Data Science HW #5 - Max Greene

Load libraries:

library(tidyverse)

## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts -------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(nycflights13)

Import data from “salaries.csv”.

if(isTRUE(file.exists("Salaries.csv"))){  
salaries <- read.csv("Salaries.csv")  
}else{  
salaries <- read.csv("https://raw.githubusercontent.com/jmgraham30/UoSDataSci/master/code/baseballSalaries/Data/Salaries.csv")}

# Question #1

Create a data.frame hp\_df from salaries consisting of the highest paid players from each team for each year.

hp\_df <- top\_n(group\_by(salaries,teamID,yearID),1)

## Selecting by salary

# Question #2

flights\_df <- mutate(flights,  
 true\_dt = floor(dep\_time/100)\*60 + dep\_time%%100,  
 true\_sch\_dt = floor(sched\_dep\_time/100)\*60 + sched\_dep\_time%%100)

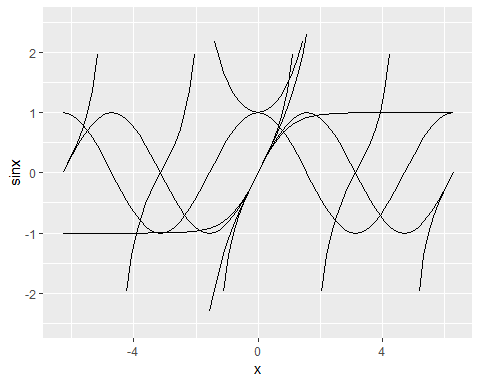
# Question #3

R offers regular trig functions, sin(),cos(),tan(), as well as hyperbolic trig functions, sinh(), cosh(),tanh().

x <- seq((-2\*pi),(2\*pi),by = pi/20)  
sinx <- sin(x); cosx <- cos(x); tanx <- tan(x)  
sinhx <- sinh(x); coshx <- cosh(x); tanhx <- tanh(x)  
  
ggplot(data = data.frame(sinx,cosx,tanx,sinhx,coshx,tanhx)) +   
 geom\_line(mapping = aes(x = x, y = sinx)) +   
 geom\_line(mapping = aes(x = x, y = cosx)) +  
 geom\_line(mapping = aes(x = x, y = tanx)) +   
 geom\_line(mapping = aes(x = x, y = sinhx)) +  
 geom\_line(mapping = aes(x = x, y = coshx)) +  
 geom\_line(mapping = aes(x = x, y = tanhx)) +  
 scale\_y\_continuous(limits = seq(-2.5,2.5,5))

## Warning: Removed 60 rows containing missing values (geom\_path).

## Warning: Removed 62 rows containing missing values (geom\_path).



# Question #4

5 different ways to assess the typical delay characteristics of a group of flights.

1. A geom\_freqpoly can be used to visualize the dep\_delay column for each group of flights, possibly fit into a normal curve or distribution.
2. The percent of flights delayed more than a certain time amount. (e.g. what percent of flights are delayed by more than 15 minutes?)
3. The amount of time at which a certain percent of flights are delayed by more than. (e.g. by how many minutes are 50% of flights delayed on average? somewhat like a S.D.)
4. The percent of flights not delayed at all.
5. Simply the average flight delay.

In reality, the arrival delay is more important to the passenger. Some flights make up time on their flight, relative to the scheduled time of arrival, so the departure delay is not necessarily how late the flight will be.

byflight <- group\_by(flights\_df,flight)  
byflight$year <- byflight$month <- byflight$day <- byflight$carrier <- byflight$tailnum <- byflight$origin <- byflight$dest <- byflight$air\_time <- byflight$distance <- byflight$hour <- byflight$minute <- byflight$time\_hour <- byflight$dep\_time <- sched\_dep\_time <- NULL

# Question #5

## Problem Description and Objectives

### Background

Cryptocurrency is a digital form of currency that has become increasingly popular in the past few years. Market prices are determined by trade values between individuals, which can vary with quite high volatility as supply and demand change. Data had become available on trade history and market prices over the past few years. The transparancy of the market value and trade prices make the price changes somewhat predictable over time. That said, it is possible to build a predictive model, trained on this historic data and the market listings, that can accurately predict future price changes.

### Problem Description

Due to the emotional nature of trading any assets, such as stocks, FOREX, etc. it can be decievingly difficult to make a profit off of initial investments. I believe it is possible to avoid this problem by constructing a predictive model based solely on historic market prices and market listings.

### Objectives

To build a predictive model of cryptocurrency prices using the crypto R package that provides historic and live cryptocurrency prices for a wide variety of coins.

## Data Description

The data provided by the crypto package is 24h trade volumes, open, close, high and low values for the specified market since 2010.

# Question #6

My final project folder can be found on my [GitHub Page](https://github.com/maxwellgreene/DataScience/tree/master/FinalProj_ETH) I do not have a data sub-folder because the crypto package obtains the live and historical data straight from the web.