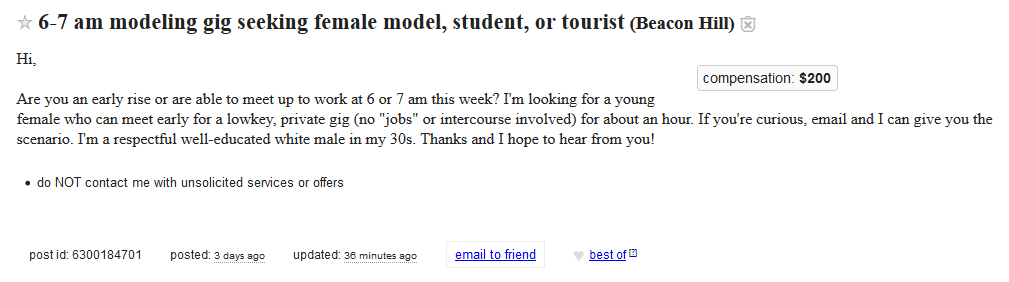
**Exploration**

My first step was to look at the craigslist page to figure out the kind of data that I would be scraping and potential pitfalls.

The first thing to note is that on most of the gigs posted, there is a <span> that contains compensation. Some of them have a dollar amount listed so it is straight forward. Others list a hourly wage, which is also straight forward, but I would have to do some finagling to estimate how much the particular gig would pay. Lastly, there are some that say "negotiable" or something to that effect. I'm thinking that these would be the edge cases that I can't really do anything about for the sake of the coding exercise.



**Data collection**

The layout of the Craigslist gigs page is simple. There is the main listings view that gives clickable urls for individual listings. 120 listings are available per page, and passing a url variable (eg. s=120) will get to the next page. I scrape through all these pages to make a master list of all listing urls. Then, I loop through the list of urls to scrape each individual listing page and scrape the data I need (title, posting date, compensation). This is written to an excel file.

**Looking at the data**

Identify potential challenges:

* some compensations use "k" to mean thousand. I will need to interpret these correctly.
* some compensations list a range instead of a single number. I will take the average of the range that was provided
* some are commission based gigs and say "X% of sales". Need to assume these are undefined instead of taking the numerical part of the string
* some of the listings are seeking "attractive females". They sounded a bit sketchy and could potentially involve illicit activity. I should figure out a way to flag these in my final dataset.

**Data cleaning**

Using regex, I parse out the numerical amount in the compensation field. Also, I determine if the compensation amount is a weekly, monthly, etc. rate by looking for the relevant keywords. If unspecified, I assume it will be a flat one-time fee.

Using the NLTK package, I extract out the stem words from the title of the listing and look for certain keywords. I mark "girls only" if I find language suggesting as such and "sketchy" if I find certain other words.

**Final answer**

At this point, I have my data in a clean excel format so I can use excel functions and filtering. I do a quick manual check to see if the data makes sense. I filter for compensation amounts that seem way too high (greater than $3000 a day) and double check their validity.

I add a column for "adjusted pay". This divides the compensation by 7, if it was a weekly rate and multiples by 8 (8 hours in a typical work day) if it was an hourly rate.

excel formula:

=IF(F2="hourly",E2\*8,IF(F2="weekly",E2/7,IF(F2="monthly",E2/30,E2)))

I filter by a single day to get all of the gig posting for that day and sum up the adjusted pay to get my final answer.

**Avoiding Getting IP Blocked**

I was able to scrape the first 500 or so listings without issue. However, when I tried for a larger corpus of data, I had my IP blocked. If I were to make this a more stable tool, I would implement two methods to avoid getting blocked again.

* Enforce wait times between requests
* Make use of private proxies to make requests.