**BIG DATA ANALYTICS FOR CLAIMS MANAGEMENT**

Report

Ridhi Shah

**EXECUTIVE SUMMARY**

The main objective of the Workers’ Claim Management project is to find out powerful insights from the past data of the Claims. These insights can help overcome the challenges that are faced by the Worker’s Claim Management team, which we described earlier in the project midterm. The work was started by understanding all the data (variables) that can be used for the prediction. Secondly, what needs to be predicted that can help reduce the challenges in the process. Some new variables were derived in the data from the existing columns, which can give some knowledge to the data, and act as the independent variable for prediction.

New variables derived are Number of Days To Process a Claim, Employee Our of Office days, Number of days to process a particular Claim, Total Incurred Cost, Type of Injury Nature, HighRisk Claims depending on Amount. After deriving new variables, the team started with Logistic Regression. All the variables that were found to be informative with the outcome were used. The outcome of the Logistic Model RiskType, that is predicting whether the Claim is High Risk or Low Risk depending on the data provided. This outcome can help the Management team to decide which Claims are of high risk with respect to the Claim Amount, which can help reduce the cost and time of the management team.

The Logistic Regression was built using the following independent variables: Injury

Nature of the Claim, Gender of the Claimant, Age of the Claimant, Type of the Claim(Medical, Idemnity, Reports), No of transactions per Claim, the Processing time of Claim, and Number of days the Employee was out of office. All these variables highly affect whether the Claim is High Risk or How Risk. After taking a threshold of 0.5, the computed Accuracy of the model is 74.7%. The above-mentioned variables were significant and gave an inflicting outcome.

Coming to the Visualization part, we found key findings that can help the team with various recommendations and insights. Some of the major key findings are as follows: Injury as Strain and Contusion has higher Claims and is also predicted as the one with High-Risk Amount. In the Claimant Type, Medical Claims are the one which has High Risk with respect to the Total Incurred cost. Also, High Risk is directly proportional to No of days took to process a Claim, No of days Employee was out of office, No of transactions per Claim.

The strategic recommendations from the prediction are that Injury Nature and Body Part Region affected is the most significant part of the outcome, Proactive intervention in the Claim process in the early stages will help the employee to get the best medical treatment as soon as possible, which will reduce the time of the teams as well as employee to get recovered from the injury and get back to work as early. This will also reduce the Incurred cost of Claim. Additionally, there are lots of claims that occur in the third quarter of the year, which should be taken into consideration and safety and health conditions of the employee should be taken into care by doing drug screening in the company which can detect and prevent the bad health conditions of the employees.

**INITIAL HYPOTHESES:**

Based on our groups research and data exploration we have made below hypothesis by considering HighRisk as the outcome variable. This column can be computed based on Total\_incurred\_cost\_per\_claim. If Total\_incurred\_cost\_per\_claim is greater than Median value

(Based on the box plot) then HighRisk is ‘1’ (considering that claim as the High risk claim) and if it is less than Median value then HighRisk is ‘0’ (considering that claim as the Low risk claim).

The initial hypotheses obtained from data exploration and groups research that affect the claims as HighRisk are as follows:

1. When the ‘Number of days out of office’ is high then it is a HighRisk claim. (

‘Number of days out of office’ is computed by taking the difference between ReturnToWorkDate and IncidentDate.)

1. We have 53 distinct values in InjuryNature. For injuries like Burn, Poisoning and Physical Injuries we can assume it to be a HighRisk claim.
2. When ClaimantType= “Report Only” and Days to process claim(difference between ClaimantClosedDate and ClaimantOpenedDate) is less then we assume that particular claim as low risk.
3. When ClaimantType= “Indemnity” and Days to process claim(difference between ClaimantClosedDate and ClaimantOpenedDate) is more then we assume that particular claim as high risk.
4. When the Number of Transactions for a ClaimID are more then we assume that claim as a HighRisk claim.
5. A litigated claim will fall into High Risk Category
6. A subrogated claim will fall into Low Risk Category

**DERIVED VARIABLES:**

Before removing the outliers we have 108263 records with the claimant status as open, reopen and closed. We have filtered the dataset for only claimant status as closed and observed 103897 records.

Following are the Derived Variables:

1. **Days\_To\_Process\_Claim :**

Days\_To\_Process\_Claimgives the information about the Number of days it took to process a claim. Claim Processing Time is derived from two columns namely ClaimantClosedDate and ClaimantOpenedDate.The values in

Days\_To\_Process\_Claim is numeric.

Days\_To\_Process\_Claim= ClaimantClosedDate - ClaimantOpenedDate

1. **OOO\_Days:**

The OOO\_Days gives the information about the number of days the employee is out of office. It is derived from the columns IncidentDate and ReturnToWorkDate.

OOO\_Days = ReturnToWorkDate–IncidentDate

1. **Notification\_time\_frame :**

Notification\_time\_frame column gives information whether the claim can be processed are not. It is derived from two columns EmployerNotificationDate and IncidentDate.

Notification\_time\_frame = EmployerNotificationDate – IncidentDate

1. **Total\_incurred\_cost\_per\_claim :**

Total\_incurred\_cost\_per\_claim column gives the information about the total cost that is incurred per each claim. It is derived from two columns TotalPaid and TotalRecovery.

Total\_incurred\_cost\_per\_claim = TotalPaid–TotalRecovery

1. **InjuryNature\_Type :**

InjuryNature\_Type gives the information of how severe is the injury with three values Low and high. It is derived from InjuryNature column.

1. **Notification\_Category :**

Notification\_Category gives the information of how risk the claim is with four values

Low\_Risk, Normal, Medium\_Risk, High\_Risk.It is derived based on theNotification\_time\_framecolumn.The Normal value corresponds to Notification\_time\_frame below 31 days, the Low\_Risk corresponds to Notification\_time\_frame between 31 and 180 days, the Medium\_Risk corresponds to Notification\_time\_frame between 181 days to 365 days and the High\_Risk corresponds to Notification\_time\_frame above 365 days.

1. **High\_Risk\_Claim\_By\_DollarAmount :**

High\_Risk\_Claim\_By\_DollarAmount will be the outcome variable and it is derived fromtheTotal\_incurred\_cost\_per\_claimcolumn.WhentheTotal\_incurred\_cost\_per\_clai

m is greater than the median of the values in it then

High\_Risk\_Claim\_By\_DollarAmount is said to be a High risk with value 1 and when Total\_incurred\_cost\_per\_claim is less than or equal to the median of the values in it then the High\_Risk\_Claim\_By\_DollarAmount is said to be Low Risk with value 0.

We have calculated the outliers for the variable Total\_incurred\_cost\_per\_claim and removed them. The total observations we are using after removing the outliers are

61536. The median that we took after removing the outliers from the Total\_incurred\_cost\_per\_claim is 325.52$ to compute HighRisk column.

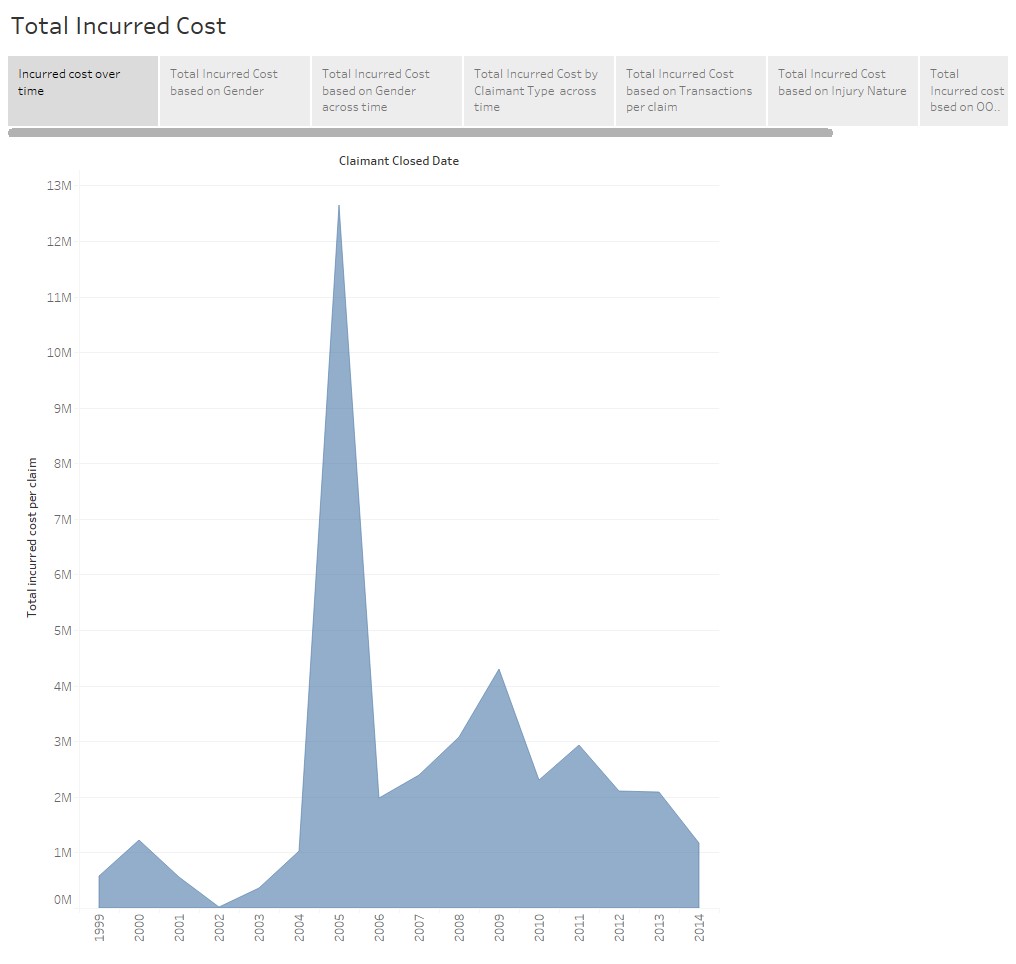
**TABLEAU VISUALISATIONS:**

**Visualization For Total incurred cost**

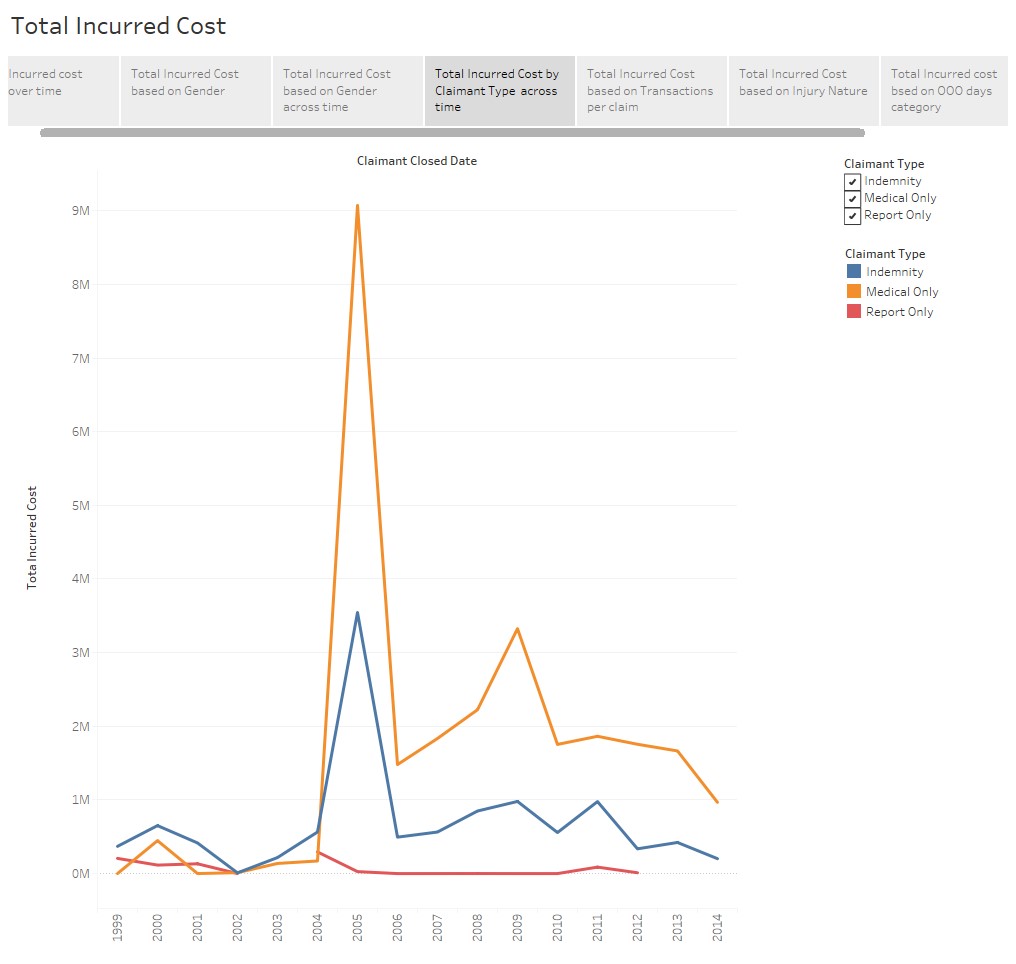
**1**

**. Incurred cost over time**

**:**

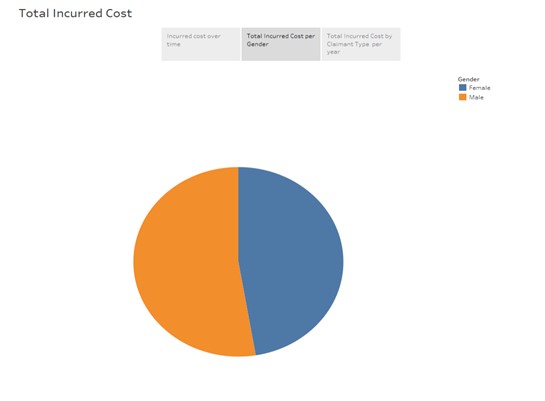


This visualization shows how the total incurred cost varies over time. We can see some peaks and valleys however there is no specific pattern or trend followed. In the year 2005 we can notice a sudden high peak. The claims management team can further investigate the claims data for the year 2005 and look at other variables to identify the reason for the high incurred cost in that year. **2. Incurred cost based on claimant type**



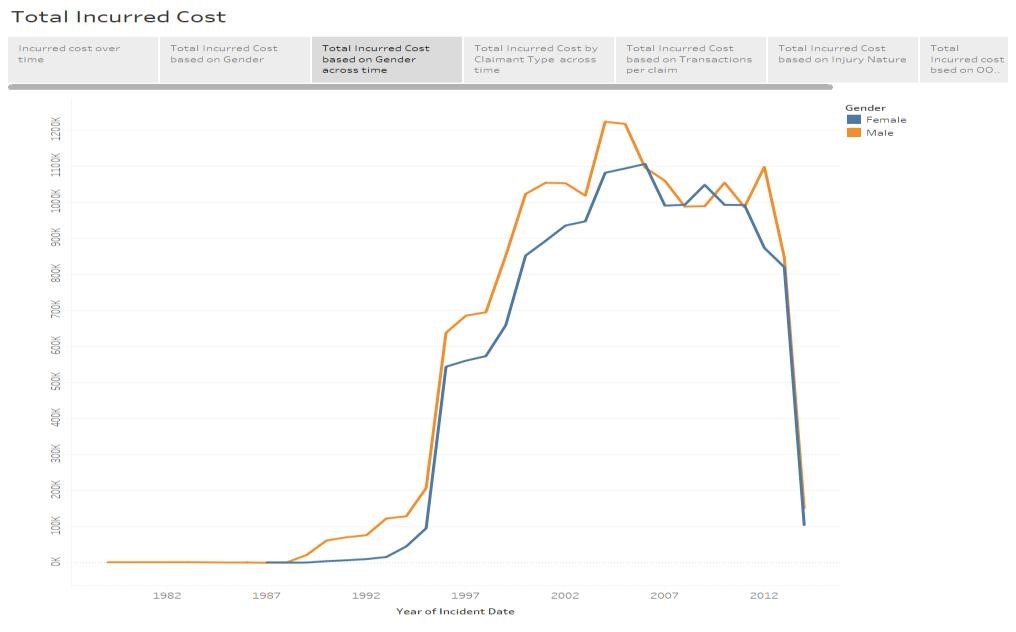
This visualization shows how the incurred cost varies based on the claimant type across time. We can see that medical claims have high total incurred cost per year followed by Indemnity. Report only has the least incurred cost per year. Also, we can see that Medical only claimant type is the major contributor for the total incurred cost in the year 2005.

1. **Incurred cost by Gender:**



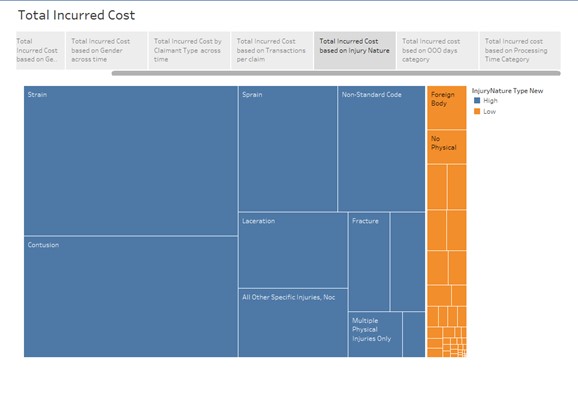
This visualization shows how the incurred cost varies over gender. We can see that more cost is incurred towards male claims over female claims.

1. **Incurred cost by Gender across time:**



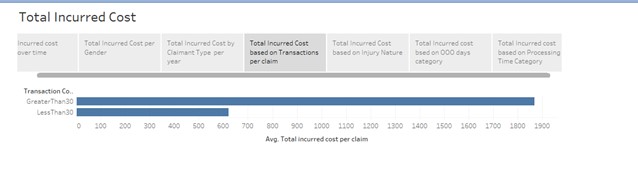
This visualization shows how the incurred cost varies over gender across time. Again here we can see that for most years more cost is incurred towards male claims over female claims.

1. **Incurred cost based on injury Nature:**



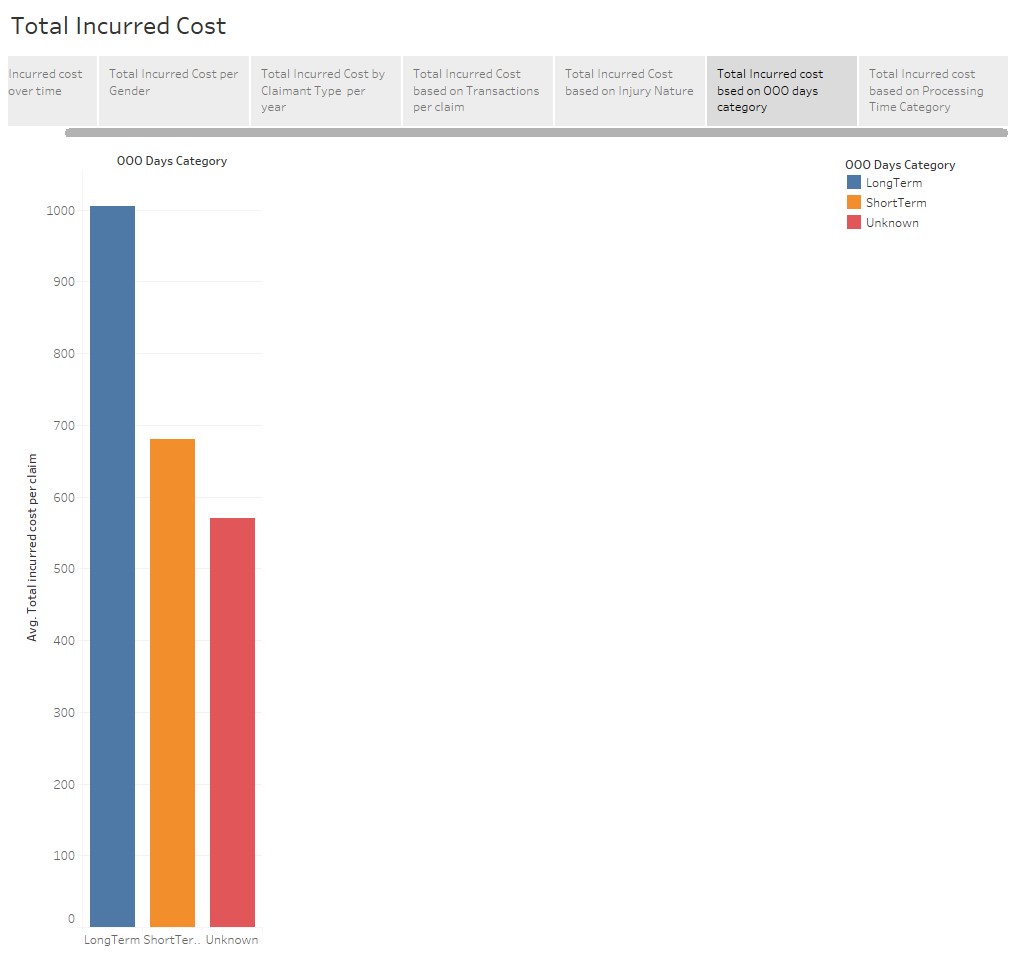
This visualization shows the distribution of claims cost across injury nature type and injury nature. We can see that claims of injury type high incurs towards more cost than injury nature type low. Also, among injury nature type high we can see that injury type Strain and Contusion have high total incurred cost. The claims processing team can make sure that employees with injury nature type of Strain and Contusion get better treatment to ensure faster recovery of the employee which may help reduce overall cost and reduced employees return to work day.

1. **Incurred cost based on Transactions per claim:**



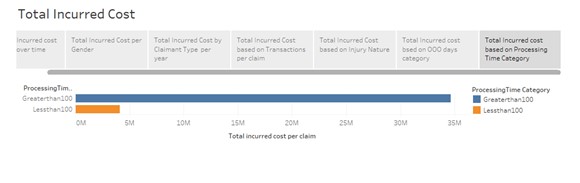
This visualization shows how the incurred cost varies based on the number of transactions per claim. We can see that the greater the number of days to process the claim, greater is the incurred cost.

1. **Incurred cost based on out of office days category:**



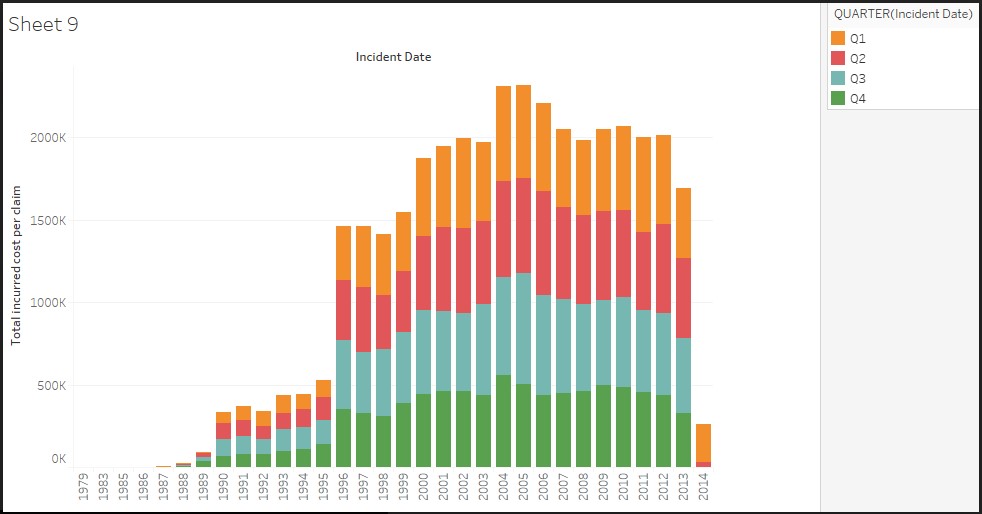
This visualization shows how the incurred cost varies over out of office days category. We can see that long term out of office category incurs high cost.

1. **Incurred cost based on Processing Time:**



This visualization shows how the incurred cost varies over claims processing time. We can see that the claims processing time greater than 100 days incurs very high cost.

1. **Incurred cost across the years grouped by quarters in the year:**



This visualization shows how the incurred cost across the years which is grouped by the quarters in the year. Most of the incidents are observed to be in the quarter-3.

**LOGISTIC REGRESSION MODEL:**

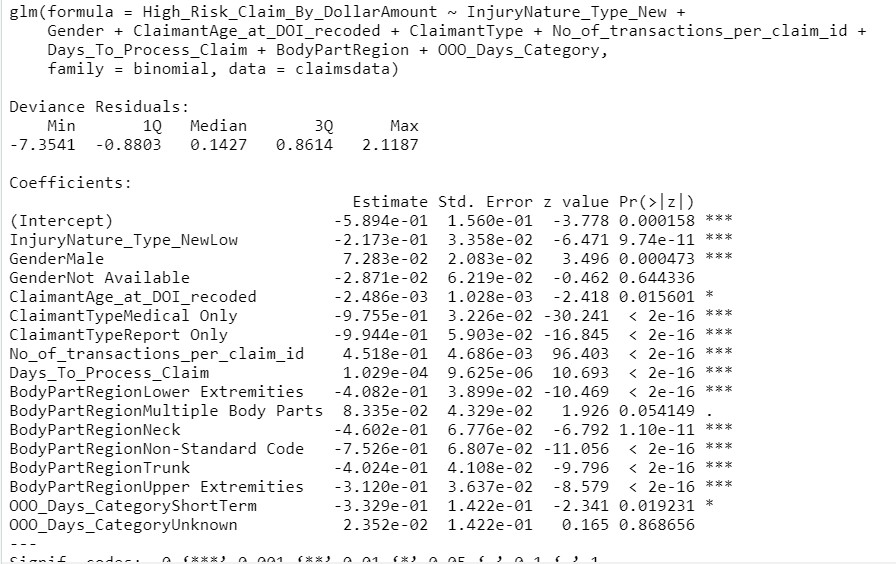
The Independent variables that we took to run the logistic regression model are InjuryNature\_Type, ClaimantAge\_at\_DOI\_recoded, Notification\_Category,

BodyPartRegion, ClaimantType, Is\_Litigated, Transaction\_Count\_Category, Gender, ProcessingTime\_Category, OOO\_Days\_Category. But many variables are insignificant .So we have dropped the insignificant variables and reran the model until all the variables are significant in the model.

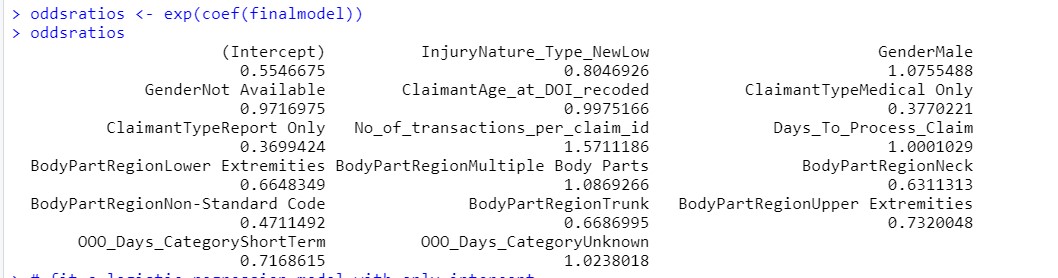
**FinalRun:**

After removing the insignificant variables one by one from the previous runs. The final independent variables that are used for the FinalRun are InjuryNature\_Type, Gender,ClaimantAge\_at\_DOI\_recoded,ClaimantType,No\_of\_transactions\_per\_claim\_id,Days\_T o\_Process\_Claim,OOO\_Days\_Category.

**Step 1**: After rerunning the logistic regression on the above variables the outcome is as follows.



**Step 2**: Computing the oddratios for the model:



**Interpreting the OddsRatio:**

InjuryNature\_Type\_NewLow:

For InjuryNature\_TypeLow the oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is smaller by a factor of 0.8046than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for InjuryNature\_TypeHigh.

GenderMale:

For GenderMalethe oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is higher by a factor of 1.0755 than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for GenderFemale.

GenderNot Available:

For GenderNot Available the oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is lesser by a factor of 0.9716 than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for GenderFemale.

ClaimantAge\_at\_DOI\_recoded:

For a unit increase in ClaimantAge\_at\_DOI\_recoded the odds of High\_Risk\_Claim\_By\_DollarAmount being 1 decreases by a factor of 0.9975.

ClaimantTypeMedicalOnly:

For ClaimantTypeMedicalOnlythe oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is lesser by a factor of 0.377than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for ClaimantTypeIndemnity.

ClaimantTypeReportOnly:

For ClaimantTypeReportOnly the oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is lesser by a factor of 0.3699 than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for ClaimantTypeIndemnity.

No\_Of\_Transactions\_Per\_Claim\_Id:

For a unit increase in No\_Of\_Transactions\_Per\_Claim\_Id the odds of High\_Risk\_Claim\_By\_DollarAmount being 1 increases by a factor of 1.5711.

Days\_To\_Process\_Claim:

For a unit increase in Days\_To\_Process\_Claim the odds of High\_Risk\_Claim\_By\_DollarAmount being 1 increases by a factor of 1.0001.

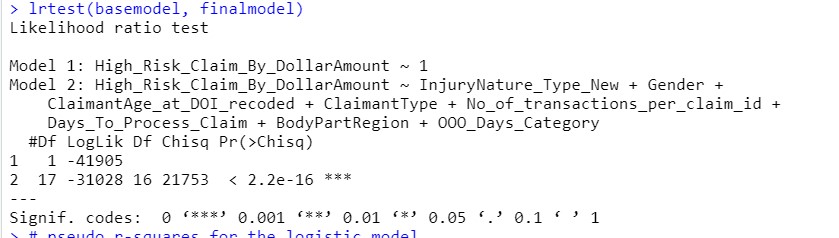
OOO\_Days\_CategoryShortTerm:

For OOO\_Days\_CategoryShortTermthe oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is lesser by a factor of 0.6648 than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for OOO\_Days\_CategoryLongTerm.

OOO\_Days\_CategoryUnKnown:

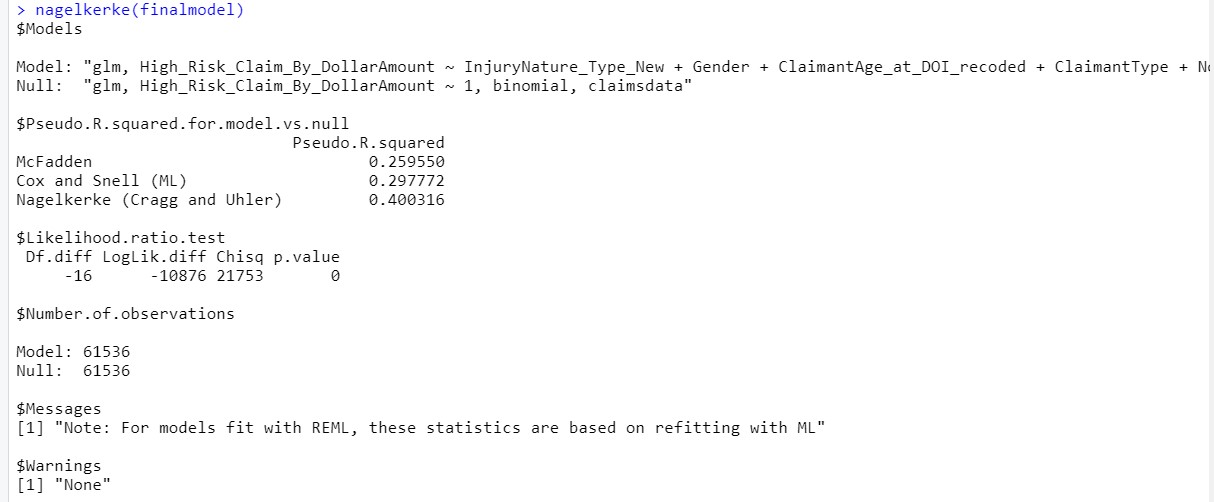
For OOO\_Days\_CategoryUnKnown the oddsratio of High\_Risk\_Claim\_By\_DollarAmount being 1 is higher by a factor of 1.0238than the High\_Risk\_Claim\_By\_DollarAmountbeing 1 for OOO\_Days\_CategoryUnKnown.

**Step 3:** Likelihood ratio of the Model:



The Log Likelihood of the final model has decreased when compared to base model which indicates that the model is good. Also the P-value of the final model is less than 0.05 which indicates the model is fit for the data.

**Step 4:** Finding the pseudo r-squared values of the final model through nagelkerke function from rcompanion package:



**Step 5:** Predicting the probabilities for individual cases from the model and assigning it to predicted\_probabilities.

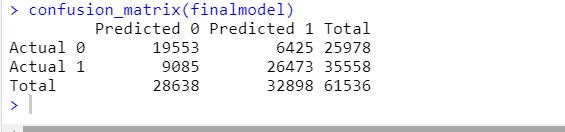
predicted\_probabilities<- predict(finalmodel, claimsdata, type= "response")

**Step 6**: Assigning the threshold as 0.5, if predicted\_probabilities are more than 0.5 then the prediction is 1 else 0.

**Step 7:** Combining the claims data with the predictions made.

claimsdata<- cbind(claimsdata, predicted\_probabilities, predictions)

**Step 8:** Building a confusion\_matrix for the final model to calculate the accuracy of the model.



Accuracy = Total correct predictions/ All cases

Accuracy= True Positive + True Negative/ All cases

Accuracy = 26473 + 19553/61536

Accuracy = 0.747 = 74.7%

The model we built is 74.7% accurate.

**STRATEGIC RECOMMENDATIONS:**

A proactive and aggressive medical treatment and intervention during early stages of the claim when the Injury Nature is attributed to one of the following values:

1. Strain
2. Contusion
3. Sprain
4. Non-Standard Code
5. Laceration
6. Fracture
7. All other Specific Injuries, Noc
8. Puncture
9. Multiple Physical Injuries Only
10. Burn

Injuries tend to me more during the third quarter (July, Aug, Sep) compared to other quarters. This can be an indication that it might be due to improper safety measures including equipment malfunction, sanitization, number of labor hours to meet the demand, sleep deprivation, employee drug screening, etc. With more information around these assumptions, we might be able to pinpoint the reason more accurately.

When a claim is associated with one of the following body part values, a proactive approach in getting advanced and appropriate medical help that is efficient and cost effective will help employees recover in time along with saving the dollar amount spent on these claims:

1. Lower Extremities
2. Multiple Body Parts
3. Neck
4. Non-Standard Code
5. Trunk
6. Upper Extremities

Standards should be in place to set a threshold on treating a claim as “Short Term Disability” vs “Long Term Disability” and process it accordingly. Based on the data on hand, it is not clear if there is a process put in place.

Encouraging employees to get yearly physical checkups, periodic drug screening and safety work environment setup along with educating the employees on the same will help mitigate most of the injuries and claims and to most extent prevent them.