Workers' Compensation

Incorporating Data Analytics in Claims Management

Akshata Kanumuri, Dikshya Mohanty, Lei Huang, Muralidharan Singaravel

DSBA 6100 Fall 2018 – Group Project 2

SUMMARY

In this second part of the paper, we studied the variables and the data in the wrangled dataset, developed some initial hypotheses, identified and added few derived variables, discussed how Multiple Linear regression (MLR), Decision tree and Association Rule Mining modeling techniques can be applied to better understand claims payments and processing time. We built and ran the MLR model on claims amount and also provided few strategic recommendations for the claims management company on the steps to better manage the business.

Few initial hypotheses which we identified are i) The young workers (age <24) are at a greater risk for injury at work. ii) The insurance company states that more 70 percent of its claims are settled within 180 days of filing the claim. iii) The non-indemnity claim amount is dependent on the Worker's wage and no. of days the worker was away. iv) The most common workplace injury is Sprains and Strains (including back injuries).

We added the following derived variables - Time To Process, Day Of Incident, Age Group and No Of Days Away which we felt will be helpful in building the predictive models.

From MLR, we identified that given Average weekly wage, No of days the worker was away after the injury, time to process a claim and count of transaction of per claim, we can predict the non-indemnity claim amount for a lower back injury resulting from strain or sprain.

Few steps by which the claims can be better managed are by understanding the factors which can affect the claims amount, tackle missing data, identify the claims which need to be processed faster and moving to electronic claims submission which will lead to faster and effective claim processing.

C.1 Study the variables and the data in the wrangled dataset and develop some initial hypotheses (at least 4) regarding the expected relationships between the independent variables in the dataset and claims outcomes. Please note that this part is based on your group's research and data exploration done for midterm without the benefit of actually running any analytics modeling.

Hypothesis 1: The young workers (age <24) are at a greater risk for injury at work and therefore be charged higher insurance rates.

Hypothesis 2: The insurance company states that more 70 percent of its claims are settled within 180 days of filing the claim.

Hypothesis 3: The non-indemnity cost per claim is dependent on the Worker weekly wage and Number of days the worker was out of work because of the injury.

Hypothesis 4: If an employee is out of work for more than six months, they have less than a 50% chance of ever returning to work in any capacity.

Hypothesis 5: The most common workplace injury is Sprains and Strains (including back injuries).

Hypothesis 6: Male workers have a higher risk of injury compared with females when performing the same job.

C.2. To the extended claims dataset, add a minimum of 3 new derived independent variables which can be used to examine payments/processing time from different dimensions. [Hint: Good candidates to derive new variables are date/time variables, from

which you can derive the day of the week, month, etc. You can also recode categorical independent variables and create new variables.]

1. TimeToProcess – The processing time per claim, which is the difference between ClaimantClosedDate and ClaimantOpenedDate.

2. DayOfIncident – Day of the week when the incident occurred at the workplace.

3. AgeGroup – Worker age group based on the following values –

   18 Years or Younger

   19 - 24 years old

   25 - 54 years old

   55 - 64 years old

   65 Years or Older

4. NoOfDaysAway – Number of days worker was out of the work place following the date of the incident, calculated by subtracting ReturnToWorkDate and IncidentDate.

C.3. Add a binary dependent variable to classify claims as critical and non-critical based on total payments or based on processing time. The group can determine the appropriate cut-off to create the classification. This variable can be used for predictive modeling that requires binary outcome.

Critical or non-critical claims will be classified based on processing time or Claims Settlement Cycle Time metric which is defined as, the number of days required to settle an insurance claim from the time the claim is reported. Claims processed within 180 days can be classified as critical-claims.

C.5. Based on the understanding among your group members of the various analytics models discussed in class (i.e., linear regression, logistic regression, decision-trees, cluster analysis, association mining) , discuss how any 3 of the modeling techniques can be applied to better understand the drivers of claims payments and processing time. For each modeling technique, discuss the dependent and independent variables you would choose, the rationale for your choice, and how the results from the model could be used. In addition, discuss the limitations of each of the modeling techniques in addressing the business problem. [Important: For this part, DO NOT actually run the models.]

1. **Multiple Linear regression on Worker's compensation claims data:** Multiple Linear regression (MLR) model can be used here to predict the insurance amount paid per claim based on various predictors like Age, Injured Body Part, Fatal Injury or Not, Average Weekly Wage, Days Lost, Count of transactions per claim and Time To Process. All these independent variables can have statistically significant influences on claim costs.

   In this dataset, the MLR model variables are:

   - Dependent Variable – OtherPaid (excluding Indemnity)
   - Independent Variables – Claimant_Age_DOI, BodyPartRegion, IsFatality, AverageWeeklyWage, NoOfDayAway, TimeToProcess, CntTransPerClaim

   The results from this regression model can be used to:

- Successfully identify and predict medical cost drivers within Workers' Compensation claims (i.e., specific diagnoses that result in costlier claims).
- Settle claims early and more effectively.
- Use to more accurately set reserves for an individual claim.

Some of the limitations of this MLR model are:

- One of the key assumptions is that there is linear relationship between dependent and independent variables, and normal distribution of residuals to be satisfied to adequately model the underlying data. Before fitting the model, each of the independent variable is to be evaluated to test this assumption.
- There are many independent categorical variables in this dataset like Gender, BodyPartRegion, InjuryNature which has to be first transformed and create dummy variables for each categorical variable.
- Multiple Linear Regression is sensitive to Outliers, so we need to first identify and remove outliers before fitting the model.

2. **Decision Tree Model on Worker's compensation claims data:** Decision tree can be used to predict if the worker will be away from work for more than 180 days based on the injured employee's age, gender, nature of injury, cause of injury, body part injured, Wage and the nature of their job.

   In this dataset, the decision tree variables are:

   - Dependent Variable – 180DaysAway : a binary variable with Yes or No

- Independent Variables – AgeGroup, Gender, BodyPartRegion, InjuryNature

The results from this decision tree model can be used to:

- Measure efficacy of treatments through the number of lost work days for each case, and identify treatments that correspond to speedier resolution.
- Evaluate providers' 'performance' and cost utilization across the patient populations they treat in a fair, risk-adjusted way.
- Identifying which provider would be more effective at treating a specific injured worker, given the age, gender, and diagnoses.

Some of the limitations of this decision tree model are:

- Decision trees are prone to over fitting and are not well suited when most of the variables in the training set are correlated. So, we first need to set constraints on model parameters and pruning.
- It is not fit for continuous variables. While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories. In this case, decision tree cannot be used to predict how many days the worker will be away from work based on the independent variables identified above.

3. **Association Rule Mining on Worker's compensation claims data:** In Workers Compensations insurance, understanding the cause of injury is crucial for preventing repeat injuries. Association Rule analysis can be used to reduce both claim frequency and

severity. This technique enables Insurers. Employee and Employers to understand the pattern of circumstances related to the injuries.

Lower Back injury is the most common type of injury in the claims data set. So, the input to this model can be BodyPart, DayofIncident, InjuryNature, IncidentDescription, Shift and TypeofWork (we do not currently have this in the dataset).

The Output from this model:

- Can help the safety manager understand the circumstances (what/where/when) associated with the back injuries and company would benefit from implementing safety programs to prevent lower back injuries.
- Gain interesting and useful insights about the circumstances surrounding the claims and these insights can be used to minimize future incidents leading to such claims

Some of the limitations of this decision tree model are:

- Association Rule analysis will also yield rules that may be trivial. For example, there may be a rule that tells us that the dayshift workers suffering strain by lifting on Tuesdays are 100% associated with strain. This is trivially true: strain implies strain.
- Association rule mining algorithms normally discover a huge quantity of rules and do not guarantee that all the rules found are relevant.

- Some rule may be non-actionable for example a rule tells us that 40% night-shift injuries occur on Mondays. This rule does not provide a clear explanation of what is going on here and does not suggest a course of action.

C.6. Choose one of the modeling technique from above, build and run the model on one of the outcome variables. Present your model results. Interpret the results for the claims management company and present the significant findings in business language (i.e., avoid focusing only on F-value or p-values or odd-ratios). Present your conclusions on the key factors impacting the outcome variable you have chosen.

We ran Multiple Linear Regression Model with the following dependent and independent variables to predict the amount paid per claim (excluding Indemnity) for a lower back injury claims.

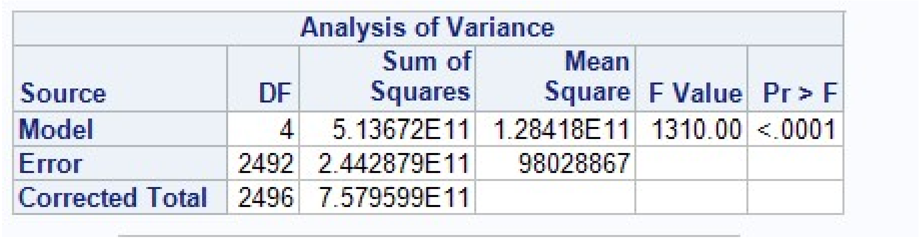
- Dependent Variable – OtherPaid (excluding Indemnity)
- Independent Variables – AverageWeeklyWage, NoOfDayAway, TimeToProcess, CntTransPerClaim

Input data is modified to meet the following conditions:

1. We have taken only the claims which are 'Closed'.
2. Body Part injured is 'Lower Back Area' – which is the most commonly injured body part.
3. Injury Type is 'Strain' or 'Sprain' – which is the most common type of injury in 'Lower Back Area'.
4. No of Days away and Average weekly wage are not missing.

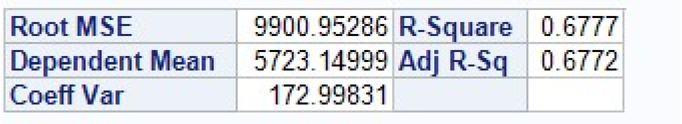
Multiple Linear regression model results:

1. Model is statistically significant as p-value from the f test is <0.0001 which rejects the null hypothesis saying all the β_0, β_1, …, β_k are zero



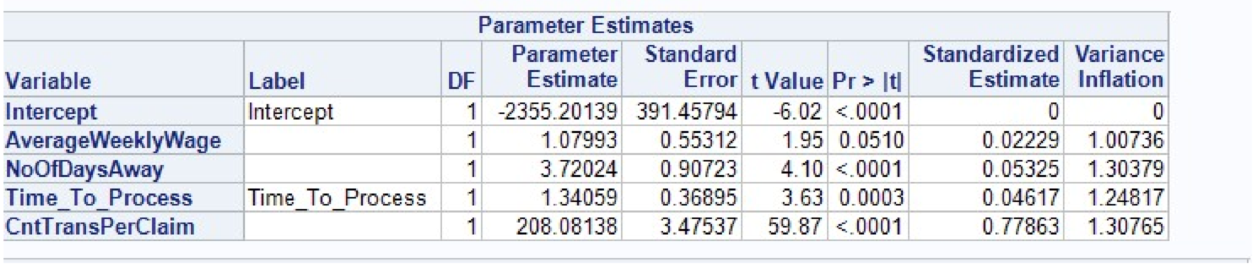
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5.13672E11	1.28418E11	1310.00	<.0001
Error	2492	2.442879E11	98028867		
Corrected Total	2496	7.579599E11			

2. R-Square and adjusted R-Square are both close to 68% which means that the model can explain 68% of the variability of the response data around its mean.



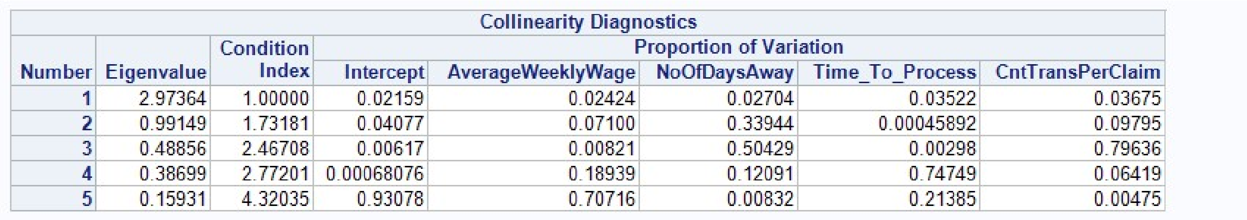
Root MSE	9900.95286	R-Square	0.6777
Dependent Mean	5723.14999	Adj R-Sq	0.6772
Coeff Var	172.99831		

3. All the inputs variables are statistically significant as the p-value is < 0.05.



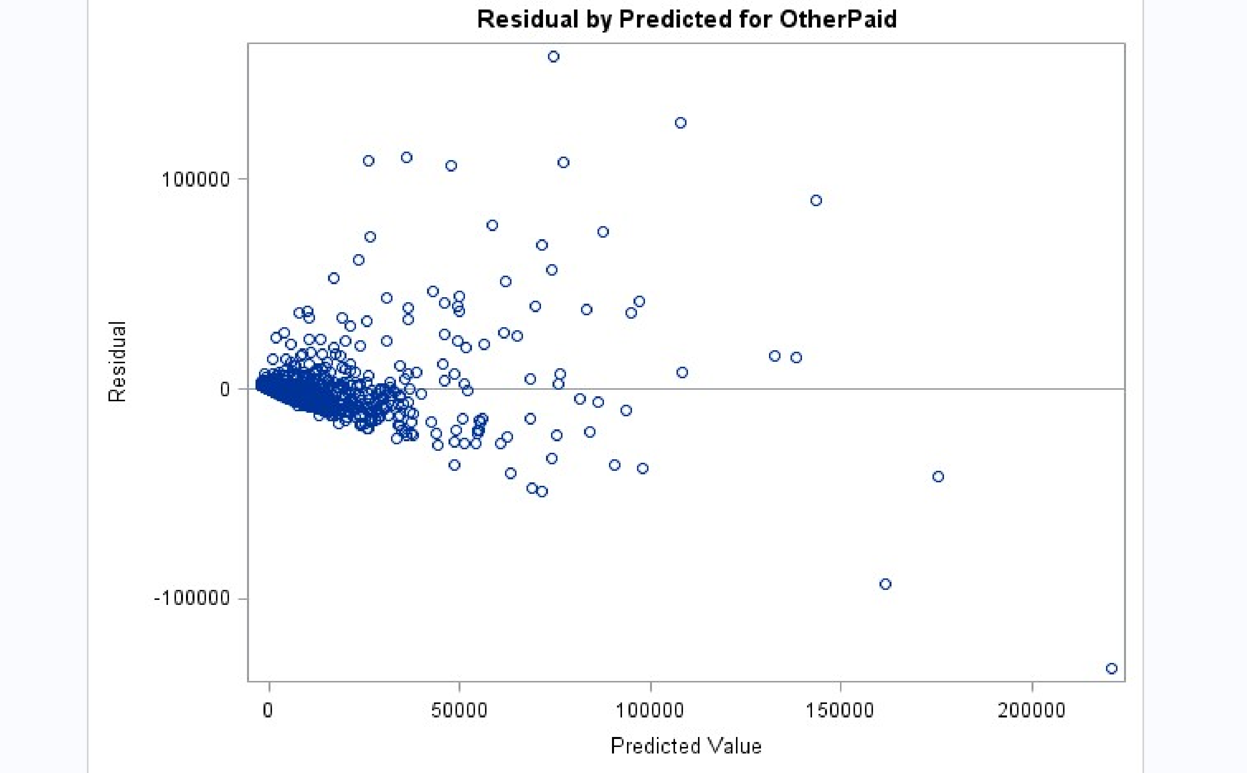
Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	-2355.20139	391.45794	-6.02	<.0001	0	0
AverageWeeklyWage		1	1.07993	0.55312	1.95	0.0510	0.02229	1.00736
NoOfDaysAway		1	3.72024	0.90723	4.10	<.0001	0.05325	1.30379
Time_To_Process	Time_To_Process	1	1.34059	0.36895	3.63	0.0003	0.04617	1.24817
CntTransPerClaim		1	208.08138	3.47537	59.87	<.0001	0.77863	1.30765

4. None of the independent variables have multi-collinearity,

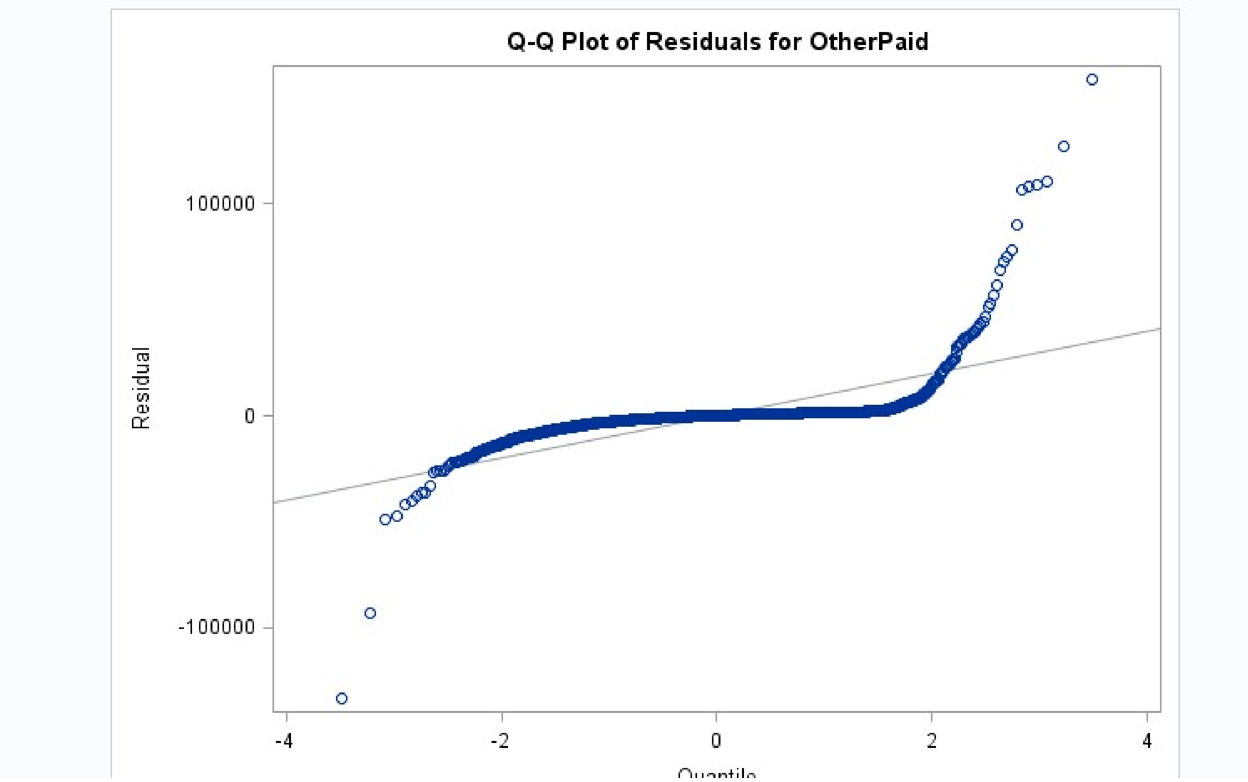


Collinearity Diagnostics							
			Proportion of Variation				
Number	Eigenvalue	Condition Index	Intercept	AverageWeeklyWage	NoOfDaysAway	Time_To_Process	CntTransPerClaim
1	2.97364	1.00000	0.02159	0.02424	0.02704	0.03522	0.03675
2	0.99149	1.73181	0.04077	0.07100	0.33944	0.00045892	0.09795
3	0.48856	2.46708	0.00617	0.00821	0.50429	0.00298	0.79636
4	0.38699	2.77201	0.00068076	0.18939	0.12091	0.74749	0.06419
5	0.15931	4.32035	0.93078	0.70716	0.00832	0.21385	0.00475

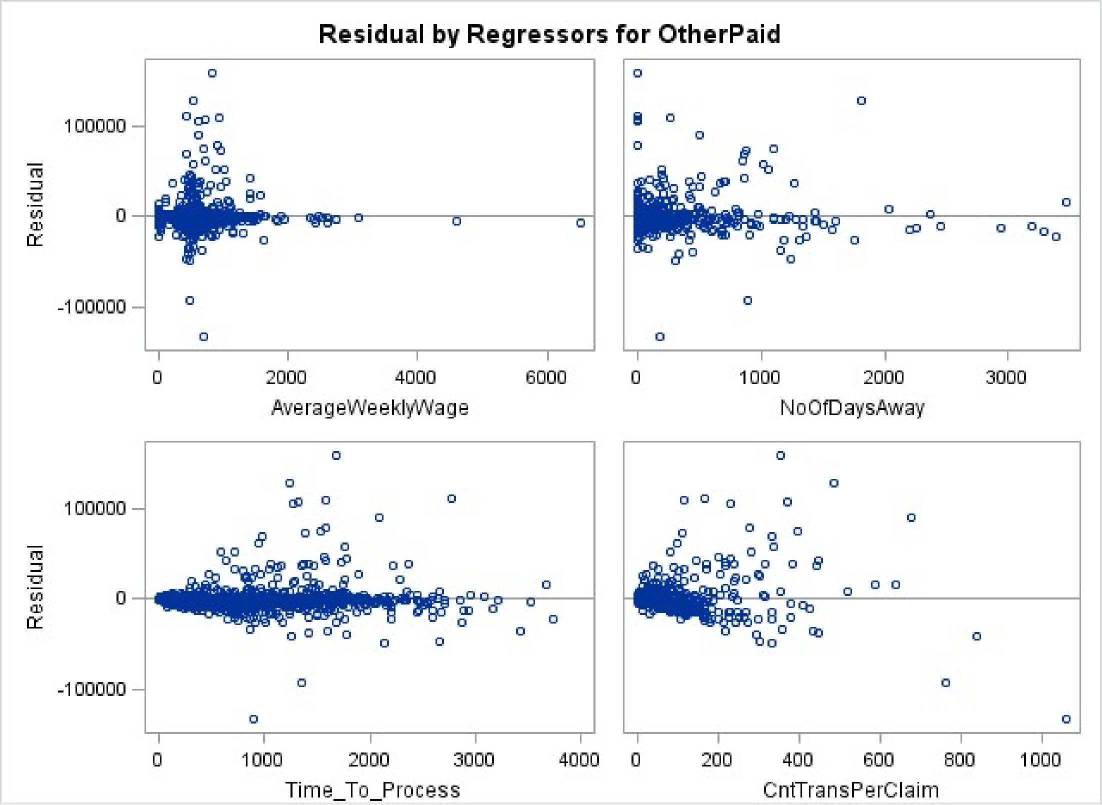
5. Residual vs predicted plot: residuals are observed somewhat evenly on both sides of the reference zero line, and the spread of the residuals is fairly constant for lower claims amounts (claim amount less than $25000.00) and there are outliers.



6. From the Q-Q plot we can see that it is heavy-tailed or outlier-prone distribution.



7. Residual plot for each of the independent variables shows that the residuals are centered on zero and appear to be symmetric about zero with the presence of outliers.



8. From the above model results, we can say that the multiple linear regression model is statistically significant for non-indemnity claim amount of less than $25000.00 (after we remove the outliers) with the predictors variables of AverageWeeklyWage, NoOfDayAway, TimeToProcess and CntTransPerClaim.

   That is, given the Average weekly wage, No of days the worker was away after the injury, time to process a claim and count of transaction of per claim, we can predict the non-indemnity claim amount for a lower back injury resulting from strain or sprain.

9. The Linear regression equation is

   Claim Amount = -2355.201 + 1.080* AverageWeeklyWage + 3.720* NoOfDayAway + 1.341 * TimeToProcess + 208.08* CntTransPerClaim

10. Interpreting the above linear regression equation:

    a. The intercept is -2355.201 which corresponds to the estimated response variable when all the input variables equal zero. In this model, the intercept corresponds to an estimated claim of $-2355.201 for a worker with zero AverageWeeklyWage, NoOfDayAway, TimeToProcess, CntTransPerClaim.

    b. All the coefficients are positive which represents positive relationship between dependent and independent variables which says that there is an increase in dependent variable caused by each one-unit increase in the independent variable.

    c. The coefficient for AverageWeeklyWage is 1.080 which means that for every one year increase in a person's age, the claim amount is expected to increase by 1.08 times i.e the higher the worker's wage the higher will be the medical claim costs.

    d. The coefficient for NoOfDayAway is 3.270 which means that for every one day increase in a NoOfDayAway, the claim amount is expected to increase by 3.270 times i.e the more number of days the worker is absent because of injury, the more will be the medical claim costs which seems to be obvious.

    e. The coefficient for TimeToProcess is 1.341 which means that for every one day increase in a TimeToProcess, the claim amount is expected to increase by 1.341 times i.e the more number of days it takes to be process a claim, the more will be the medical claim costs, may be the increased time is because there is a need to inspect higher medical claims.

f. The coefficient for CntTransPerClaim is 208.08 which means that for every one unit increase in a CntTransPerClaim, the claim amount is expected to increase by 208.08 times i.e the higher the count of transactions per claim the higher will be the medical claim costs which seems to be obvious.

C.7. Provide a set of strategic recommendations for the claims management company on the steps it can take to better manage its business. Each recommendation should be justified by your data exploration and model results.

1. **Understanding the Many Factors Affecting Claims:** There is a tremendous amount of data that are being generated constantly and it is important to understand the factors affecting the claim. From the multiple linear regression model we saw that one can predict the claim amount using Average weekly wage, No of days the worker was away after the injury, Time to Process and Count of Transactions per Claim that will allow for more responsive service, faster and more effective claims settlement which will in turn improve the claims management.
2. **Tackling Missing Data:** For any effective predictive model building, the data has to be reasonably accurate and complete that is, essentially free of missing values. It can either be handled at the data collection stage by training the data entry personnel to enter all the fields that are identified as mandatory/critical (example employee average weekly wage, employee return to work date which were identified as significant predictors) or at the data wrangling stage by replacing with Mean/Median/Mode like we did for above mentioned variables.

3. **Identifying the Critical Claims:** When the insurance claims process offers customers consistent service across multiple delivery channels, the company and the customer benefit. For faster claims processing, it is essential to identify the critical claims either based on the claims amount or the employee/employer information or injury type or processing time. Once the claim is identified as critical, it can be processed faster. Submitting an insurance claims means the customer is dealing with something stressful that has happened. When the insurance claims process relieves some of this stress by faster and accurate claims processing, customers are likely to become more loyal.
4. **Moving to electronic claims:** E-claims are more accurate and offer faster processing than paper submission, submitting claims electronically reduces mistakes that are caused when paper claims are converted to an electronic format like invalid worker's age in this dataset and some of the information about the employer/employee can be pre-populated like Average weekly wage which had close to 63% missing values.