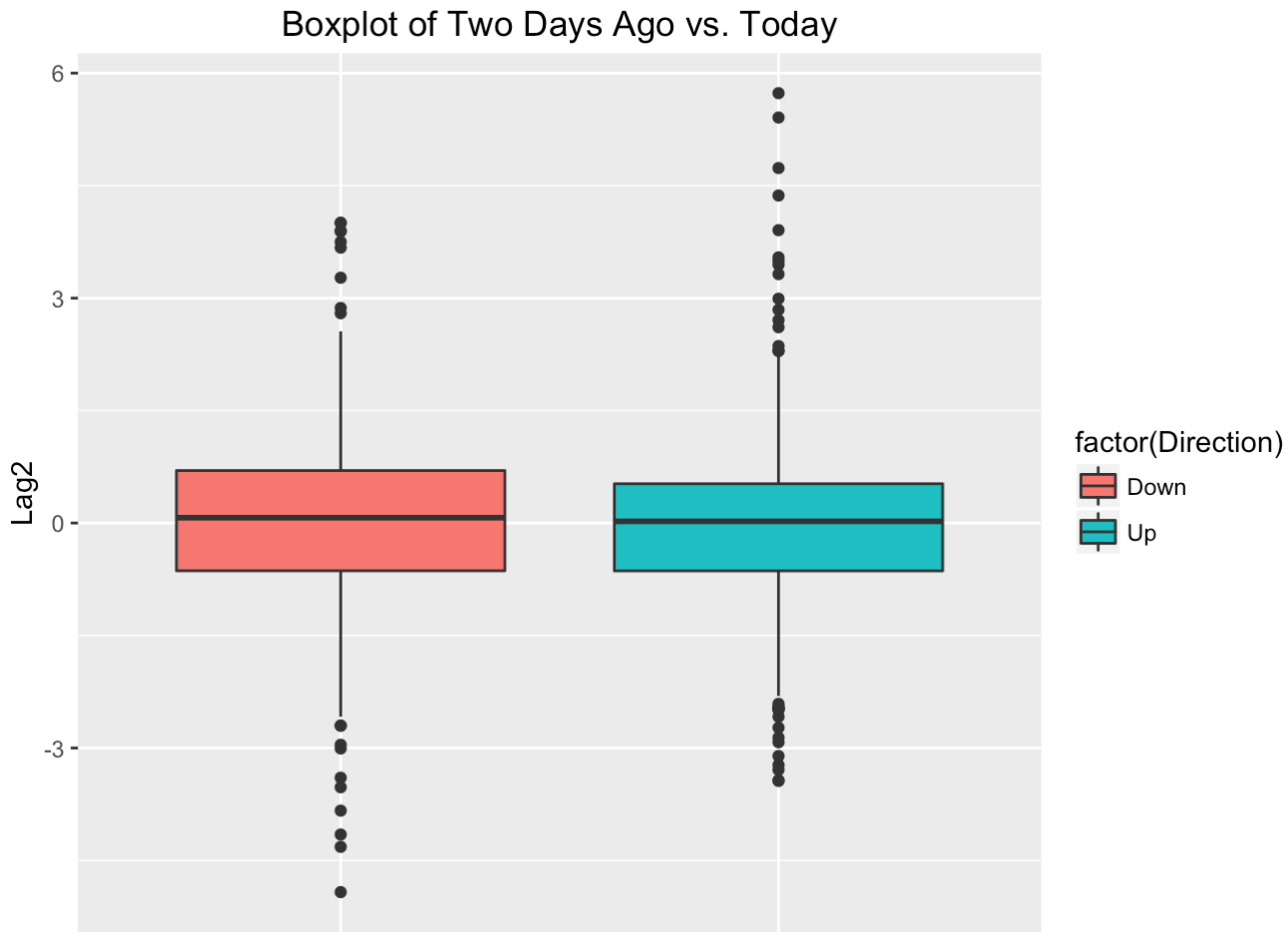
Time to visualize in some boxplots.

- The first chart displays two boxplots of the previous day's percentage changes in the stock index, one accounts for the days the mareket has went up and the others for when the market has gone down.
- The remaining two visualizations display boxplots for the percentage changes 2 and 3 days previous to today, similarly indicate little association between past and present returns.

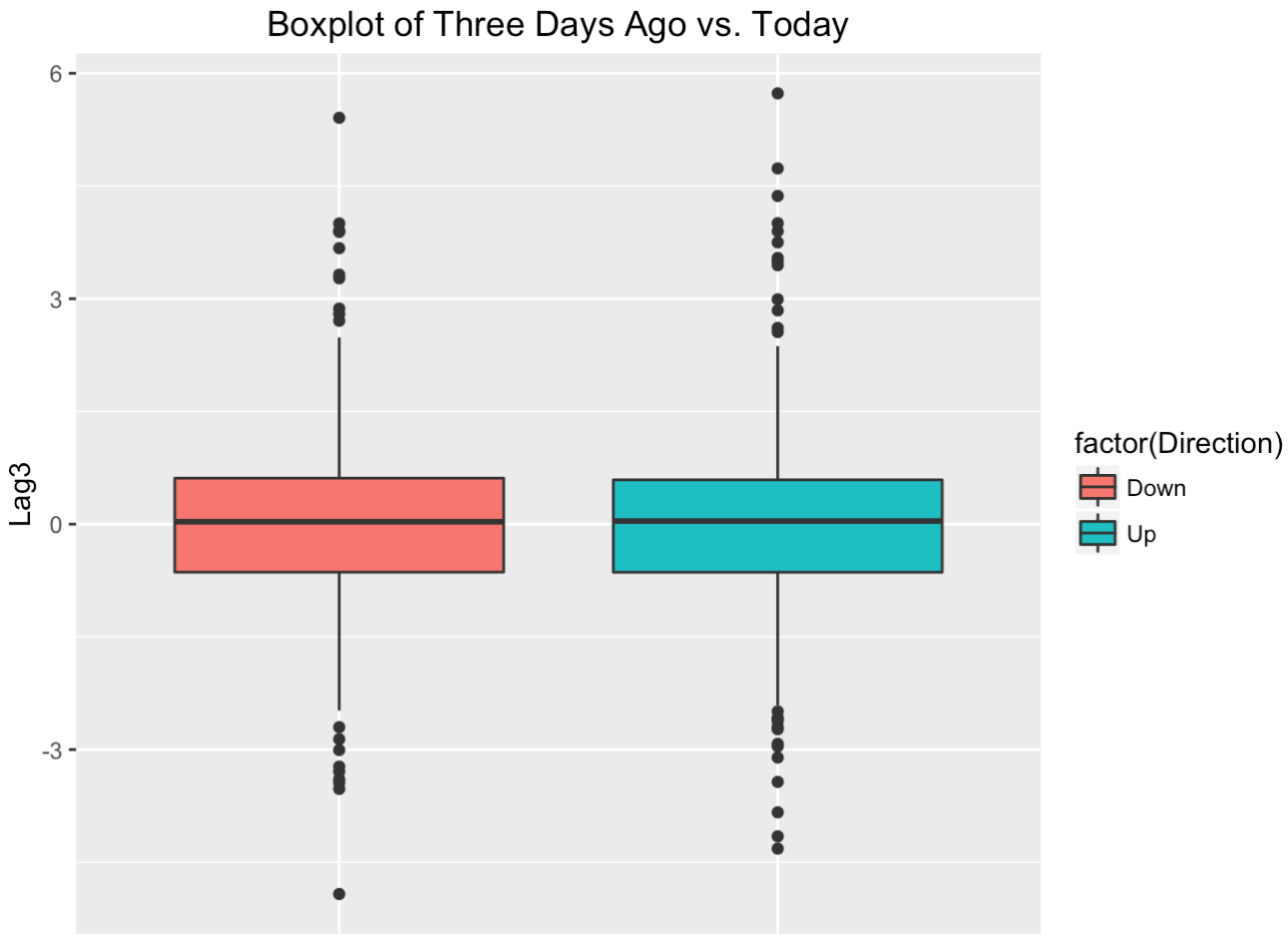
```
ggplot(data = Smarket,
        aes( x = factor(Direction), y = Lag1)) +
        geom_boxplot(aes(fill = factor(Direction), xlab = "Direction")) +
        ggtitle("Boxplot of Yesterday vs. Today") +
        theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=el
ement_blank())
```




```
ggplot(data = Smarket,
        aes( x = factor(Direction), y = Lag2)) +
        geom_boxplot(aes(fill = factor(Direction), xlab = "Direction")) +
        ggtitle("Boxplot of Two Days Ago vs. Today") +
        theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=el
ement_blank())
```




```
ggplot(data = Smarket,
        aes( x = factor(Direction), y = Lag3)) +
        geom_boxplot(aes(fill = factor(Direction), xlab = "Direction")) +
        ggtitle("Boxplot of Three Days Ago vs. Today") +
        theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=el
ement_blank())
```



What is experimental design? Why is experimental design so important?

- Experimental design refers to a plan for assigning experimental units to treatment conditions
  - **Causation** allows the experimenter to make causal inferences about the relationship between independent variables and a dependent variable
  - **Control** allows the experimenter to rule out alternative explanations due to confounding effects of extraneous variables
  - **Variability** is reduced by experimental design wihtin treatment conditions, making it easier to detect differences in treatment outcomes
- Types of Experimental Design
  - **Completely Randomized Design** objects or subjects are assigned to groups completely at random. One standard method for assigning subjects to treatment groups is to label each subject, then use a table of random numbers to select from the labelled subjects.
  - **Randomized Block Design** experimental subjects are first divided into homogeneous blocks before they are randomly assigned to a treatment group. Then, within each level, individuals would be assigned to treatment groups using a completely randomized design. In a block design, both control and randomization are considered.
  - **Matched Pair Design** A matched pairs design is a special case of a randomized block design. It can be used when the experiment has only two treatment conditions; and subjects can be grouped into pairs, based on some blocking variable. Then, within each pair, subjects are randomly assigned to different treatments.

What is experimental design? Why is experimental design so important?

- Given the User interface design and implementation example in (reference 5), user flows are something I live in on a day to day performing data science in the advertising world.
- Its always good to re-iterate the importance of t test (which asks whether the mean of one condition is different from the mean of another condition) and ANOVA (which asks the same question when we have the means of three or more conditions). Their definition in combination with the conversation around significance levels using p-value, confidence intervals and single or double tailed tests are integral to intro statistics, and immediately serves as a reminder that though statistics relies on “hard” math, that there is still a vast amount of relativity involved.
- In our work we additionally use experimental design usefulness in digital media channels. We control for certain groups, often performing randomized block design across funnel and affinity categories, and then perform ANOVA testing to determing if our advertising dollars has been effective in boosting sales.

Sources:

1. **An Introduction to Statistical Learning**, *Chapter 1: Introduction*, Gareth James, Daniella Witten, Trevor Hastie, Robert Tibshirani
2. Stattrek Experimental Designlink (http://stattrek.com/experiments/experimental-design.aspx?Tutorial=AP)
3. Stattrek Matched Pairslink (http://stattrek.com/statistics/dictionary.aspx?definition=Matched%20pairs%20design)
4. Stats at Yale Experimental Designlink (http://www.stat.yale.edu/Courses/1997-98/101/expdes.htm)
5. MIT User Interface and Statistics link (https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-831-user-interface-design-and-implementation-spring-2011/lecture-notes/MIT6_831S11_lec15.pdf)