Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

# PROJECT PROPOSAL

## i. DATASET

| | | |
|---|---|---|
| Dataset Name | : | Severeinjury |
| Available Format | : | Comma Separated Values, Available as Severeinjury.csv |
| Source Link | : | https://www.kaggle.com/jboysen/injured-workers/data |

## ii. DATASET DESCRIPTION

**Description:**

The "Severeinjury" dataset contains the records of an OSHA Inspection, which informs us about the occurrence of the industrial accident in United State of America. It also elucidates when, where, how the accident occurred with the cause and nature of it.

**Summary of the Dataset:**

```
> summary(severeinjury)
       ID                 UPA             EventDate          Employer          Address1          Address2
 Min.   :2.015e+09   Min.   : 892735   Length:21578       Length:21578       Length:21578       Length:21578
 1st Qu.:2.015e+09   1st Qu.:1003217   Class :character   Class :character   Class :character   Class :character
 Median :2.016e+09   Median :1060610   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :2.016e+09   Mean   :1064256
 3rd Qu.:2.016e+09   3rd Qu.:1127405
 Max.   :2.017e+09   Max.   :1219296
 NA's   :1819
     City              State               Zip               Latitude        Longitude       Primary NAICS
 Length:21578       Length:21578       Length:21578       Min.   :-15.78   Min.   :-170.71   Min.   :    21
 Class :character   Class :character   Class :character   1st Qu.: 32.20   1st Qu.: -95.40   1st Qu.:311411
 Mode  :character   Mode  :character   Mode  :character   Median : 38.77   Median : -87.66   Median :333120
                                                          Mean   : 36.72   Mean   : -87.69   Mean   :393922
                                                          3rd Qu.: 41.09   3rd Qu.: -80.62   3rd Qu.:491110
                                                          Max.   : 61.29   Max.   : 145.75   Max.   :999999
                                                          NA's   :91       NA's   :91        NA's   :3
  Hospitalized      Amputation        Inspection       Final Narrative        Nature      NatureTitle
 Min.   :0.000   Min.   :0.0000   Min.   : 837147   Length:21578       Min.   :   7   Length:21578
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1077550   Class :character   1st Qu.: 111   Class :character
 Median :1.000   Median :0.0000   Median :1121269   Mode  :character   Median :1311   Mode  :character
 Mean   :0.808   Mean   :0.2698   Mean   :1122091                      Mean   : 887
 3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1170130                      3rd Qu.:1312
 Max.   :3.000   Max.   :9.0000   Max.   :1231162                      Max.   :9999
                 NA's   :2        NA's   :13756
  Part of Body   Part of Body Title      Event        EventTitle           Source      SourceTitle
 Min.   :   6   Length:21578       Min.   :  20   Length:21578       Min.   :  10   Length:21578
 1st Qu.: 320   Class :character   1st Qu.: 642   Class :character   1st Qu.:3230   Class :character
 Median : 513   Mode  :character   Median :4331   Mode  :character   Median :4154   Mode  :character
 Mean   :2283                      Mean   :3905                      Mean   :4811
 3rd Qu.:4422                      3rd Qu.:6252                      3rd Qu.:7124
 Max.   :9999                      Max.   :9999                      Max.   :9999

 Secondary Source Secondary Source Title
 Min.   :  10     Length:21578
 1st Qu.:2214     Class :character
 Median :4418     Mode  :character
 Mean   :5211
 3rd Qu.:8621
 Max.   :9999
 NA's   :15766
```

| | | |
|---|---|---|
| **No. of Tables** | : | 1 |
| **No. of Records** | : | 21578 |
| **No. of Attributes** | : | 26 |
| **Source** | : | Occupational Health and Safety Association OSHA |

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

**Attributes:**

**Data Table:**

| *Attributes ID* | *Type of Data* |
|---|---|
| ***ID*** | Integer |
| ***EventDate*** | Character |
| ***Employer*** | Character |
| ***Address1*** | Character |
| ***Address2*** | Character |
| ***city*** | Character |
| ***state*** | Character |
| ***zip*** | Character |
| ***Latitude*** | Numeric |
| ***Longitude*** | Numeric |
| ***Primary NAICS*** | Integer |
| ***Hospitalized*** | Numeric |
| ***Amputation*** | Numeric |
| ***Inspection*** | Integer |
| ***Final Narrative*** | Character |
| ***Nature*** | Integer |
| ***NatureTitle*** | Character |
| ***Part of Body*** | Integer |
| ***Part of Body Title*** | Character |
| ***Event*** | Integer |
| ***EventTitle*** | Character |
| ***Source*** | Integer |
| ***SourceTitle*** | Character |

1. **ID** – Represented as “ID” in the raw dataset, gives the Case Number of the Accident occurred, issued by OSHA.

2. **Event Date –** Represented as “EventDate”, gives the Date of occurrence of the Accident.

3. **Employer –** Represented as “Employer” in the raw dataset, gives the name of the company, where the accident occurred**.**

4. **Address** – Represented as “Address1” & “Address2”, give the address of the company, where the accident occurred.

5. **City –** Represented as “city”, gives the Company’s City Location.

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

6. **State** – Represented as "state", gives the State of the respective City.

7. **ZIP Code** – Represented as "zip", gives the Zip Code of the respective company's area.

8. **Latitude & Longitude** – Represented as "Latitude" & "Longitude", give the earth's co-ordinate position of the Company.

9. **Primary NAICS –** Represented as "Primary NAICS", gives the code according to North American Industry Classification System, which issues Codes based on the operation performed by the companies.

10. **Hospitalized** – Represented as "Hospitalized", gives the number of Employees/Personals Hospitalized after the accident.

11. **Amputated** – Represented as "Amputated", gives the number of Employees/Personals, who had their Body Parts amputated after accident.

12. **Inspection Number** – Represented as "Inspection", gives the unique Inspection ID provided by OSHA, while examining about the accident.

13. **Final Narrative** – Represented as "Final Narrative", gives the exact explanation of the accident's occurrence.

14. **Nature Code & Title** – Represented as "Nature" & "Nature Title", give the generalized category to which the effect of injury of the Employee/Personal has occurred.

15. **Part of Body's Code & Name** – Represented as "Part of Body" & "Part of Body Title", give the part of the body which was amputated, if the "Amputated" attribute has a record in it.

16. **Event Code & Title** – Represented as "Event" & "EventTitle", give the brief information about the accident.

17. **Source Code & Title** – Represented as "Source" & "SourceTitle", give the exact Source for the accident occurred.

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

## iii. OBJECTIVES

1. To develop a Predicting model for OSHA to predict the occurrence of an accident and conduct an inspection to recommend the safety measures.

2. To recommend a sales proposal for production of a required Personal Protective Equipment (PPE) for any protective equipment manufacturing company.

**Objective 1:**

**a. Sampling:**

We have used Simple Random Sampling to find the occurrences of accidents' location, industry and study the trend of them in it. The probability of an injury occurring has equal probability with the injury not occurring.

**b. Aggregation:**

Removed the attribute "Secondary Source", as it had negligible Data Entries, Merged the attributes "Address 1" & "Address 2" to "Address", for a clear representation of the location of the company.

**c. Data Cleaning:**

Checking the quality of our dataset, removed the attribute "Secondary Source", as it had negligible Data Entries. Created a function in R, which could convert Text in the Attributes to "Characters" and format it as "Capitalizing each Word". Example: TEXAS aviation → Texas Aviation.

**d. Subset Selection:**

For the objective 1, we had to take the whole dataset, eliminating Address, Final Narrative, Source, Secondary Source and UPA, as these attributes were irrelevant to the mentioned objective. Then, the values of records from 2015 to 2016 to study the trend of injuries.

**e. Feature Creation:**

In "Primary NAICS" attribute, the code was a six-digit entry. The First Two Digits represent the sector of the Industry. To process the data in sector-vise, we had to strip out the two digits to a new attribute "Sector". Example: 922140, 339999, 237120 → 92, 33, 23 (92 – Agriculture, 33 – Manufacturing, 23 Construction).

**f. Variable Transformation:**

Using the available info in the data, missing values were filled out. Example: Using Latitude & Longitude info, filled out the Missing City, State and Zip Code values. The "Event Date" attribute was in "Character" & "Date Format" format. We had to choose a format to proceed further, so, it was assigned to "Date" format.

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

**Objective 2:**

**a. Sampling**:

We performed Stratified Sampling for proceeding into Objective 2 and to find the probability of choosing an injury from different amputation levels.

**b. Aggregation:**

Same as Objective 1. (Refer above paragraphs)

**c. Data Cleaning:**

Same as Objective 1. (Refer above paragraphs)

**d. Subset Selection:**

Created a new subset with Data Objects, which have an Amputation Level of greater than zero, i.e., separated injury records, had at least one amputation, with some severity level.

**e. Feature Selection:**

Same as Objective 1. (Refer above paragraphs)

**f. Variable Transformation:**

Same as Objective 1. (Refer above paragraphs)

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

## iv. DATA ANALYTICS METHODS

### Data Analytics Method for Objective 1:

The attributes we are interested in Objective 1 are City Location, State, Industrial Sector plotted against the Injury Records. We create a predictive model and test its fitness to predict the possible occurrence of another Injury.

Various predictive models to be used are:

1. Support Vector Machines (SVM)
2. Random Forest
3. Naive Bayes
4. Linear Regression
5. Logistic Regression
6. Decision Tree

After performing the analysis, we find the fitness of models predicted by each technique and select a suitable model with a higher efficiency, using training, testing and validation sets. These datasets are derived from the original dataset in the basis of 60:20:20 with equal split.

### Data Analytics Methods for Objective 2:

The attributes we are interested in Objective 2 are Location, Amputated and Hospitalized records to predict which part of Body gets amputated and requires a protective equipment. This data could be used as a sales proposal by a PPE manufacturer to start his production of the highly demanded PPE.

We use,

1. Ridge Regression
2. Jackknife Regression

We use these regression models, as we have a non-linear relationship between the values in the interested attributes. Also, Ridge Regression is a robust method and is less subject to over-fitting. While, Jackknife Regression can work fine with independent and non-corelated values and is easier to implement and interpret.

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

## v. EVALUATION AND PERFORMANCE ANALYSIS APPROACH

### Objective 1:

After performing the analysis using mentioned techniques (as mentioned in iv. Data Analytic Methods for Objective 1), we analyze and select a suitable model with a higher efficiency, using training, testing and validation sets and use the Forward Feature Selection technique, to find the best set of attributes, defining the predictive model.

### Objective 2:

| | **AMPUTATION** | **HOSPITALIZED** |
|---|---|---|
| **LATITUDE** | 0.048254360 | -0.052482043 |
| **LONGITUDE** | -0.012158626 | 0.013200845 |
| **ZIP** | 0.006381465 | -0.006841417 |
| **INDUSTRY CODES** | -0.064016694 | 0.046724795 |

As we can see from the correlation matrix that the attributes we are interested in do not have a linear correlation, so we go for Jackknife Regression and Ridge Regression and we plot the Receiver Operating Characteristics (ROC) Curve for the mentioned Analytic Methods and select the model with a larger AUC (Area Under Curve)/ having the highest curve nearing absoluteness.

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

## vi. EXPECTED RESULTS AND INTERPRETATION

### Summary of the Dataset after Pre-Processing:

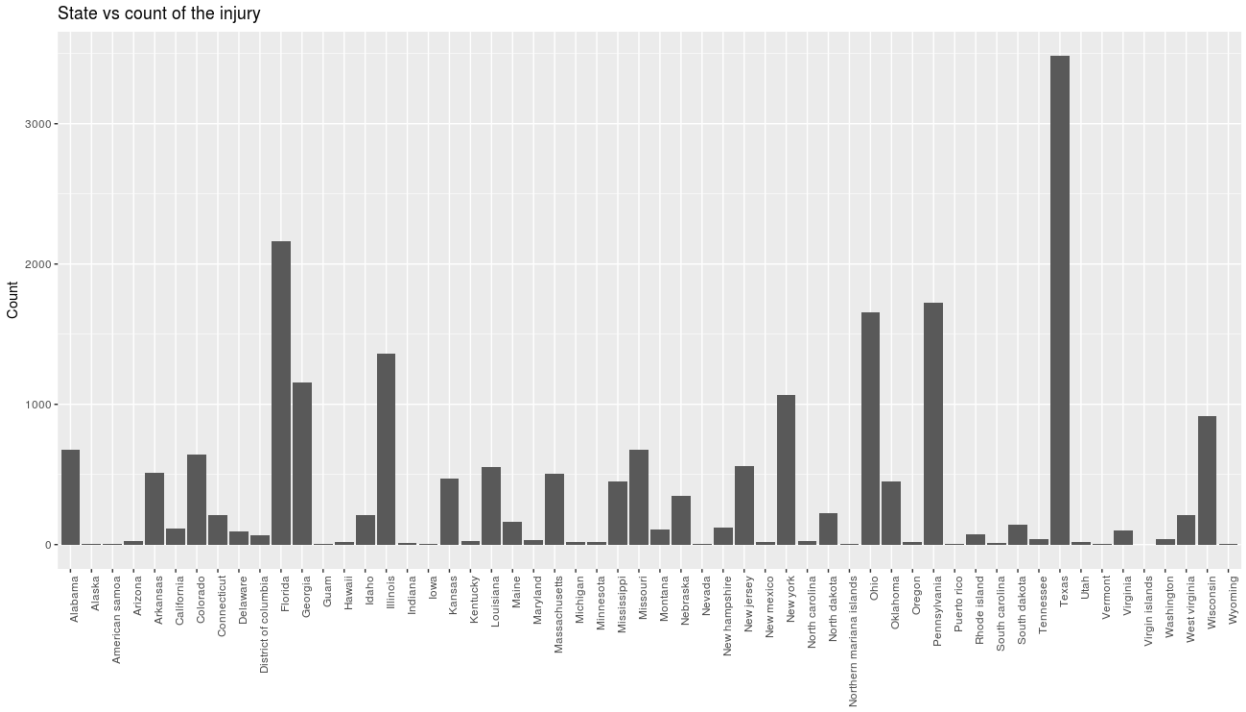
```
> summary(preprocessed_data_csv_preprocessed_data_csv)
       X1              ID                 UPA            EventDate             Month             Day        
 Min.   :    1   Min.   :2.015e+09   Min.   : 892735   Length:21578       Min.   : 1.000   Min.   : 1.00  
 1st Qu.: 5395   1st Qu.:2.015e+09   1st Qu.:1003217   Class :character   1st Qu.: 3.000   1st Qu.: 8.00  
 Median :10790   Median :2.016e+09   Median :1060610   Mode  :character   Median : 6.000   Median :15.00  
 Mean   :10790   Mean   :2.016e+09   Mean   :1064256                      Mean   : 6.135   Mean   :15.58  
 3rd Qu.:16184   3rd Qu.:2.016e+09   3rd Qu.:1127405                      3rd Qu.: 9.000   3rd Qu.:23.00  
 Max.   :21578   Max.   :2.017e+09   Max.   :1219296                      Max.   :12.000   Max.   :31.00  
                 NA's   :1819                                                                             
      Year        Employer            Address              City              State                Zip        
 Min.   :2015   Length:21578       Length:21578       Length:21578       Length:21578       Min.   :  802  
 1st Qu.:2015   Class :character   Class :character   Class :character   Class :character   1st Qu.:30315  
 Median :2016   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :45216  
 Mean   :2016                                                                             Mean   :47501  
 3rd Qu.:2016                                                                             3rd Qu.:72127  
 Max.   :2018                                                                             Max.   :99901  

    Latitude        Longitude       Primary.NAICS     Industrynames         Sector          Hospitalized   
 Min.   :-15.78   Min.   :-170.71   Min.   :    21   Length:21578       Length:21578       Min.   :0.000  
 1st Qu.: 32.20   1st Qu.: -95.40   1st Qu.:311411   Class :character   Class :character   1st Qu.:1.000  
 Median : 38.77   Median : -87.66   Median :333120   Mode  :character   Mode  :character   Median :1.000  
 Mean   : 36.72   Mean   : -87.69   Mean   :393922                                         Mean   :0.808  
 3rd Qu.: 41.09   3rd Qu.: -80.62   3rd Qu.:491110                                         3rd Qu.:1.000  
 Max.   : 61.29   Max.   : 145.75   Max.   :999999                                         Max.   :3.000  
 NA's   :91       NA's   :91        NA's   :3                                                             
   Amputation       Inspection      Final.Narrative       Nature      NatureTitle        Part.of.Body   
 Min.   :0.0000   Min.   : 837147   Length:21578       Min.   :   7   Length:21578       Min.   :   6  
 1st Qu.:0.0000   1st Qu.:1077550   Class :character   1st Qu.: 111   Class :character   1st Qu.: 320  
 Median :0.0000   Median :1121269   Mode  :character   Median :1311   Mode  :character   Median : 513  
 Mean   :0.2698   Mean   :1122091                      Mean   : 887                      Mean   :2283  
 3rd Qu.:1.0000   3rd Qu.:1170130                      3rd Qu.:1312                      3rd Qu.:4422  
 Max.   :9.0000   Max.   :1231162                      Max.   :9999                      Max.   :9999  
 NA's   :2        NA's   :13756                                                                        
 Part.of.Body.Title     Event          EventTitle       
 Length:21578       Min.   :  20   Length:21578       
 Class :character   1st Qu.: 642   Class :character   
 Mode  :character   Median :4331   Mode  :character   
                    Mean   :3905                      
                    3rd Qu.:6252                      
                    Max.   :9999                      
```

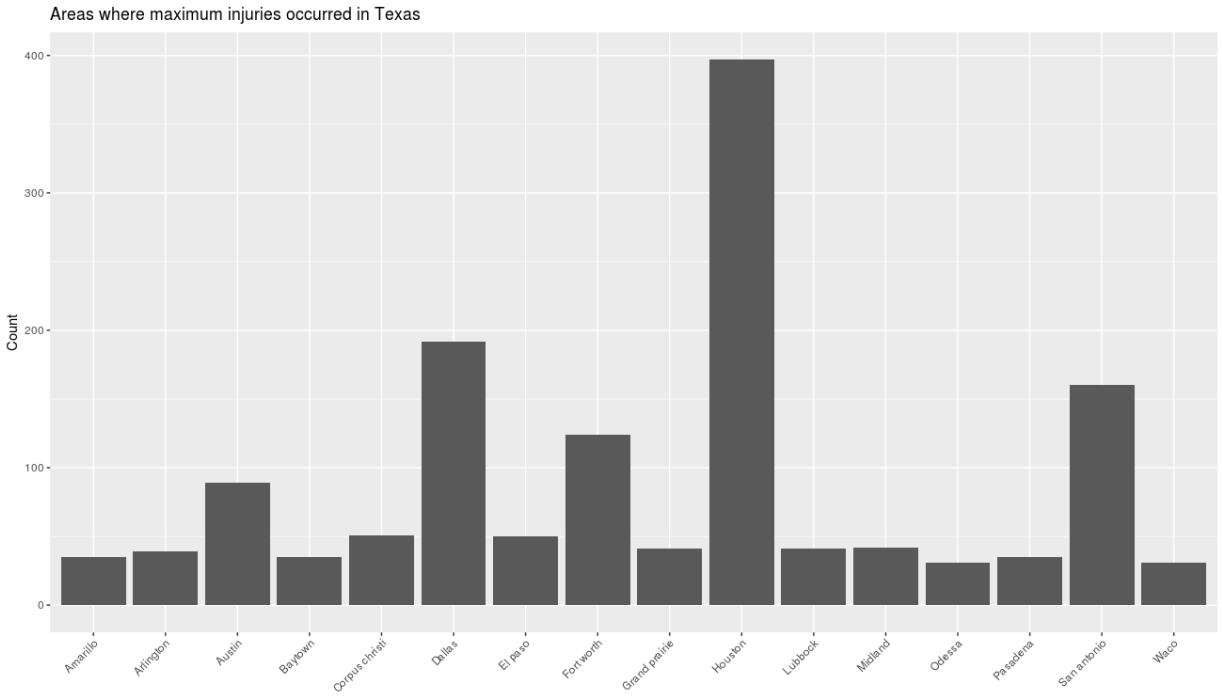
Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

## Objective 1:

1. To find the state with maximum number of injuries, we plotted the graph and Inferred that Texas had the most number of accidents occurred.

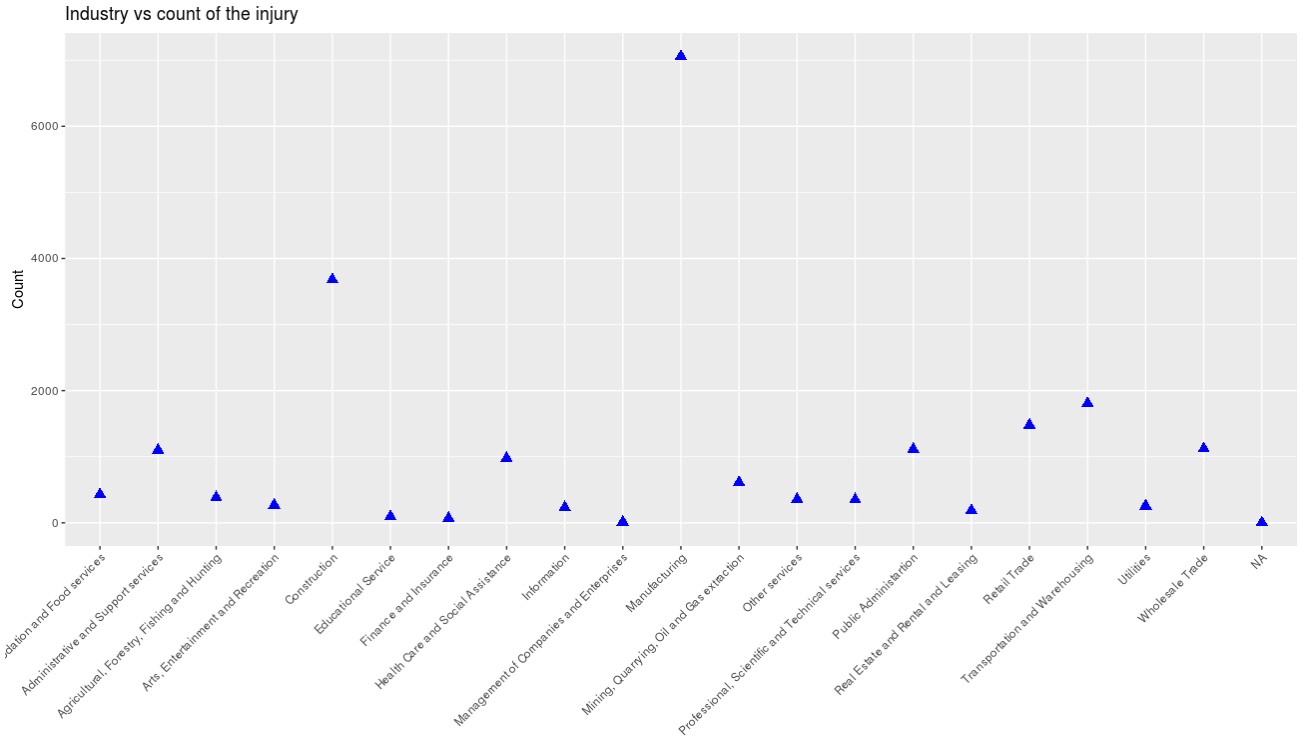



2. To find the City in Texas, where the injuries occurred, we plot the graph and infer that Houston had the most accident records.



Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

**3.** Finally, to find the sector which had the most injuries, we plot the graph and infer that the Manufacturing Sector had the most of Amputated and Hospitalized level of Injuries and accidents.

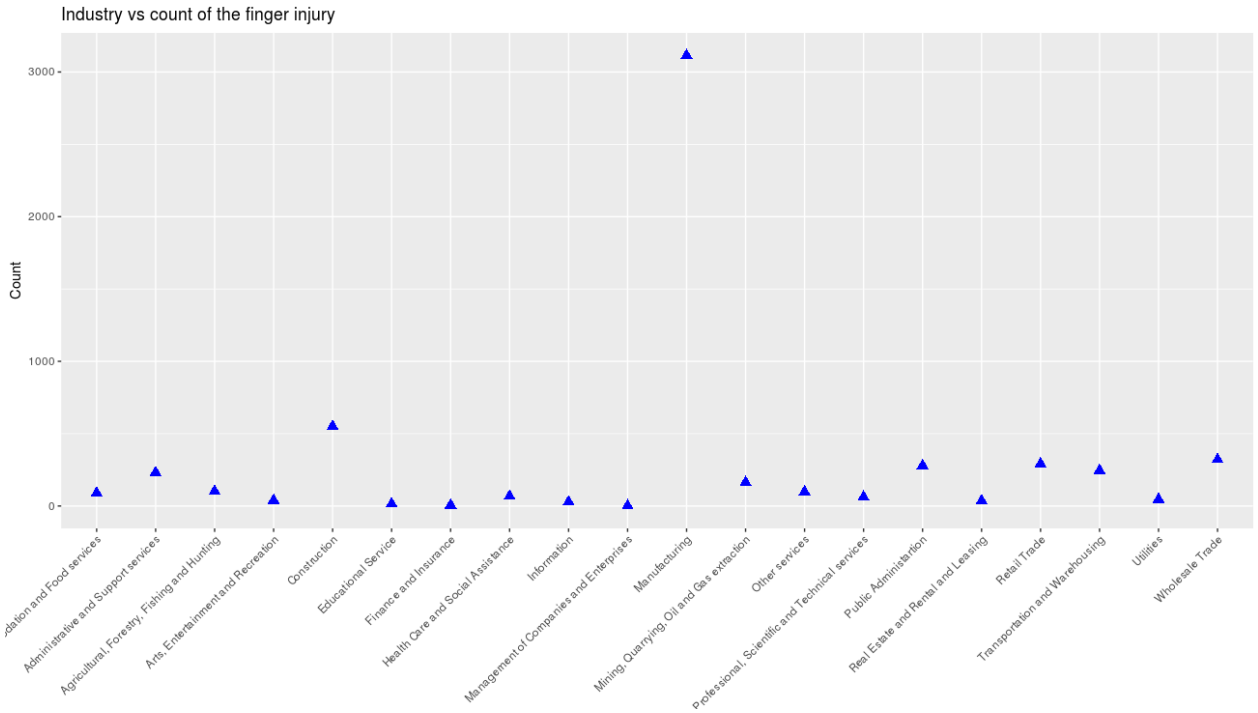



We would further proceed, creating predictive models using various techniques, mentioned in previous sections and select the best model to predict the occurrence of Injury for OSHA to recommend safety measures.

Aravind Sri Saravanan Sevvilam Parithi
Ananth Chandrasekaran
Kishore Ramakrishnan

## Objective 2:

For Objective 2, we created the subset new subset with Data Objects, which have an Amputation Level of greater than zero and plot the number of Finger Injuries, as we recorded values of Amputated from 1 to 4, with a higher frequency, which correspond to the finger injury from Code provided by OSHA. We also find that Manufacturing industries involve high risk of Personals' fingers amputated.




We further perform the Analysis and find a model, which would provide the Sales Proposal for any PPE Manufacturing company to increase its production of PPEs for Hand and Finger Safety.