Homework 8

Due by 9PM November 14th

Please submit R code, explanations, and / or plots for any section marked with the **[To submit]** heading by emailing jakeporway+itp@gmail.com.

REFERENCES

These aren't mandatory but might help if you get stuck:

- Go back and watch Lecture 9 online (the one I sent an email out about).
- Section 4.1 in http://brainimaging.waisman.wisc.edu/~perlman/R/Chapter%20four%20Descriptive%20statistics%20and%20R%20graphics.pdf
- An introduction to linear regression. http://scienceblogs.com/goodmath/2008/03/27/introduction-to-linear-regress/

For next time:

- Go watch this awesome intro to machine learning by Hilary Mason - http://www.hilarymason.com/presentations-2/devs-love-bacon-everything-you-need-to-know-about-machine-learning-in-30-minutes-or-less/

ASSIGNMENT

Predicting The Future!

All right, last lecture we learned a bit about how to talk about statistical distributions and saw how Galton came up with the idea of regression, our first statistical modeling technique. Linear regression describes relationships between variables that we believe to be linear (an increase in one predicts an increase in the other by some constant value) and R gives us awesomely easy functions to explore all of these things. Let's hop back over to our NYPD data to see what we can do.

Describing Distributions

If you load up the following file:

http://jakeporway.com/teaching/data/snf_5.csv

you'll be ready to go. A quick names() or head() of the data will show you that we've added some new variables to our dataset, namely "period_obs", "period_stop", "feet", and "inches". Let's explore these variables more closely.

[To submit]

1. Make a "height" column for your data that is the total number of inches tall each person is.
2. Plot and describe the variables "height", "weight", "period_obs", and "period_stop". Use terms that we learned this lesson – talk about the centers of the distributions, their shapes, and whether they're skewed or not. If they're very

skewed, try plotting a smaller subset of the data and describing that (e.g. all values less than 50).

Slippery Slopes

Let's try to find some relationships in our data. Period_obs and period_stop are the number of minutes that someone was observed by the police and how long they were subsequently stopped. Since these are very skewed, let's just look at the data where they're each less than 40 minutes. From here, we'll try to model their relationship to one another.

[To submit]

1. Create a subset of the data where period_obs and period_stop are less than 40.
2. Create a jittered() scatterplot of the data. What do you see?
3. Build a linear model predicting the period_stop variable from period_obs. What is the slope of your model? Based on your intuition, would you say this is a good model?
4. Using your model, predict how long you expect someone to be stopped if they're observed for 5 minutes.
5. Using your model, predict how long you expect someone to be stopped if they're observed for 60 minutes. Even though we built the model only on data for periods < 40, we do have some data for when people were observed for 60 minutes. Compute the mean for those period_stops where period_obs = 60.

Neat, let's do the same for the other continuous variables we have in our dataset.

[To submit]

1. Create a scatterplot of the height and weight variables. Jitter() or use transparency() so we can see where the bulk of the data lies.
2. Trim your data to exclude extreme height or weight values. Write down what threshold you used.
3. Run a linear model predicting weight from height. What is the slope of that model?
4. How much do you expect someone who's 6' 0" to weight?