Data Analysis & Data Visualization- Using Python & Tableau

Overview

The business has come to you with questions regarding recent claim activity in our Workers' Compensation Program. Your goal will be to analyze the data set given to provide the business with insights to make future decisions. Things to keep in mind:

- As you work through the tasks, provide your answers as if you are presenting them to a decision-maker.
- Support your conclusions with either charts or visualizations.
- It is common to review and clean data sets before deriving your conclusions. Feel free to transform the data set in any way you deem fit. Just note what you altered, and why.
- Use your preferred tool to complete the challenge, e.g. MS Excel, programming language, statistical package, etc.

You will be evaluated on the steps you took to derive your insights, the insights themselves and your supporting work.

Data Cleaning: Data Cleaning is done in Python and a clean csv was saved locally. Dataset is analyzed in Python and visualized using Tableau.

Data Cleaning steps explained in Python File- [Data_Challenge_2022_Data_Cleaning.ipynb]

Data Visualization- Tableau Files- [DataChallenge2022_tableau.twb, DataChallenge2022_Part2_tableau.twb]

Hypothesis Test- Excel File- [cleaned_data_challenge_Analysis]

Task 1

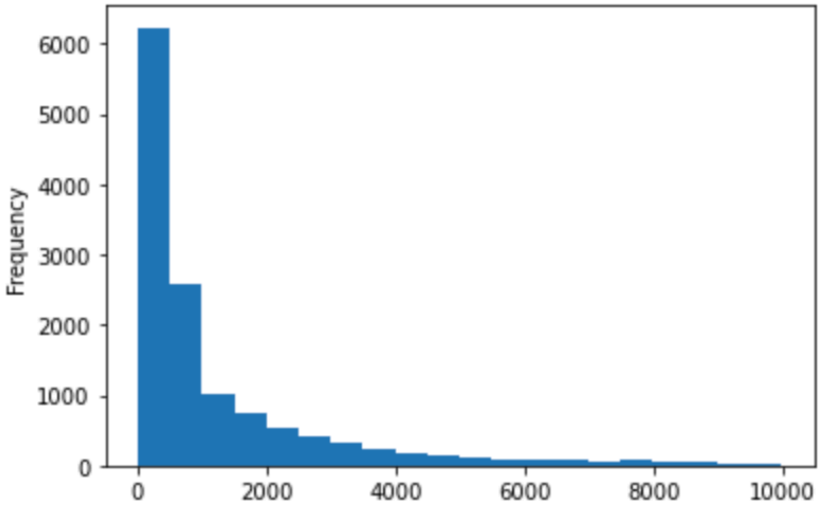
Provide a summary table which includes the following metrics per calendar year:

- Total number of claims
- Total Claim Cost
- Average Claim Cost
- Median Claim Cost
- Maximum Claim Cost
- Minimum Claim Cost

Injury_Year	Claim_Cost					
	Total number of claims	Total Claim Cost	Average Claim_Cost	Median	Minimum	Maximum
2018	6933	22600000.00	3263.959174	584.14	0	475562.01
2019	7120	25000000.00	3506.438191	594.005	0	692147.78

Follow-up Question:

- Is there a difference between Average Claim Cost and Median Claim Cost? Why might that be?
  There is a difference between Average Claim Cost and Median Claim Cost. Most of values in the claim cost falls within 0-2000 range making histogram right skewed as shown below.
  In this case, Median (50%) of the values in the dataset are within 0-2000 and mean has a higher value as the maximum Claim Cost for few values goes up to 692K
  Title: Claim Cost



Task 2

Provide a Five-Number Summary on age of injured worker by gender, include average age.

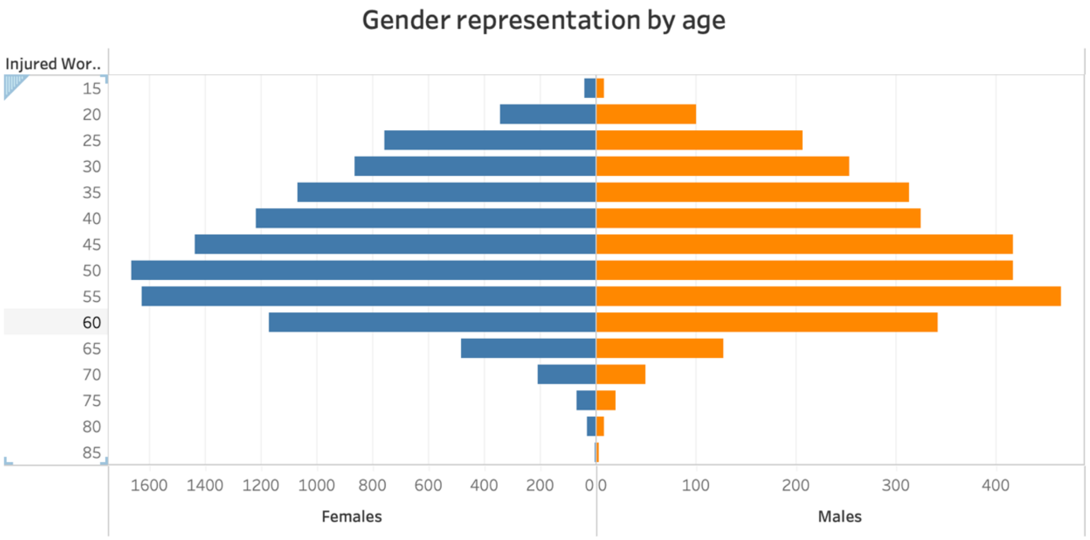
Summary Table:

	Mean	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Injured Male Worker Age	47	17	38	49	57	85
Injured Female Worker Age	47	17	38	49	57	89

Follow-up questions:

- Which visualization do you recommend to best show the age distribution by gender?

The best visualization to show the age distribution by gender is pyramid chart as shown below. This graph helps us to demonstrate distribution of ages between male and female in a dataset or sample. From the dataset, we can figure out that 1600 Females within 50-55 years of age tends to have more injury than 400 males in the same age category.



- A manager hypothesized that there is no correlation between gender and age of an injured worker, is that correct?

  Performing t-Test: Two-Sample Assuming Unequal Variances, we can infer that mean value of Injured_Male_Worker_Age and Injured_Female_Worker_age is same but we don't have enough evidence to prove that there is no correlation between gender and age of an injured worker. The level of statistical significance (p-value) for Injured_Female_Worker_age and Injured_Male_Worker_Age is greater than 0.05. Thus, we cannot reject null hypothesis and cannot prove alternate hypothesis that there is no correlation between gender and age of an injured worker. Result in Excel sheet-3 (Hypothesis-test)

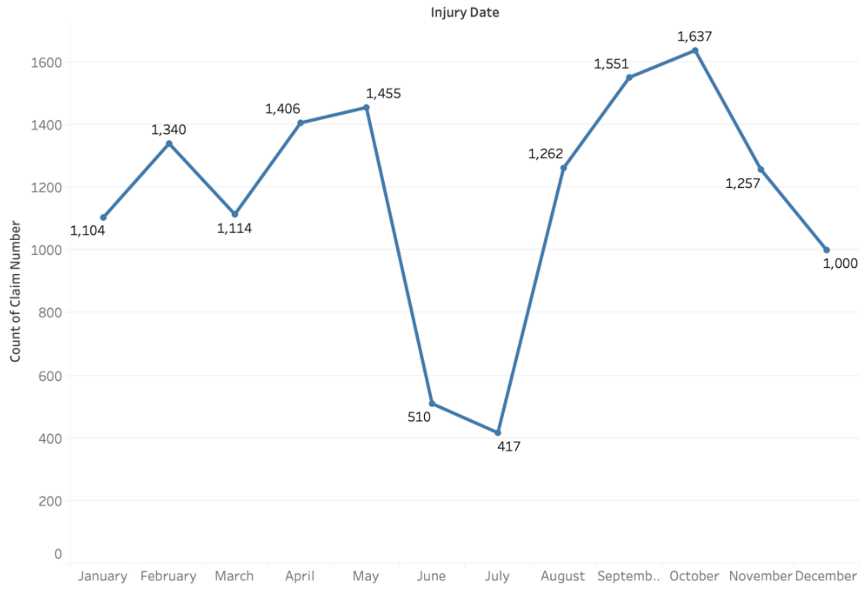
Task 3

The Risk Management's Loss Prevention department plans to hold a series of training events throughout the year and need recommendations on the best effective strategy. They would like recommendations on the following:

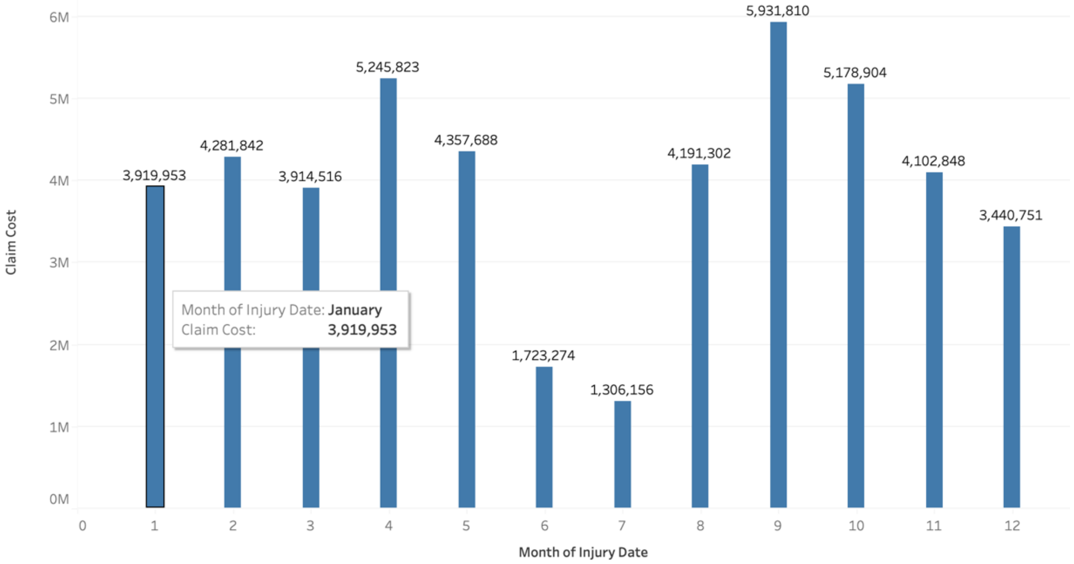
- What time of year is best to hold trainings to *prevent* claims?

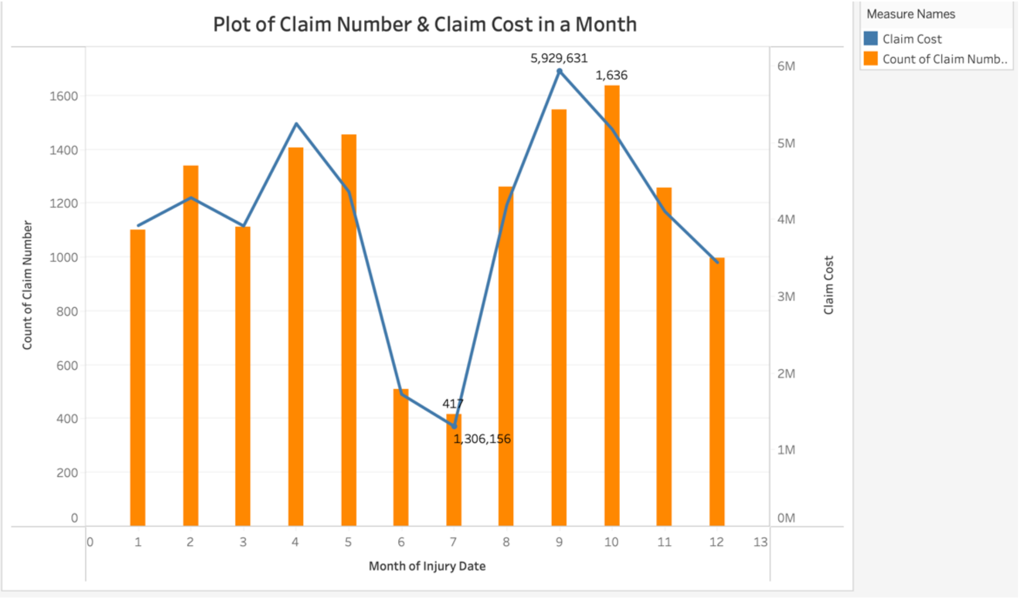
  The best time of year to hold trainings to prevent claims is 'August'. The maximum count of Claim Number is for the month of October and Maximum Claim Cost is for September. With this plot, we can observe that July has the lowest claims (Cost & Number) and there is an increasing trend up to October. Hence, as per my recommendations, it would be good to hold trainings for the workers in the month of August. Also, we need to analyze further to assess a sudden increase in claims after July.

Plot of Claim Number Count in a month- October has the highest and July has the lowest



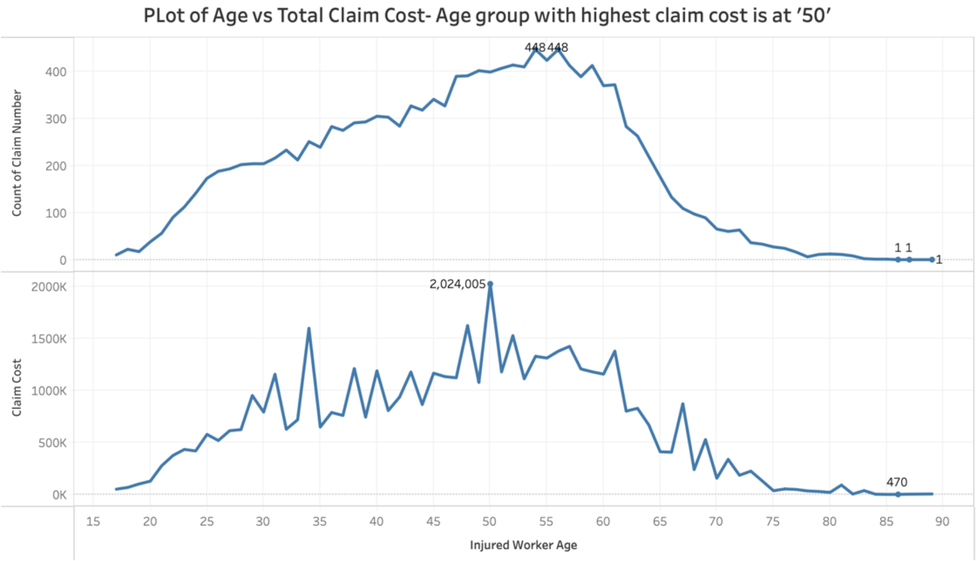
Plot of Sum of Claim Cost for every month - September has the highest and July has the lowest

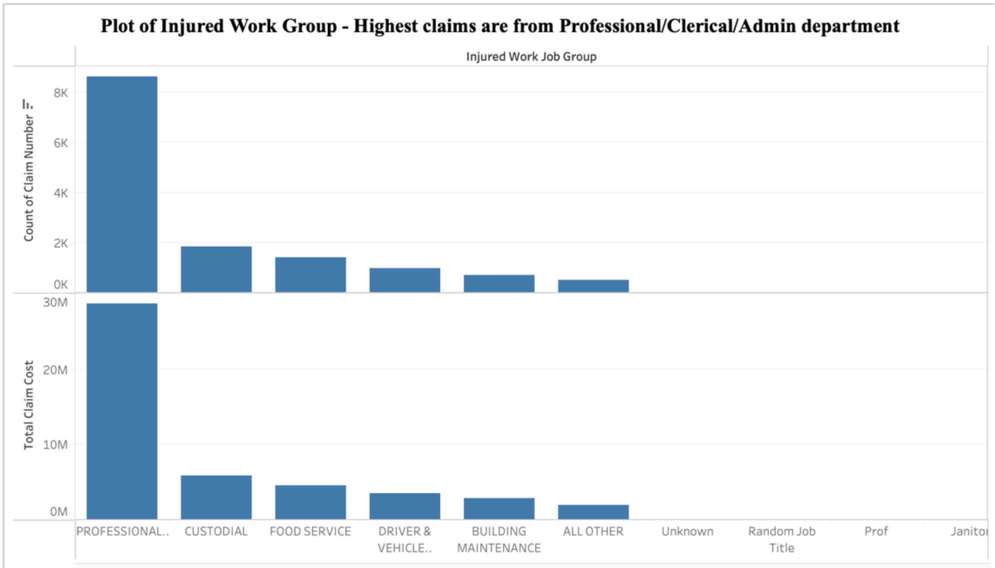




- What age and/or work group should they target during a training?

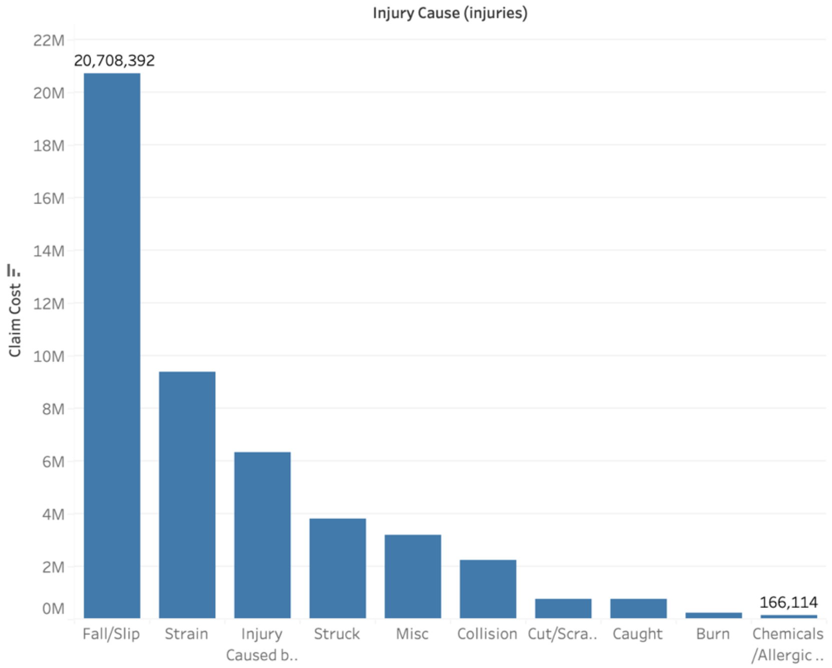
Age group '50-55' should be targeted during a training program as the maximum Claim Cost and highest number of claims occurred for this group.
Also, the maximum claims in Injured work group is from 'Professional/Clerical/Admin department. Hence, my recommendation would be to provide training to people who falls under '50-55' age group and works in Professional/admin/Clerical department. All the other work groups have significantly less claims.

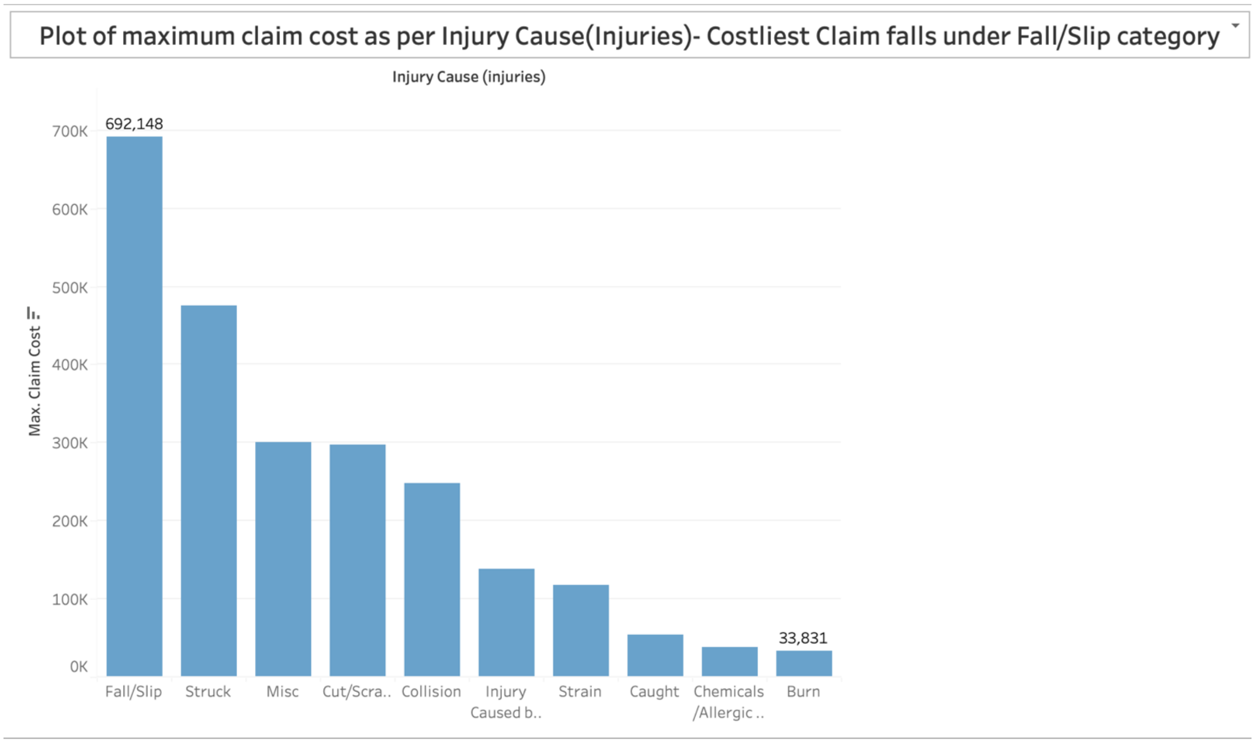



Plot of Injured Work Group - Highest claims are from Professional/Clerical/Admin department

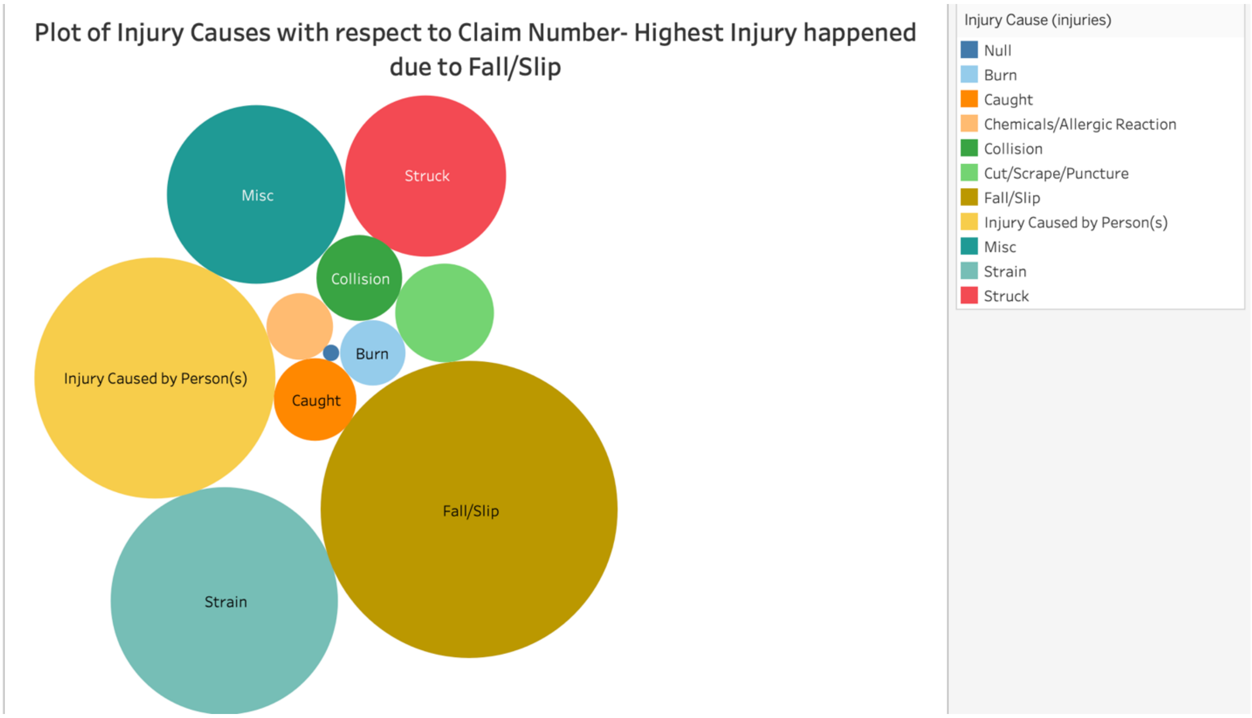
- What injury topics should be addressed?

By plotting Claim Number and Claim Cost, maximum injuries occurred due to Fall/Slip. Thus, while addressing the injuries caused by Fall/Slip it's important to understand the actual reason behind it. Both Histogram and Bubble plot shows the maximum claims is happening because of Fall/Slip. The costliest claim is in Fall/Slip category with a total of $692,148. The other common injuries are Strain and Injuries caused by person(s) which can be addressed during the training program.


Plot of total claim cost as per Injury Cause(Injuries)- Maximum claims are under Fall/Slip category



Plot of maximum claim cost as per Injury Cause(Injuries)- Costliest Claim falls under Fall/Slip category



Plot of Injury Causes with respect to Claim Number- Highest Injury happened due to Fall/Slip

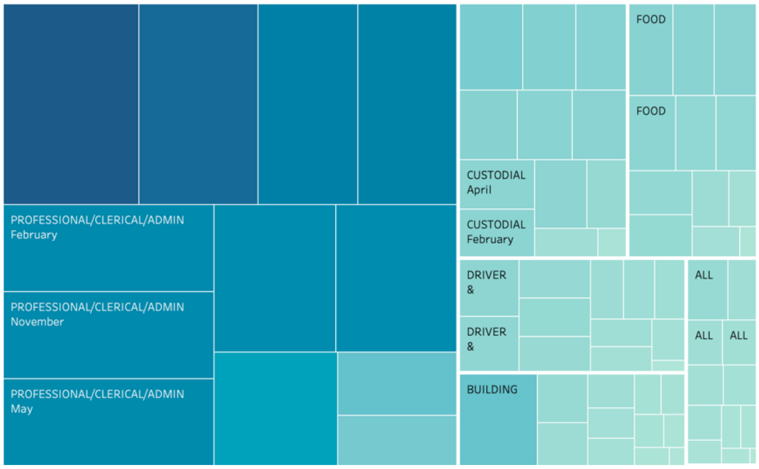
Recommendation: Combining, all the instances, department should focus to provide trainings to people from Professional/Clerical/Admin work group who had injury due to Fall/Slip, Struck and Injury caused by Person(s) within an age group of '50-55' in the month of August in a year to prevent claims.

Department should also try to understand the cause of Fall/Slip in Professional/Clerical/Admin work group.

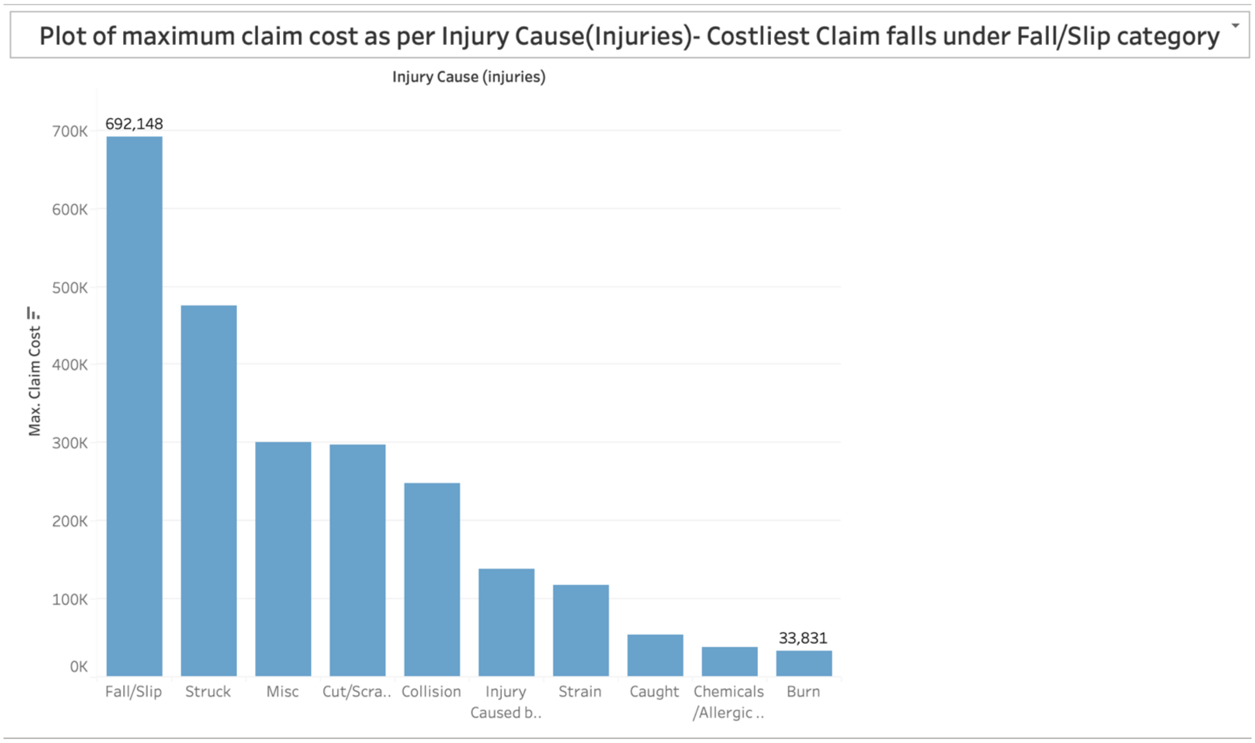
In 500-1000 words, present your recommendations for the above questions and provide at least two visualizations which support your ideas.

Some questions you might consider as you develop your recommendations:

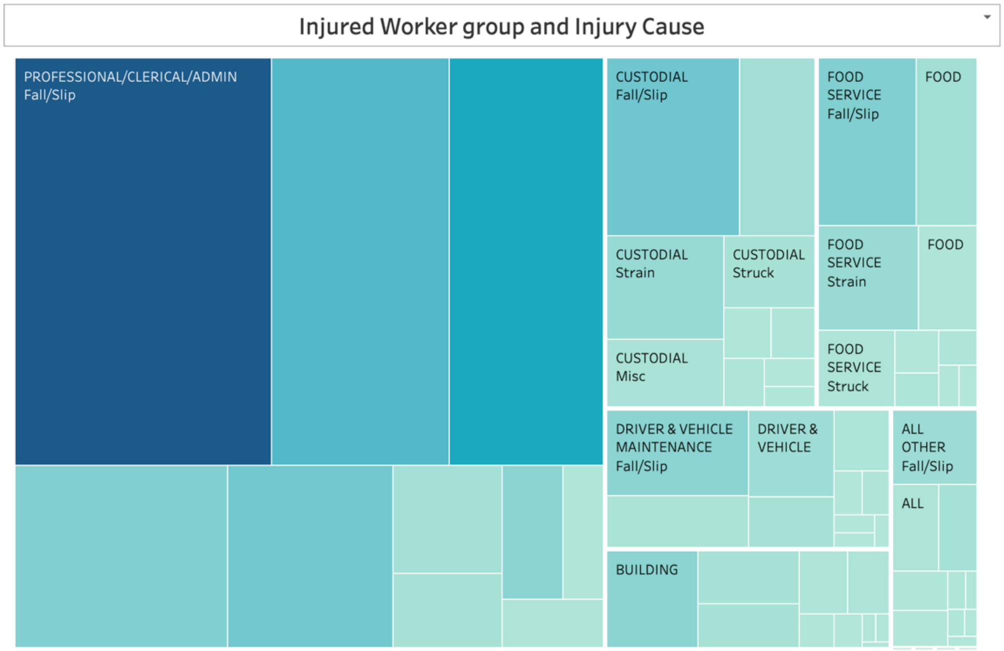
- *Which months do we see the highest volume of claims?* **October**
- *Do all work groups experience the same claim volume distribution throughout the year?*
  No, all work groups do not experience the same claim volume distribution throughout the year as shown below. Different colors represent different work groups claims in different months.



- *What are the most common injuries and which ones are the costliest?*
  Most common injuries are Fall/Slip, Strain and Injuries caused by person(s). Fall/Slip and Struck are the costliest.

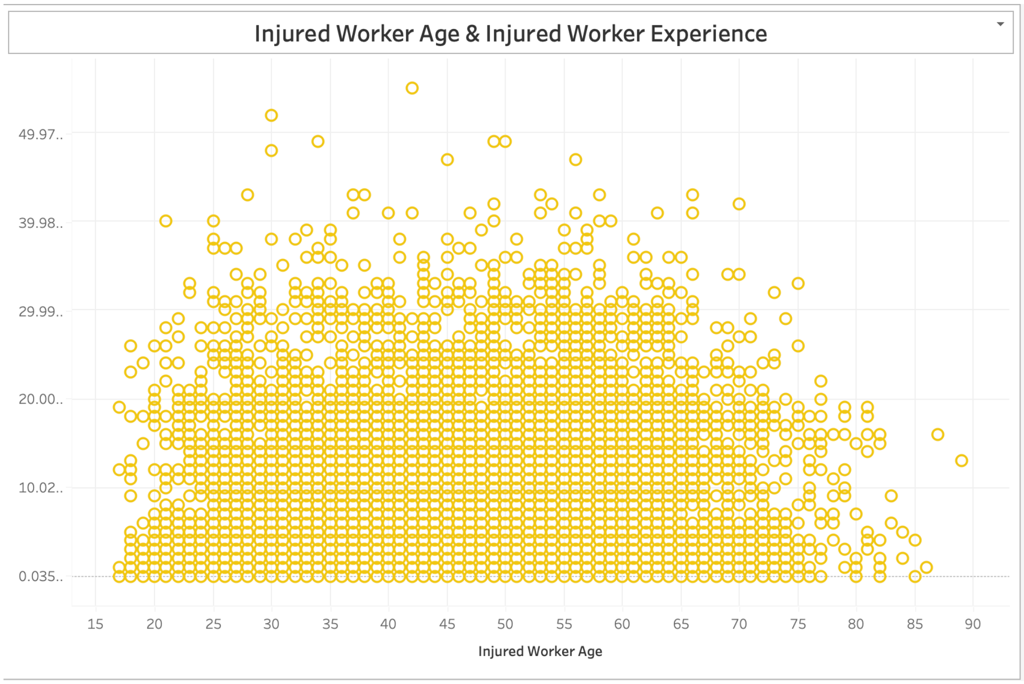

Plot of maximum claim cost as per Injury Cause(Injuries)- Costliest Claim falls under Fall/Slip category

- *Do all work groups experience the same type of injuries?* No, different work groups experience different injuries. Work group Professional/Clerical/Admin experience more injuries by Fall/Slip.


Injured Worker group and Injury Cause

- *Is there a correlation between an injured worker's experience and their age?*

  In Dataset for few rows, Years of experience is more than the Injured Worker’s age which makes their relationship abnormal. Thus, to establish correlation between them would be difficult.



Task 4

What other background information or pieces of data might have been helpful to expand your analysis?

With the information given in the dataset, if there would have been data about number of employees working in the organization, details of the employees to verify if the same person is claiming more than once in a year, would provide more relevance to the dataset.

Also, having employee id information would have helped to uniquely define the dataset.

Dataset of 5 years would have been useful to determine time series pattern by recognizing trend or seasonality in the dataset. These are the most relevant pieces of information which I would need to expand my analysis to recommend departments with better results.