U.S. Workplace Injury Analysis

IRON HACK

By: Yu Ting Hu Wu
Date: 30-July-2022

Table of Contents

Project Overview

Within this project we will analyze data of **work-related injuries and illnesses** from different employers in United States.

The objective is to explore the officially reported data from **OSHA** to have insights that can serve to **spread awareness and contribute to improve workplace health & safety**, and eventually save people lives.

Companies may also take these insights into consideration to:

- ✓ see **relative level of injuries and illnesses among different industries**
- ✓ understand **why employees are suffering from injuries** (which sometimes cause fatalities)
- ✓ determine **problem areas and progress in preventing** work-related injuries and illnesses

Data source

Data is extracted from **OSHA (Occupational Safety and Health Administration)**, which is the federal agency of the United States, part of the United States Department of Labor, that regulates workplace safety and health: https://www.osha.gov/Establishment-Specific-Injury-and-Illness-Data

Incident Rate

As per **US Bureau of Labor Statistics**, an **incidence rate** of injuries and illnesses is computed from the following formula:

$$TCR = \frac{(Number\ of\ injuries\ and\ illness)\ x\ 200{,}000}{Employee\ hours\ worked}$$

$$DART = \frac{(Number\ of\ injuries\ and\ illnesses\ with\ days\ away\ from\ work,\ job\ transfer,\ or\ restriction)\ x\ 200{,}000}{Employee\ hours\ worked}$$

Notes:

- The 200,000 hours in the formula represents the equivalent of 100 employees working 40 hours per week, 50 weeks per year, and provides the standard base for the incidence rates).
- Hours worked should not include any nonwork time, even though paid, such as vacation, sick leave, holidays, etc.

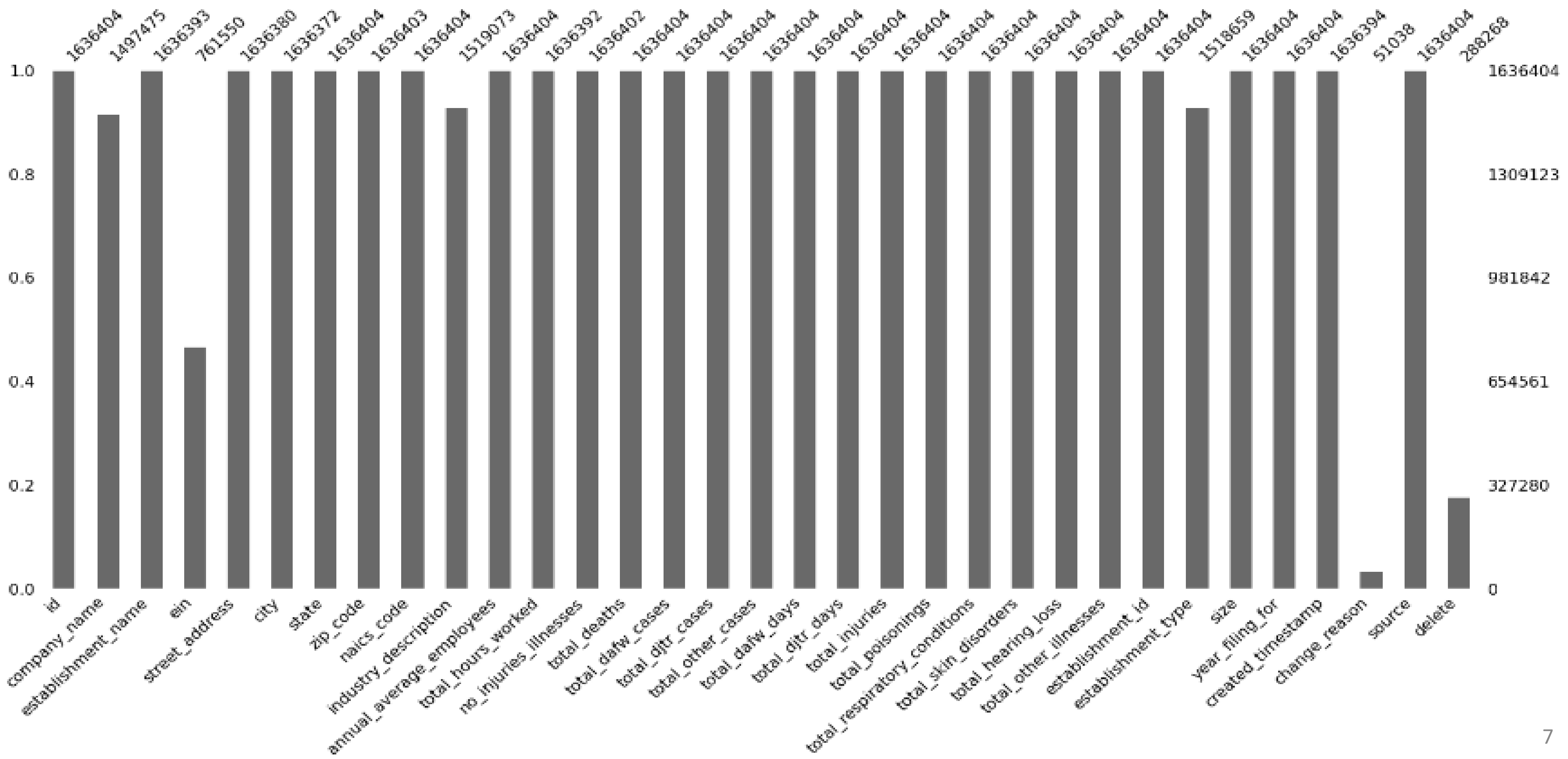
Raw Data Exploration

```
Original Dataframe size: Number of rows = 1636404; Number of Columns = 33
```

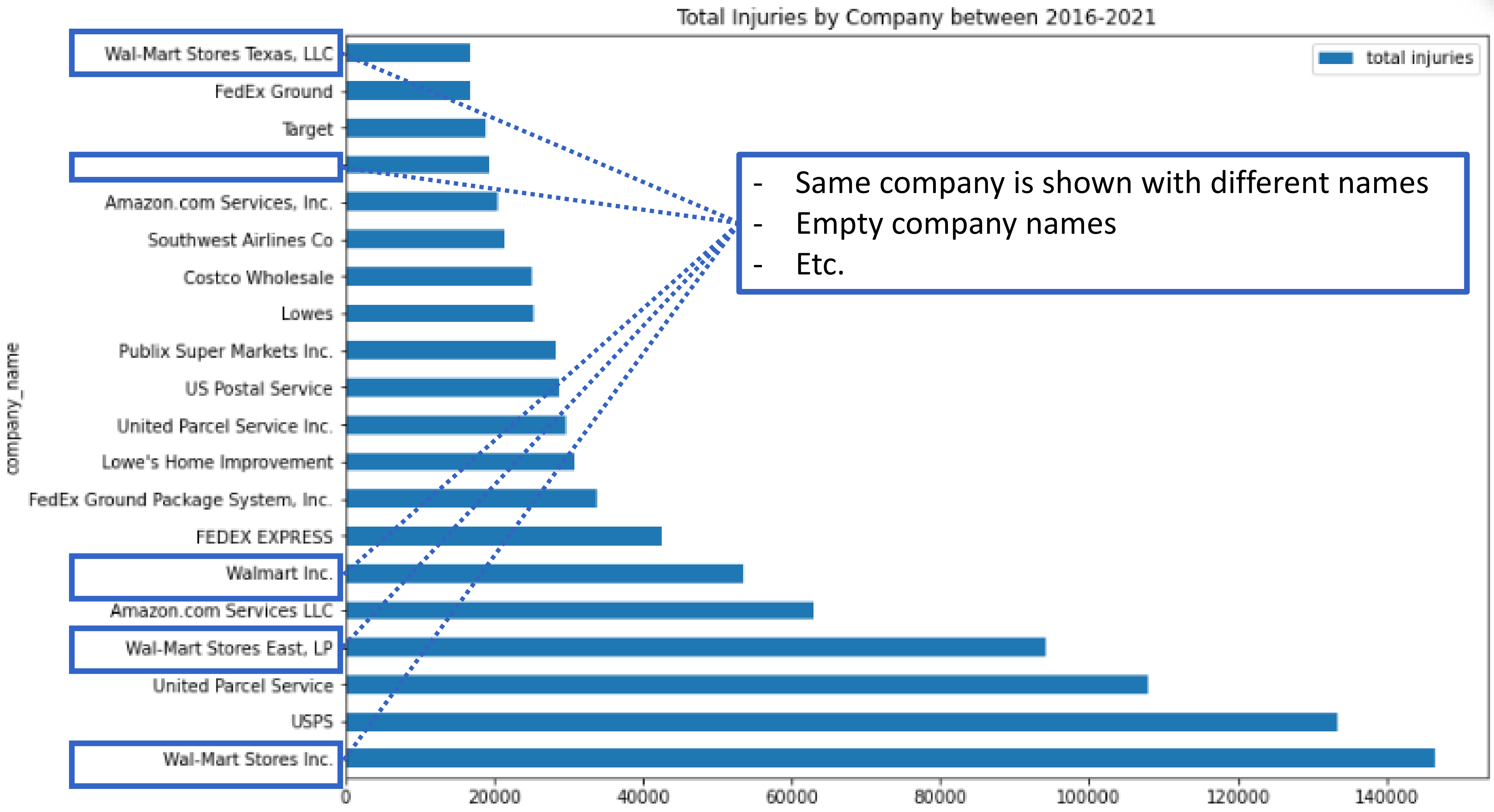
	id	company_name	establishment_name	ein	street_address	city	state	zip_code	naics_code	industry_description	annual_average_employees	total_hours_worked
0	4.0	McKamish, Inc.	McKamish, Inc.	NaN	50 55th Street	Pittsburgh	PA	15201.0	238220.0	Heating, ventilation and air-conditioning (HVA...	280.0	579688.0
1	5.0	The Talaria Company, LLC	The Hinckley Company	NaN	40 Industrial Way	Trenton	ME	4605.0	336612.0	Pleasure boats manufacturing	246.0	501578.0
2	6.0	Williamsburg Manufacturing	Williamsburg Manufacturing	NaN	408 Maplewood Ave	Williamsburg	IA	52361.0	336370.0	Motor vehicle metal parts stamping	273.0	619945.0
3	7.0	The Talaria Company, LLC	Morris Yachts, LLC	NaN	27 Ramp Road	Trenton	ME	4605.0	336612.0	Pleasure boats manufacturing	33.0	75794.0
4	8.0	The Talaria Company, LLC	Hunt Yachts, LLC	NaN	1909 Alden Landing	Portsmouth	RI	2871.0	336612.0	Pleasure boats manufacturing	43.0	114734.0

Raw Data Exploration

```
# visual plot of null values per column to see density of dataframe
msno.bar(osha_df_raw);
```



Raw Data Exploration

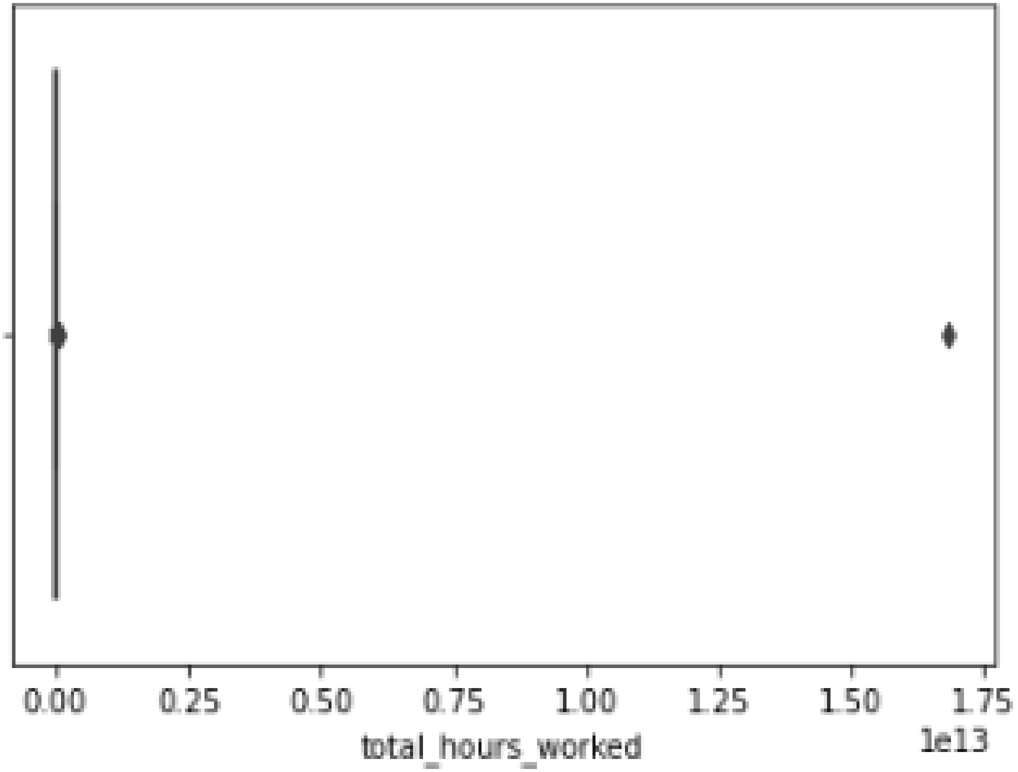


- Same company is shown with different names
- Empty company names
- Etc.

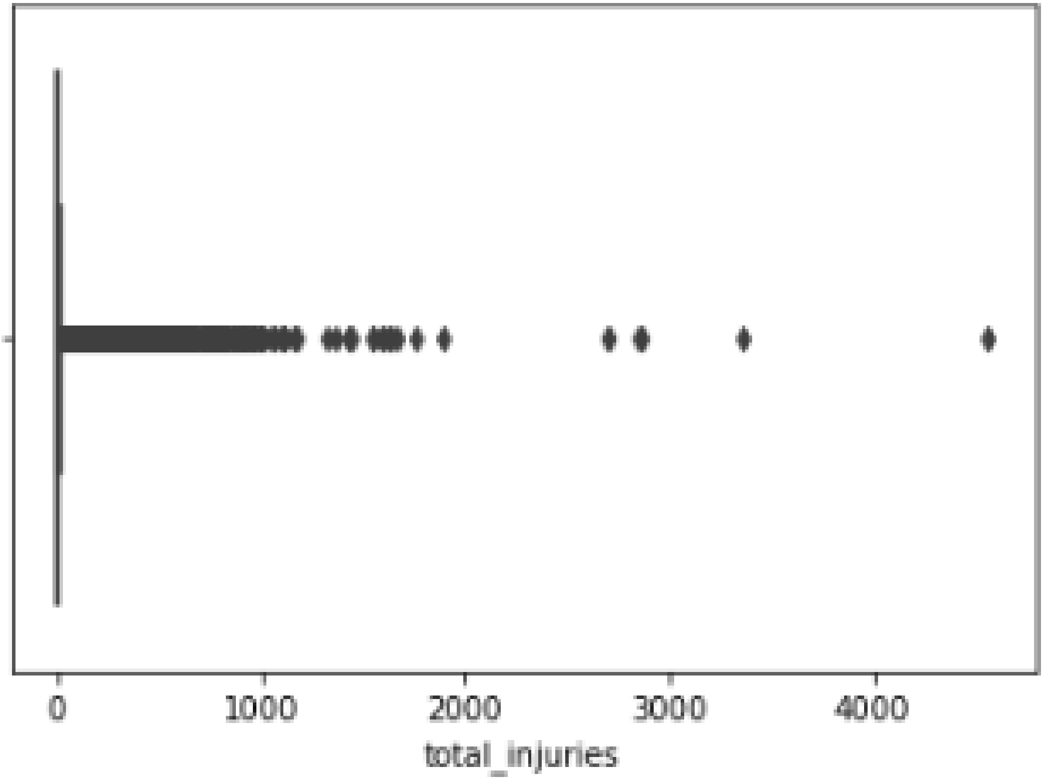
Raw Data Exploration

Outlier values

```
# box plot to see distribution of total_hours_worked
sns.boxplot(x="total_hours_worked", data=osha_df_raw);
```




```
# box plot to see distribution of total_injuries
sns.boxplot(x="total_injuries", data=osha_df_raw);
```



Data Processing

In any data science project, **data wrangling** is a very important step, since it removes the risk by **ensuring data is in a reliable state** before it is analyzed and leveraged, making it to be a critical part of the analytical process. Thus, despite dataset quality is good, it is highly advised to process and format the data.

- ✓ **Clean the data**: by keeping only the columns that are meaningful for the analysis; and use statistics to detect and remove outliers with reference to its interquartile range.
- ✓ **Format the data**, such as data conversion, remove negative numeric values, filling empty values, removing special characters, standardize company name, etc.

Exploratory Data Analysis

After processing the raw data, we will use the **clean dataset output** to perform EDA (Exploratory Data Analysis), to plot our variables in order to extract **meaningful insights and conclusions about our data**, by means of visual explorations.

EDA is based on graphical and descriptive techniques whose objective is to:

- ✓ gain intuition about the data
- ✓ detect outliers
- ✓ extract important variables
- ✓ discover underlying structures in the data
- ✓ It also allows organizing the data, detecting failures and evaluating the existence of missing data

Technology Stack

- Python (Jupyter)
- Web Scrapping / Requests / BeautifulSoùp
- Numpy, Pandas, Pickle
- Data wrangling. Main functions:
  - import_data(), combine_csv()
  - df_clean_format(**kwargs parameters)
- EDA and descriptive statistics techniques
- Matplotlib, Seaborn
- PowerBi
- …

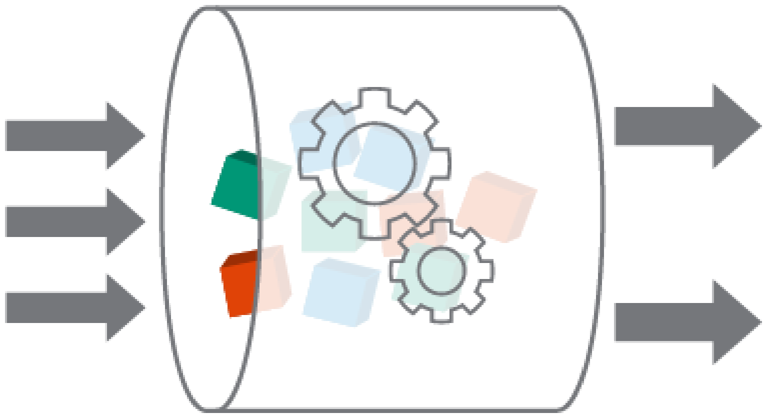
Processed Data Results

Input

```
osha_df_raw.isnull().sum()

id                                0
company_name                 138929
establishment_name               11
ein                          874854
street_address                   24
city                             32
state                             0
zip_code                          1
naics_code                        0
industry_description         117331
annual_average_employees          0
total_hours_worked               12
no_injuries_illnesses             2
total_deaths                      0
total_dafw_cases                  0
total_djtr_cases                  0
total_other_cases                 0
total_dafw_days                   0
total_djtr_days                   0
total_injuries                    0
total_poisonings                  0
total_respiratory_conditions      0
total_skin_disorders              0
total_hearing_loss                0
total_other_illnesses             0
establishment_id                  0
establishment_type           117745
size                              0
year_filing_for                   0
created_timestamp                10
change_reason               1585366
source                            0
delete                      1348136
dtype: int64
```

Processing



Output

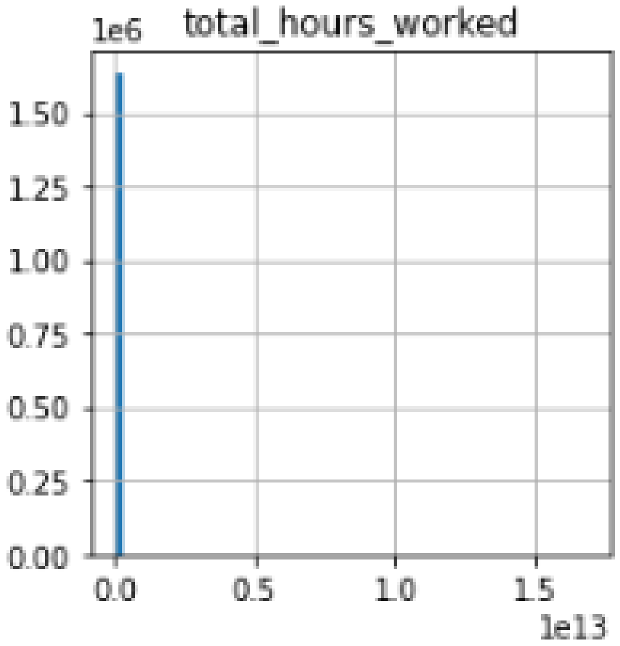
```
osha_df_raw.isnull().sum()

id                              0
company_name                    0
establishment_name             10
state                           0
naics_code                      0
total_hours_worked              6
injury_illness                  1
total_deaths                    0
total_dafw_cases                0
total_djtr_cases                0
total_other_cases               0
total_dafw_days                 0
total_djtr_days                 0
total_injuries                  0
size                            0
year_filing_for                 0
naics_industry_description   3216
TCR                             0
DART                            0
dtype: int64
```

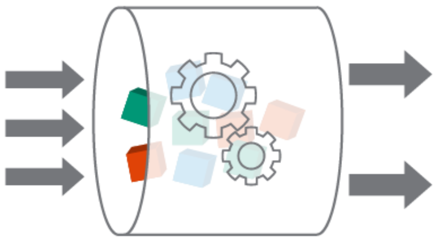
Almost no null values!

Processed Data Results

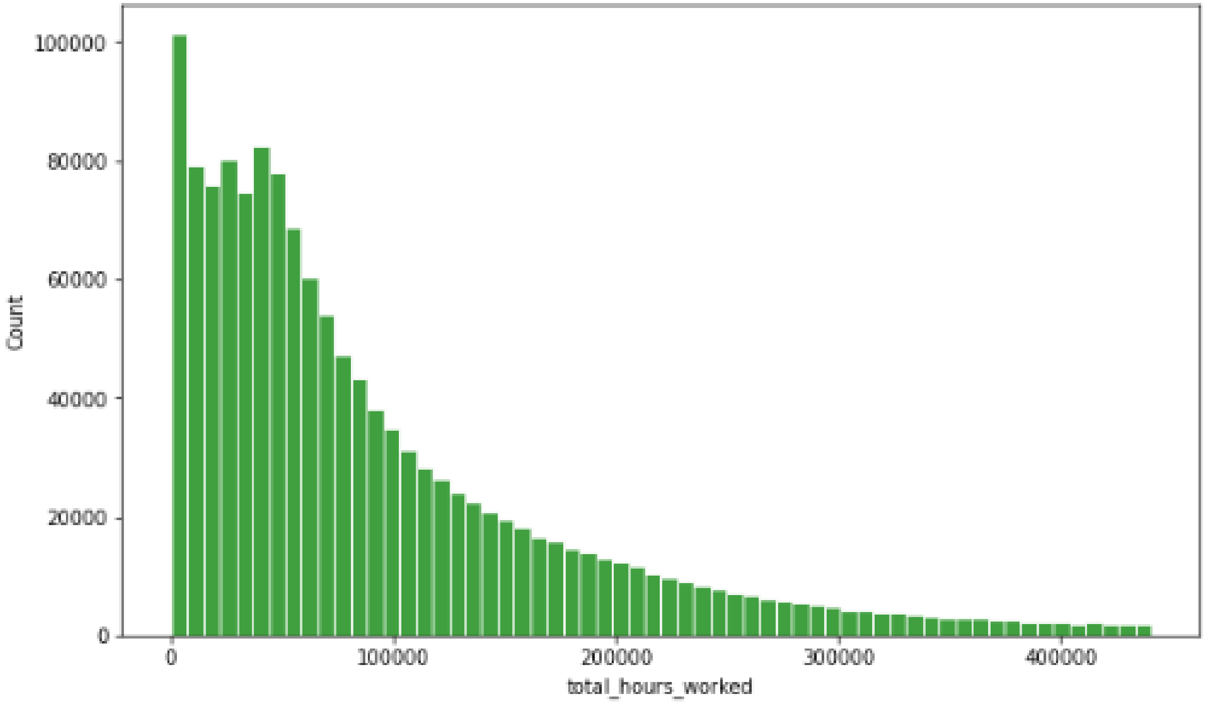
Input



Processing



Output



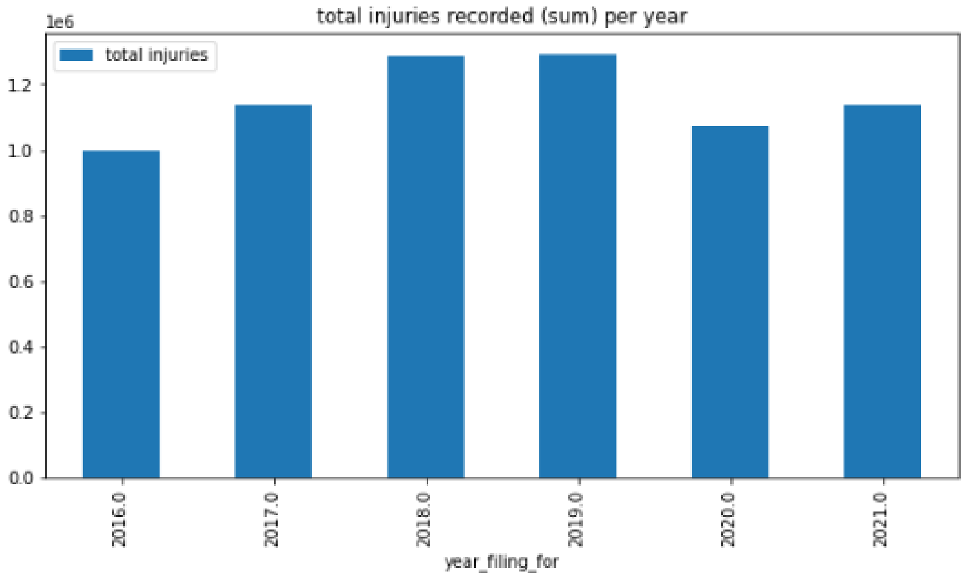
Same histogram plot is now much more meaningful

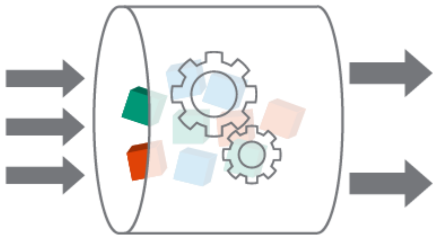
Processed Data Results

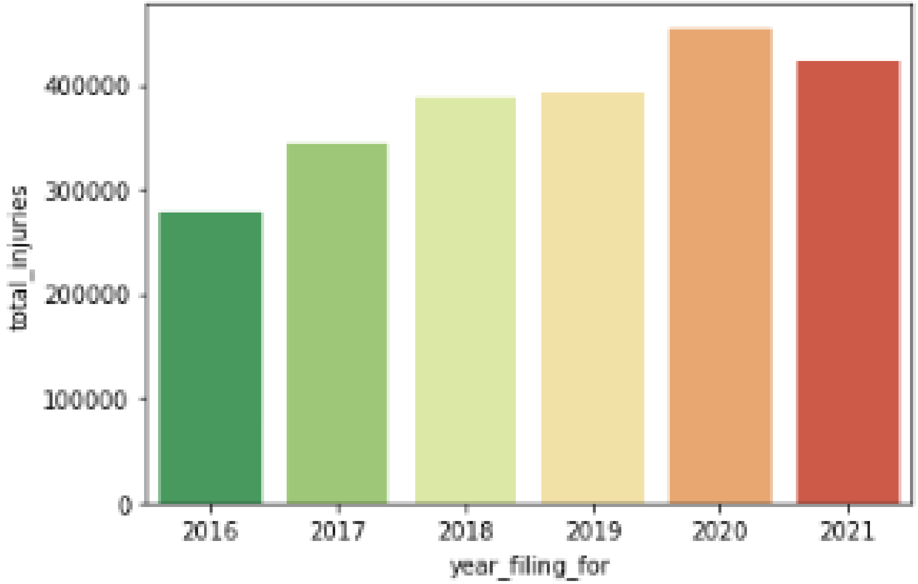
Input

Processing

Output





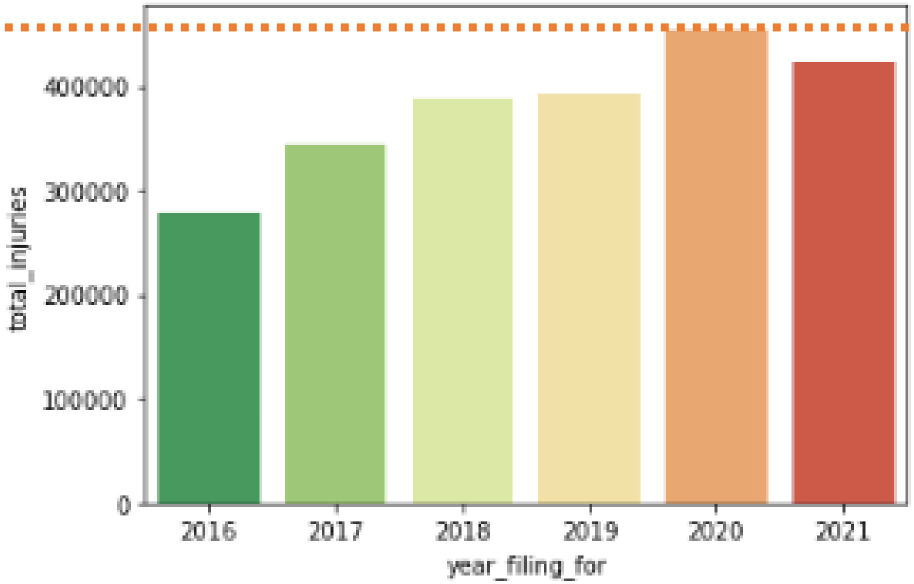


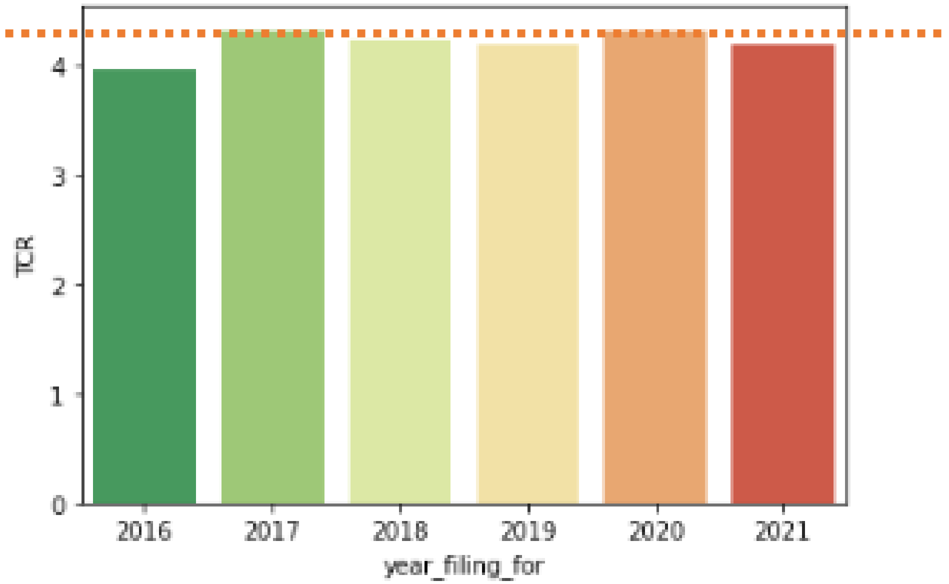
After processing the data, 2020 is the year with highest number of recorded injuries

Processed Data Results

However, if we compare recorded injuries per year **based on TCR incident rate**, we can see highest TCR is maintained on year 2020, but with **very slight difference between one year and another**.

```
Average TCR by year: year_filing_for
2016    3.963683
2017    4.300468
2018    4.222766
2019    4.180378
2020    4.311314
2021    4.189305
Name: TCR, dtype: float64
```



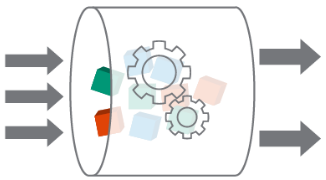


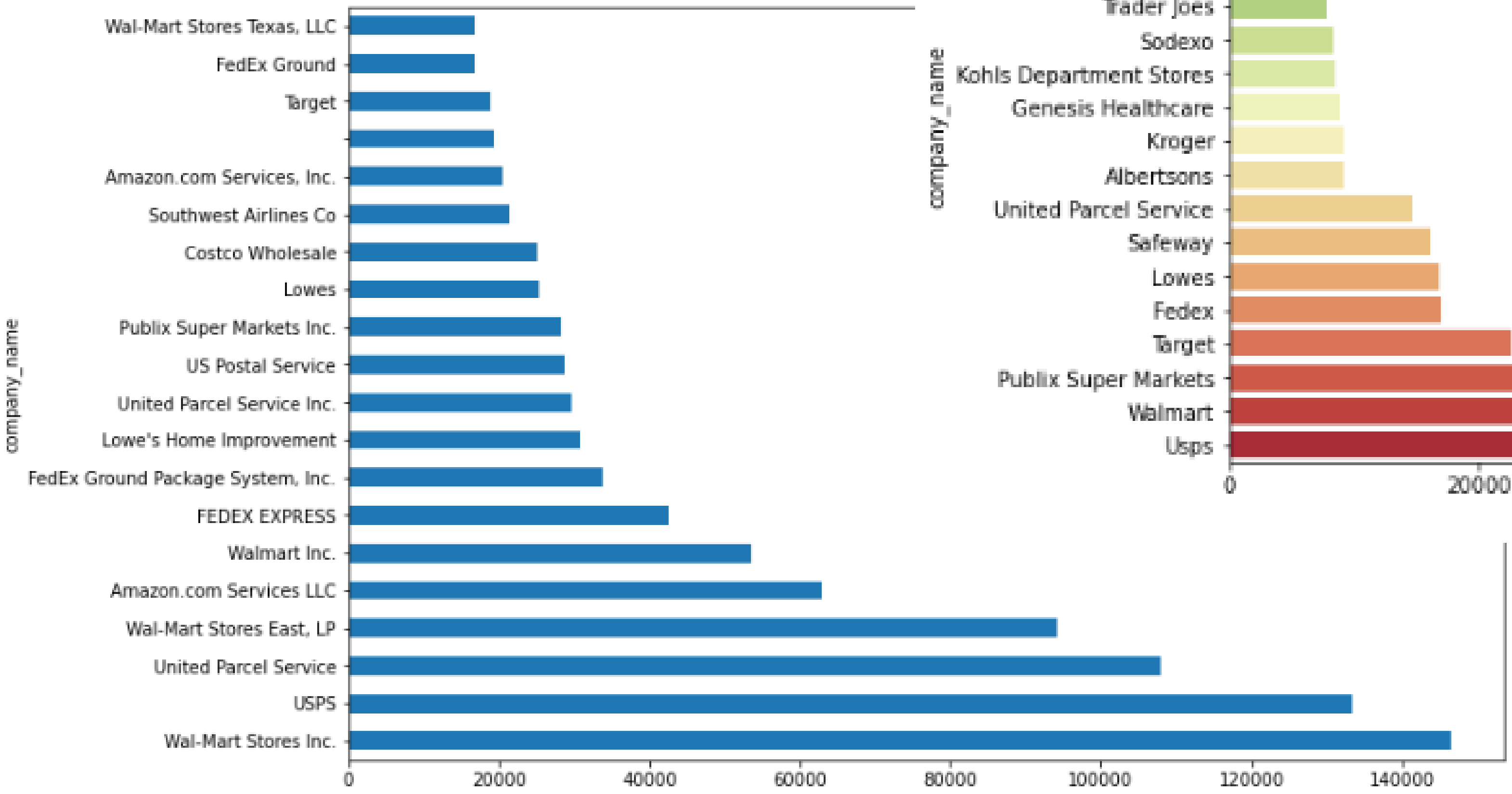
Processed Data Results

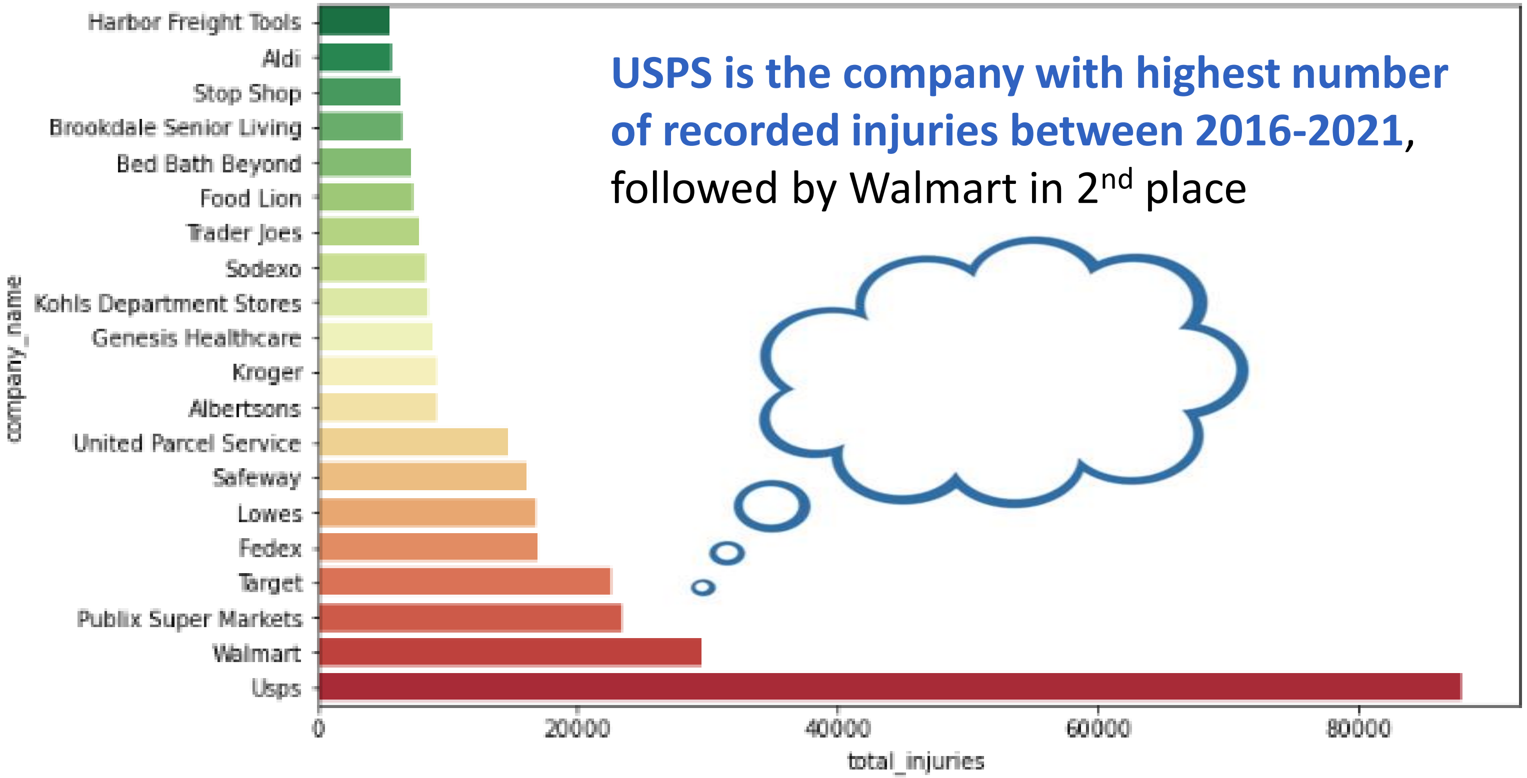
Input

Processing

Output



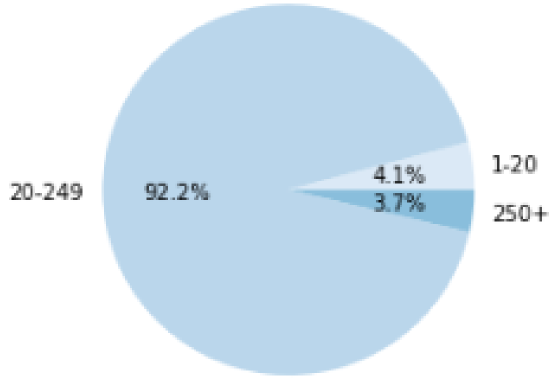


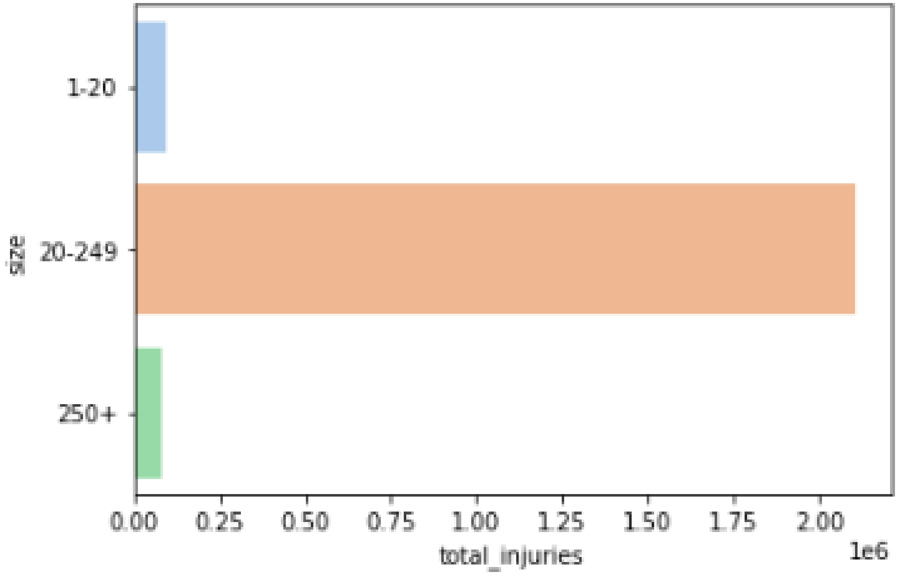


USPS is the company with highest number of recorded injuries between 2016-2021, followed by Walmart in 2nd place

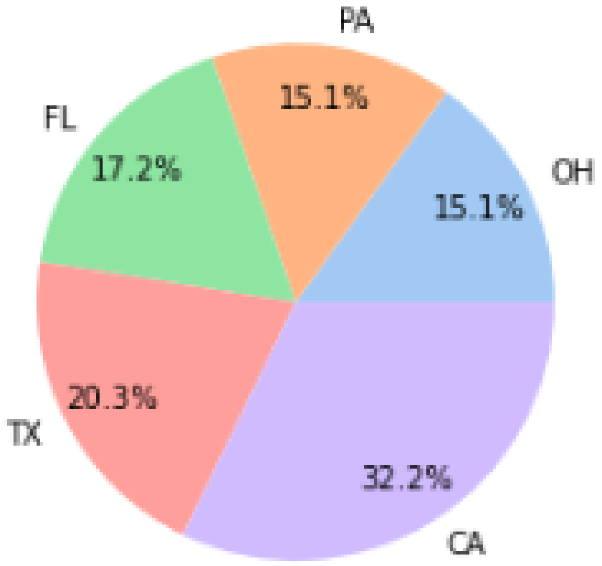
Processed Data Results

92.2% of the companies with recorded injuries are SME (Small Medium Enterprises) with 20-249 employees

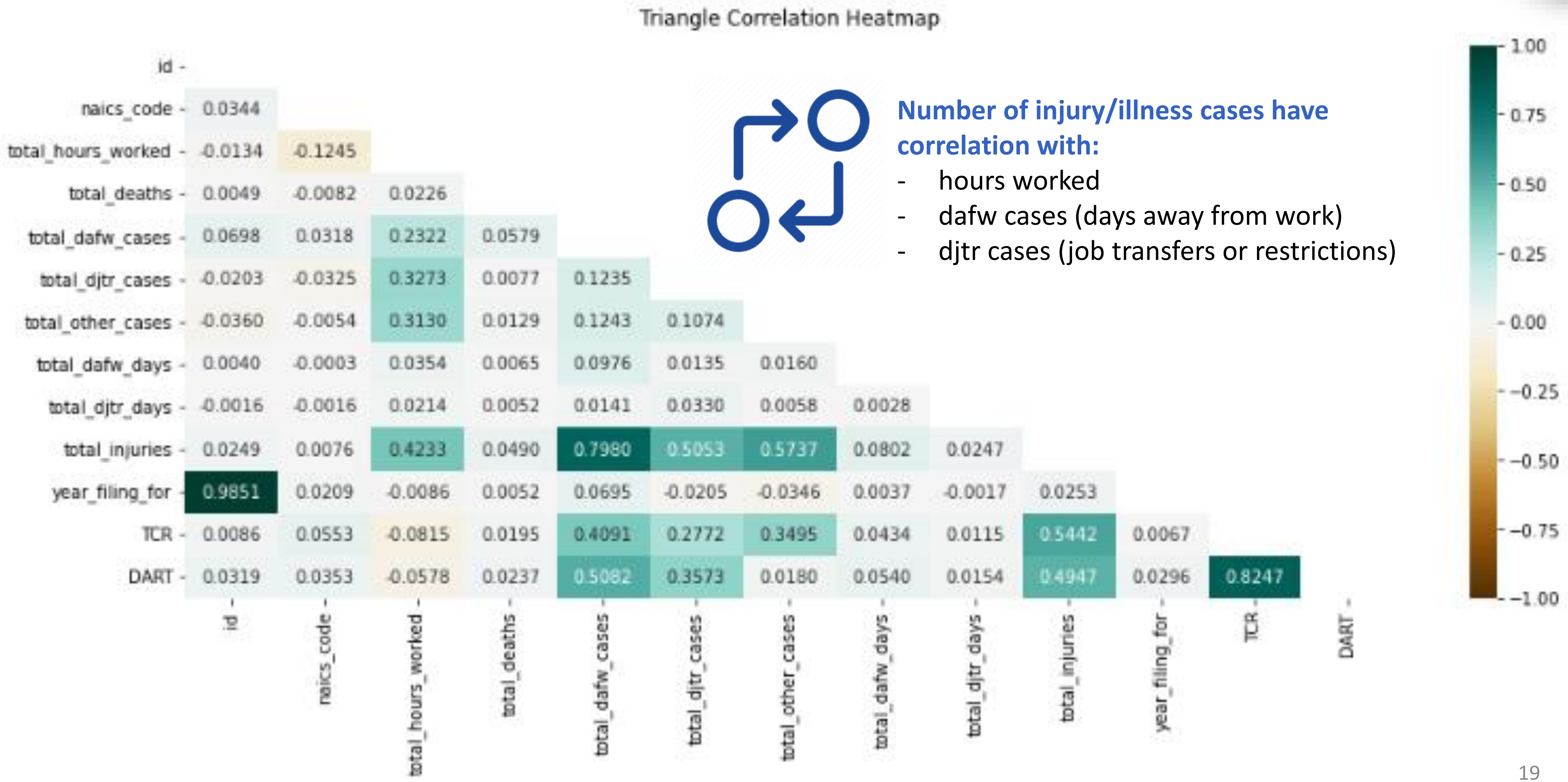




53.5% of injuries/illnesses occur only in the states of **California and Texas**



Processed Data Results

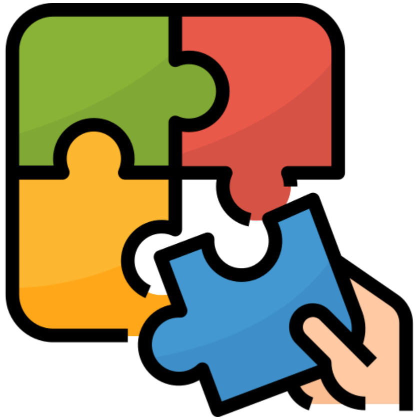


Number of injury/illness cases have correlation with:

- hours worked
- dafw cases (days away from work)
- djtr cases (job transfers or restrictions)

Conclusions

Wrapping up:



1. 2020 has the highest number of recorded injury/illness and TCR.
2. TCR is very similar for all recorded years.
3. USPS is the company that registered highest number of total injuries, followed by Walmart.
4. 92.2% of the companies are SME (Small Medium Enterprises) with 20-249 employees
5. 53.5% of injuries/illnesses occur only in the states of California and Texas
6. There is correlation between injury cases and hours worked, dafw cases (days away from work) and djtr cases (job transfers or restrictions)

... but the most important one:

Data Cleaning and Data Pre-processing part is essential.
Fighting with your data, will make it to confess their secrets.
Correct insights can only be extracted if you have good data.

Thank you!