**Lab 2: Checking data for randomness**

**MSDS 6370**

**Objectives:**

* Introduce the student to the checking data for randomness.
* Give the student insight to subtleties in checking for randomness through demonstration.

**Introduction:**

Even though a sample design may follow all the principles of sampling statistics, the selection of the sample may not be implemented in a manner that provides a random sample. Usually the errors in implementation are inadvertent and not intentional. Even when every effort is made to implement the data collection in a rigorous many, mistakes can be made. There are some methods for checking for randomness that can be done before proceeding with the analysis.

#### Draft lottery data

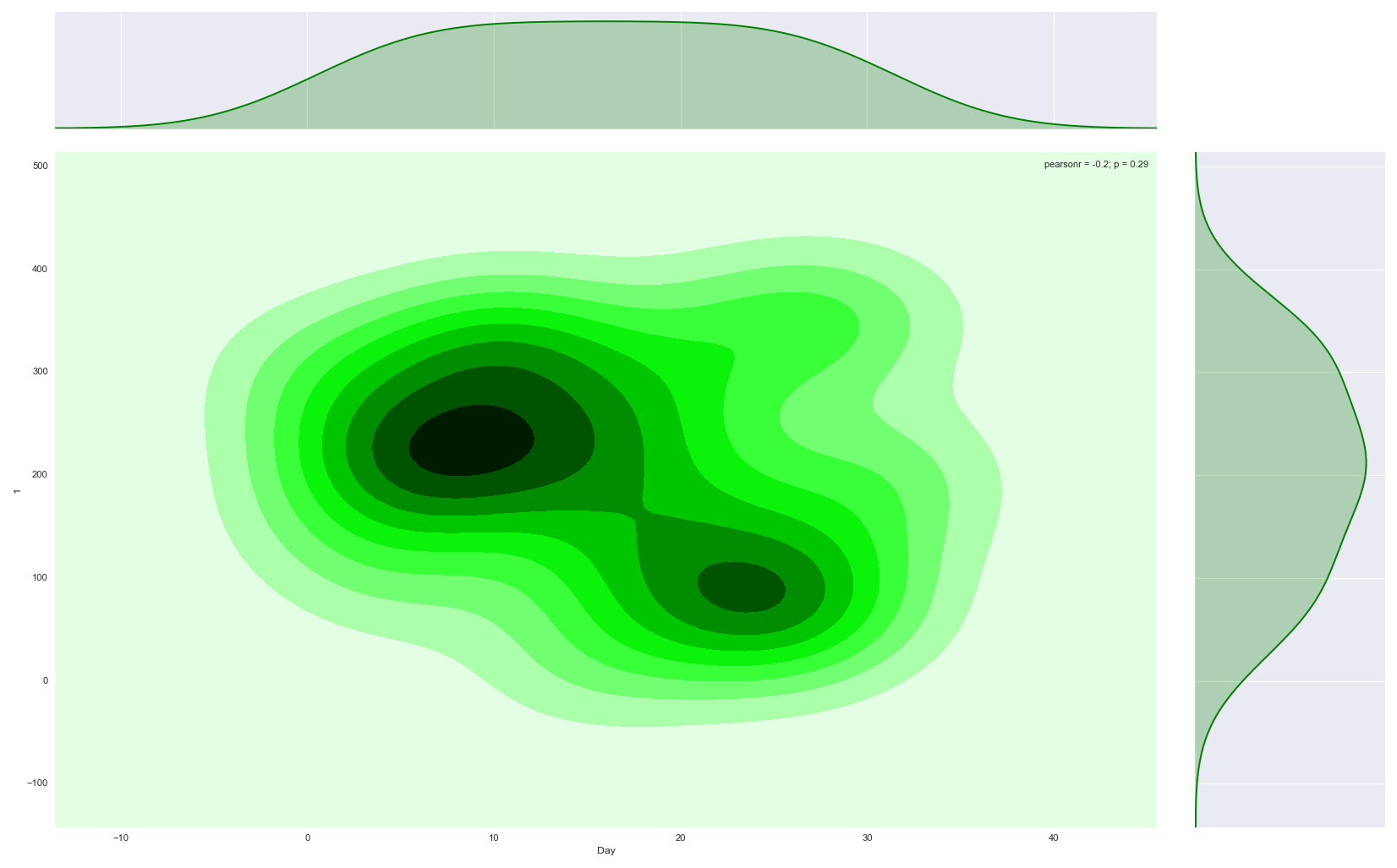
Case Study 3.1 on page 64-65 in your text *Applied Survey Sampling* discusses the U.S. Selective Service System lottery to select the numbers determining the order in which men born 1944 to 1950 would be drafted for military service beginning in 1970. Each day of the year was placed in a plastic capsule, and all the capsules were placed in a box where they were mixed up. Next the capsules were transferred to a deep glass jar. The capsules were selected one by one and assigned a draft number in order of their selection.

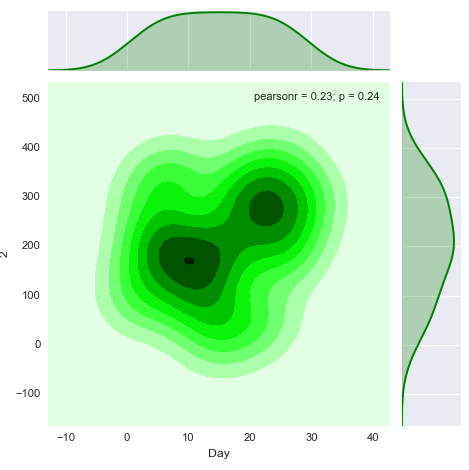
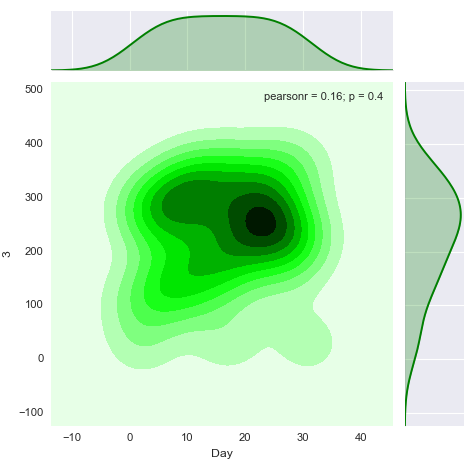
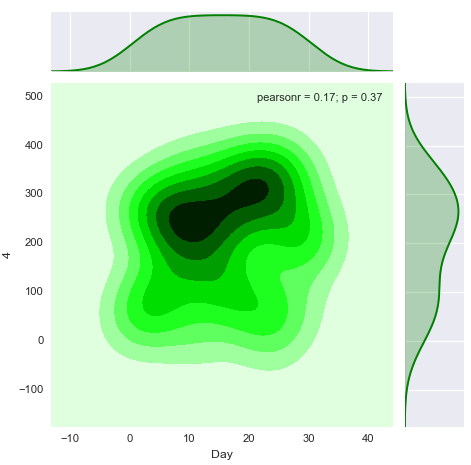
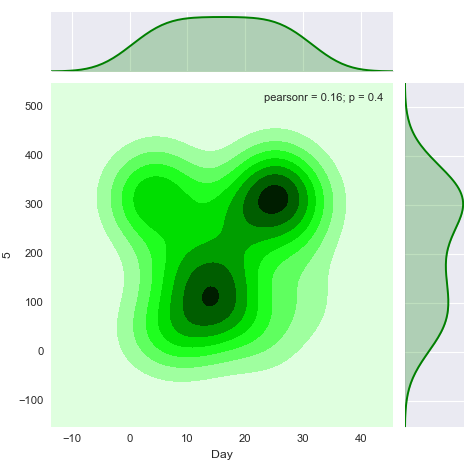
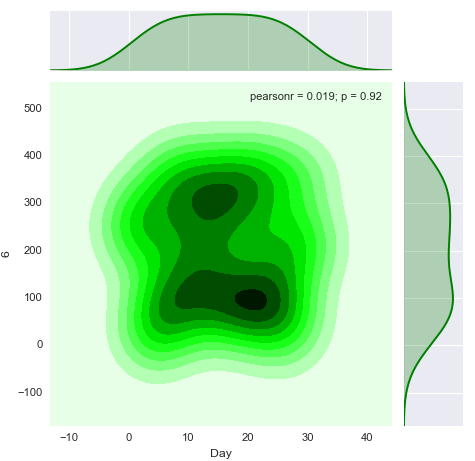
There were complaints that the lottery was not random, but the results were allowed to stand. However, the procedure changed for the selection of draft numbers for the order in which men born in 1951 would be drafted for military service beginning in 1971. The new procedure used 2 containers, one with plastic capsules, the each with one of the 365 days of the year to represent the birthdays and a second with another set of capsules with the numbers 1 to 365 to represent the order in the draft. The capsules in each container were mixed and then drawn in pairs to produce the day of the year and its order in the draft.

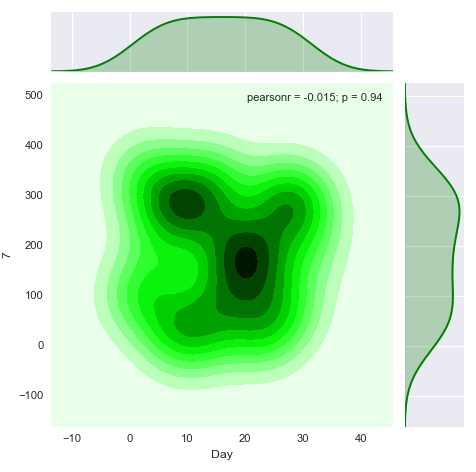
**Checking data for randomness**

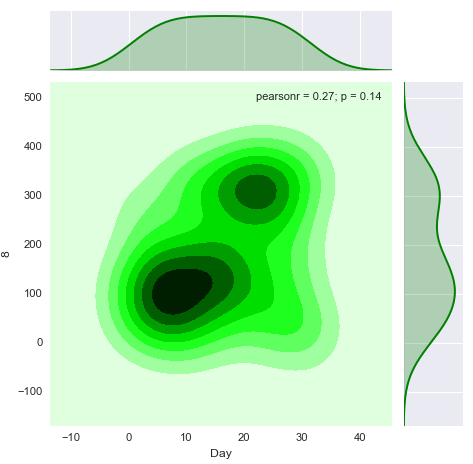
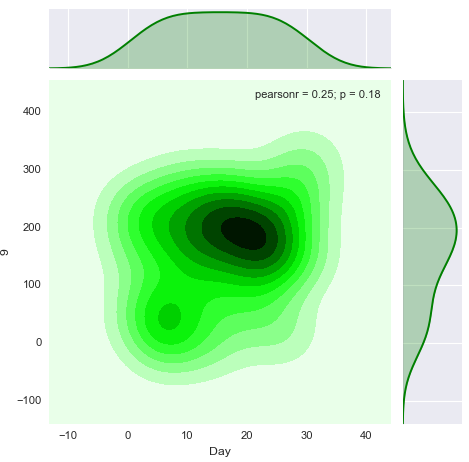
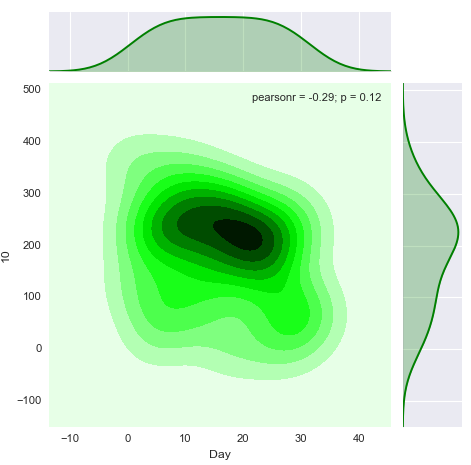
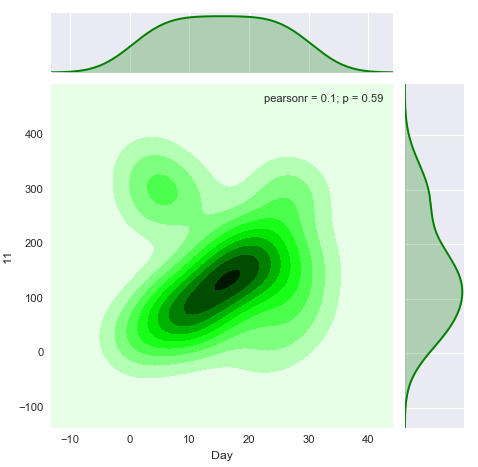
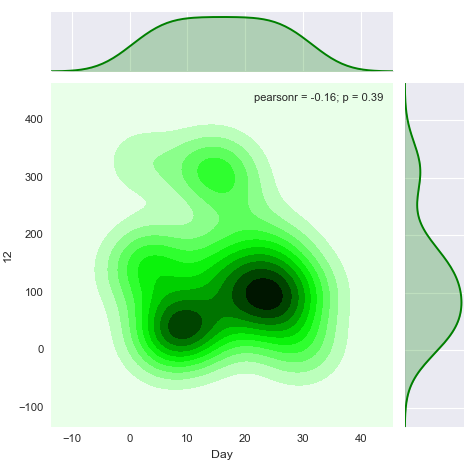
There is no formula for checking that guarantees that data were selected in a random manner. However, there are several basic checks that you can do to check for randomness. A few of these techniques:

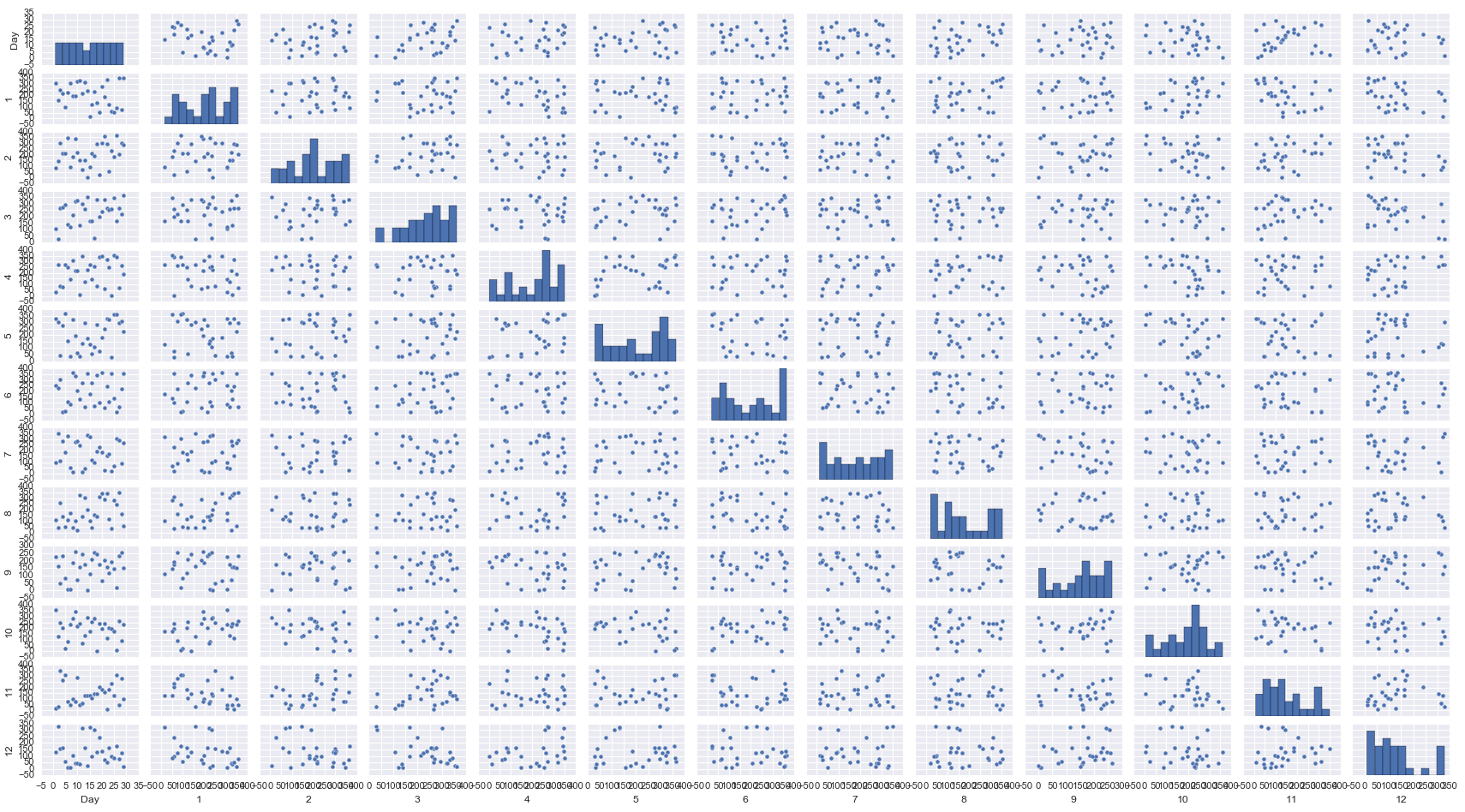
* Construct a scatter plot and calculate correlation coefficient.
* Examine distributional properties using categories of interest.
  + Plot the means within the categories to see if they are approximately equal (assuming they should be).
  + Plot the medians within the categories of interest to see if they are approximately equal (assuming they should be). Keep in mind that a median is robust to unusually large or unusually small observations while mean is not.
  + For continuous data that are expected to be normally distributed, construct a normal probability plot (not covered in this lab).







**Lab 2. Results and Exercises**

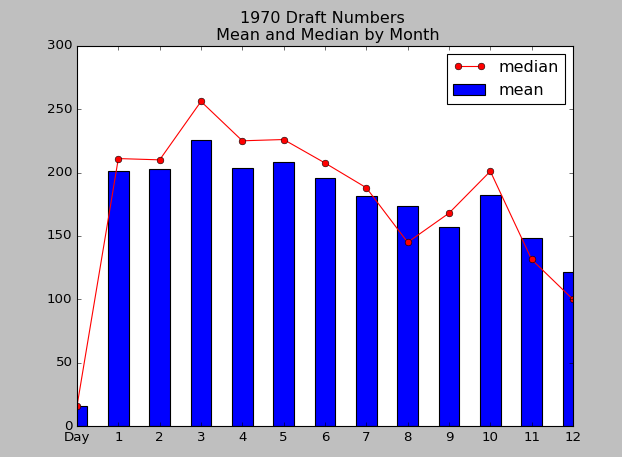
**Exercise 1 - 1970 draft numbers**

1.Using the 1970 draft numbers, calculate the following and construct the corresponding plots listed below and paste them on the Results page.

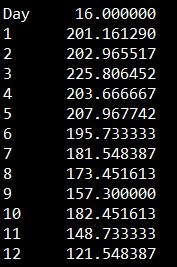
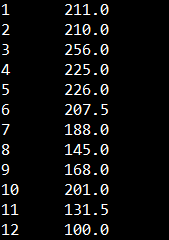
a) Construct a scatterplot of the 1970. The scatter plot should have the days of the year on the horizontal axis and the draft numbers on the vertical axis. Calculate correlation coefficient. Paste the scatterplot.

**Correlation Coefficient - 0.055**

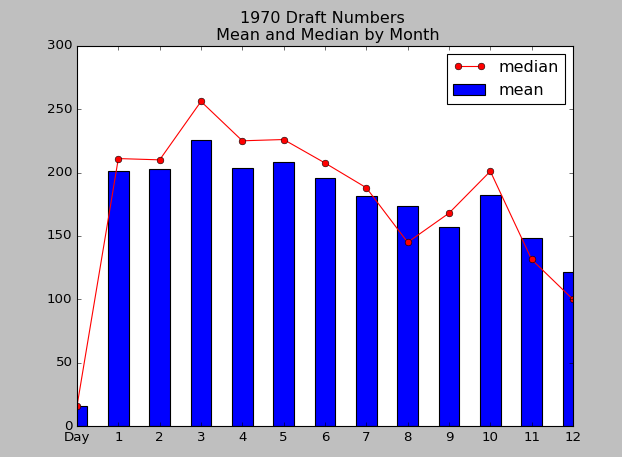
b) Calculate the mean of the draft numbers for each month. Plot these where the vertical axis is the mean and the horizontal axis is the month.



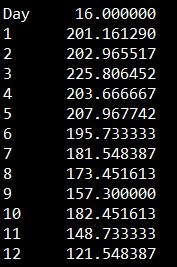
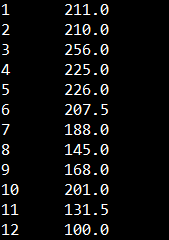
MEAN MEDIAN

c) Calculate the median of the draft numbers for each month. Plot these where the vertical axis is the median and the horizontal axis is the month.

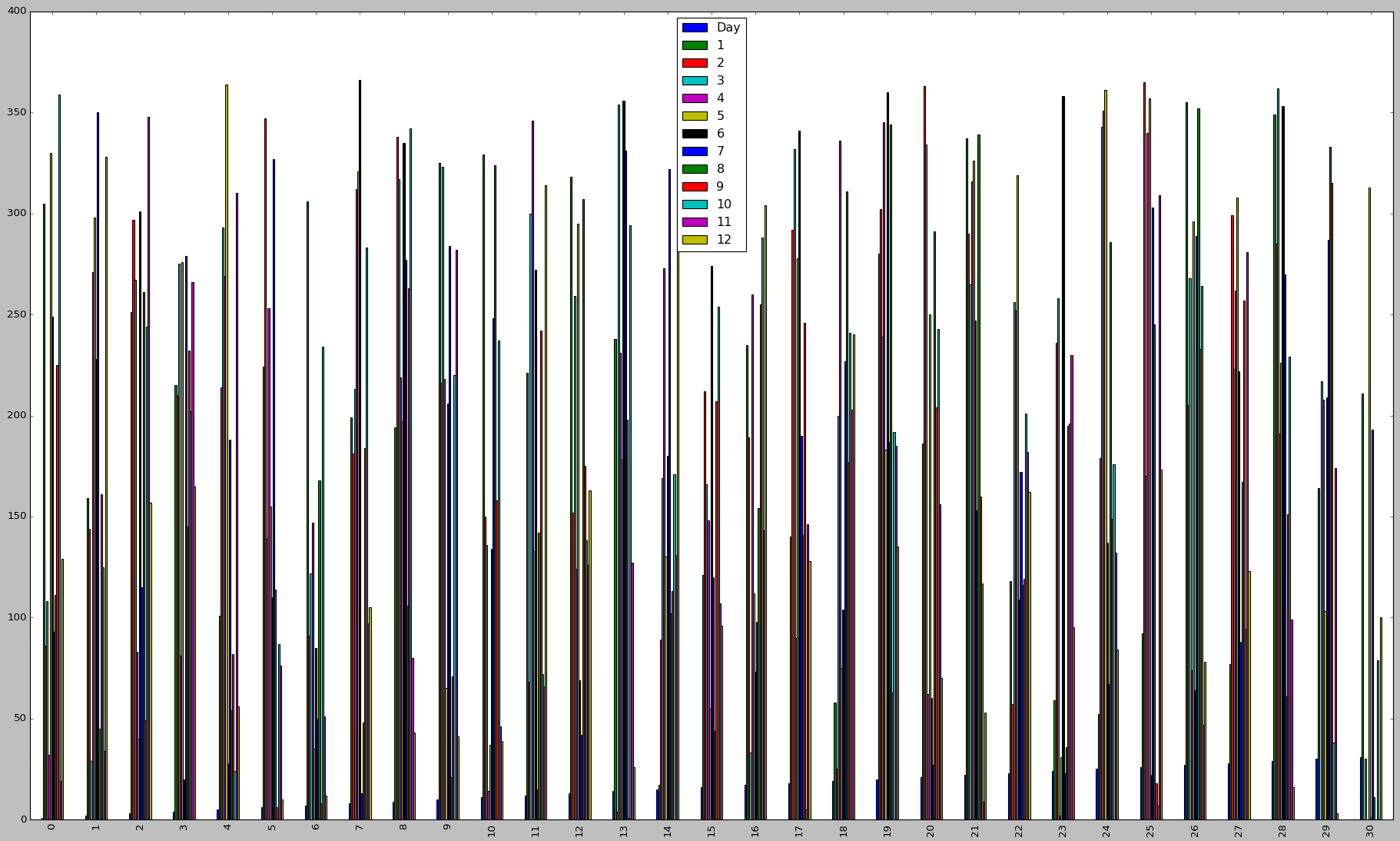


MEAN MEDIAN

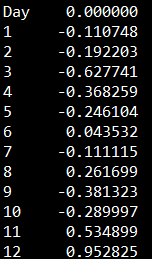
2. Decide whether the 1970 selection of draft numbers was random. Make your case using the calculations of mean and median month with corresponding plots.

Distributions broken down month by day:



A decreasing trend does not support random sampling at all. As you can even see in the skew, the distributions become increasingly one sided. This would be extremely hard to support that this draft was purely random.

SKEW

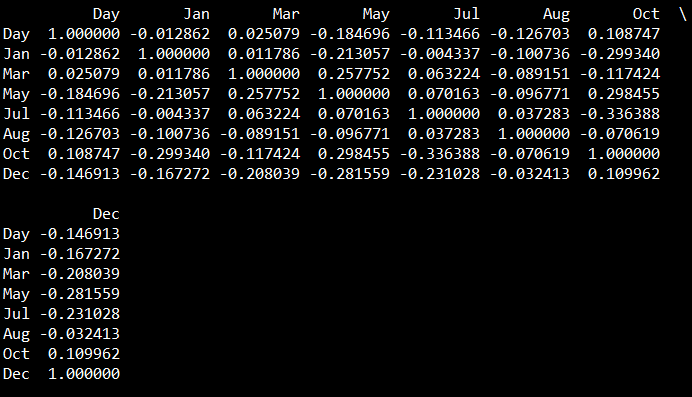


**Exercise 2 - 1971 draft numbers**

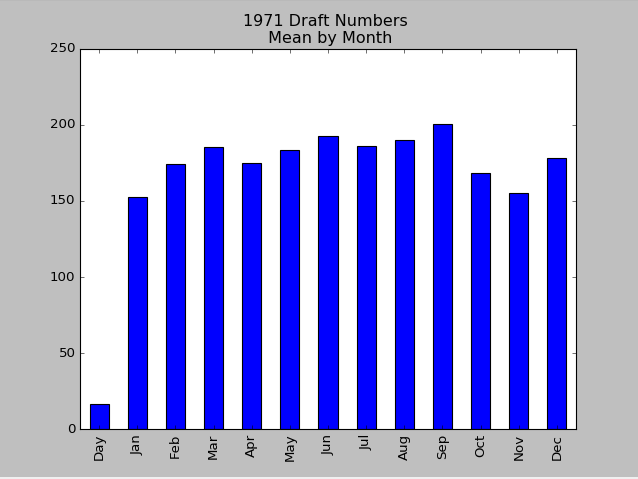
1. Construct a scatterplot of the 1971 draft numbers. The scatter plot should have the days of the year on the horizontal axis and the draft numbers on the vertical axis. Paste the scatterplot on this page. Calculate the correlation coefficient.

**Correlation Coefficient = -0.03471**

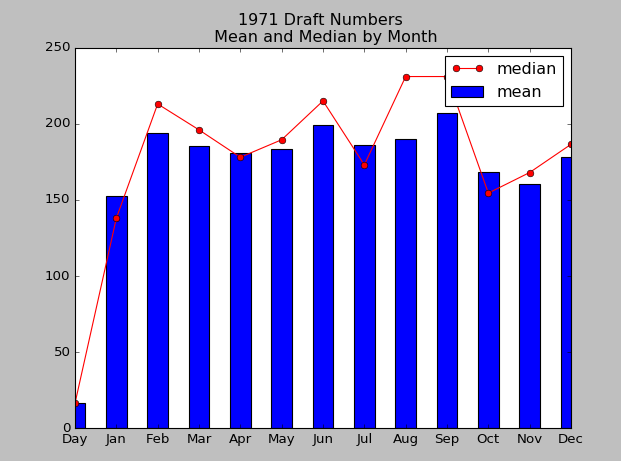
**Correlation by Month:**



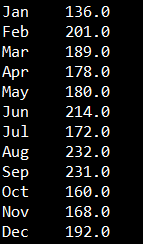
2. Sketch (or use a software package of your choice and cut-&-paste) a plot of the mean of the 1971 draft numbers by month where the vertical axis is the mean and horizontal axis is the month.



3. Sketch (or use a software package of your choice and cut-&-paste) a plot of the median of the 1971 draft numbers by month where the vertical axis is the median and horizontal axis is the month.

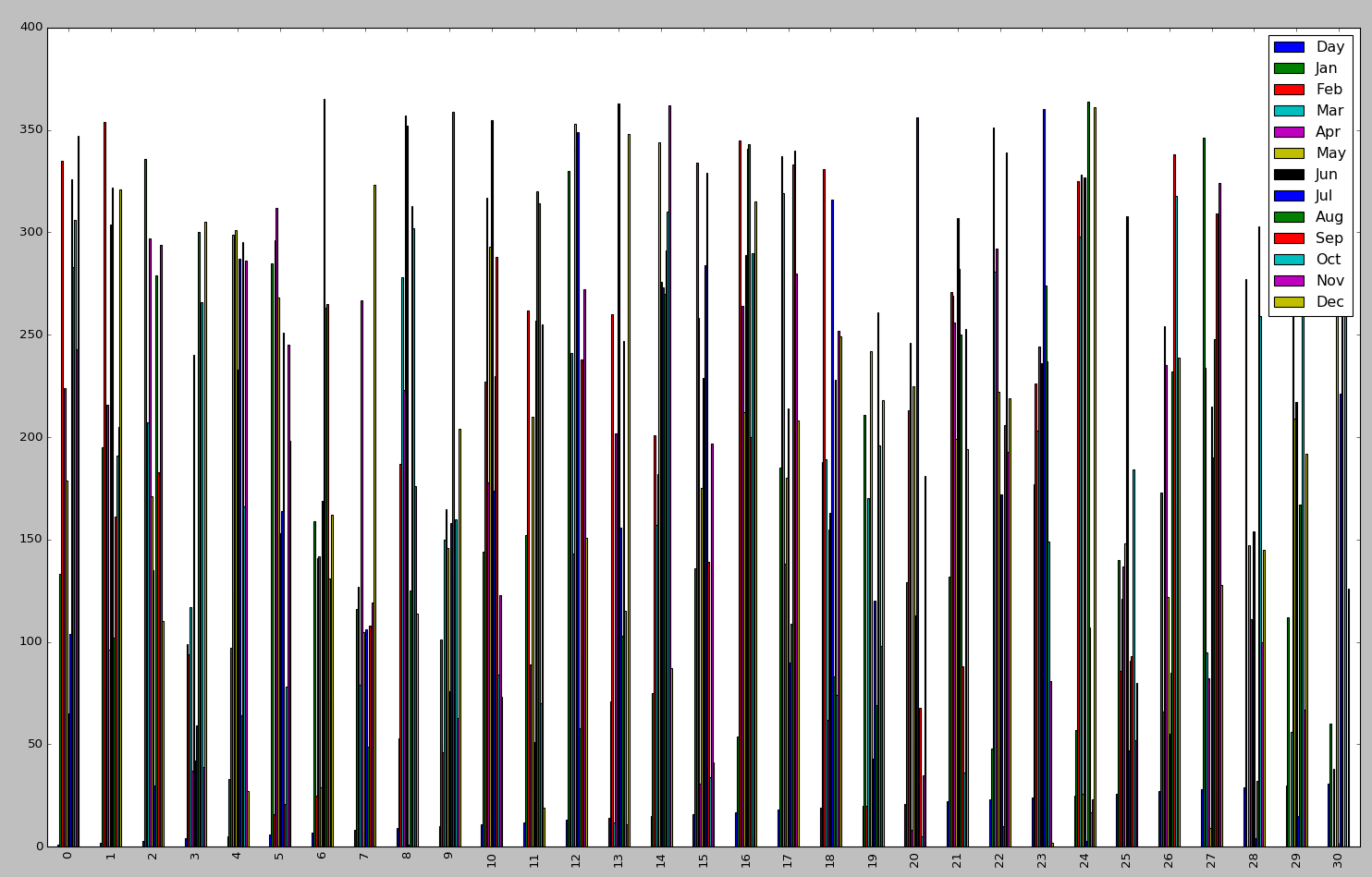


MEAN MEDIAN

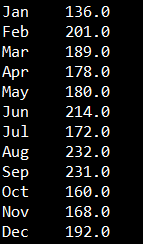


4. Decide whether the 1971 selection of draft numbers was random. Make your case using the calculations of mean and median by month with corresponding plots and the scatterplot of 1971 data.

Month Distributions Overall by each day:



MEAN MEDIAN SKEW

I don’t believe the draft was 100% random given the Skewness of the distribution compared to other months. Given the nature of selecting at random, we should observe a consistent measure of skew around zero in which we don’t always observe.

