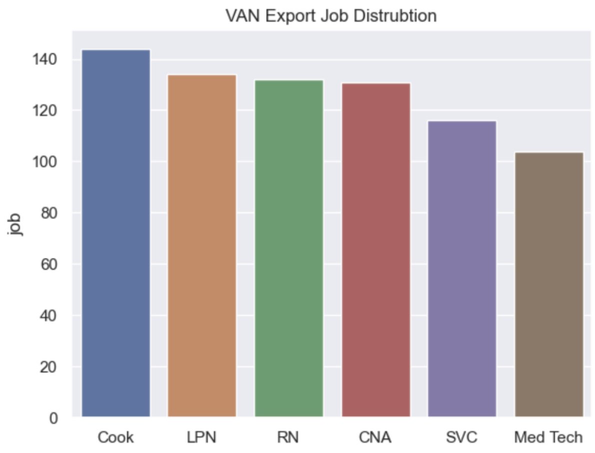
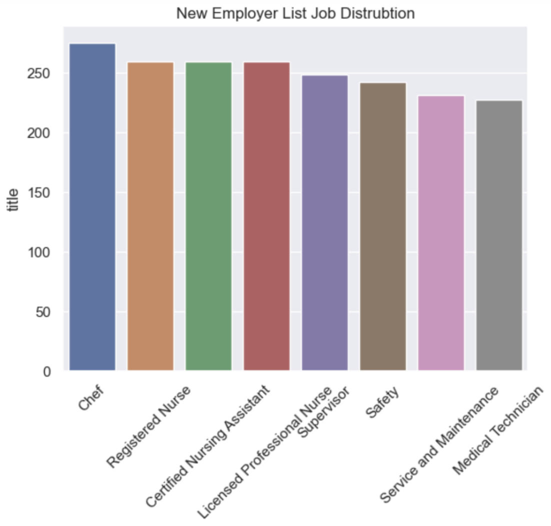
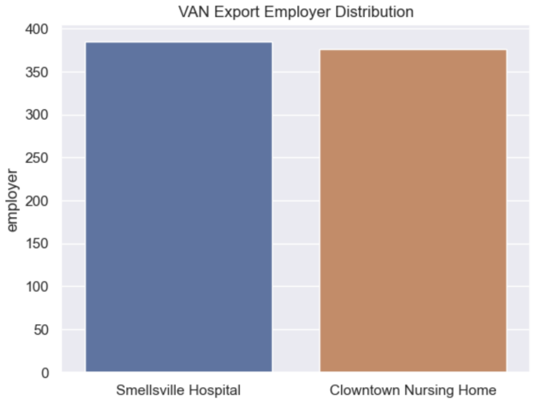
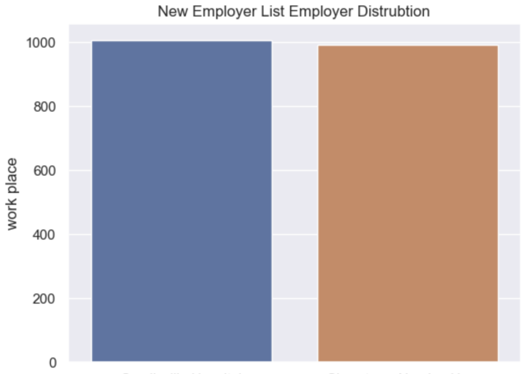
See preliminary exploration code here

1. Preliminary Exploration:
    a. Each file contains fields that hold corresponding data
        i. All columns have formatting differences between the two files
            1. VAN Export “name” : first name, last name
            2. New Employer List “name” : first initial, last name
            3. Variations in column names
                a. “job” = “title”
            4. Addresses in New Employer List do not have zip codes
            5. Job titles in New Employer List are not abbreviated
            6. New Employer List only has one field for phone numbers rather than having a distinction between “cell” and “home”
        ii. VAN Export contains unique VAN IDs
            1. Matching workers to their VAN IDs is critical for keeping updated records
        iii. New Employer List contains two job titles that are not accounted for in Van Export
            1. Supervisor
            2. Safety
            3. These roles must have been created after the last VAN upload as there are no existing VAN IDs for any worker with this job title
        iv. No missing data for “name”, “job”/”title”, or “employer”/”workplace”
        v. Distributions of employer and job title seem to be very similar for both files
        vi. Distributions of job titles seem to be similar for both files


VAN Export Job Distrubtion


New Employer List Job Distrubtion


VAN Export Employer Distribution


New Employer List Employer Distrubtion

2. We can identify which workers in the New Employer List file match our workers in VAN by matching an abridged full name (“First Initial, Last Name”) using a VLOOKUP formula in Excel
    a. Considerations:
        i. New workers will not match to a VAN ID
        ii. Names are not unique, and cause multiple workers to receive the same VAN ID
        iii. Combining first initials with last name will reduce duplicates
        iv. Duplicated names in the New Employer that do not exist in VAN Export (ie. R. Brakespear) do not pose an issue, as neither record has a VAN ID
        v. Inconsistent records in fields such as phone, address, and email, make it impossible to create a unique identifier for the following duplicated names:
            1. D. Cummings
            2. D. Thorsby
            3. Given the small amounts of mismatches, these errors can be quickly manually edited
            4. With more time, a more robust formula that accounts for these variations could be created

1. Immediately after uploading these records in bulk, I would create a list by narrowing the entire VAN database to only include the records of workers who work in Smellsville Hospital or Clowntown Nursing Home. I would name this based on the data it represents and when it was created (eg. Smellsville_Clowtown_03/2022), and I would enable appropriate users and groups to have access. I would also save this subset as a “search”, to allow others to recreate and explore this list. As we would receive more data on these medical facilities, I would repeat this process with an updated nomenclature (eg. Smellsville_Clowntown_06/2022). In addition, using visualization tools such as Tableau can assist organizers in tracking metrics like support percentages

1. The best way to regularly match and upload records is by having routine checks in place to ensure our databases are as up to date as possible. Coding programs (Python, R, etc.) that automate the ETL process are a great aid in this effort. While there is no guarantee on the format of the data we will receive, after some preliminary manual formatting, having automated programs that can perform cleaning, matching, uploading, and reporting procedures builds a lot of capacity. Specifically, the Python library “Parsons” allows for efficient code to be written that allows for easy communication between the VAN API and the APIs of other data tools (Hustle, TargetSmart, Mailchimp, etc.).