BGT technical document

## Burning Glass Background

Burning Glass Technologies is a Boston-headquartered labor market analytics firm that uses artificial intelligence to collect and host a massive repository of workforce and employment data. The company makes use of text-parsing and predictive matching methods to extract and aggregate real-time job market data. Their unparalleled view of the job and candidate landscape allows them to partner with government agencies to understand regional employment trends, recruiters to serve candidates with meaningful insight on career exploration, and research organizations to conduct studies on changing employment trends such as skills gap analysis.

Founded in 1999, the company focuses its products primarily on six target markets: Australia, Canada, New Zealand, Singapore, the UK, and the US. Last year, Burning Glass acquired Tabulaex and expanded its reach across continental Europe. Tabulaex spun off from the University of Milano-Bicocca’s CRISP research institute (centro di ricerca interuniversitario per i servizi di pubblica utilità) and specializes in digital data methods for the private and public sector.

Burning Glass is led by CEO Matt Sigelman, who has actively voiced the following opinions. [[INSERT STUFF FROM HIS ARTICLES HERE.]]

Sources:

* burningglass.com: <https://www.burning-glass.com/blog/burning-glass-technologies-acquires-tabulaex/>
* linkedin.com: <https://www.linkedin.com/company/burning-glass-technologies/about/>
* tabulaex.com: <http://www.tabulaex.com/en/about-us/>

Matt Siegalman’s Articles: (CEO)

* <https://thehill.com/blogs/congress-blog/labor/268415-why-tpp-isnt-the-real-problem-for-american-jobs>
* <https://www.insidehighered.com/views/2016/02/08/debate-over-liberal-arts-vs-vocationalism-lazy-one-essay>
* <https://nebhe.org/journal/do-employers-value-the-bachelors-degree-too-much/>

#### Literature Background?

TBD #### Company Background

#### Data Inventory

At the data inventory stage, potential data sources (identified in the data discovery stage) are further screened to determine if they would support the research questions. Specifically, the screening process involves assessing data sources on six factors including purpose, method, description, timeliness, selectivity, and accessibility.

* Purpose: Burning Glass collects this data primarily for commercial purposes and secondarily for research purposes. The company markets both the data as a product as well as their consulting expertise on labor market questions to customers across a variety of industries, such as higher education, local and regional government, recruiting and staffing agencies, and other corporate firms.
* Method: Burning Glass scrapes over 40,000 websites to collect data from online job postings. The data is cleaned and deduplicated to present a national view of the labor market landscape across time.
* Description: The data is centered around individual positions, with one-to-many related tables for requirements for skills, certification, and education. While the data includes various types of data, the bulk of this data is text-based.
* Timeliness: The data is collected in real-time and covers 2007 and 2010-2017. [How soon after collection is data available?]
* Selectivity: The data is intended to represent the universe of all US jobs posted online over the years mentioned above.
* Accessibility: The data is accessible via file-transfer protocol, but a data-sharing agreement may fetter its accessibility outside of the lab staff.

#### Data Profiling

This section profiles the Burning Glass data, reviews its quality, and seeks to determine its useability on the following metrics.

* Completeness.
* Value validity.
* Consistency.
* Uniqueness.
* Duplication.

Here, we focus on the postings themselves which represent the central table of the data.

This first table summarizes the postings dataset. Of note, this shows that 99.2894199 % of the data are complete cases. The structure of this data includes 8 discrete columns and 5 continuous ones.

|  |  |
| --- | --- |
| Metric | Value |
| rows | 63610 |
| columns | 13 |
| discrete\_columns | 8 |
| continuous\_columns | 5 |
| all\_missing\_columns | 0 |
| total\_missing\_values | 904 |
| complete\_rows | 63158 |
| total\_observations | 826930 |
| memory\_usage | 9902352 |

This second table breaks down the postings dataset by column using a superficial check for blanks. Of note, this shows that only two columns have blanks: the FIPS county and FIPS value columns are both missing 452 observations. It is important to note, however, this table qualifies missing values as blanks or NA only.

|  |  |  |
| --- | --- | --- |
| Column | # Missing | % Missing |
| bgtjobid | 0 | 0.0000000 |
| jobdate | 0 | 0.0000000 |
| occfam | 0 | 0.0000000 |
| occfamname | 0 | 0.0000000 |
| employer | 0 | 0.0000000 |
| city | 0 | 0.0000000 |
| state | 0 | 0.0000000 |
| county | 0 | 0.0000000 |
| fipsstate | 0 | 0.0000000 |
| fipscounty | 452 | 0.0071058 |
| fips | 452 | 0.0071058 |
| lat | 0 | 0.0000000 |
| lon | 0 | 0.0000000 |

We would like also to capture invalid values in addition to the missing ones. This last table captures blanks, NA, and ‘na’ text values (Burning Glass entered). Here we see that more missing values are captured for the County column and new invalid values are captured for the Occupation Family code, Occupation Family Name, Employer, and City columns.

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Blanks | % Blank | NA |
| bgtjobid | 0 | 0.0 | 0 |
| jobdate | 0 | 0.0 | 0 |
| occfam | 0 | 0.0 | 0 |
| occfamname | 0 | 0.0 | 0 |
| employer | 0 | 0.0 | 0 |
| city | 0 | 0.0 | 0 |
| state | 0 | 0.0 | 0 |
| county | 452 | 0.7 | 0 |
| fipsstate | 0 | 0.0 | 0 |
| fipscounty | 452 | 0.7 | 452 |
| fips | 452 | 0.7 | 452 |
| lat | 0 | 0.0 | 0 |
| lon | 0 | 0.0 | 0 |

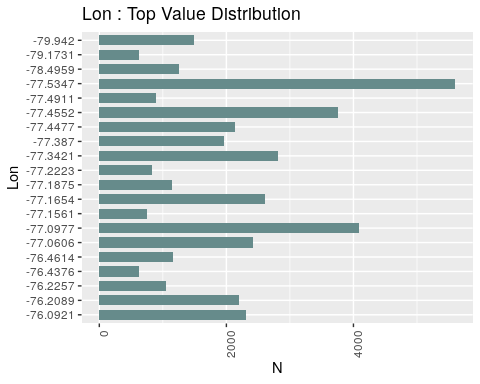
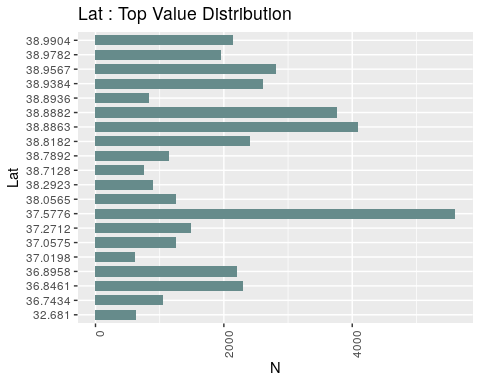
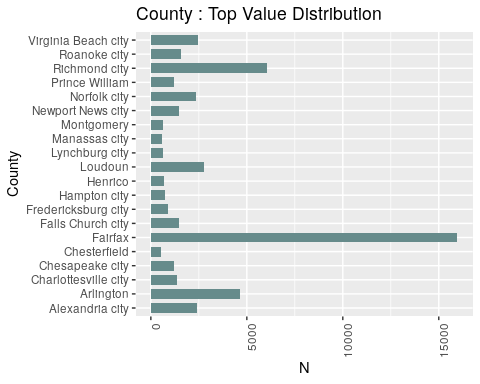
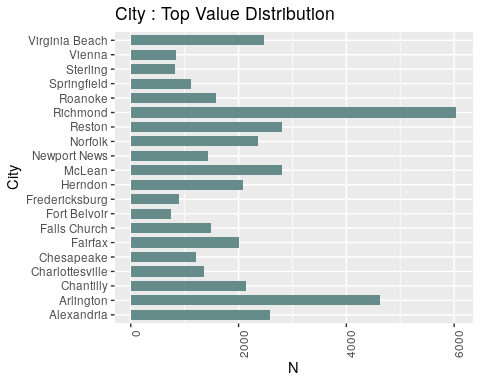
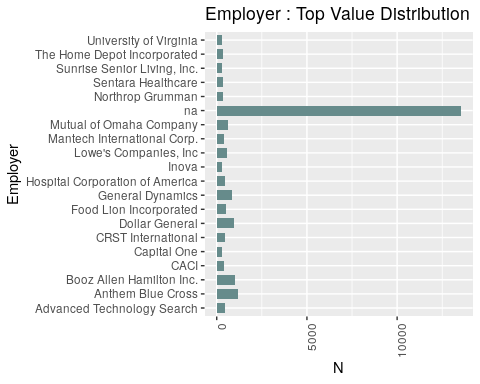
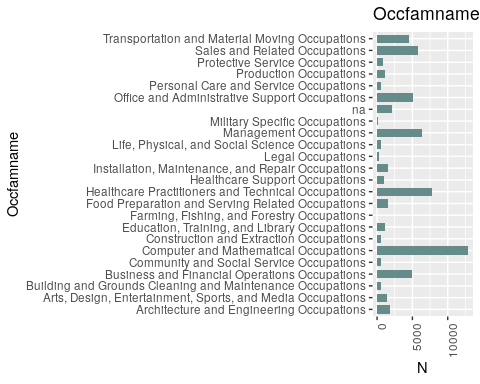
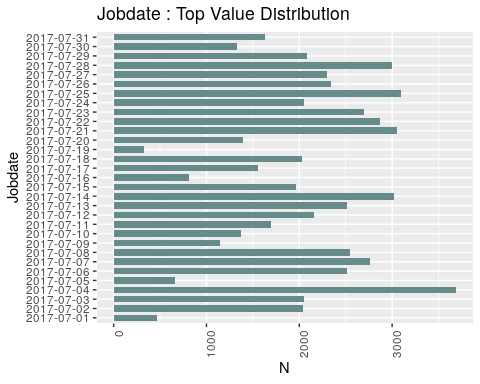
We also perform a quick check for consistency, uniqueness, and duplication, and find that 0 have multiple dates associated with them. We can be reasonably confident the data is consistent and unique.

## Exploratory Data Analysis

This section walks through the data, to get ourselves acquainted iwth the data before we perform.. something

#### Histograms

These histograms show the top value distribution for each column in the dataset (except the identifier column). It may be interesting to note that the top three coordinates given in the last two plots for the Latitude and Longitude columns, (37.5776, -77.5347), (38.8863, -77.0977), and (38.8882, -77.4552), respectively represent Richmond, Arlington (Clarendon), and Chantilly respectively.



Employer, occfam, and occfamname are text fields so let’s look for empty and missing strings

### Certification   
emp <- main$employer  
(sum(is.na(emp)) + sum(emp == 'na') +sum(emp == '')) / length(emp)

[1] 0.2139555

occ <- main$occfam  
(sum(is.na(occ)) + sum(occ == 'na') +sum(occ == '')) / length(occ)

[1] 0.03377244

occname <- main$occfamname  
(sum(is.na(occname)) + sum(occname == 'na') +sum(occname == '')) / length(occname)

[1] 0.03377244 Employer is around 21% missing (filled with ‘na’), occfam and occfamname are around 3% missing

#### Maps

#### Validation

## Questions

* If occfam is missing, is occfamname always missing as well?
* what naming standardizations have been done to the employer column?
* How does burning glass handle subsidiaries?

Data structure. Refers to the way in which the data are structured and organized. Data can be messy and not always conducive to statistical analysis so the structure of the data needs to be assessed to identify any issues that might need to be cleaned or transformed/restructured. Metadata and provenance. Metadata refers to information about the data. Serving the purpose of providing relevant information pertaining to a particular data element or object, metadata captures elements like unit definitions, unit attribute definitions, semantic confusion, multiple attribute names, and inconsistent attribute formats. In addition, metadata can provide a history of the data, or provenance. Provenance refers to where the data originated, what the data are, and the history of access, transmission, and/or modification in terms of when and by whom. Data Preparation In prior stages, issues with the data were identified, in the data preparation stage, determinations on what to fix and how are made. Cleaning and transformation. Data cleaning refers to the process of fixing or removing data that is incorrect, incomplete, improperly formatted, or duplicated. Data transformation refers to the mapping of original data values into values that are in a more useful format. Restructuring. Refers to the process of creating multiple new datasets from the data source that can be more easily analyzed. Data Linkage In the data linkage stage, elements of a dataset are linked to the same corresponding elements in another dataset (e.g., ID numbers, geolocation, index variables). Data Exploration In the data exploration stage, the data are analyzed by summarizing main characteristics and depicting them using visual tools (e.g., tables, graphs, dashboards). Fitness-for-Use Assessment