**Executive Summary**

The objective for this project was to determine the feasibility of analytics models for the workers’ compensation industry and to provide strategic recommendations for the claims processing organization. The most significant problems to be addressed using analytics were reducing the cost per claim and the time it takes to process and close claims. In order to do so, we had to identify the most influential drivers in claims.

Our decision tree model was constructed using three comprehensive variables, centered around classification of the injury – BodyPart and InjuryNature and the TotalTime of the claim from open to close. Using just three variables helps to ensure that we do not overfit with our model. The decision tree was used to classify claims as Critical or not, and we found that cases were most likely to be considered critical when their open time was greater than 276 days. Each of our three variables were deemed to be useful in predicting Critical injuries based on calculations in R. We also found that injury types of strains and contusions, despite not always being from significant body parts, would often be classified as significant. Our visualization of the total compensation paid for each injury type supported this claim, as they were the two highest paid injury types in the data.

From our derived variable of Total Loss, we were able to see some of the most significant body parts affected by the injury based on the loss incurred by the company. Understandably, injuries affecting multiple body parts generated the greatest loss by the company. This will enable the claims processors to identify cases with great potential for loss ahead of time so that steps can be taken to reduce that cost.

When looking at Total Time vs. Total Compensation Paid, by injury nature, we see that Cancer, Severance and Mental Disorders were injuries that caused employees a great deal of time away from work and gain the most money from the claim. These injury types are more permanent in nature than others on the list, so it’s logical that they cost more for the insurance company.

Recommendations for the company largely deal with changing the culture of decision making from the top down, becoming highly data driven. This requires a lot of effort, including investing in dedicated analytics executives, analysts who are skilled in text mining to observe trends and gather information from the unstructured data littered through every claim and machine learning to develop new models that require little maintenance by humans. These efforts are all purposed to increase processing speed, with the goal of closing claims much quicker since it is shown that longer claims tend to cost more for the company over time. Once the time taken to close claims has been reduced, the monetary value will be reflected on the insurance company’s bottom line.

1. **Worker’s Compensation: An Introduction**

Worker’s Compensation is a form of insurance for employees that has been used in the United States for over 100 years. In its most general definition, worker’s compensation is a sum of money paid to an employee for any costs incurred or wages lost from an injury or illness directly caused by performing his or her duties. Because work injuries are so common and because each injury is unique, submitting and fulfilling a claim can be difficult and time consuming. This first section will serve to inform the reader of the details of filing a worker’s compensation claim and the intricacies involved

I. Claim requirements

There are three criteria that need to be met for an employee to be eligible to receive worker’s compensation. First, the company must be one that is required provide worker’s compensation to its employees, based on the nature of business (for profit or non-profit), number of employees, and type of work performed. Usually the work is physical in nature, although most farm labor is not included, and volunteers are usually not eligible to receive benefits. However, volunteer firefighters are almost always eligible to receive benefits. Secondly, the person receiving benefits must be an employee of the company that is paying the benefits. Seasonal or temporary workers, independent contractors, domestic (in home) workers, and undocumented workers are usually not eligible to receive benefits for an injury sustained on the job. Lastly, the injury or illness must be a direct result of performing a job’s duties. While it may seem straightforward, this last requirement is the one adjusters have trouble with the most. For instance: If an employee sustains an injury while on lunch break, but is running an errand for his boss, is he eligible to receive benefits? The answer depends on how the employee, the employer, and the insurance company all view the situation. It is important to note that all three requirements vary by state, as worker’s compensation laws are made at the state level.

II. Filing a Claim

If an injury or illness is sustained by an employee, the employee must first seek emergency medical attention if needed. A claim should be filed to the employer as soon as possible. To receive fair and proper benefits, the employee must provide as much documentation, written, video, audio or otherwise, as possible: type of injury and body parts affected, date, time and location of injury, parties involved in the accident, how the accident occurred, and any medical treatment received.

III. Receiving Benefits

Injuries and illnesses fall into four main categories: Temporary disability, permanent partial disability, permanent total disability, and death. Insurance payouts to an employee will depend on severity of injury and wages lost and costs incurred from the injury. If the injury is non permanent and will not cause the employee to lose income, typically the workers compensation extends to just paying medical bills to get the employee back to work as quickly and safely as possible. If the injury or illness causes an employee to temporarily miss substantial work time, the worker will receive checks to cover the loss in pay, which will stop immediately once the employee returns to work. Failure by an employee to quickly and properly document the claim could result in loss of compensation. Failure by an employer to provide proper and adequate compensation could result in fines, criminal charges, and lawsuits. Employers may not punish an employee for filing a workers comp claim, so it is in the employer’s best interest to comply with all valid claims.

**Big Data and Worker’s Compensation**

As with many industries, insurance companies today are working to leverage the use of Big Data in order to reduce costs associated with worker’s compensation pay outs. In order to do so, an effort must be made to reduce the time to close claims, which is often very long and inefficient. There is often a significant amount of unstructured data involved with worker’s compensation claims, which is time-consuming to parse through manually. As with other issues needing to be solved by Big Data, it is necessary to know which issues need to be addressed by the data before applying any strategies. In the context of worker’s compensation, companies can utilize Big Data analytics strategies to increase the speed of processing claims, as well as reducing their costs for claims and identifying fraudulent claims that are being made.

I. Time Reduction

Often times, the ability to make a simple decision based on specific claims will allow for a speedier closing of a claim. Insurers need to consider the areas that require the most effort in addressing a claim, such as manual data reviews. Text mining and analytics strategies will help in this effort by identifying key phrases that commonly occur for certain injury or claim types to be able to predict future claims that may contain similar phrases, thus limiting the need for manual reviews.

Combining the need to reduce time to process a claim and the total cost, it would be possible to settle some claims instantly. This could prove to be costly if the automatic payment is too high. By analyzing claims histories, an insurer will be able to set limits on the maximum instant payment allowed to be made to a single claim or grouping past claims that were paid immediately. By shortening the claim time for some cases, it could improve customer satisfaction and reduce labor costs by allowing adjusters to work less on any single case.

II. Clustering

Clustering similar claims together can help prioritize claims and properly assign to correct adjusters. More experienced adjusters can be automatically assigned the cases that are predetermined to be more complex or adjusters who have previously handled similar cases can be assigned claims from that same cluster. Less time taken to reassign claims will lead to less time taken to close a claim.

III. Cost Reduction

In order to reduce costs, it is beneficial to plan for a certain number of injuries and the associated costs based on the most common injuries encountered. Steps may be taken to reduce these common injuries, such as employee training or workplace improvements. Reduction in occurrence of the most common injuries will inevitably lead to the reduction of costs overall, stemming from the identification of the causes of the most common injuries by clustering similar past claims together. In the effort to better prepare for future costs, insurers will be able to estimate future claims costs, including litigation costs for cases that are disputed, by comparing losses of similar claims.

Text analytics may also enable a more efficient subrogation process, where it is identified that there is a third party that should be responsible for payment of the claim beyond the claimant or the insurance company. Subrogation cases are often characterized by unstructured data, such as the description of how the incident occurred. Pinpointing subrogation cases will maximize loss recovery and reduce loss expenses by properly identifying the correct culprit for each claim.

IV. Fraudulent Claims

To further reduce costs, it will be increasingly important to be able to identify fraudulent claims. The National Insurance Crime Bureau estimates that 10% of all casualty insurance claims are fraudulent. Currently, most companies identify fraudulent cases by a certain set of rules. Again, this must be investigated manually. However, predictive analytics can use a combination of those set rules, predictive modeling, text mining, database searching and exception reporting to identify fraud soon and more effectively at each stage of the claims cycle.

V. Data Cleansing

Basic data cleansing will always be an effective way to improve efficiency of claims processing, as with any big data problem. Claims data can come in unclean by missing specific fields, including misspelled words, having duplicated data and possibly information that is incorrect altogether. Cleansing these problems will lead to a standardized dataset that will allow for more appropriate clustering and predictions.

**2. Initial Hypotheses**

1. Exploring the most common injuries

In manual labor jobs, knee and back injuries seem to be the most common. They can be overused when and bent the wrong way, both of which can lead to injury. We want to explore the count of body parts injured and see where knees and back injuries rank.

b. Body Part will be driver of Compensation Paid

We believe the body part affected could have a big effect on Total Paid amount and whether the claim is critical or not. For example, brain, neck, back and knee injuries probably require the most health care costs. Additionally, injuries that occur more often will require more pay out costs from the company.

c. Total Time will be a driver of Compensation Paid

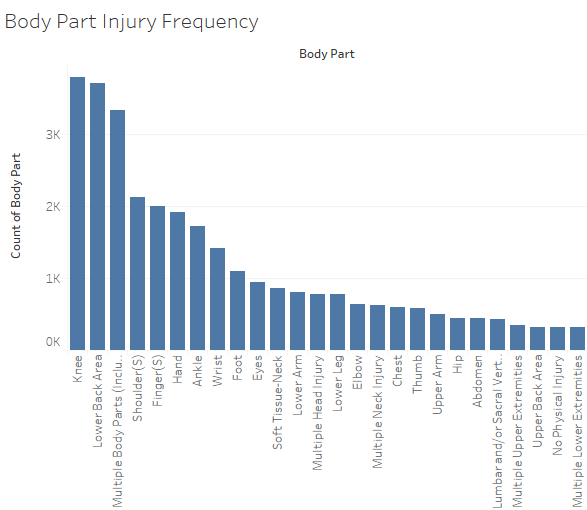
Claims that take a long time to process probably involve major injuries, and will require more healthcare costs to treat them. The company and insurance company might want to look at the relation between the two variables.

d. Body Part will be a driver of Average Total Loss

Total loss is a derived variable, equal to Number of Weeks \* Average Weekly Wage- Total Paid. A negative total loss indicates the company spent more on compensation than they would have if the employee was working their average hours at their average wages. A positive total loss indicates the employee lost out on money by not being able to work, and their worker’s compensation did not make up for it. We want to explore which body parts have the lowest average total loss, indicating which body part injuries are costing the company the most money when compared to a worker without an injury.

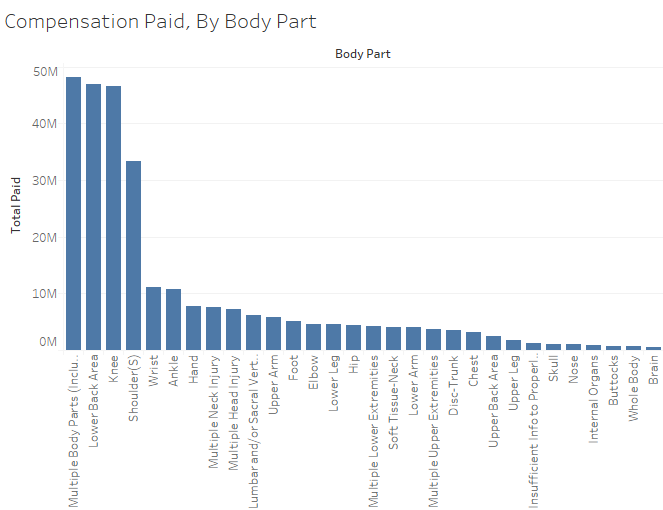
**3. Visualizations and Additional Hypotheses**

Initial Visualizations:

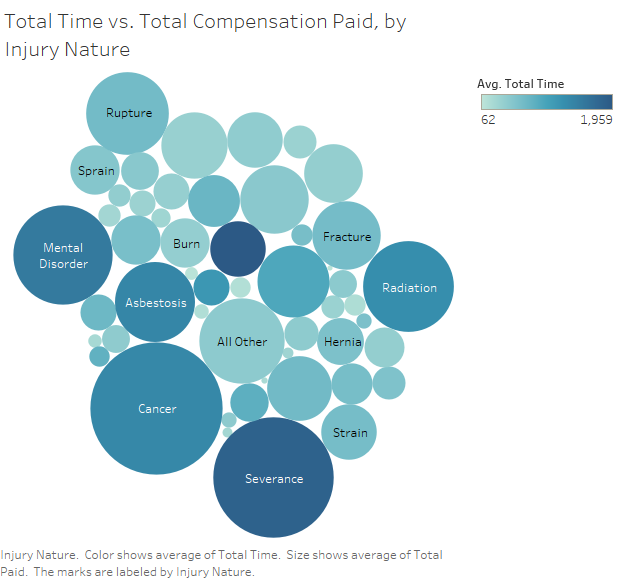
1. Exploring the most common injuries: Injured body parts with count greater than 300

Knee and lower back are the two most common injuries.

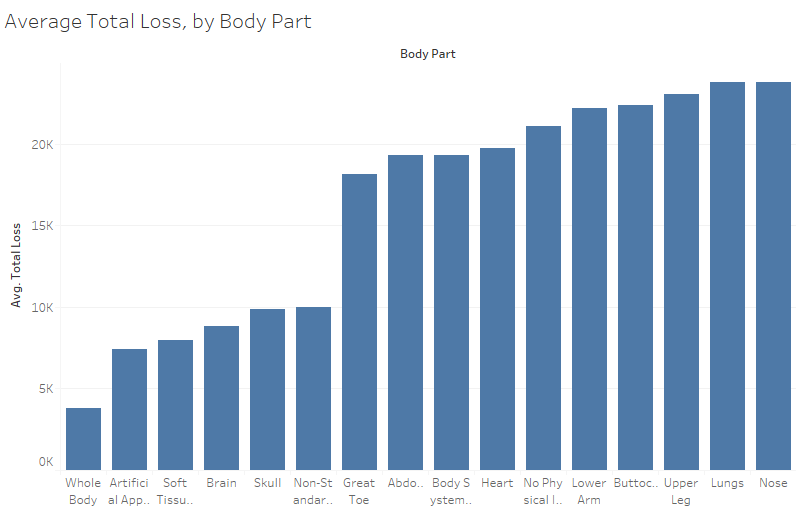
b. Body Part versus compensation paid



Multiple body parts, lower back, knees and shoulders are the clear top four.

c. Total Time versus Compensation Paid by Injury nature

d. Body part versus Average Total Loss



The company could be losing money on whole body, artificial appliances, soft tissue and brain and skull injuries.

Additional Hypotheses:

1. Injury nature could be a driver of cost

Whether due to frequency or severity, we want to explore if the nature of the injury (break, strain, bruise, etc) will drive the costs paid out by the company.

b. Exploring the occurrence of injury nature by body part

After noticing how certain injury natures (strains, sprains) and body parts (knee, back) can drive costs, we want to see how often certain injury natures coincide with certain body parts.

c. Exploring Total time versus total compensation paid, by body part.

After gaining some key insights from looking at a similar visualization regarding injury nature, we wanted to recreate this visualization but by body part.

**6).** There are a few different modeling techniques we can use to identify cost drivers in the Claims data. Since insurance companies care about the bottom line rather than processing time, we decided to classify the cases in the top 25% in TotalPaid as critical. From there we can use the models to see what drives cost and criticalness. In all cases, the dependent variable is “Critical”. We want to consider all possible indicator variables at first, then decide which are significant in driving Criticalness. So the predictors will be all other variables in the data set, with the exception of TotalPaid and all columns relating to it, such as IndemnityPaid.

**Logistic Regression:** Since there are many variables that could drive cost, it would be appropriate to use logistic regression over linear regression. This can be done in R by using the glm function, and using all variables as mentioned above to predict Critical. After running the model we can discard the insignificant variables, as they are not shown to be drivers of Critical. Once the model is set, predicting if a case is critical or not is as simple as plugging in the information from the case to the model’s respective coefficients. The model will return a number between 0 and 1. The closer it is to 1, the better chance there is at being considered critical. Regression is good for specifically determining whether a case is critical or not. However, logistic regression carries the chance of overfitting the data, where the model is accurate for the data being currently used, but may not be accurate for future claims data. Additionally, management may not understand a logistic regression model or what statistical significance is, and these concepts can be hard to explain.

**Clustering:** Clustering can be a quick way to determine the chances of a case being Critical. The goal here is to have cluster members in one cluster be similar to each other, and more clusters to be as distinct as possible from other clusters. R will make clusters based on the most significant drivers of Critical, display the characteristics of each cluster, and determine the odds of a cluster member being critical or not. Once the model is complete, a company can see what variables were used to form clusters, and match a case to a cluster. For instance, one cluster could consist of women older than 30 with head injuries, and this cluster could be 85% critical cases. If a new case occurs where a woman older 30 gets a head injury, an adjuster can look at her cluster and predict if the compensation will be critical or not. Clusters can be overfit, where each case is its own cluster, and they can be underfit, where the clusters are not distinct from each other. Generally, prediction is not a feature of cluster analysis. Clusters are better for determining groups with similar attributes, and looking for reasons why some clusters are more critical than others.

**Decision Tree:** Decision trees are visual models where the cases are separated into mostly homogeneous groups (branches) to determine the chances that members of a branch are critical or not. Similar to both clustering and logistic regression R can determine what variables and values to branch on. Seeing whether a case is critical or not is as simple as following the tree from the top down based on the attributes of the case. For example, it could initially split on gender, then on the Men’s branch split on Body Part, then on the Leg branch split on Injury type. Each end of a branch contains a group and the chances of that group being critical. An overfit decision tree has a branch for every case, and an underfit decision tree does not have enough branches to make accurate predictions. We decided to make a decision tree model for data, as we felt it was the best of both worlds between clustering and logistic regression.

**7. Model and Results**

Since we decided on the decision tree model, we had to figure out the variables to use. We landed on BodyPart, InjuryNature, and TotalTime, since these variables are comprehensive enough to give us what we need but the inclusion of just 3 is succinct enough to not overfit. Based on our tree, it seems that a critical injury would ideally be more than 276 days. For body part, it seems that there are a list of significant body part injuries that tend to be more critical, but even for injuries that are not in this list there are still critical ones as well. For the significant body parts, it seems that a TotalTime of less than 366 days would produces critical injuries, so the ideal range is between 276 and 366. For the non-significant body parts, it seems that there are a list of significant injuries that would produce critical injuries. For example, the most paid injuries (strain and contusion) are in this list.

**8. Analytics 3.0**

The shift from Analytics 1.0 and Analytics 2.0 to Analytics 3.0 is characterized by companies having their data and analytics strategies more embedded into the company as a whole. The purpose of big data is now not only to drive improvement of internal business decisions, but to also driving more value to customers in the products and services provided. It also should have a trickle-down effect on all employees, from leadership to frontline employees, such as individual claims workers in this organization. To become more of an Analytics 3.0 company, the claims processing organization should start by having a C-level executive devoted to analytics. This person will create and drive all analytics decisions. For companies to be purely driven by analytics, there is a need to have dedicated executive leadership. This will provide a shift in general decision making for the organization. Data-driven decisions allow for greater certainty when taking action, which is paramount when analyzing thousands of individual claims. If decisions are to be made automatically in large volumes, the organization must have confidence in what is predicted by the models. This certainty will be fueled by extensive formal testing on the difference between causation and correlation. Using methods such as decision trees to classify claims and drive decision making, the analytics division of the organization will be able to identify what truly drives claims cost, whether that is the time of the claim or injury nature. The nature of claims processing is highly reliant on unstructured data, the analysis of which is necessary to become an Analytics 3.0 company. Analytics 3.0 is also characterized by the emphasis on prescriptive analytics, specifying optimal options and actions for the future. This branch of analytics is vital to the worker’s compensation industry, because it revolves around making decisions on and prioritizing claims. Thus, analytics must be embedded in all processes and planning. It is no longer enough to simply be information providers using analytics, but it is now necessary to become insight providers as well. To increase speed to market for all decisions, it is important in the Analytics 3.0 environment to also place high importance on machine learning capabilities. An emphasis on machine learning within the worker’s compensation industry will allow for greater scaling out of models and increase the possible number of models to be utilized as well, with limited human effort. In conclusion, this organization needs to devote resources to analytics from leadership down to the teams devoted to extracting, analyzing and prescribing action upon the data from claims.

**Recommendation**s:

* Hire a C-level executive to lead data analytics division
* Invest in text analysis capabilities for the large volume of unstructured data - focus on identifying key phrases and words that will help in classifying claims and fueling prescriptive analytics
* Design formal tests to determine causation for critical claims (identifying most influential variables) using decision tree models
* Identify and implement machine learning methodologies to automatically maintain predictive models with limited human effort

**Appendix**

**Data Cleaning**

Total Loss and Total Cost per Day variables were created.

NA values and age < 17 were removed for Claimant Age.

**Variables Created:**

**For Analysis and Prediction:**

TotalTime = ClaimClosed - ClaimOpened

Longer claim times probably correlate with more costly major injuries

TotalLoss=Number of Weeks \* Average Weekly Wage- Total Paid

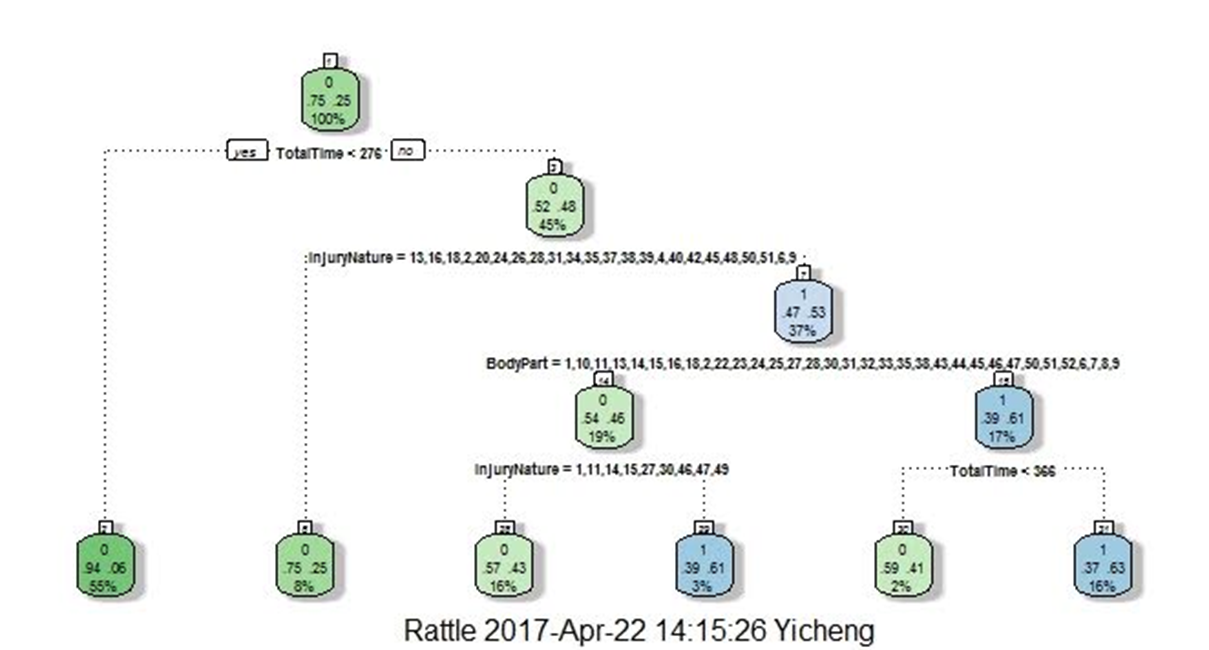
If TotalLoss is negative, the company paid the employee more than they would receive from working regular hours. A highly negative TotalLoss could indicate a major, critical injury

TotalCostPerDay=TotalPaid/TotalTime

A variable created to see the average payout over the length of the claim

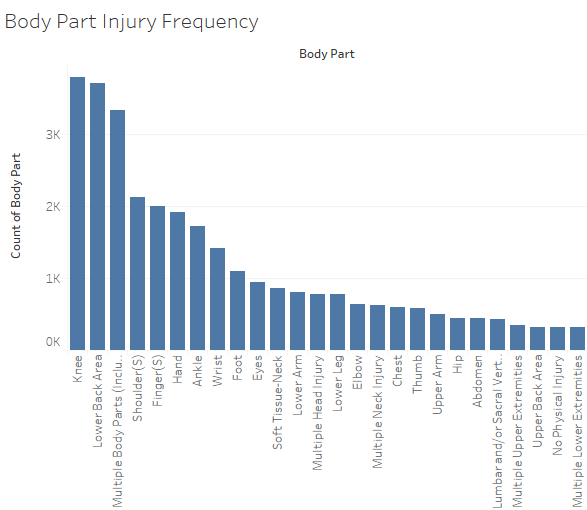
**Binary Dependent Variable**

We named our dependent variable “Critical”. If a claim is in the top 25% of all claims in terms of amount paid out (TotalPaid) by the insurance company, that case will be classified as critical. We explore if all other variables (besides TotalPaid and other variables that relate to TotalPaid) are significant, and use these to determine if future claims are Critical or not, with 1 being critical and 0 being non critical.

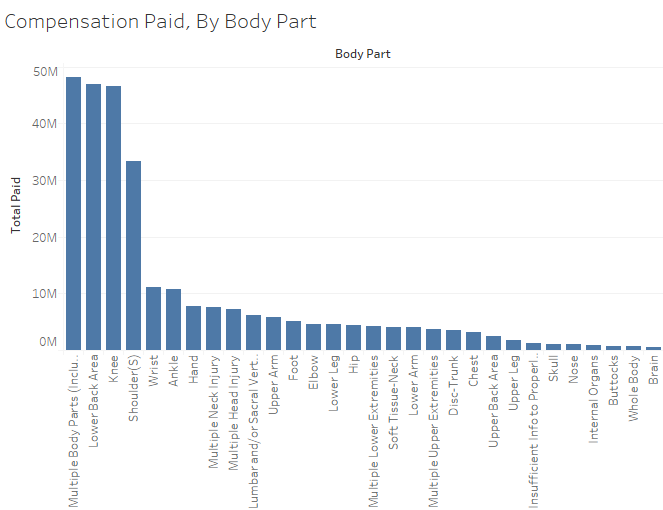
**Output of Model**

We used a decision tree. R calculations determined that TotalTime, InjuryNature, and BodyPart are all useful in determining if a claim will be Critical or not.

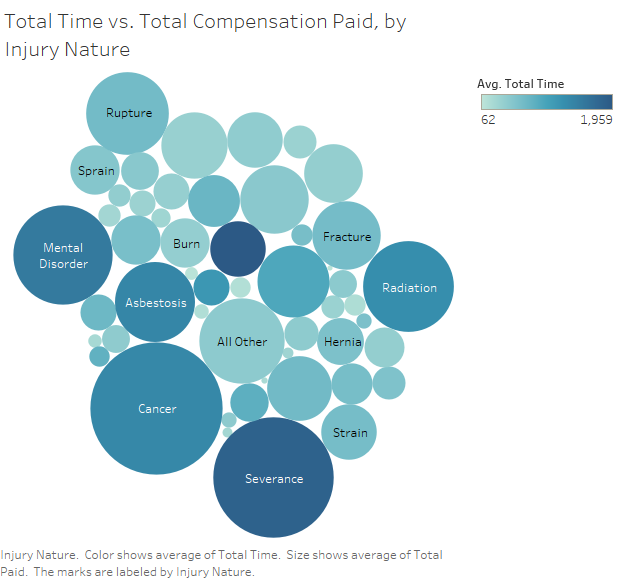
**Visualizations and Explanations**



1. Initial exploration to see which body parts are injured the most frequently. Knee, lower back, and multiple body part injuries stand out as occurring the most often.

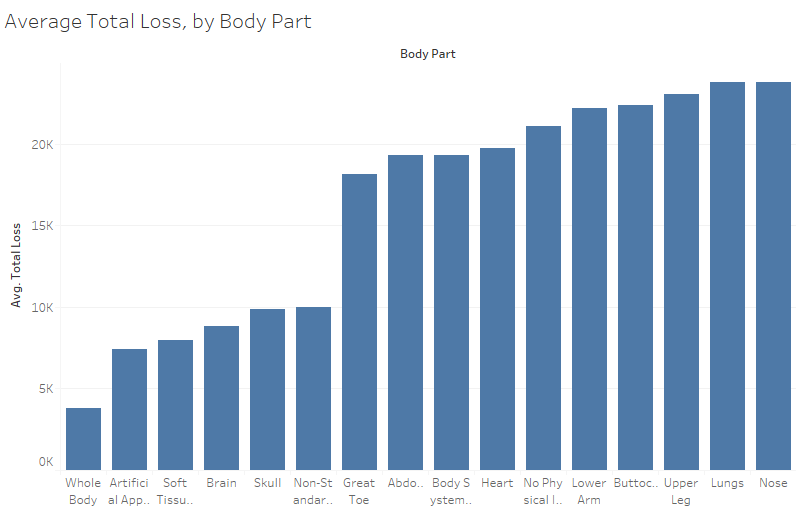


2. We looked at what injured body part companies were paying for the most, in terms of total dollars spent. As expected, the most common injuries also incur the most costs.

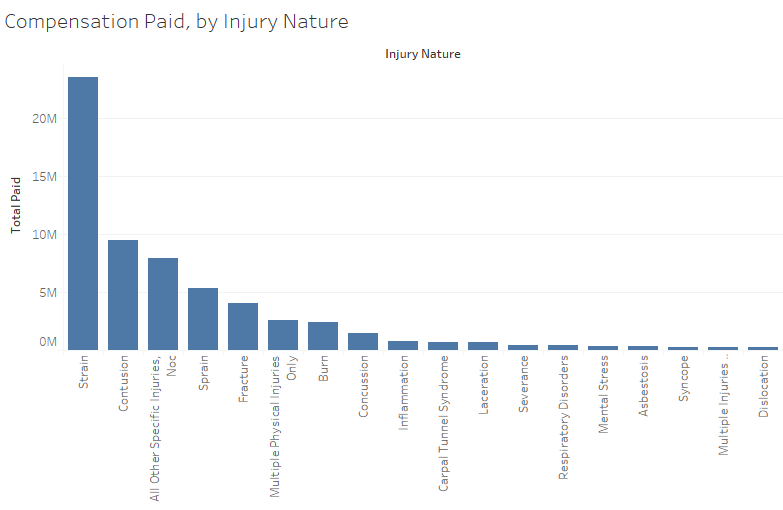


3. This bubble chart gives the information with regards to the total paid by the company and the total time the employee was out of the workstation because of the injury. Bubble size is used to represent the average of total paid by the company and the color of the bubble is used to represent the time the employee was out of work.

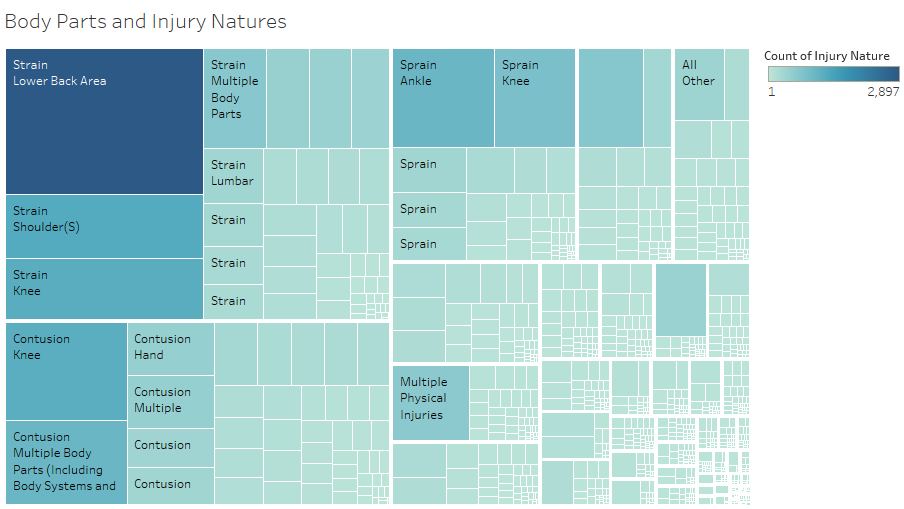
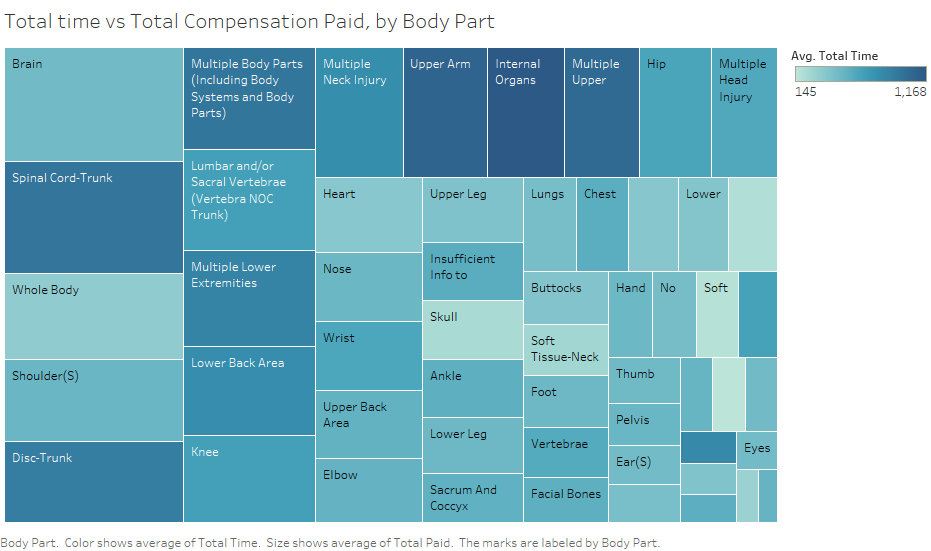
Observation: Cancer is the injury nature to which the company paid the most and other injuries are the ones which kept the employees away from work for a longer duration followed by severance.



4. Major injuries, like whole body and brain and skull, have a low average total loss, which means the company could be losing more money than expected on these types of injuries.



5. Exploring which injury types required the most total compensation. Strains clearly account for the most compensation paid. This is probably due to employees straining their backs, knees, and shoulders.

6. Exploring the relationship between injury type and body parts. From this visualization, sprains, strains and contusions stand out, and back, knee and shoulder injuries stand out. 

7. Similar to the above bubble chart, we want to explore how time and body part affected compensation paid. Once again, brain, neck, back and knee injuries stand out as having high compensation.