A Study on Workplace Accidents

Ethan Tan Wee En  
*p2012085*DAAA/FT/2A/03  
Diploma in Applied A.I. and Analytics

Singapore PolytechnicDover Road, Singapore  
ethantan.20@ichat.sp.edu.sg

*Abstract*

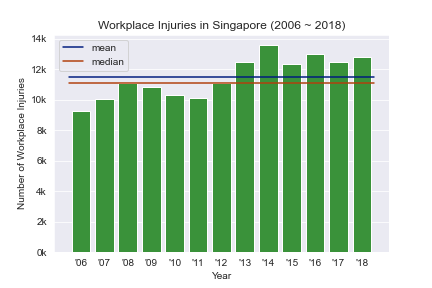
*Workplace accidents are frequent, and are, more often than not, benign. However, this does not mean that severe or even fatal accidents do not occur. From Manufacturing to Social Services, no industry is safe from harmful accidents. This paper seeks to investigate the most dangerous and hazardous types of workplace incidents and aims to outline the steps involved in designing a relevant machine learning model to predict the outcome of any such accident. The content comprises the general steps taken to build, train, score and evaluate the machine learning model. The data is obtained from the data.gov.sg website. The task is a classification problem, and the target of the machine learning model's predictions is the outcome of an accident at work, whether it be a minor or major injury or if it be fatal, given the details of the accident.*

Keywords

Technical Paper, Classification, Workplace, Accidents, Incidents

# Introduction

Workplace accidents are not new. In fact, more than 10,000 of such local cases are reported annually.



It is of this study's interest to investigate which are the more crucial factors of such workplace incidents (i.e., which lead to the most severe outcomes).

# Related Works

This paper is not the first of its kind to be published; other similar studies and research have been conducted [2][3] prior to the writing of this paper. This paper is neither a proof, an assertion, nor does it intend to compete with the above-mentioned works or other. Please refer to these articles [2][3] for more in-depth and comprehensive compilations.

# Experiment

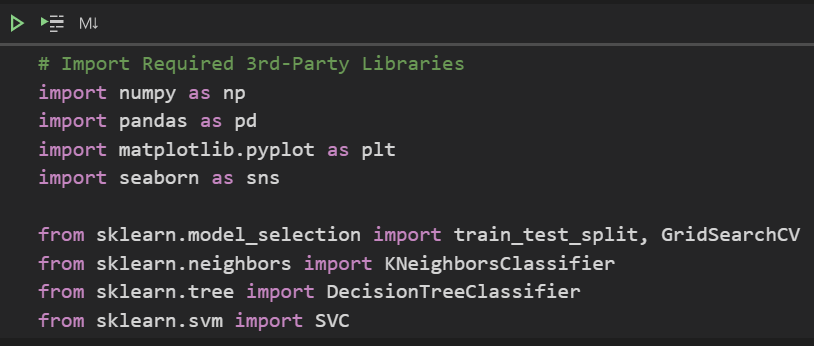
## Workspace

For this study, we will be using Python as our programming language of choice since there are numerous machine learning packages already available for use.

We will write and run all our code in a Jupyter notebook because it is convenient to observe all the output in an organized fashion.

## Dependencies

To lessen our workload and avoid reinventing the wheel, we will be utilizing third-party libraries, such as NumPy, Pandas, Matplotlib, Seaborn and Scikit-Learn.



This can be easily achieved as depicted in the figure above.

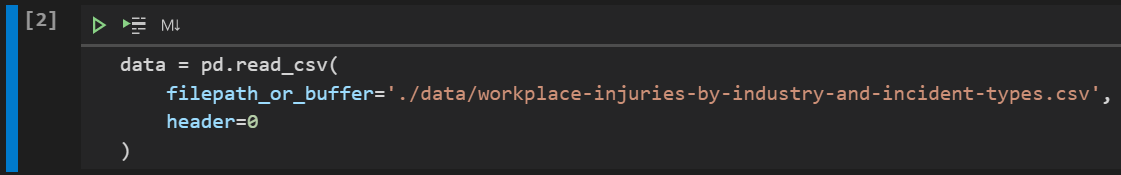
*Notes:*

1. *To keep this study simple, we will only consider the following three classification algorithms: K Nearest Neighbours, Decision Tree and Support Vector Machine.*
2. *These dependencies are not exhaustive (i.e., we might need to import other dependencies later).*

## Data

Before we begin to build any machine learning model, we need data, and, preferably, a lot of it. In our case, we have obtained our data from the data.gov.sg website [1]. The file of interest is titled "workplace-injuries-by-industry-and-incident-types", in Comma-Separated Values (csv) format.

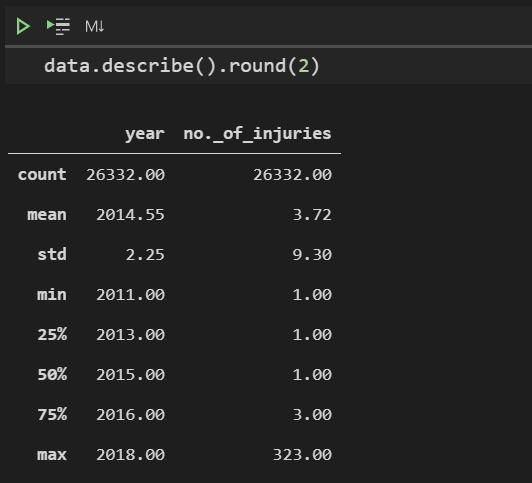
We will import our data using the *read\_csv* function defined in the *pandas* package. We will set the *header* parameter to 0 to specify that the first row of the file contains the header names.



The first thing we will do after importing the data is to inspect it. By the *head* method of the *pandas* *DataFrame* data structure, we can view the top 5 rows of our raw dataset.

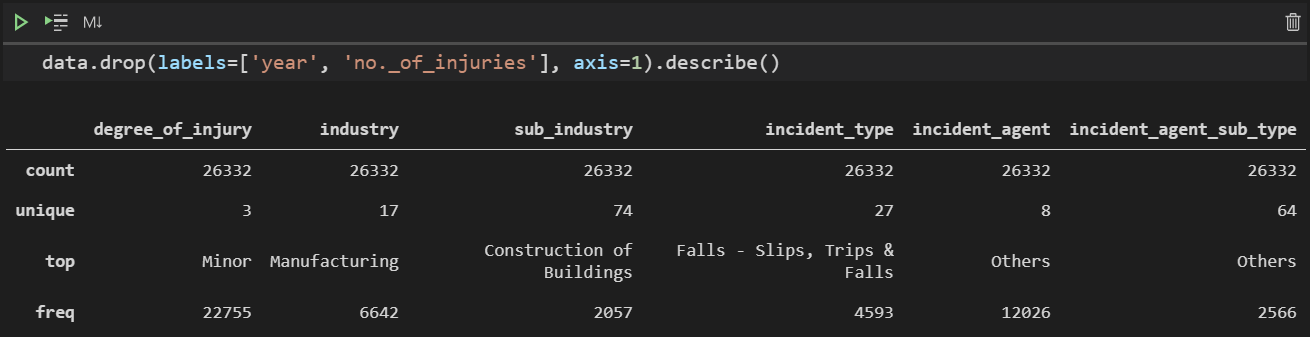


Next, we will take a look at the numerical summaries of the feature variables in the dataset. We will use the *describe* and *round* methods to generate the numerical summaries for the numerical variables in the dataset precise to 2 decimal places.



From this, we can tell that the dataset contains data from the year 2006 to 2018, and that there are only two numerical variables in the data.

We can use the same *describe* method to generate numerical summaries for the non-numerical variables. However, we need to drop the *year* and *no.\_of\_injuries­* columns temporarily as we want to focus on the non-numerical variables.

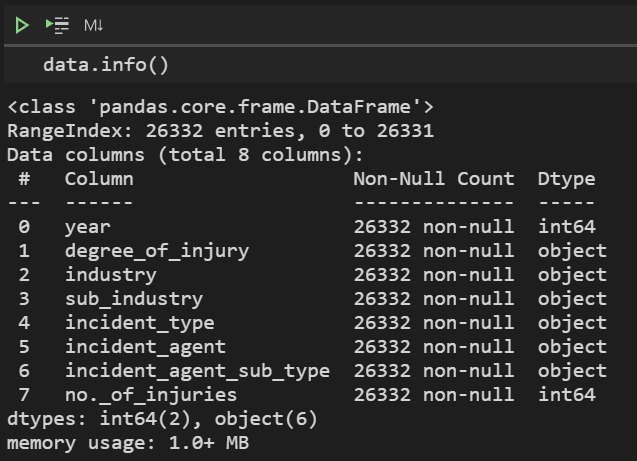


From this, we can tell that most workplace accidents are due to falls and the majority of incidents result in minor injuries. This is rather expected. There are 17 industries involved, which can be drilled down into 74 sub-industries.

Since these are currently categorical, we will have to conduct some form of encoding later on so that they are converted to numeric.

Preferably, we should combine some, if not most of the values to mitigate the *Curse of Dimensionality* when we encode the data subsequently.

We will perform one more inspection of the data using the *info* method to check if there are any missing values in the dataset.

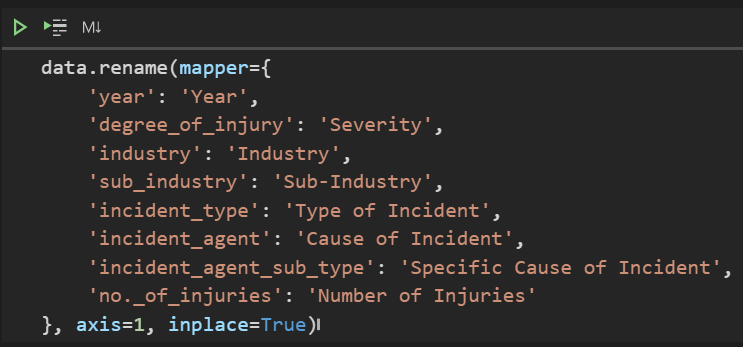


It seems there are no missing values in our data. We can thus skip any missing value imputation in the next step (data pre-processing).

## Pre-Processing

1. *Column Refactoring*

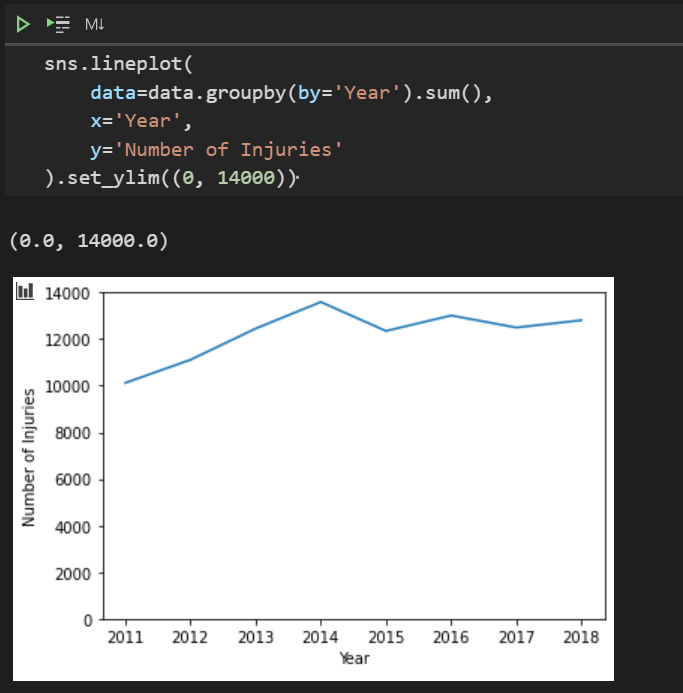
To improve readability and consistency, we will rename the columns in the *DataFrame*.



That's better. The columns have now been standardized, but the values have yet to be processed. Next, we will conduct exploratory data analysis.

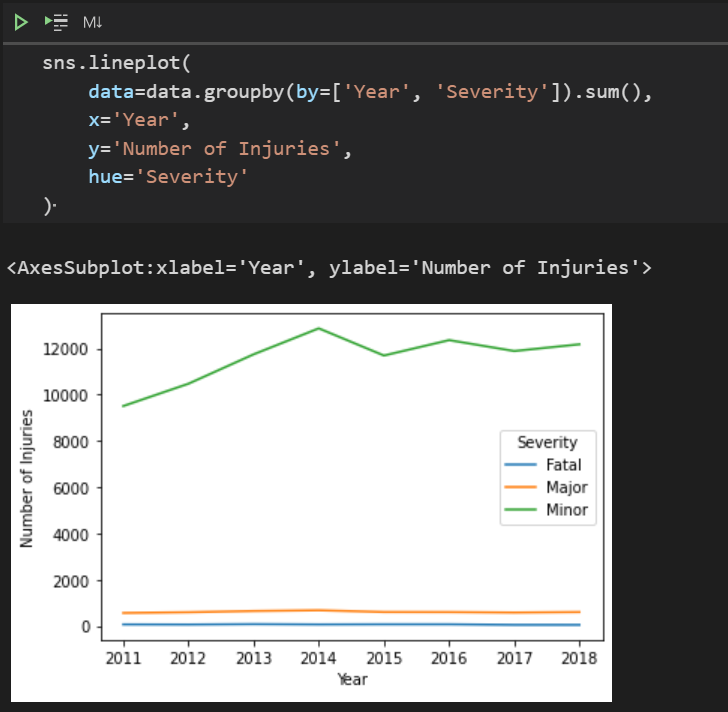
1. *Exploratory Data Analysis (EDA)*

In this step, we will explore the distributions and patterns in the data, mostly through graphs and plots.



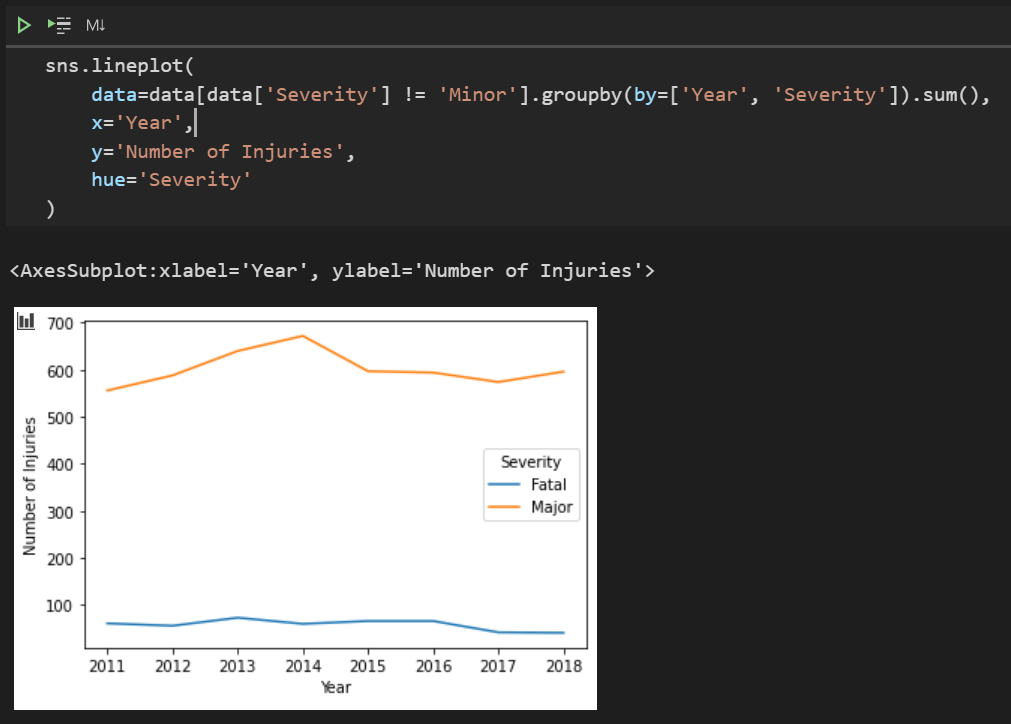
The number of workplace injuries peaked in 2014; nevertheless, number of workplace injuries seems to be on the rise.

Modifying the code above to account for the different severities of the accidents, we get the following plot:

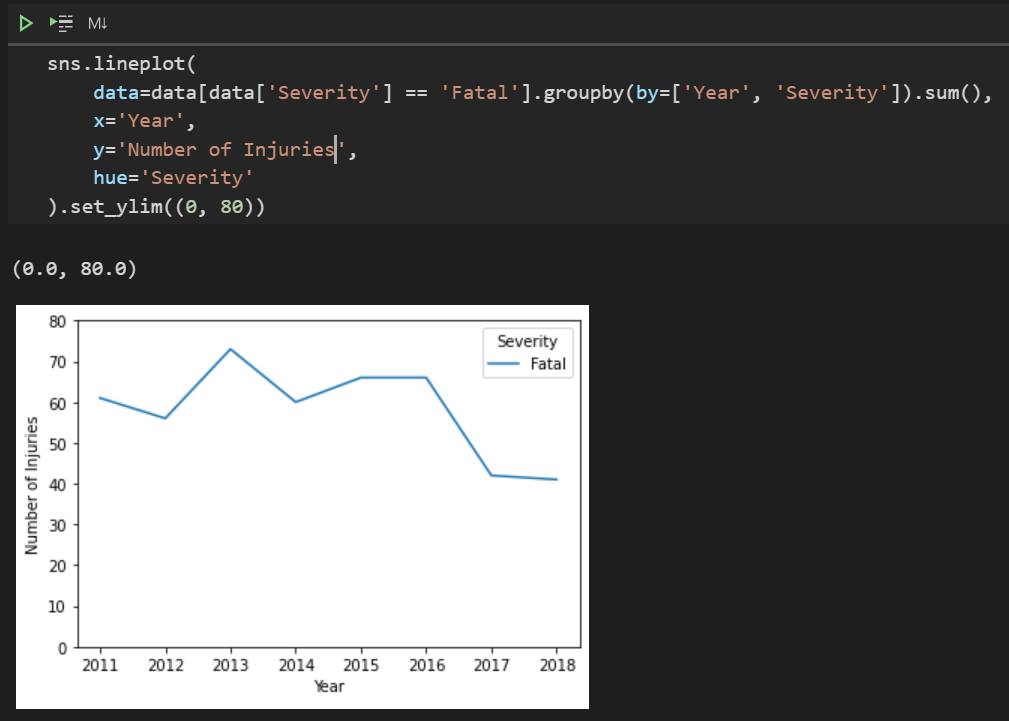


Minor injuries appear to make up most of the injuries. The overall trend seems to be represented by the trend of minor injuries.

Taking minor injuries out of consideration, we get the following plot instead:

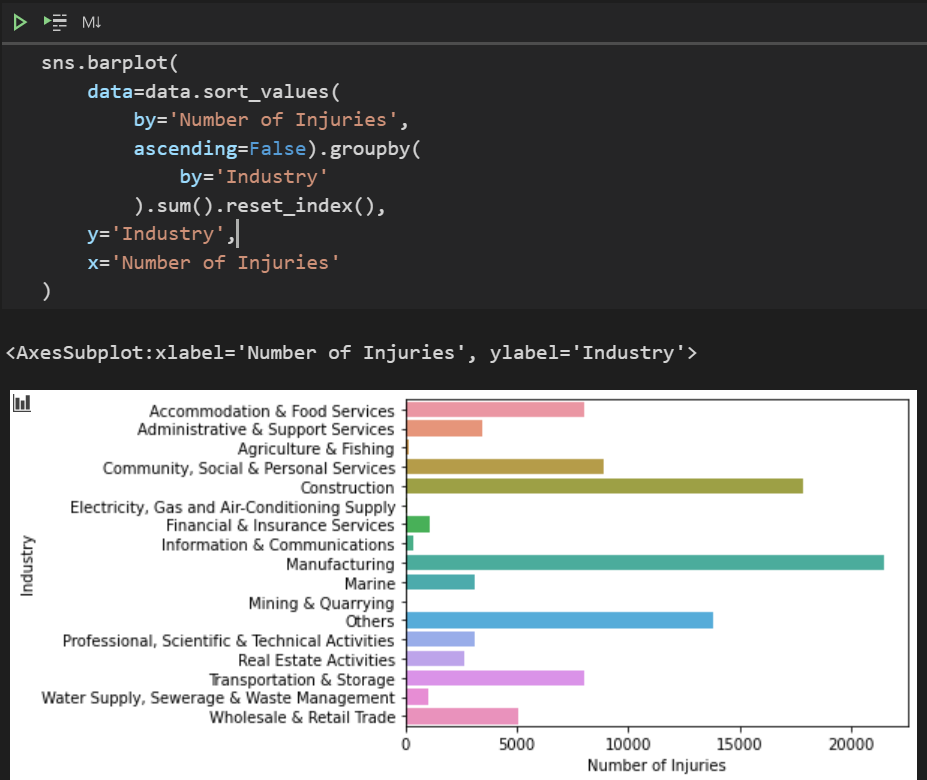


Number of major injuries are roughly the same throughout the years. Number of fatal injuries is still masked by the number of major industries, so let's focus only on fatal injuries.



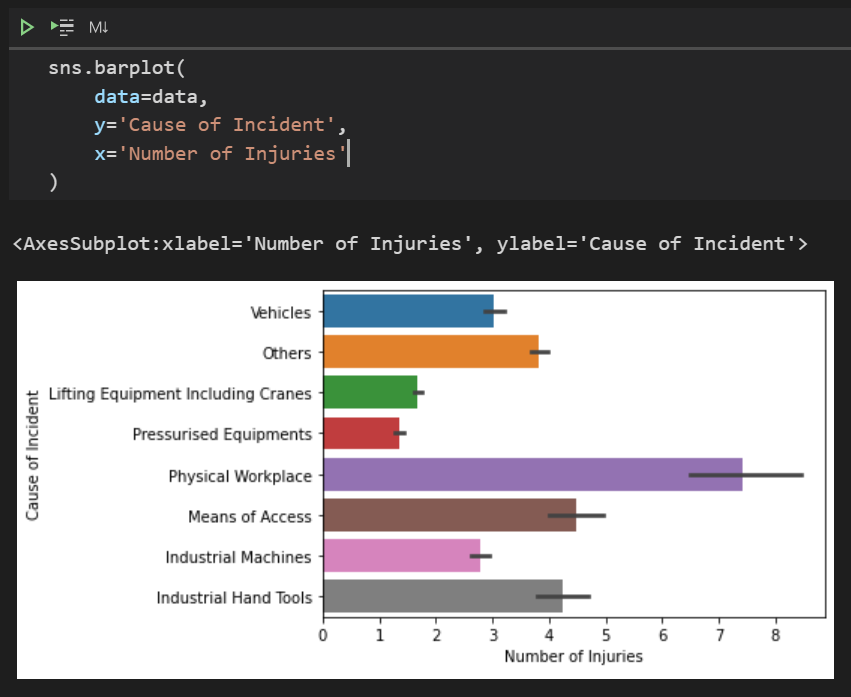
The number of fatal injuries seems to oscillate and fluctuate but follows a downward trend.

Now, let's investigate which industries do more accidents occur, and more severe ones. Merely plotting produces the following graph:



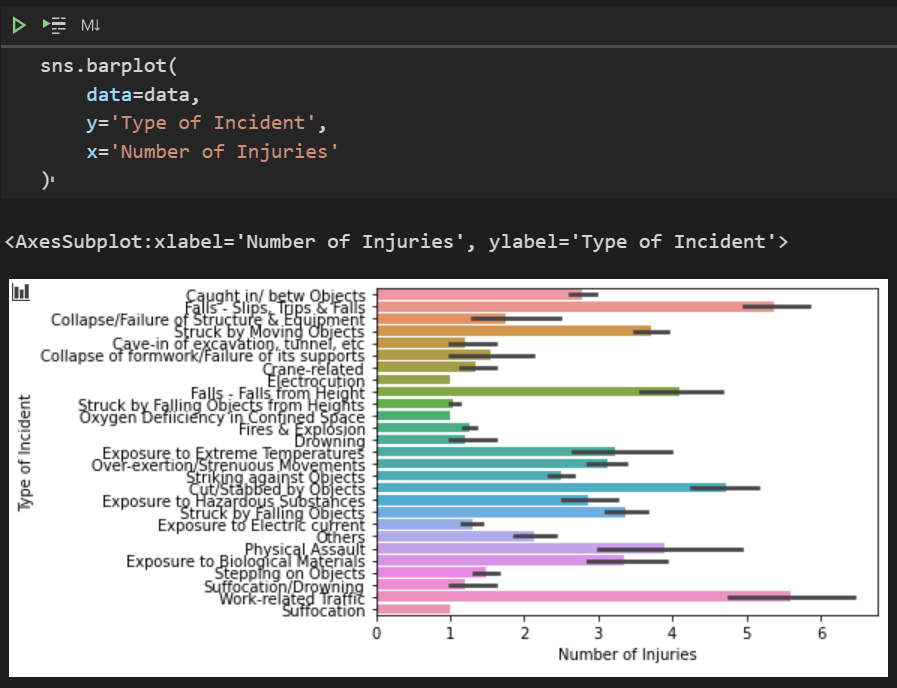
It is evident that there are too many categories for the column *Industry*. *Sub-Industry* is even more variable, with 74 possible categories. Encoding 74, or even 73, categories for one variable is definitely going to lead to a sparse matrix, which will prove challenging for our machine learning model to generalize

Plotting the *Number of Injuries* against the *Cause of Incident* reveals this:



Some of the categories for *Cause of Incident* can be grouped together too.

In addition, this is the plot for *Number of Injuries* against *Type of Incident*:

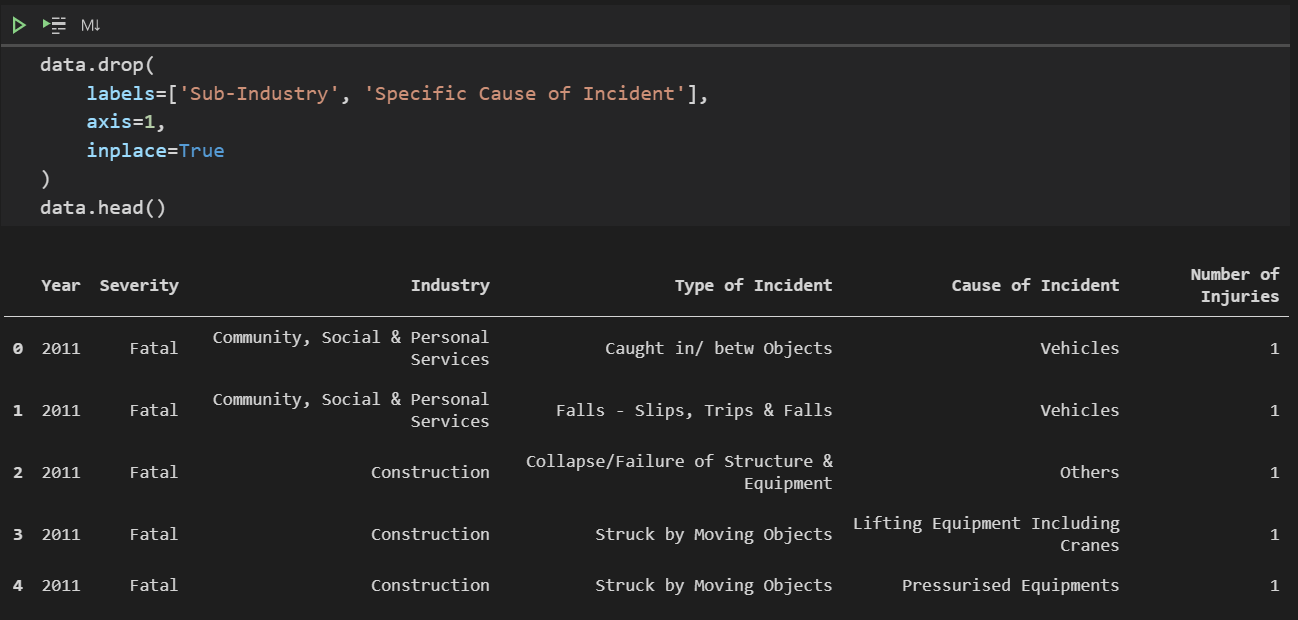


We have to merge categories for *Industry*, *Cause of Incident* and *Type of Incident*, in order to reduce the number of dimensions when we One-Hot Encode the data.

1. *Feature Selection*

Considering there are so many categories in *Industry*, *Cause of Incident* and *Type of Incident* already, we no longer need the sub-category columns, namely *Sub-Industry* and *Specific Cause of Incident*.

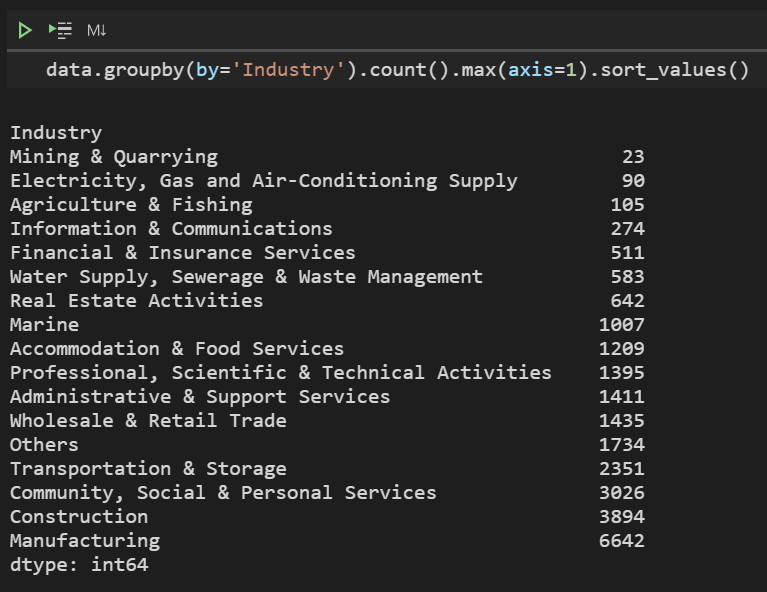
Hence, we can remove the columns using the *drop* method and specifying the removal to be performed *in-place*.



Now our dataset is more focused.

1. *Categorical Merging*

We will merge similar categories. For example, Electricity and Communications can be grouped as Utilities. To view the current possible values and their counts, we can use the *groupby* method, together with the *count* aggregate method to produce the following:



We aim to eliminate as many of the smaller categories as possible by merging them with larger ones.

1. *Scaling and Centring*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Building the Model

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Training the Model

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Scoring the Model

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

## Evaluating the Model

Hey

# Discussions

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file

# Conclusions

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. GovTech. (2019). *Workplace Injuries, Annual* [Online] Available at: <https://data.gov.sg/dataset/workplace-injuries-annual> [Accessed 29 May 2021]
2. Fatemeh Davoudi Kakhki, Steven A. Freeman, Gretchen A. Mosher. (2019). *Evaluating machine learning performance in predicting injury severity in agribusiness industries* [Online] Available at: <https://www.sciencedirect.com/science/article/pii/S092575351831107X> [Accessed 29 May 2021]
3. Zhang J.Y., Zi L.J., Hou Y.X., Deng D., Jiang W.T., Wang M.G. (2020). *A C-BiLSTM Approach to Classify Construction Accident Reports* [Online] Available at: <https://www.mdpi.com/2076-3417/10/17/5754/pdf> [Accessed 29 May 2021]
4. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
5. K. Elissa, “Title of paper if known,” unpublished.
6. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
7. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
8. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.