Open Jobs Write Up

Eirik Iversen

5/14/2019

Open Jobs Data Framework

Data Inventory

Purpose: In 2016, the Commonwealth Center for Advanced Research and Statistics (CCARS) initiated a pilot project to create an open "real-time" data set of advertised job postings in Virginia. This data set is the initial outcome of the pilot. Work on this project is reported to be ongoing and is being conducted by the Discovery Analytics Center at Virginia Tech. The intended use of this data is to "to create applications or visualizations that can help connect Virginians to job opportunities, offer insights into the needs of employers by occupation, skills, or education requirements, or create predictive models to help Virginia determine its future needs for talent!"

Method: The Discovery Analytics Center collected, cleaned, "enriched", and de-duplicated data from three sources:

- A daily feed of jobs from the National Labor Exchange made available by the DirectEmployers Association
- A snapshot of jobs in the Virginia Workforce Connection from mid-February 2016 made available by the Virginia Employment Commission
- A feed of schema tagged jobs available through an open API built by Devis for the Veterans Job Bank.

Their major steps to combine this data into a single set are listed below:

- Mapped job postings from all three sources to the job-posting schema standard
- Enriched job postings with average wage data from the Georgetown University's Center for Education and the Workforce and job title normalization assistance from Glassdoor
- De-duplicated the data using an algorithm to identify identical job postings

Description: The data contains 846,613 job postings from various regions in Virginia from 2010 – 2017 (2010 – 2013 sparsely populated). The variables available are listed below in a table along with how many observations in that column are missing.

Timeliness: The Data offers few observations from 2010 to 2013, but from June 2014 and onwards data seems to be collected either on a monthly or daily interval. For some months, most of the job postings all have the "datePosted" variable equal to the same day. We will explore this trend later on in EDA, but overall it seems that data for job postings is collected daily. Selectivity: The population of interest are online job postings in Virginia. Open Data does offer some caveats to this population: "This data set does not cover all job openings in Virginia advertised online. Not all sources of data used to create this set are "real-time." Currently, data supplied from the Virginia Workforce Connection is a snapshot of data from a point in time. Efforts are underway to explore access to these jobs in "real-time." Additionally, the schema tagged jobs pulled into this data set are limited to those jobs tagged with a "Veteran Hiring Commitment"."

Accessibility: The data is freely available to the public at this link: http://opendata.cs.vt.edu/dataset/openjobs-jobpostings. The datasets are is JSON format.

Data Profiling

Missingness

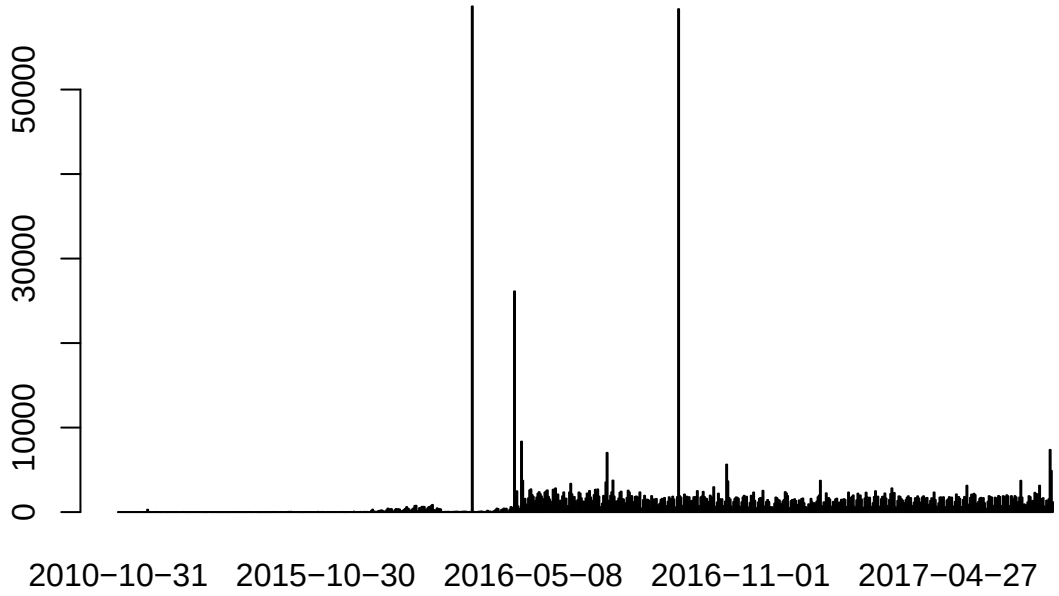
Below is a table containing the missingness and number of values missing for all the variables in openjobs:

	# of Missing or NA or Empty	Percent Missing or NA or Empty
rawdata_id	0	0.0000000
jobLocation_geo_latitude	84734	0.1000859
jobLocation_geo_longitude	84734	0.1000859
normalizedTitle_onetCode	141395	0.1670126
normalizedTitle_onetName	141395	0.1670126
datePosted	0	0.0000000
responsibilities	592184	0.6994743
experienceRequirements	667082	0.7879421
jobDescription	4	0.0000047
hiringOrg	2683	0.0031691

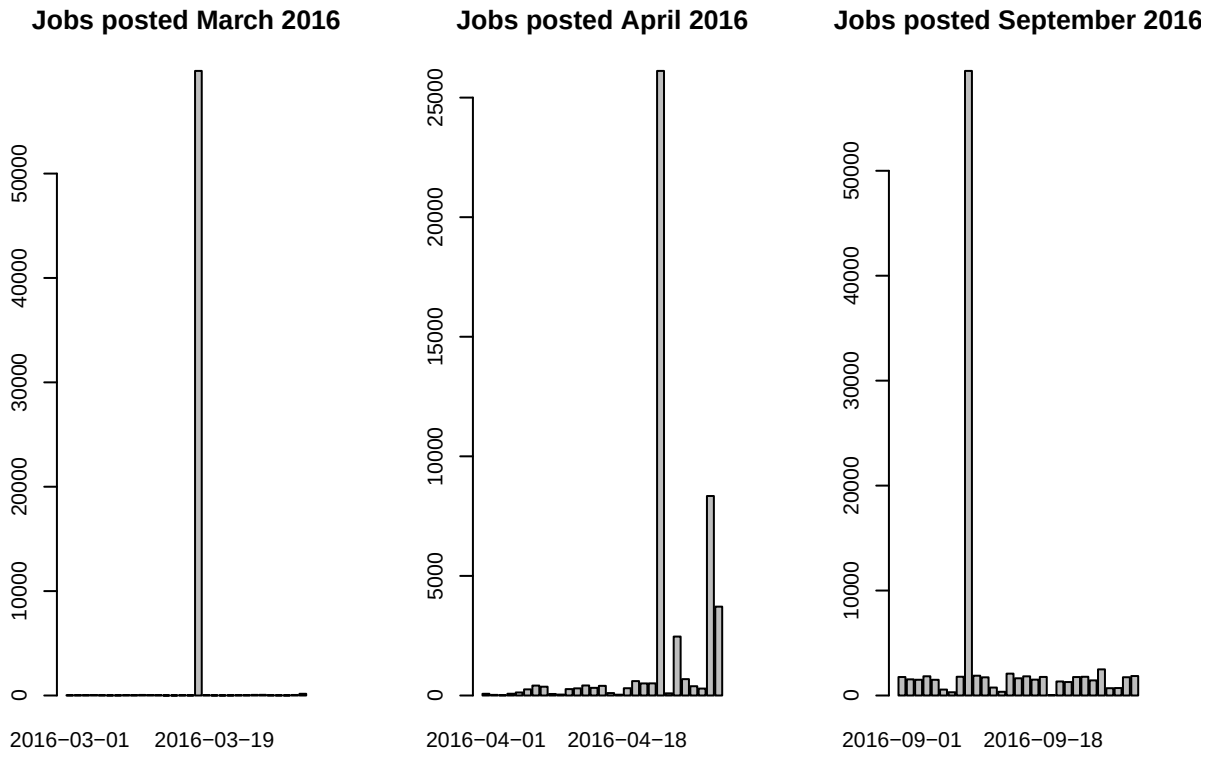
It seems that most of the fields that would be free text are the ones that are missing the most information.

Date Posted exploration

As mentioned previously, there is an interesting distribution of job postings across time:



Each bar in this plot represents a date, and the height of the bar represents how many jobs were posted on that date. This is concerning because we see there are a few dates with a suspicious amount of jobs posted. As it's hard to see what exactly is going on at this level, I have broken down the suspicious bars into their respective months to see if they're really all posted on a single day.



Sure enough, in these three months there is a day where most of the jobs are posted. My best guess for this is that there was some error in data collection that caused all of the jobs in a month to be posted on the same day, but there is no way to know for sure without contacting the original source.

Length of unique identifiers

each openjobs identifier is made of 32 characters. If we multiply that by the number of observations, then that *should* be the number of characters in the column. Let's count the number of characters in the column to double check:

```
len <- length(unique(ojobs$rawdata_id))
32*len
```

```
## [1] 27091616
```

```
32*len == (sum(as.numeric(lapply(ojobs$rawdata_id, nchar))))
```

```
## [1] TRUE
```

Duplicates

By dropping the unique identifier from the data, we can search for duplicates. A duplicate here is a row that is identical to another row in all columns other than identifier

```
data_dup <- ojobs[,-1]
data_dup <- data_dup[duplicated(data_dup)]
nrow(data_dup)
```

```
## [1] 13555
```

```r
# we see that there are 13555 duplicates overall
cleaned_data <- ojobs %>% distinct(jobLocation_geo_latitude, jobLocation_geo_longitude,
                                   normalizedTitle_onetCode,normalizedTitle_onetName,
                                   datePosted, responsibilities, experienceRequirements,
                                  jobDescription, hiringOrg, .keep_all = TRUE)
#check to see the right amount of rows removed
nrow(ojobs) - nrow(cleaned_data)
```

```
## [1] 13555
```

There are 13,555 duplicate rows in the data, which is fairly small. These duplicates could be postings for multiple openings of the same position, so we decide not to remove these.