WORKER'S COMPENSATION CALIMS PROCESSING ANALYSIS

GROUP MEMBERS:

1. Dolly Tripathy
2. Deepika Deewangan
3. Bhavya Pilli
4. Ryne Andrews
5. Pradeep Edwin Samuel
6. Haricharan Shivaram

Executive Summary

The purpose of this report was to analyze data from a company which processes workers' compensation claims and advise them on how data analytics can be used to identify major drivers of costs and time to process claims. The problem background and data exploration of the claims data were completed earlier in the part one of this project, this report is focused on below mentioned activities.

1. Data Preparation
2. Model planning
3. Model building
4. Recommendations.

FINDINGS FROM VISUALIZATION

1. High deviations in the Indemnity paid for high risk claims. Low risk indemnity paid is constant throughout time.
2. Non-Standard code has a very high average processing time and this outlier has to be taken into consideration by the insurance company.
3. Over the years, the number of claims opened keeps changing against a fairly constant trend of average total medical bills.
4. Uncategorized injury causes accounts to a huge expense as part of the total medical bills. This needs to be rectified.
5. Claimant type indemnity has more number of open and re-opened claims as compared to others.

FINDINGS FROM PREDICTIVE MODELLING

1. The unit increases in "TotalPaid_End" and "IndemnityPaid" has no effect on odds of High-Risk Claims as the odd estimate is 1.0000.
2. For Gender, the odds estimate is 0.919 which means if gender is "Female" then the odds of claim to be a high-risk claim has 0.919 times lower odds compared to the Male.
3. For Claimant Type, the odds estimate is 0.889 which means if claimant type is other than indemnity then the odds of claim to be a high-risk claim has 0.889 times lower odds compared to Claimant type as "Indemnity".
4. For Injury nature, the odds estimate is 1.333 which means if injury nature is other than strain then the odds of claim to be a high-risk claim is 1.333 times odds higher compared to "Strain".

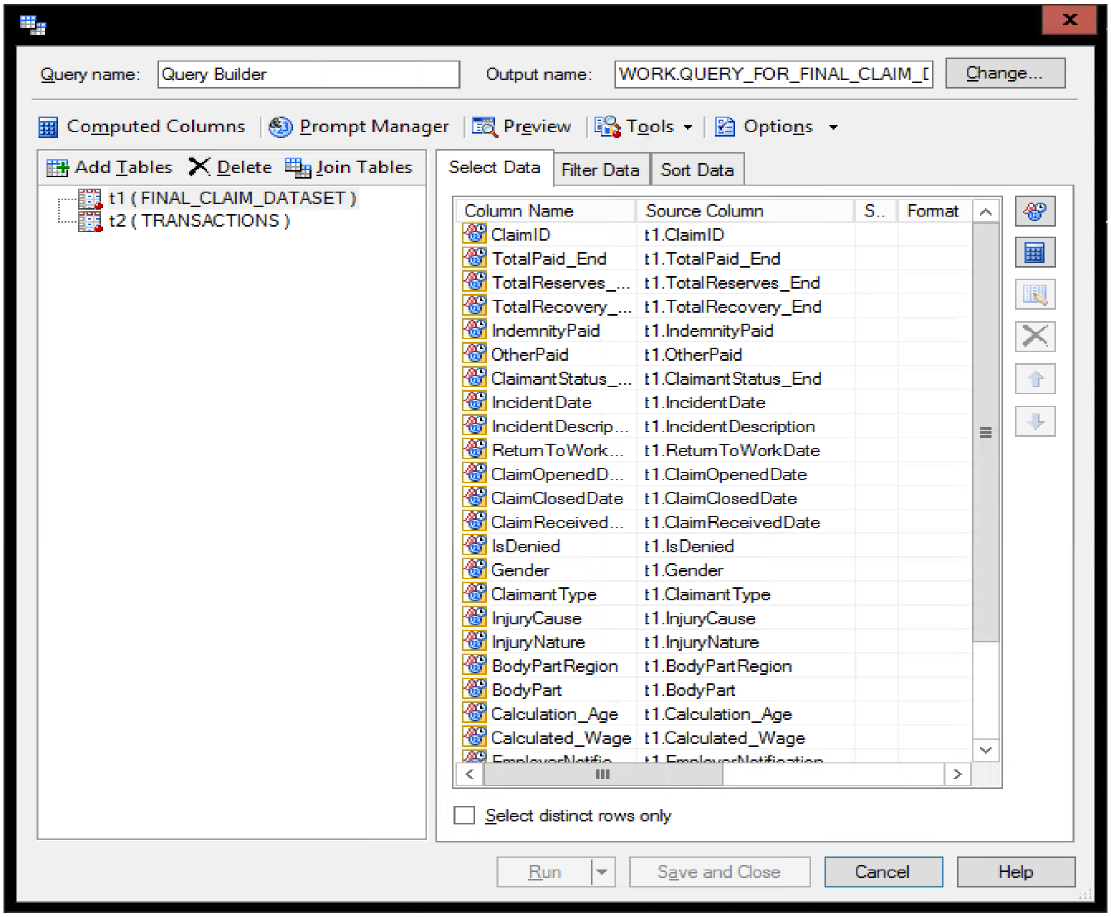
RECOMMENDATIONS

1. Regulation in the workers environment which the Insurance Company must take a firm decision on, along with the Employer.
2. Separate workflow assigned for claimant type Indemnity.
3. Mandatory input fields to put valid values or predefined values to get a valid value.

APPENDIX A – DATA PREPARATION

MERGING THE CLAIMS AND TRANSACTION DATASET

The claims dataset which was explored and cleaned in the first part of the project is merged with the transactions dataset using SAS enterprise guide. The merge of the dataset was carried on the Claims_id column which is common to both the data set.



Query name: Query Builder

Output name: WORK.QUERY_FOR_FINAL_CLAIM_[

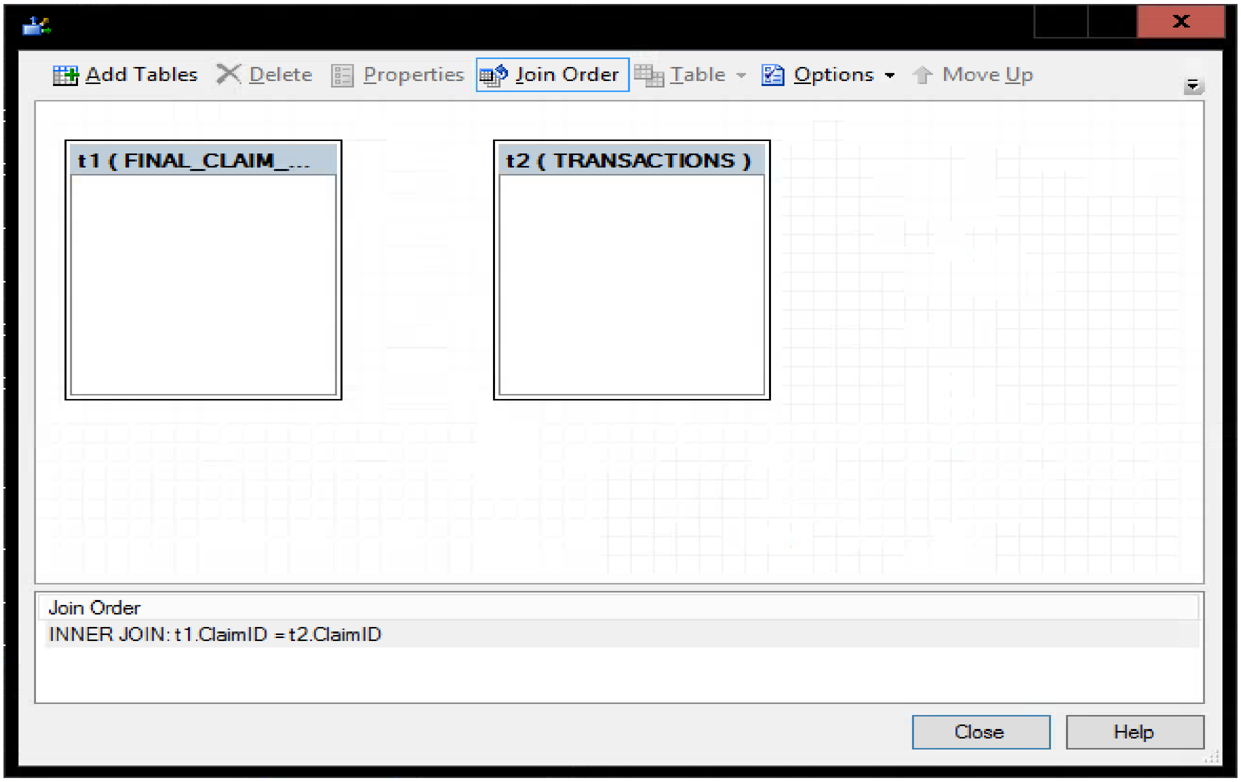
Change...

t1 (FINAL_CLAIM_DATASET)
t2 (TRANSACTIONS)

Column Name	Source Column	S..	Format
ClaimID	t1.ClaimID		
TotalPaid_End	t1.TotalPaid_End		
TotalReserves_...	t1.TotalReserves_End		
TotalRecovery_...	t1.TotalRecovery_End		
IndemnityPaid	t1.IndemnityPaid		
OtherPaid	t1.OtherPaid		
ClaimantStatus_...	t1.ClaimantStatus_End		
IncidentDate	t1.IncidentDate		
IncidentDescrip...	t1.IncidentDescription		
ReturnToWork...	t1.ReturnToWorkDate		
ClaimOpenedD...	t1.ClaimOpenedDate		
ClaimClosedDate	t1.ClaimClosedDate		
ClaimReceived...	t1.ClaimReceivedDate		
IsDenied	t1.IsDenied		
Gender	t1.Gender		
ClaimantType	t1.ClaimantType		
InjuryCause	t1.InjuryCause		
InjuryNature	t1.InjuryNature		
BodyPartRegion	t1.BodyPartRegion		
BodyPart	t1.BodyPart		
Calculation_Age	t1.Calculation_Age		
Calculated_Wage	t1.Calculated_Wage		
EmployerNotific...	t1.EmployerNotification		

☐ Select distinct rows only

FIGURE - 1

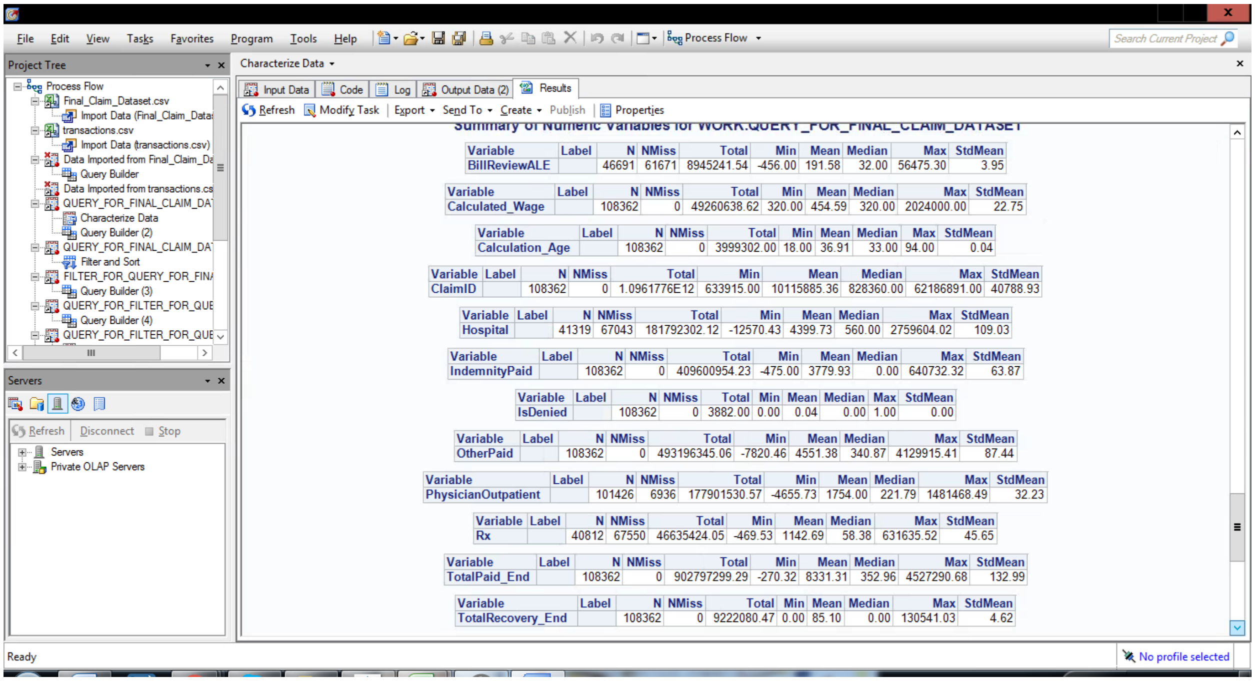


Join Order

INNER JOIN: t1.ClaimID = t2.ClaimID

FIGURE -2

Running the query fetched 108338 rows out of original 133634.Characterization of the merged dataset:



Summary of Numeric Variables for WORK.QUERY_FOR_FINAL_CLAIM_DATASET

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
BillReviewALE		46691	61671	8945241.54	-456.00	191.58	32.00	56475.30	3.95
Calculated_Wage		108362	0	49260638.62	320.00	454.59	320.00	2024000.00	22.75
Calculation_Age		108362	0	3999302.00	18.00	36.91	33.00	94.00	0.04
ClaimID		108362	0	1.0961776E12	633915.00	10115885.36	828360.00	62186891.00	40788.93
Hospital		41319	67043	181792302.12	-12570.43	4399.73	560.00	2759604.02	109.03
IndemnityPaid		108362	0	409600954.23	-475.00	3779.93	0.00	640732.32	63.87
IsDenied		108362	0	3882.00	0.00	0.04	0.00	1.00	0.00
OtherPaid		108362	0	493196345.06	-7820.46	4551.38	340.87	4129915.41	87.44
PhysicianOutpatient		101426	6936	177901530.57	-4655.73	1754.00	221.79	1481468.49	32.23
Rx		40812	67550	46635424.05	-469.53	1142.69	58.38	631635.52	45.65
TotalPaid_End		108362	0	902797299.29	-270.32	8331.31	352.96	4527290.68	132.99
TotalRecovery_End		108362	0	9222080.47	0.00	85.10	0.00	130541.03	4.62

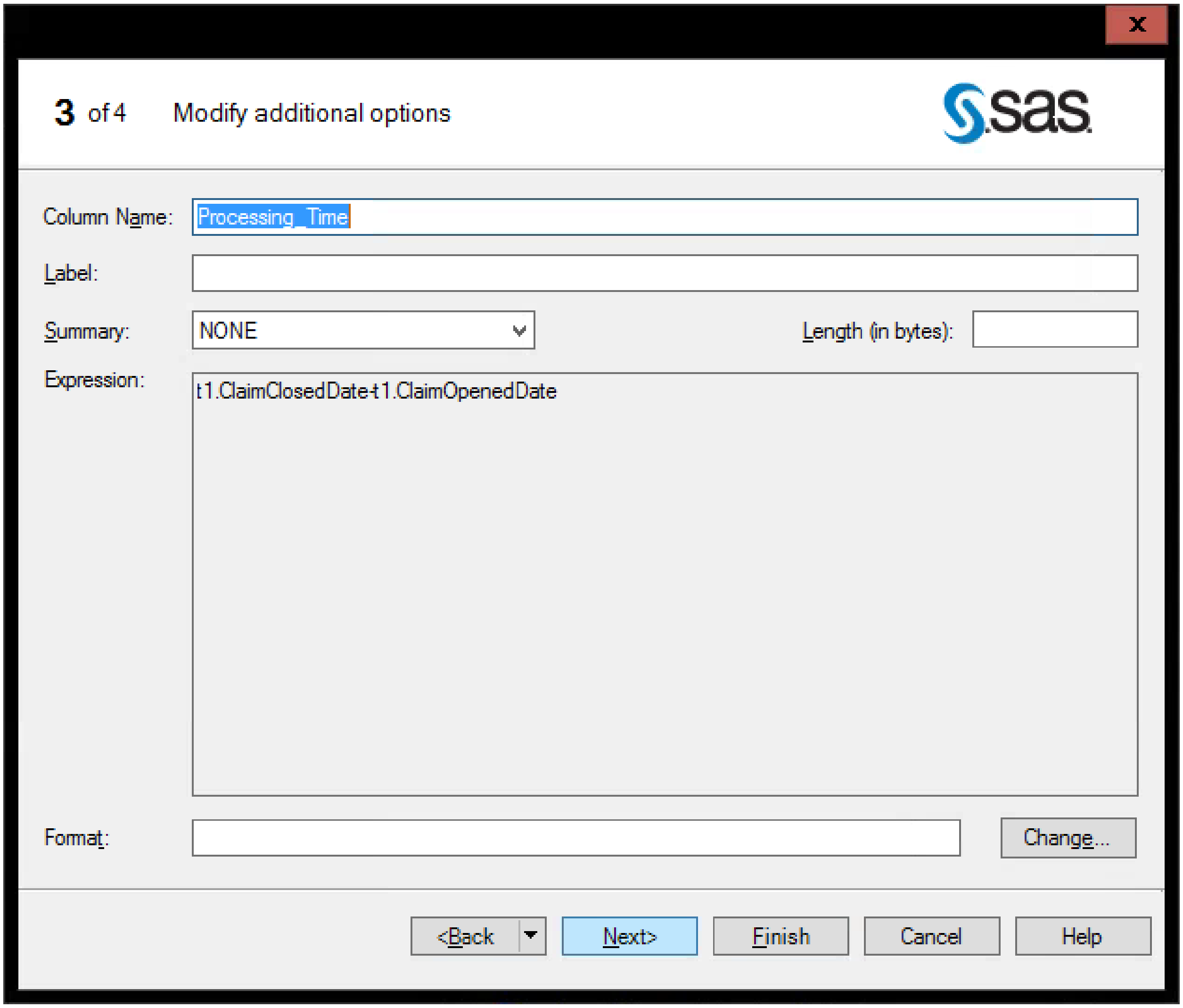
FIGURE – 3

Empty rows in the column BillReviewALE, Hospital, PhysicianOutpatient and Rx were replaced with 0

ADDING NEW VARIABLES TO THE MERGED DATASET

Below are the new derived variables:

1. Column **Processing_Time** was derived by the difference between ClaimOpenedDate and ClaimClosedDate.



3 of 4 Modify additional options

Column Name: Processing_Time

Label:

Summary: NONE

Length (in bytes):

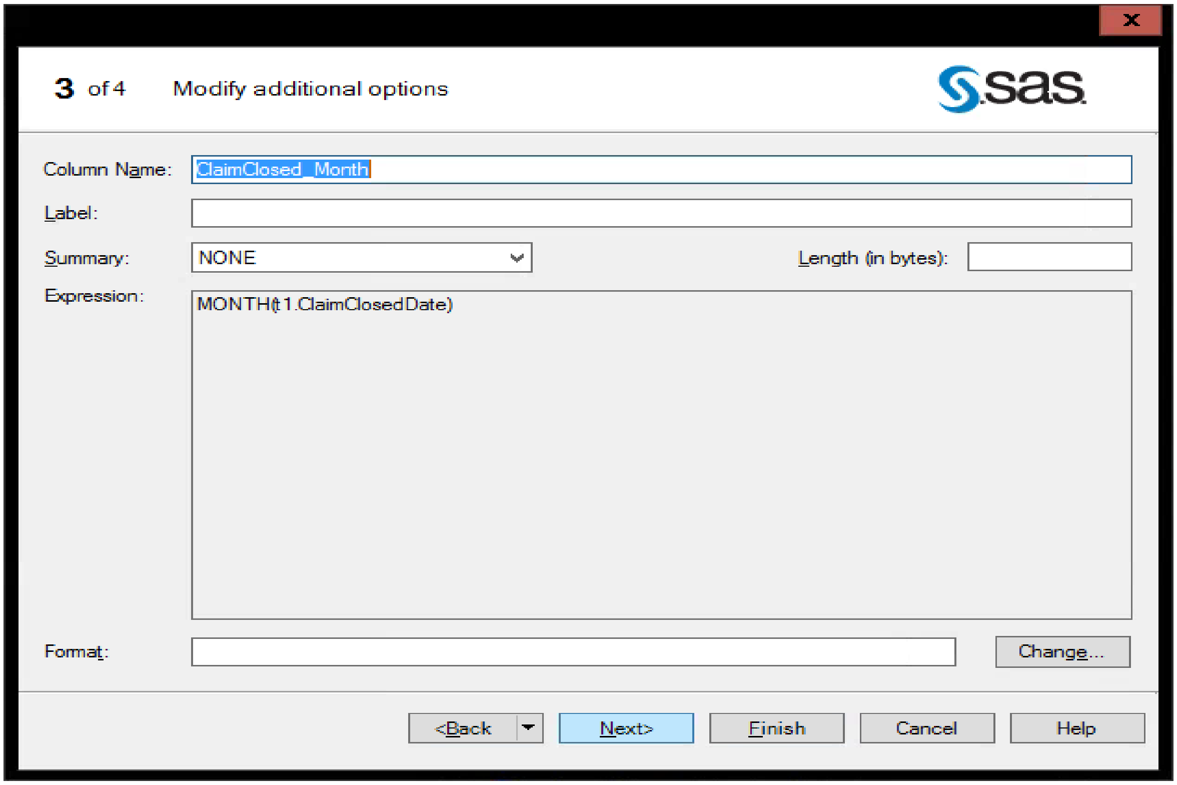
Expression:

```
t1.ClaimClosedDate-t1.ClaimOpenedDate
```

Format:

FIGURE -4

2. Columns **ClaimClosed_Month** and **ClaimClosed_Year** derived from ClaimsClosedDate.



3 of 4 Modify additional options

Column Name: ClaimClosed_Month

Label:

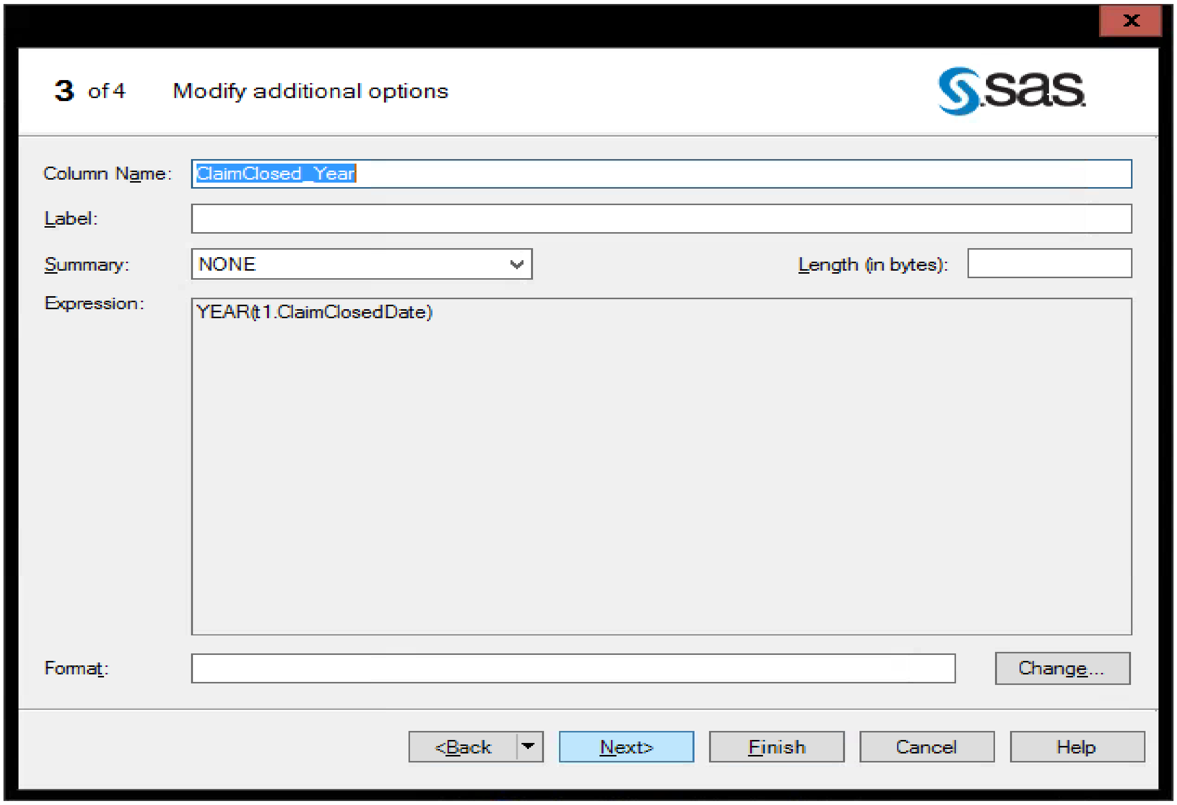
Summary: NONE

Length (in bytes):

Expression: MONTH(t1.ClaimClosedDate)

Format:

FIGURE -5



3 of 4 Modify additional options

Column Name: ClaimClosed_Year

Label:

Summary: NONE

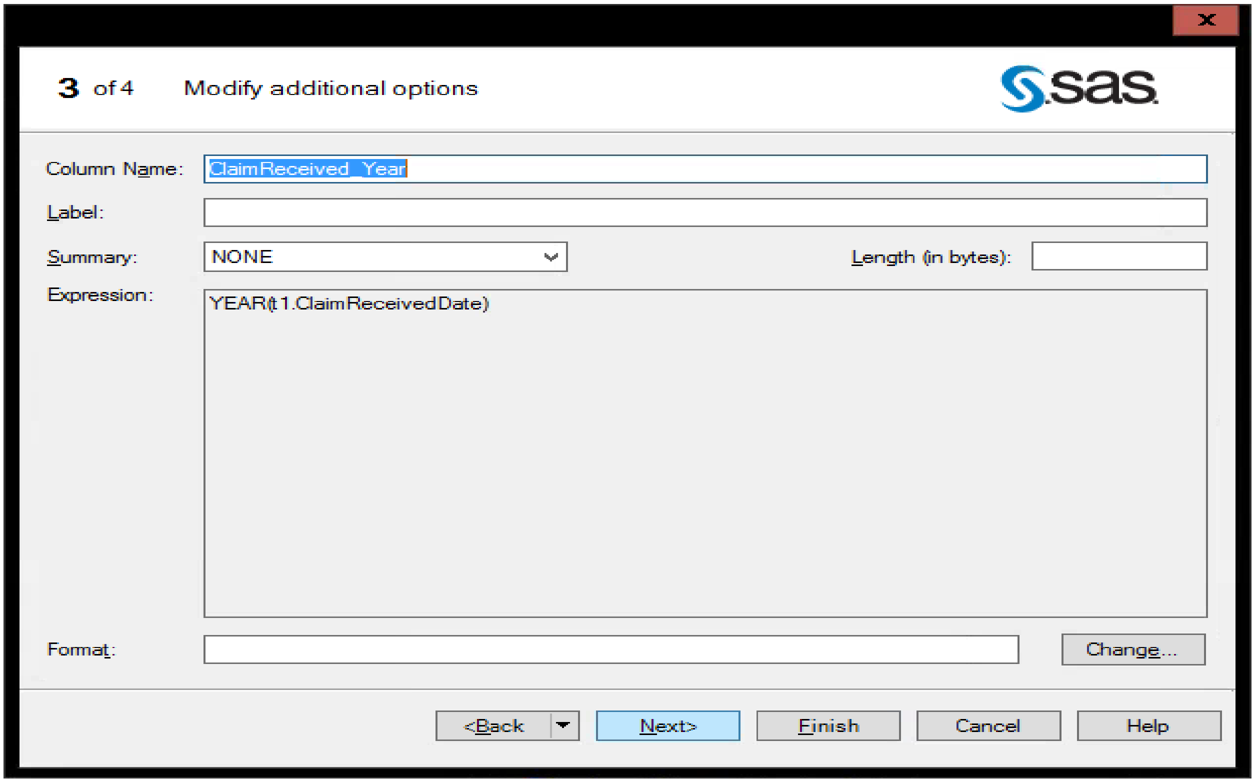
Length (in bytes):

Expression: YEAR(t1.ClaimClosedDate)

Format:

FIGURE-6

3. Columns **ClaimReceived_Month** and **ClaimReceived_Year** derived from ClaimReceivedDate.



3 of 4 Modify additional options

Column Name: ClaimReceived_Year

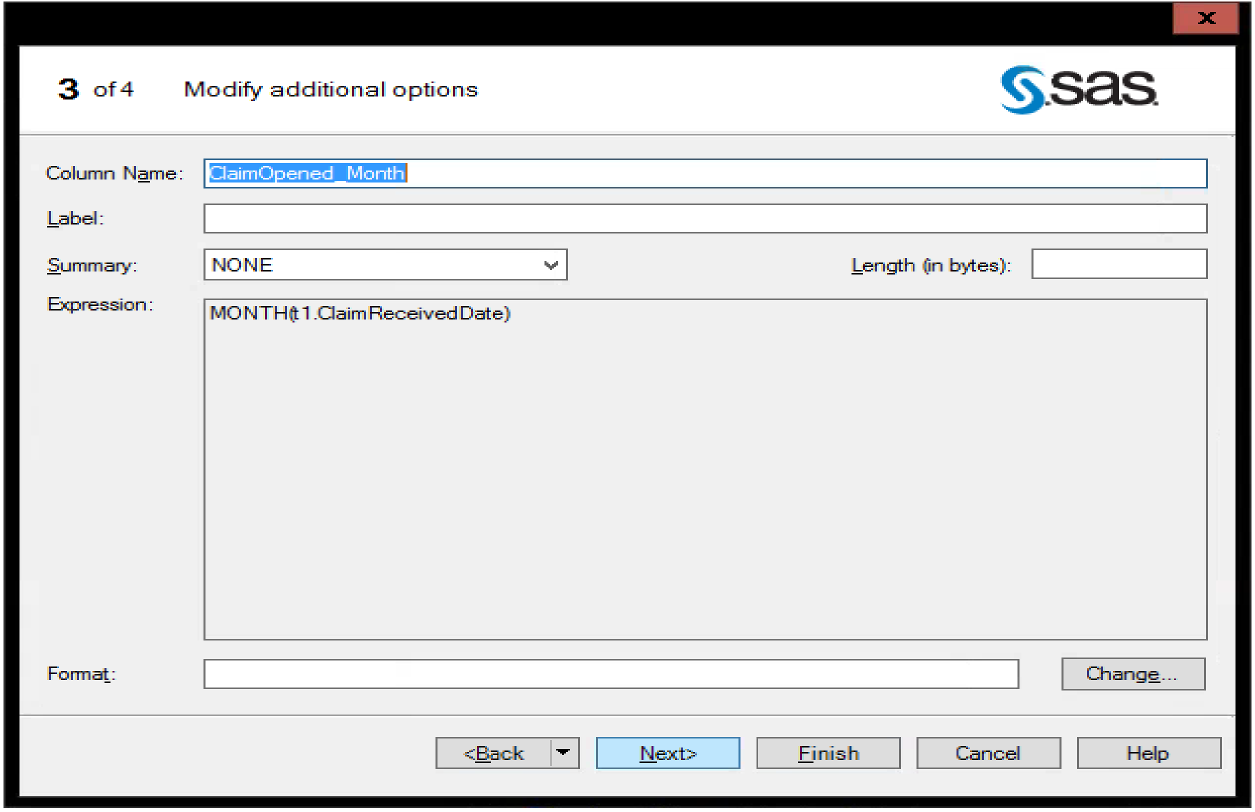
Label:

Summary: NONE

Length (in bytes):

Expression: YEAR(t1.ClaimReceivedDate)

Format:



3 of 4 Modify additional options

Column Name: ClaimOpened_Month

Label:

Summary: NONE

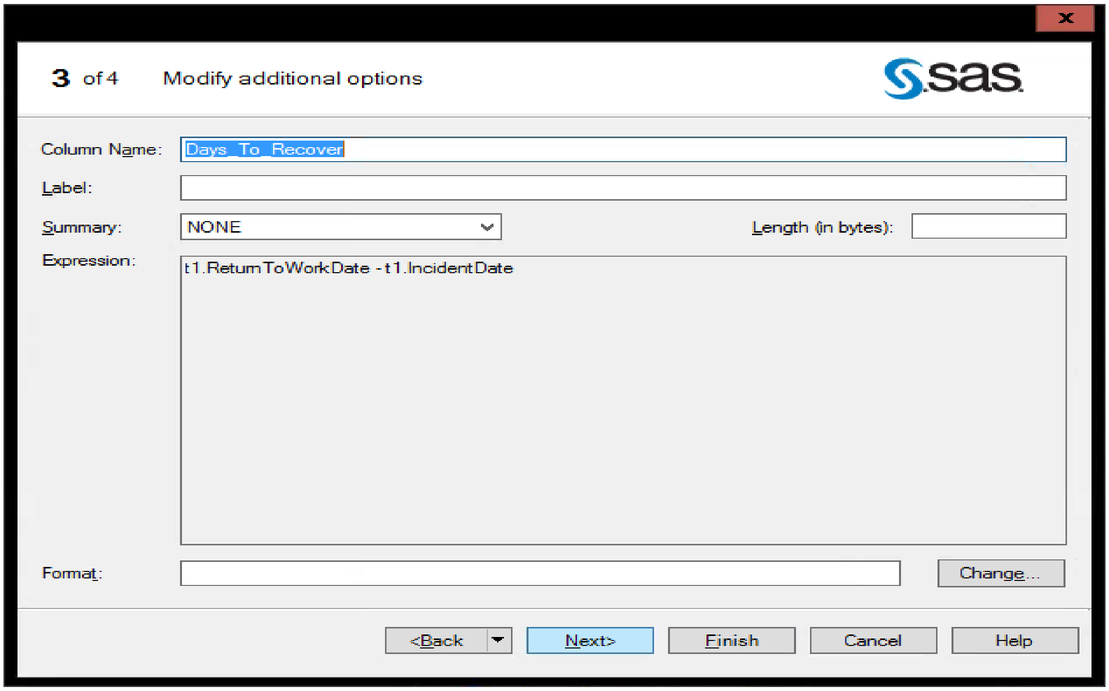
Length (in bytes):

Expression: MONTH(t1.ClaimReceivedDate)

Format:

FIGURE-7

4. Column **Days_To_Recover** derived from the difference between IncidentDate and ReturnToWorkDate.



3 of 4 Modify additional options

Column Name: Days_To_Recover

Label:

Summary: NONE

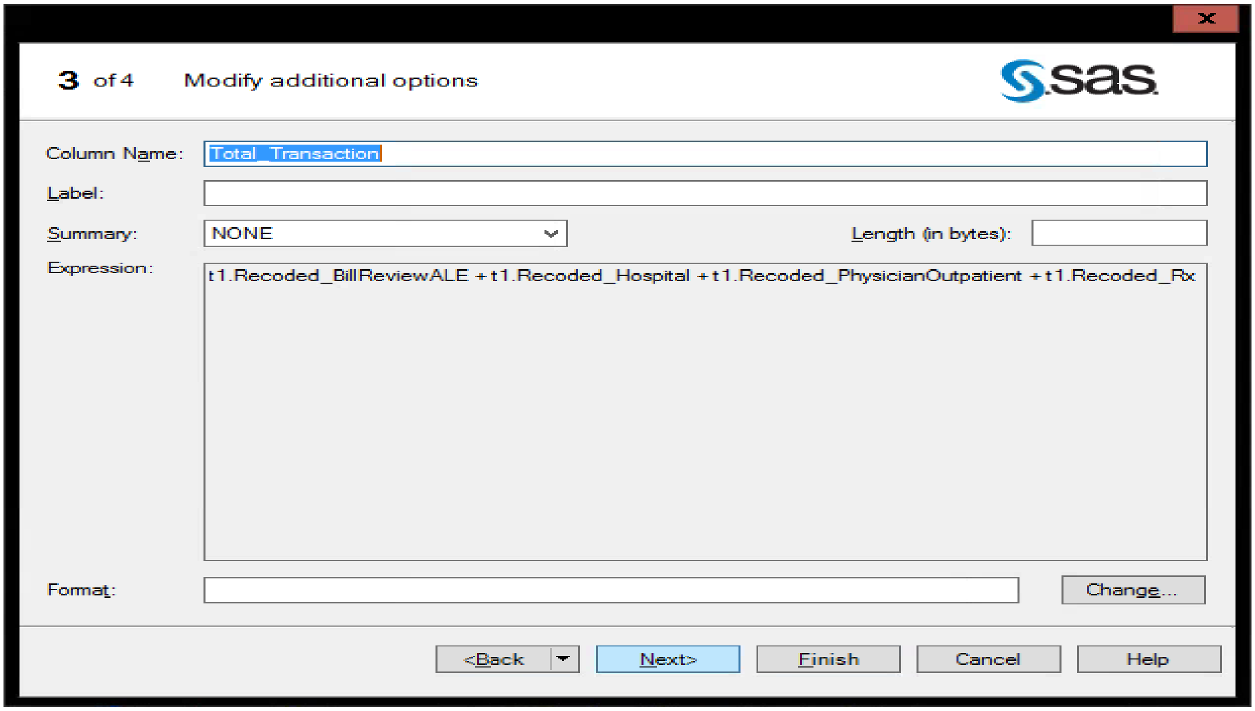
Length (in bytes):

Expression: t1.ReturnToWorkDate -t1.IncidentDate

Format:

FIGURE-8

5. Column **Total_Transaction** is total of the each row of the column in the transaction file which to our understanding does not include the indemnity paid.



3 of 4 Modify additional options

Column Name: Total_Transaction

Label:

Summary: NONE

Length (in bytes):

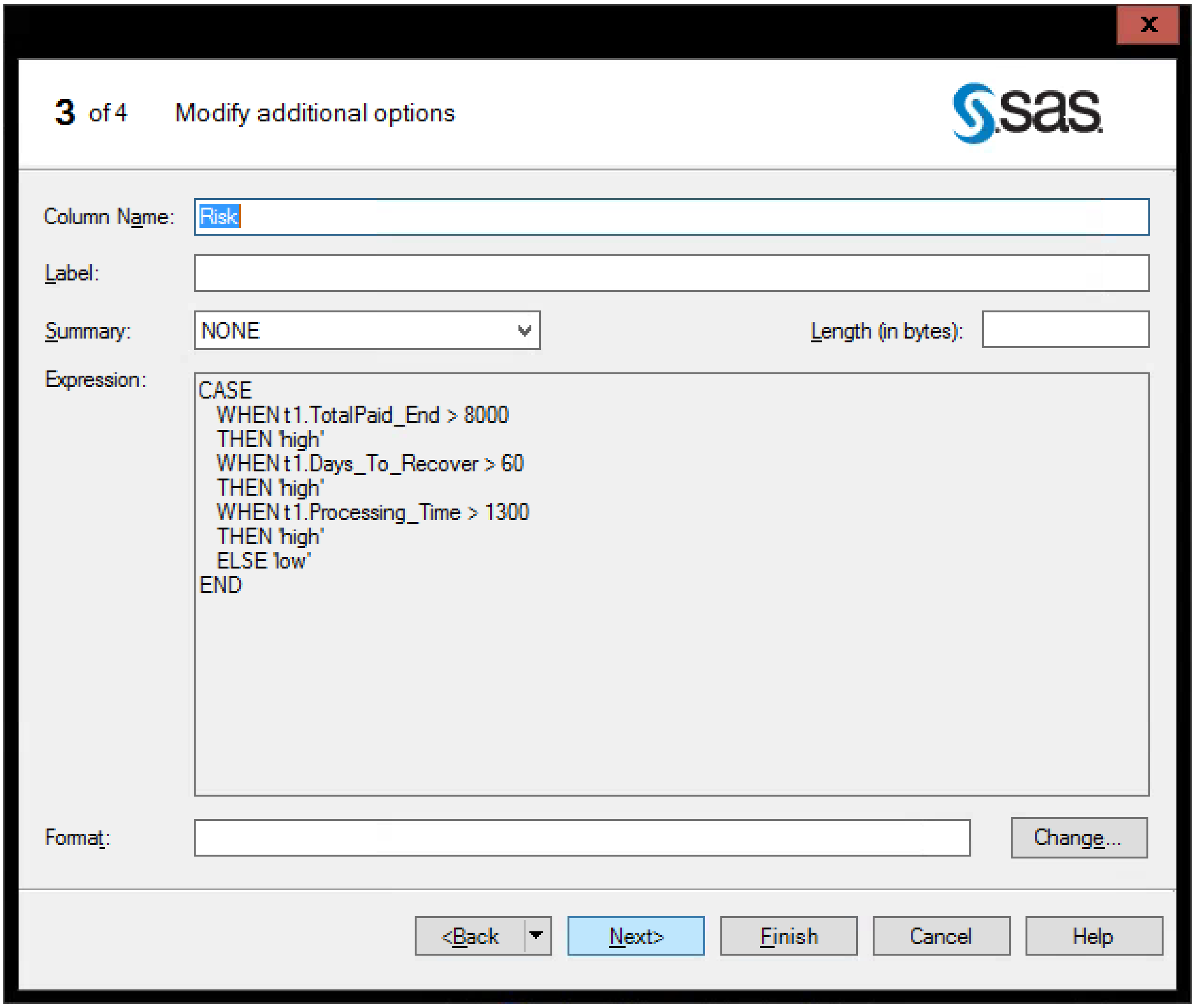
Expression: t1.Recoded_BillReviewALE + t1.Recoded_Hospital + t1.Recoded_PhysicianOutpatient + t1.Recoded_Rx

Format:

FIGURE-9

6. Column **Risk** is a binary dependent variable to indicate "high-risk" claims, satisfying the below condition:
    a. Processing_Time>1300 days
    b. Days_To_Recover >60 days
    c. ToatlPaid_end > $8000

    We have considered the mean value of each variable as the deciding values if a claim is "High" risk of it is "Low" risk.



3 of 4 Modify additional options

Column Name: Risk

Label:

Summary: NONE

Length (in bytes):

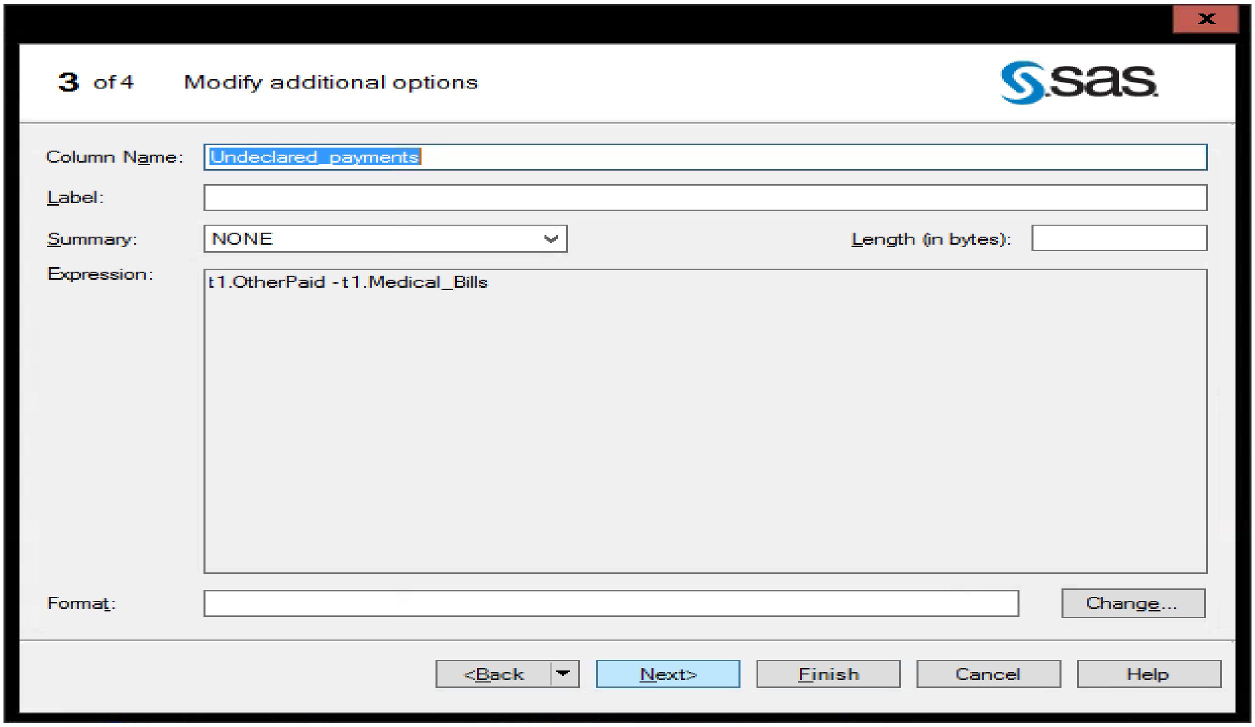
Expression:

```
CASE
  WHEN t1.TotalPaid_End > 8000
  THEN 'high'
  WHEN t1.Days_To_Recover > 60
  THEN 'high'
  WHEN t1.Processing_Time > 1300
  THEN 'high'
  ELSE 'low'
END
```

Format:

FIGURE-10

7. **Undeclared_Payments** is derived from the difference between column OtherPaid and Total_Transaction



Column Name: Undeclared_payments

Label:

Summary: NONE

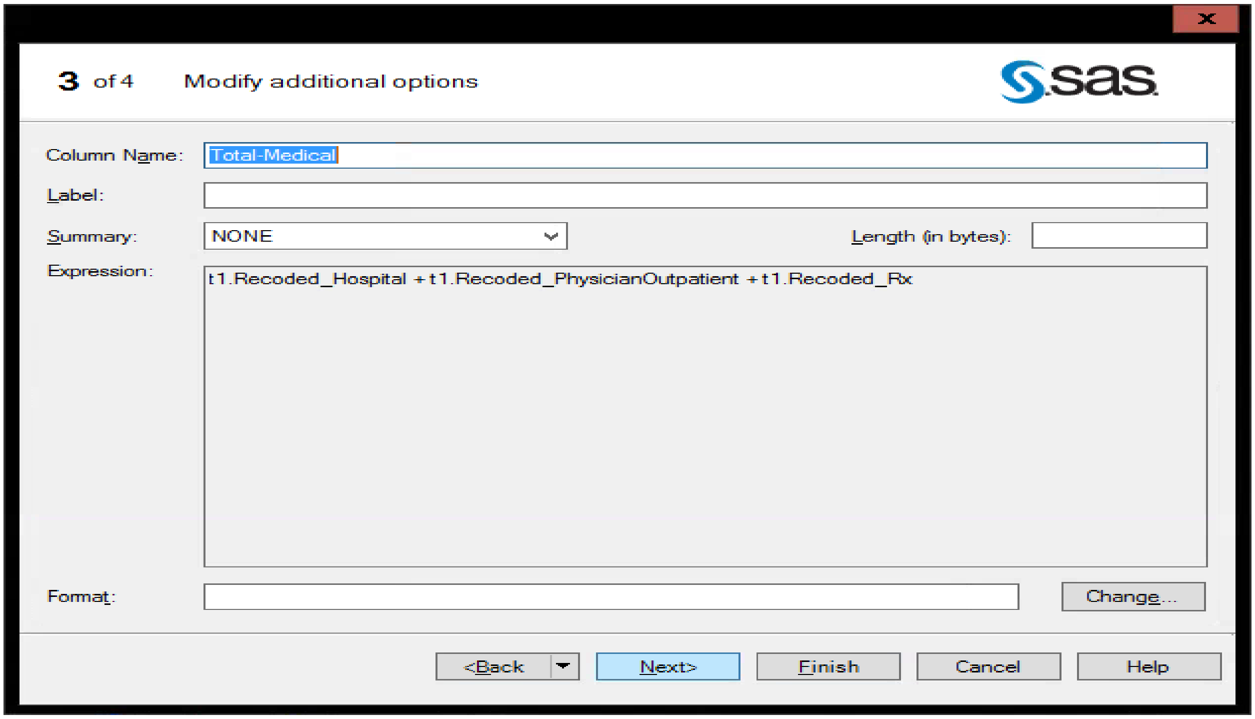
Length (in bytes):

Expression: t1.OtherPaid - t1.Medical_Bills

Format:

FIGURE- 11

8. **Total-Medical** is derived from row Hospital, PhysicianOutpatient and Rx.



Column Name: Total-Medical

Label:

Summary: NONE

Length (in bytes):

Expression: t1.Recoded_Hospital + t1.Recoded_PhysicianOutpatient + t1.Recoded_Rx

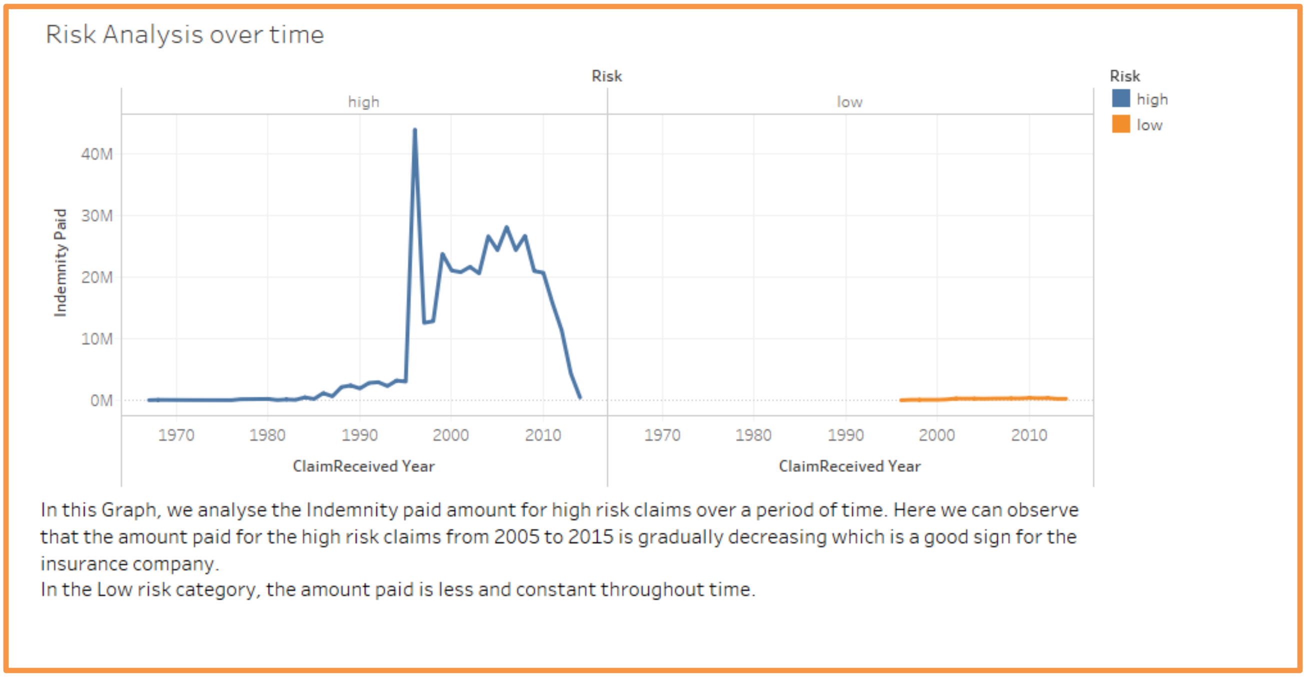
Format:

Figure-12

APPENDIX B -MODEL PLANNING

For model planning we have considered visualization of the final data set and five predictive techniques as mentioned below

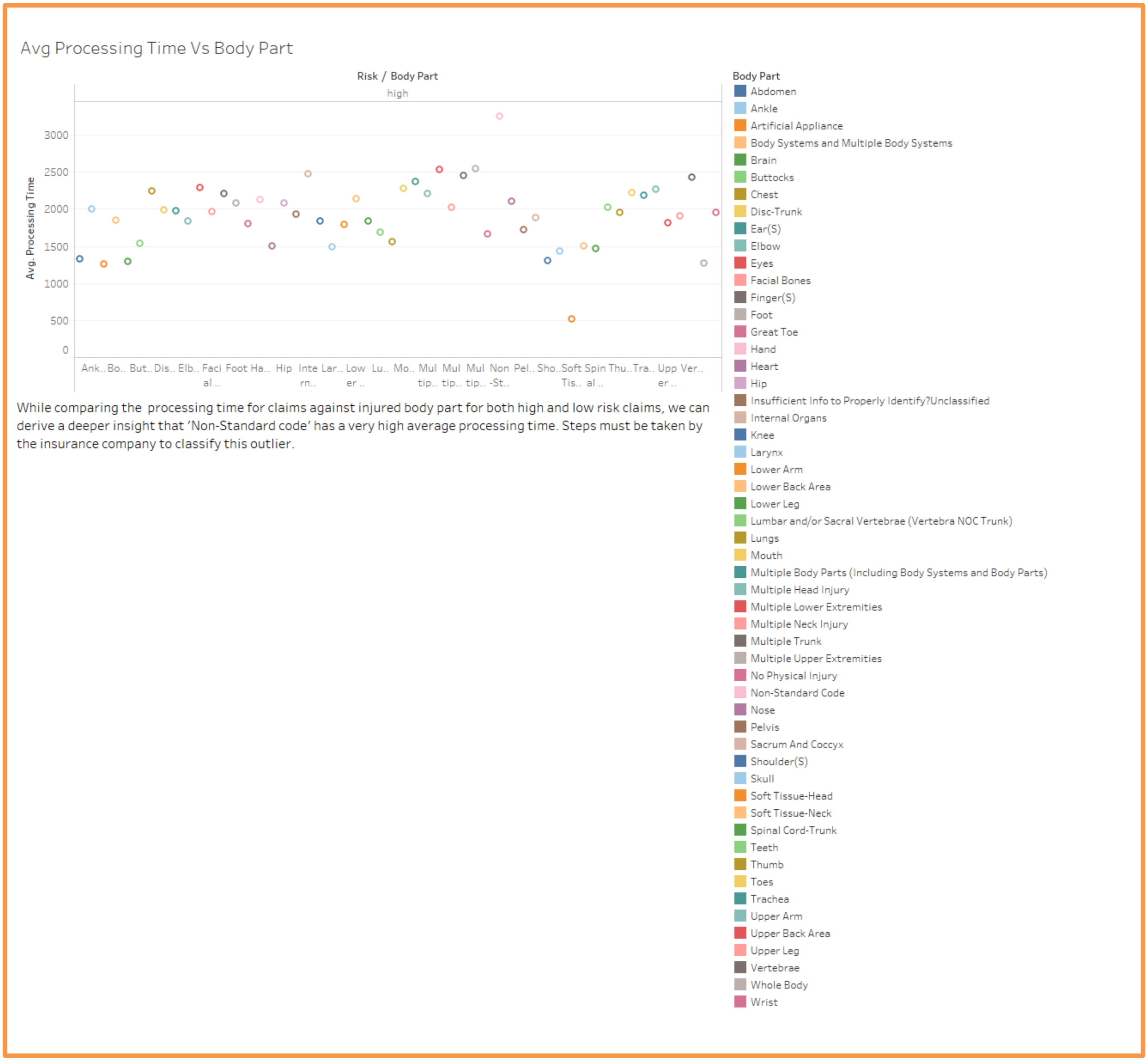
A. Visualizations



In this Graph, we analyse the Indemnity paid amount for high risk claims over a period of time. Here we can observe that the amount paid for the high risk claims from 2005 to 2015 is gradually decreasing which is a good sign for the insurance company.

In the Low risk category, the amount paid is less and constant throughout time.

FIGURE-1

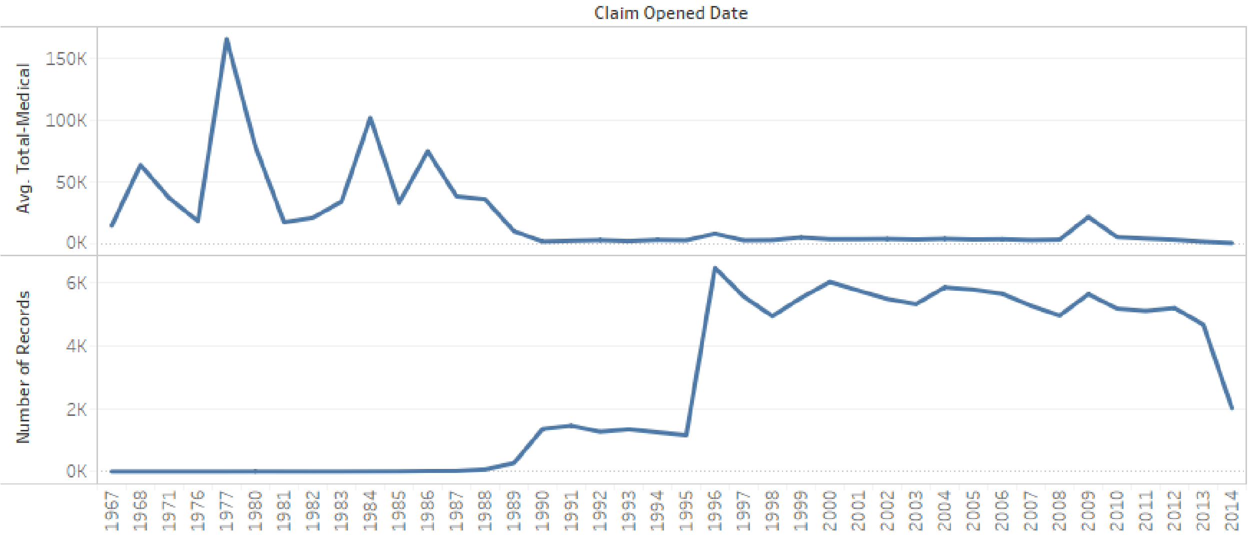


Avg Processing Time Vs Body Part

While comparing the processing time for claims against injured body part for both high and low risk claims, we can derive a deeper insight that 'Non-Standard code' has a very high average processing time. Steps must be taken by the insurance company to classify this outlier.

FIGURE-2

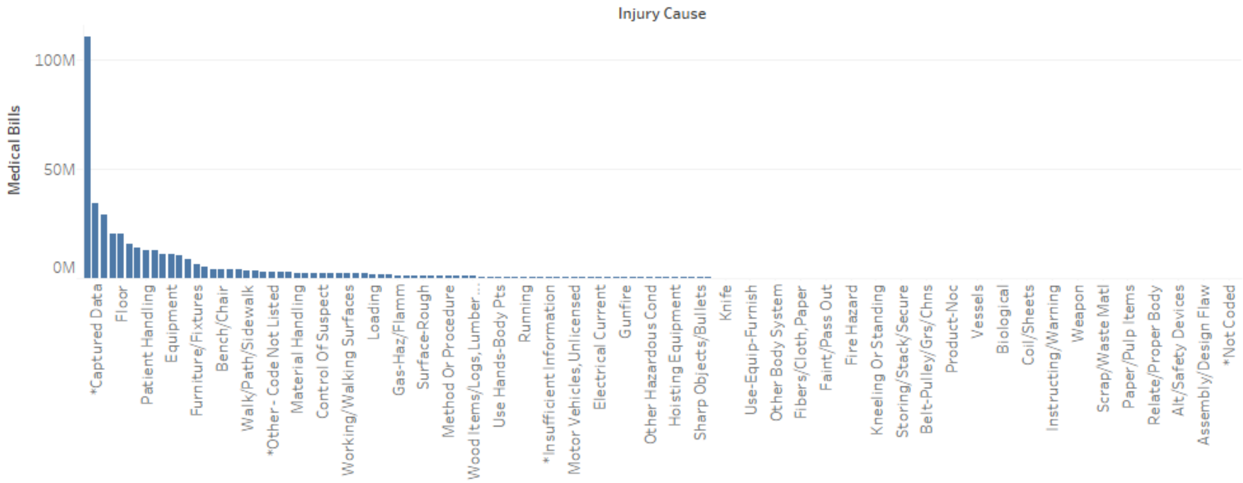
Number of claims opened Vs Total Medical bills



When observing data from 1996, the number of claims opened seems to be fluctuating; Whereas average of the total medical bills remain fairly constant.

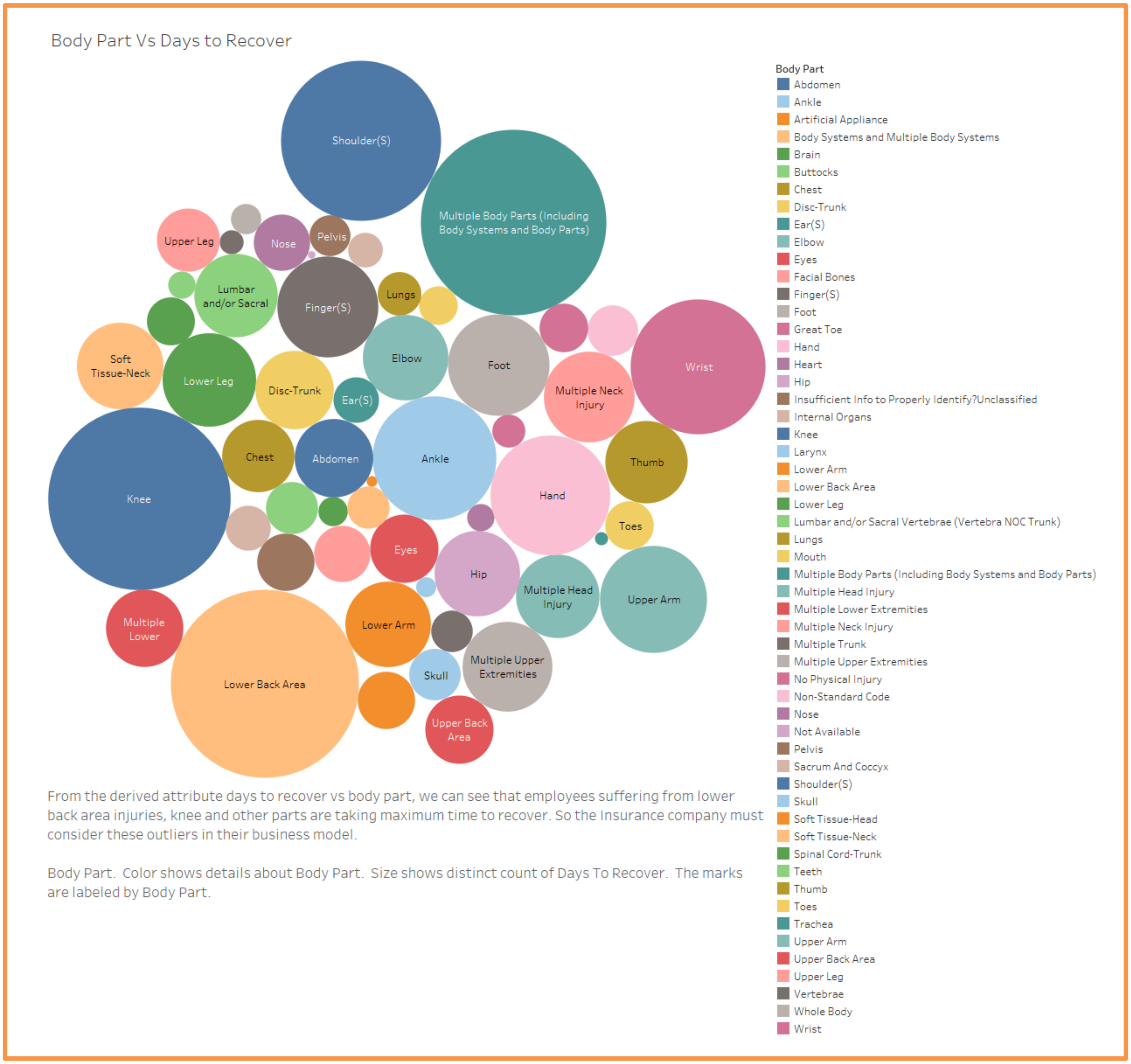
FIGURE-3

Medical bills for injury cause



Medical bills paid for various injury caused is plotted and it is seen that more than 100 million dollars has gone unaccounted. This is a huge loss for the insurance company since this is uncategorized.

FIGURE-4

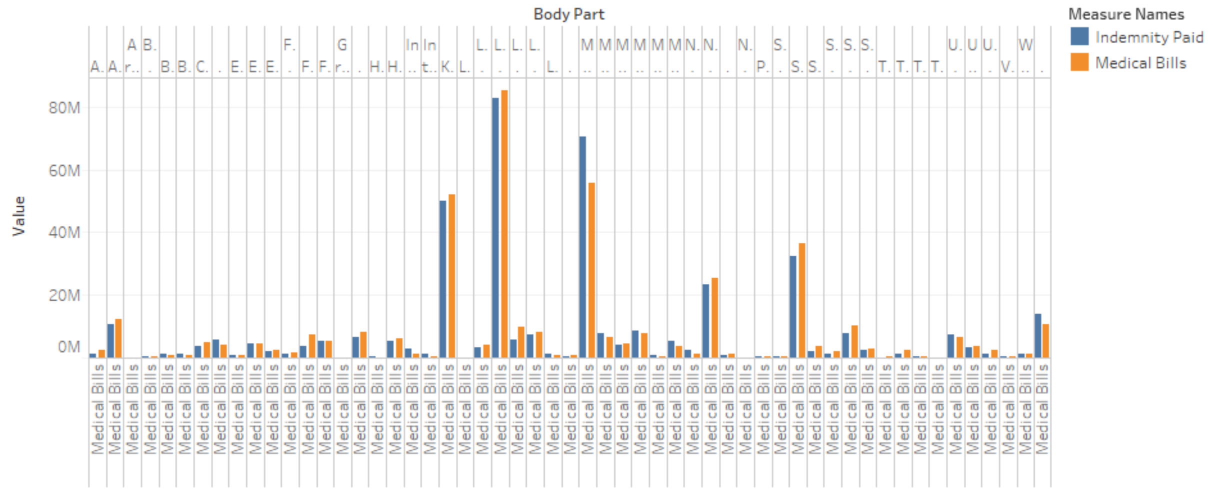


From the derived attribute days to recover vs body part, we can see that employees suffering from lower back area injuries, knee and other parts are taking maximum time to recover. So the Insurance company must consider these outliers in their business model.

Body Part. Color shows details about Body Part. Size shows distinct count of Days To Recover. The marks are labeled by Body Part.

FIGURE-5

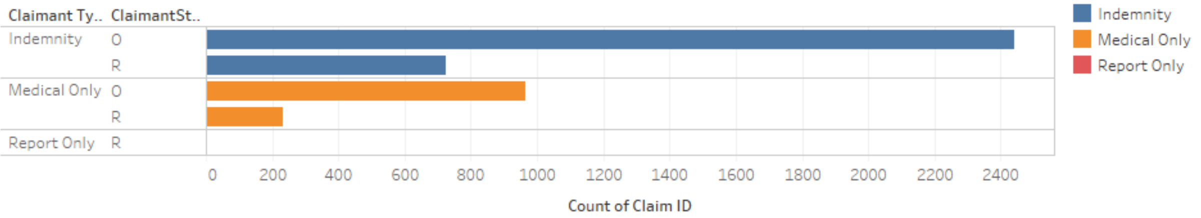
Indemnity Paid and Total Medical bills



Comparing the total medical bills with indemnity paid in body parts distribution, It can be seen that insurance company pays highest for the lower back area.

FIGURE -6

Open/Re-Open Claims and Claimant type



Count of Claim ID for each ClaimantStatus End broken down by Claimant Type. The view is filtered on ClaimantStatus End, which keeps O and R.
There more number of open and re-open claims in indemnity type than Medical Only.

FIGURE-7

B. Comparing predictive models

Based on the understanding among our group members of various predictive modeling techniques and insights from the data exploration and visualization analysis conducted, below are the 4 predictive modeling techniques with the outcome variables and independent variables that we have chosen for each model:

1. **Linear Regression:** It is a type of statistical model for estimating the relationship between a dependent variable and one or more independent variable. There are two type of linear regression- Simple and Multiple. There are certain assumptions for linear regression like linear relationship, homoscedasticity to name a few. Linear regression models are built using scatter plots to approximate a line of best fit. For our dataset, the outcome variable is the Risk and independent variables are Indemnity Paid and Claimant Type. The difficulties that could occur if we choose this predictive model are as follows:
    - In our data set, the Claimant Type is a categorical variable. So, there is a need to modify these data into numerical to fit in the linear regression model.
    - As our data set is large and the models sensitivity to outliers, the scatter plot of each independent variable against dependent variable becomes visually challenging.

2. **Decision Tree:** A decision tree is a predictive modeling technique used to create homogeneous regions, to predict the target variable of an instance by determining which segment it falls into. These decision trees are created using a root and a series of branches and nodes. The branches and nodes are split and stopped using the criteria given to the model. To classify the risk of a claim, the independent variable chosen are Days_To_Recover, Injury Nature and Gender. Below are some of the disadvantages of decision tree model:
    - When faced with data an extreme amount of data you run the risk of overfitting your tree. Overfitting is when you have too many branches that reflect anomalies due to noise or outliers.
    - Injury Nature is a categorical variable with more than two categories which could impact the branching of decision tree modeling.

3. **Logistics Regression:** When we are trying to interpret a binary dependent variable for your predictive model, linear regression will not work. To interpret a binary dependent variable a great predictive model to use is logistic regression. In a logistic model, we find what is probability is going to happen. The probability outcome is gathered by calculating the odds between variables. For our data set we have decided to choose, TotalPaid_End, Indemnity Paid, Gender, Claimant Type and Injury Nature as the independent variables that would impact the odds of claims to be a high-risk claim. There are certain limitations of Logistics Regression:
    - Along with the limitations of logistic regression only working for binary dependent variables, it is sometimes also difficult to determine which independent variables affect the dependent variable significantly.

- Another limitation is that in logistic regression with more than two categories of categorical variable, it is hard to interpret the prediction developed by the model.

4. **Association Rule Mining:** Association Rule Mining finds the affinities between two variables. Even though this type of predictive model does not produce an outcome variable, it is still a good way to see how one variable reacts to another variable. We measure interaction of variable but support and confidence. Support is the probability that an action contains either variable and confident is the probability that an action contains both variables. In the claims business, we can compare the affinities between BodyPartRegion and BodyPart from our dataset. With the observation from the calculation of support and confidence between the two variables, we will be able to tell what are the most injured parted in an associated region.
    - The limitation of this technique pertaining to our data set is that we would not be able to achieve and interpret the main event of either reducing the cost or processing time of the claim process.

5. **Time Series forecasting:** Time series forecasting is the use of a model to predict future values based on previously observed values. For our data set the outcome variable could be "average processing time" and the independent variable is year of incident. This will help us determine the nature of average processing time and evolution of the claim process over the period and depending on that it will predict the average processing time in the coming years.
    - Time series is more effective if the sequence taken is successively spaced. But in or dataset there is lot of gap in the date of incident.
    - However, the limitation with this predictive model is that from the large data set with so many different types of variables, it hard to just predict the average processing time of a claim in the coming years.

APPENDIX C - MODEL BUILDING

From the above listed predictive modeling techniques discussed and compared, we have decided to use Logistic Regression on risk which is the binary dependent variable in this modeling technique. A claim which has a processing time (difference between the ClaimOpenedDate and ClaimClosedDate) greater than and equal to 1300 days are called **High Risk Claims**. Based on the insights from visualization, understanding of the data set variables which could impact processing time of the claims and to analyze and calculate the probability of the event - "claim to be high-risk claim" based on the available past data, we have considered below independent variables:

- IndemnityPaid – Total Indemnity Paid
- TotalPaid_End – Total Amount Paid to date
- ClaimantType – Type of Claim (Medical only, Indemnity, or Report Only)
- InjuryNature – Type of Injury
- Gender – Male, Female or Not Provided

The dependent variable description:

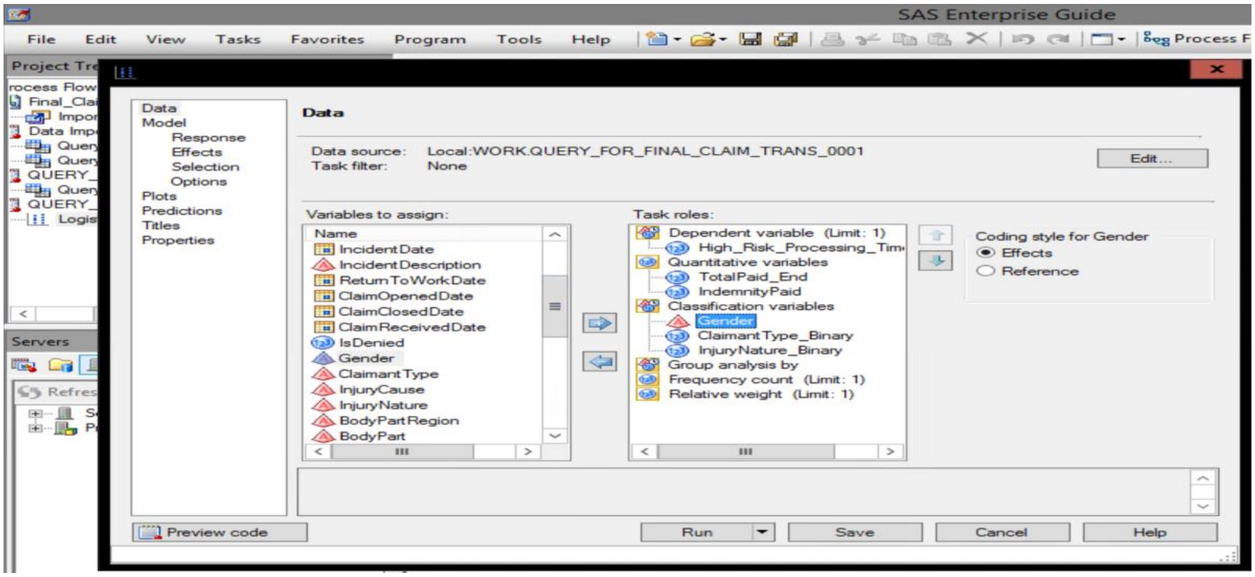
- High-Risk – Indicates high risk claims; 1 – High and 0 – low.

From the insights of our initial data exploration and visualizations is that "Strain is the most occurring injury type" and from general perception that when type of claim is indemnity, the processing could be impacted, we have decided to focus on these areas while building predictive model. Further, to develop better insights and understanding from the logistics regression modeling technique on Risk, we have modified the ClaimantType and InjuryNature to be considered as binary. The details of the modified data set before proceeding to perform logistics regression are as follows:

- ClaimantType – Indemnity has been considered as 1 and remaining are 0.
- InjuryType – Strain has been considered as 1, and other values of the InjuryType are 0.

After this modification, we have decided to remove all records where gender is not available. There were 8944 such records and after all these modification steps, we have performed logistics regression on high risk and build a predictive model. Below are the screenshots (Fig - Fig) to show the steps involved in the process :

1. Fig 1 shows the selection of dependent and independent variables to build the model. We have selected "High_Risk_Processing_Time" as a dependent variable and the independent variables has two sections:
    - Quantitative variables - "TotalPaid_End" and "IndemnityPaid" as these are numerical variables.
    - Classification variables – "ClaimantType_Binary", "Gender" and "InjuryNatrue_Binary" as these are categorial variables.



Data

Data source: Local:WORK.QUERY_FOR_FINAL_CLAIM_TRANS_0001
Task filter: None

Variables to assign:

Name
IncidentDate
IncidentDescription
ReturnToWorkDate
ClaimOpenedDate
ClaimClosedDate
ClaimReceivedDate
IsDenied
Gender
ClaimantType
InjuryCause
InjuryNature
BodyPartRegion
BodyPart

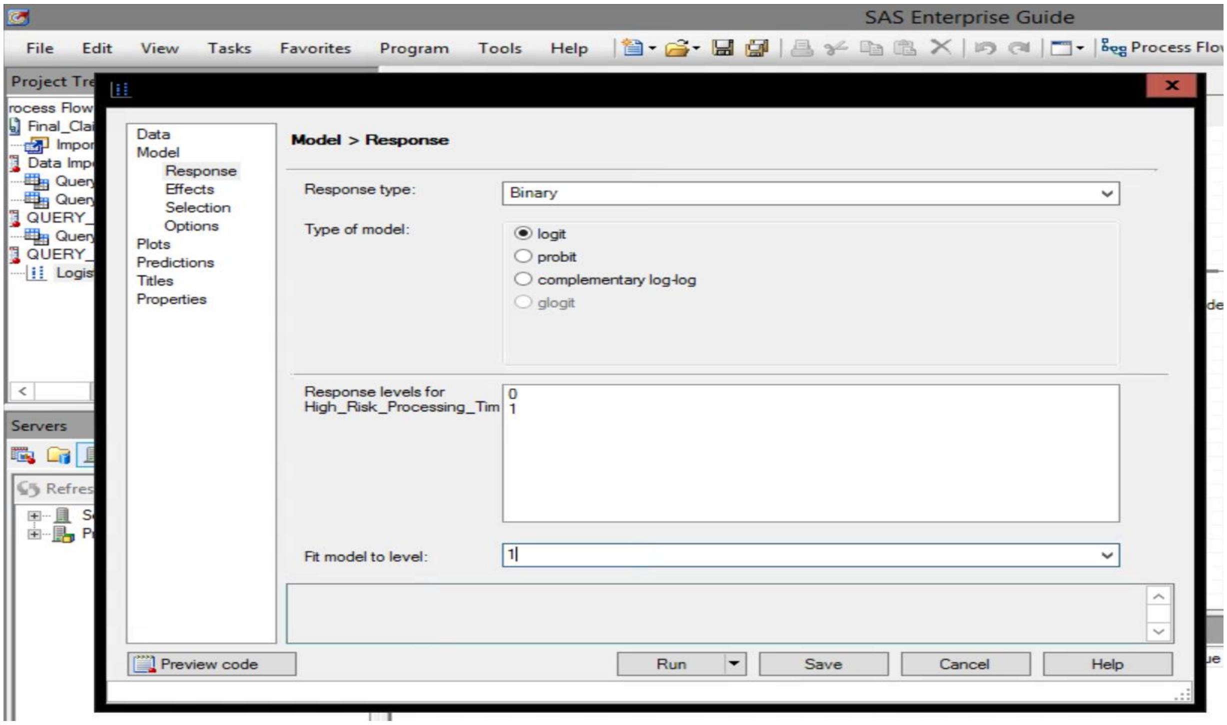
Task roles:

Dependent variable (Limit: 1)
  High_Risk_Processing_Tim
Quantitative variables
  TotalPaid_End
  IndemnityPaid
Classification variables
  Gender
  ClaimantType_Binary
  InjuryNature_Binary
Group analysis by
Frequency count (Limit: 1)
Relative weight (Limit: 1)

Coding style for Gender
● Effects
○ Reference

FIGURE-1

2. Fig 2 shows setting the model > Response section where we have:
   - Response type: Binary
   - Type of model: logit
   - Fit model to level: 1



Model > Response

Response type: Binary

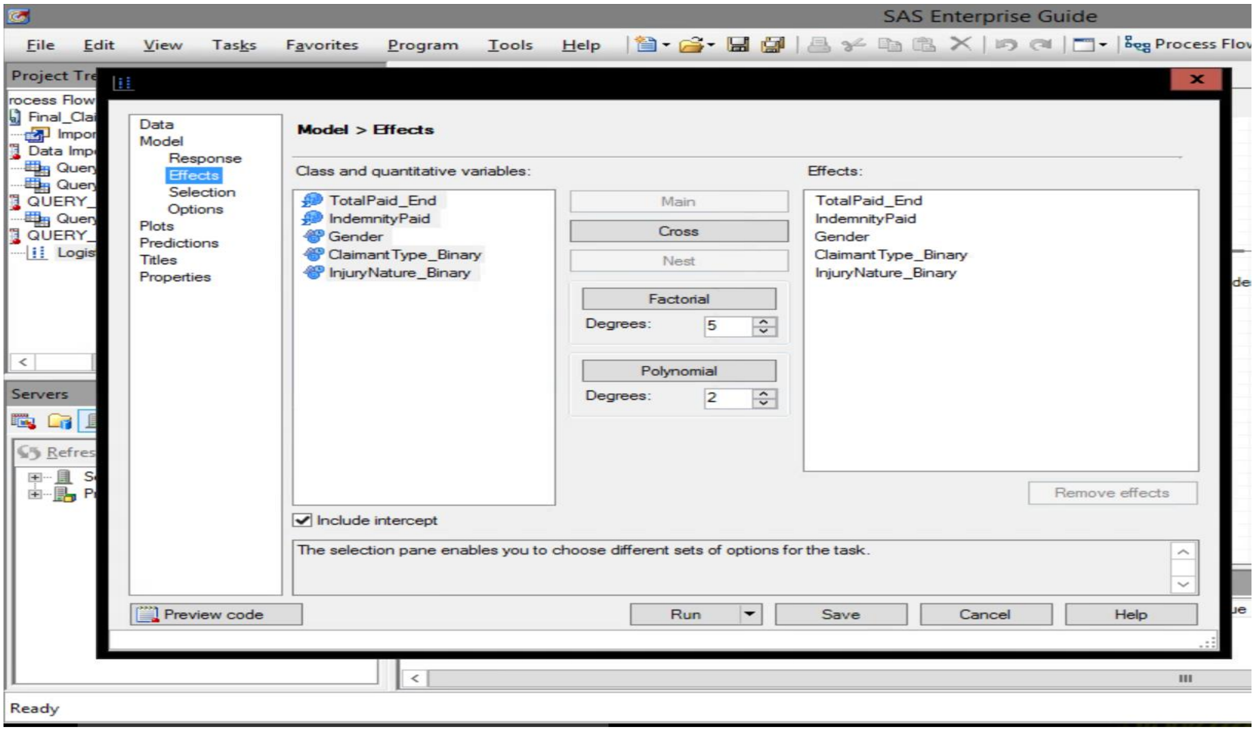
Type of model:
● logit
○ probit
○ complementary log-log
○ glogit

Response levels for High_Risk_Processing_Tim
0
1

Fit model to level: 1

FIGURE-2

3. Fig 3 shows Model > Effects. We have added all the Class and Quantitative variables as “Main” effects in the models.



Model > Effects

Class and quantitative variables:

- TotalPaid_End
- IndemnityPaid
- Gender
- ClaimantType_Binary
- InjuryNature_Binary

Effects:

- TotalPaid_End
- IndemnityPaid
- Gender
- ClaimantType_Binary
- InjuryNature_Binary

Factorial Degrees: 5

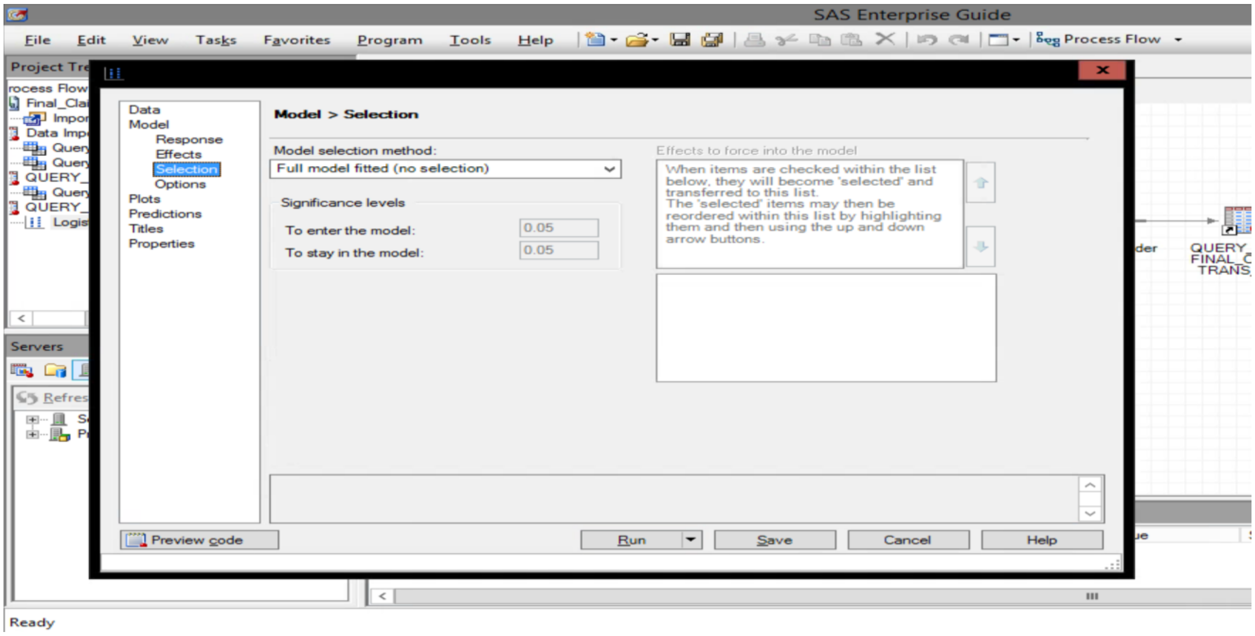
Polynomial Degrees: 2

☑ Include intercept

The selection pane enables you to choose different sets of options for the task.

FIGURE-3

4. Fig 4 shows the Model Selection Method section where we have “Full model fitted”.



Model > Selection

Model selection method: Full model fitted (no selection)

Significance levels

To enter the model: 0.05

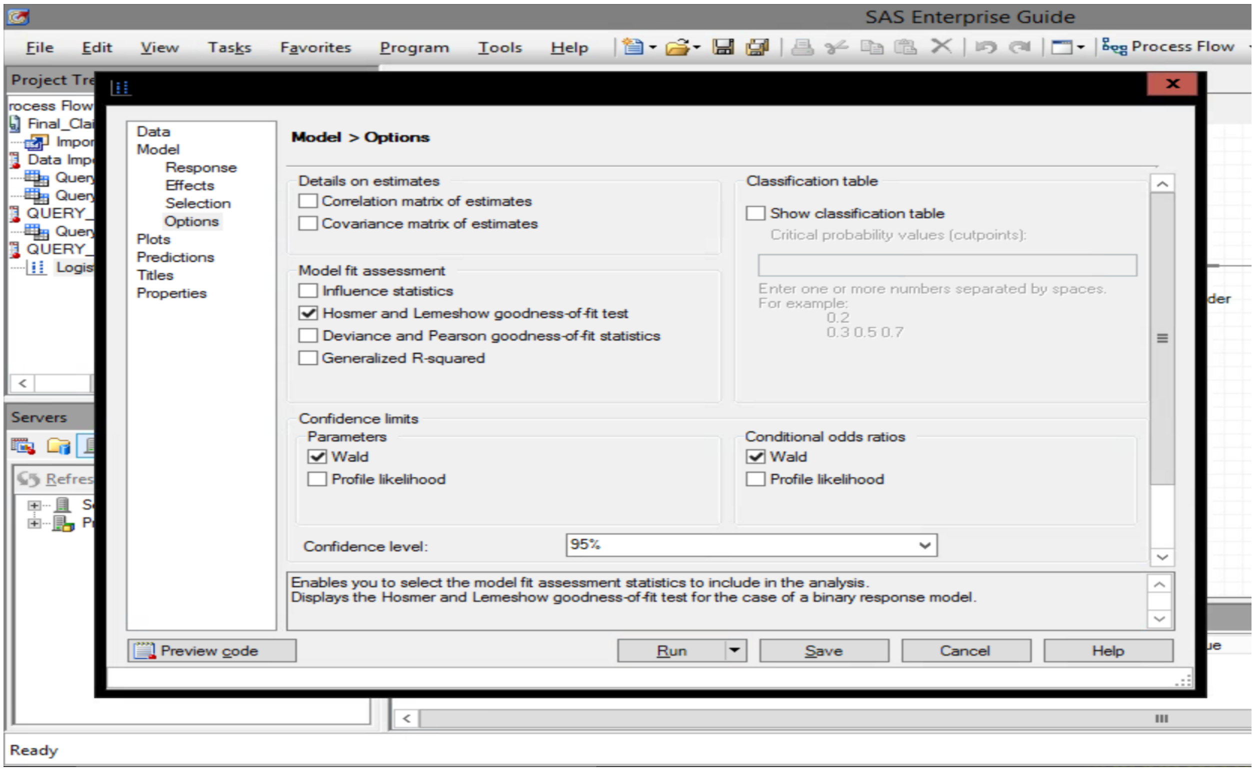
To stay in the model: 0.05

Effects to force into the model

When items are checked within the list below, they will become 'selected' and transferred to this list. The 'selected' items may then be reordered within this list by highlighting them and then using the up and down arrow buttons.

FIGURE-4

5. Fig 5 shows the Models > options selected.



Model > Options

Details on estimates
□ Correlation matrix of estimates
□ Covariance matrix of estimates

Model fit assessment
□ Influence statistics
☑ Hosmer and Lemeshow goodness-of-fit test
□ Deviance and Pearson goodness-of-fit statistics
□ Generalized R-squared

Classification table
□ Show classification table
Critical probability values (cutpoints):
Enter one or more numbers separated by spaces.
For example:
0.2
0.3 0.5 0.7

Confidence limits
Parameters
☑ Wald
□ Profile likelihood

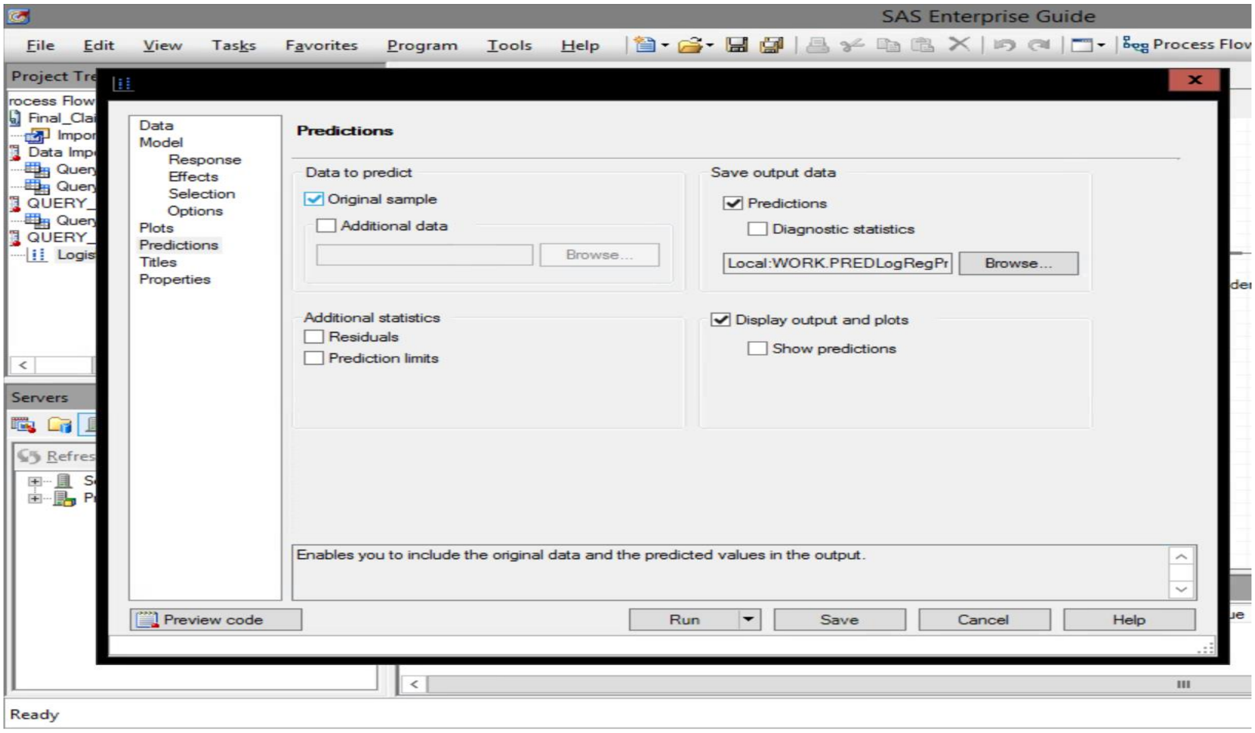
Conditional odds ratios
☑ Wald
□ Profile likelihood

Confidence level: 95%

Enables you to select the model fit assessment statistics to include in the analysis.
Displays the Hosmer and Lemeshow goodness-of-fit test for the case of a binary response model.

FIGURE-5

6. Fig 6 shows the Predictions settings:



Predictions

Data to predict
☑ Original sample
□ Additional data

Save output data
☑ Predictions
□ Diagnostic statistics
Local:WORK.PREDLogRegPr

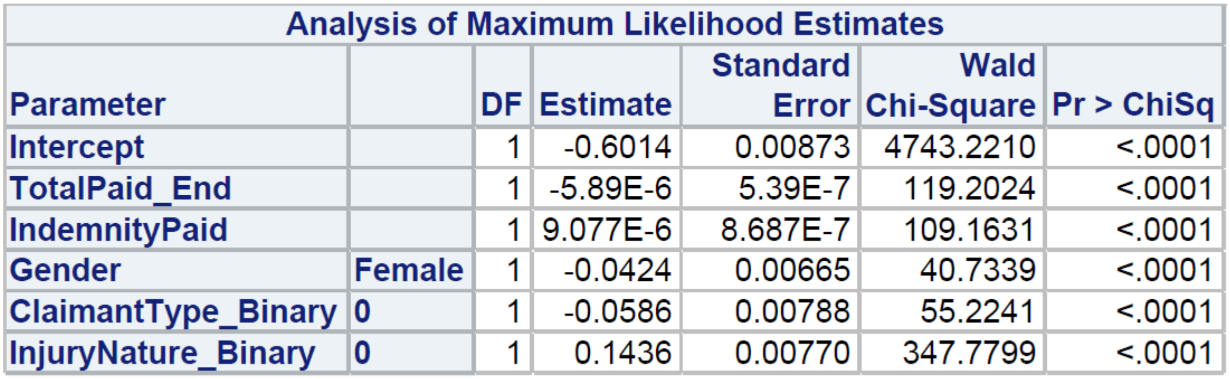
Additional statistics
□ Residuals
□ Prediction limits

☑ Display output and plots
□ Show predictions

Enables you to include the original data and the predicted values in the output.

FIGURE-6

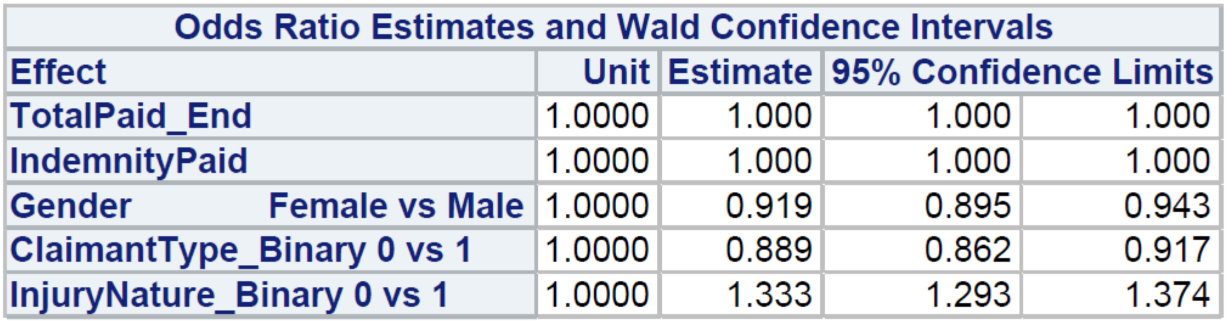
After making all the above settings, clicked on Run to generate logistic Regression Result. The final estimated coefficients are as shown below in Fig 7 and Fig 8:



Analysis of Maximum Likelihood Estimates

				Standard	Wald	
Parameter		DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept		1	-0.6014	0.00873	4743.2210	<.0001
TotalPaid_End		1	-5.89E-6	5.39E-7	119.2024	<.0001
IndemnityPaid		1	9.077E-6	8.687E-7	109.1631	<.0001
Gender	Female	1	-0.0424	0.00665	40.7339	<.0001
ClaimantType_Binary	0	1	-0.0586	0.00788	55.2241	<.0001
InjuryNature_Binary	0	1	0.1436	0.00770	347.7799	<.0001

Fig 7



Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
TotalPaid_End	1.0000	1.000	1.000	1.000
IndemnityPaid	1.0000	1.000	1.000	1.000
Gender Female vs Male	1.0000	0.919	0.895	0.943
ClaimantType_Binary 0 vs 1	1.0000	0.889	0.862	0.917
InjuryNature_Binary 0 vs 1	1.0000	1.333	1.293	1.374

Fig 8

From the above results of the logistic regression estimation, below are the interpretations of each independent variable on odds of High-Risk Claims:

- The unit increases in “TotalPaid_End” and “IndemnityPaid” has no effect on odds of High-Risk Claims as the odd estimate is 1.0000.
- For Gender, the odds estimate is 0.919 which means if gender is “Female” then the odds of claim to be a high-risk claim has 0.919 times lower odds compared to the Male.
- For Claimant Type, the odds estimate is 0.889 which means if claimant type is other than indemnity then the odds of claim to be a high-risk claim has 0.889 times lower odds compared to Claimant type as “Indemnity”.
- For Injury nature, the odds estimate is 1.333 which means if injury nature is other than strain then the odds of claim to be a high-risk claim is 1.333 times odds higher compared to “Strain”.

APPENDIX D - RECOMMENDATIONS

Based on our initial and final analysis below are the strategic recommendations we want to make to the claim processing organization:-

1) Initially we compared the Injury Type across Years, gender and total amount paid and discovered that Strain was the most high occurring injury type and almost equally distributed across genders. This calls for a regulation in the workers environment which the Insurance Company must take a firm decision on, along with the Employer. Solutions such as ergonomically sound seats, better work shifts and better nutrition/workout plans by HR etc. will help in reducing the claims.

2) From the logistic regression results in Appendix C Figure-8 suggests that odds of processing days to be higher is higher in claims of claimant type Indemnity. We recommend that in order to reduce the processing time there should be a separate workflow assigned for claimant type Indemnity to ensure few resources are directed completely to this workflow will promote faster processing time.

3) In the process of our analysis we have come across claim that have many null and invalid values which distort to get a clear picture about the injury nature and injury cause of the claims as depicted in Figure -4 of Appendix B. We recommend the organization to have mandatory fields to put valid values or predefined values to get a valid value. This will facilitate the analysis further to take strategic business decisions.

Analytics 3.0 world is an environment that combines the best of 1.0 and 2.0—a blend of big data and traditional analytics that yields insights and offerings with speed and impact. Few ways which the claim processing organization can become an Analytics 3.0 competitor are as below:

1) The company must position analytics with strategic priorities, insure leadership is advocating analytics and leading by example, and creating a Chief Analytics Officer (or equivalent role) to oversee the strategic deployment of analytics.

2) Invest in the technology needed to manage big data and provide insights quickly. This includes technologies like Hadoop, in-memory and in- database analytics and enough computing power to handle the complex calculations. In addition, it needs appropriate tools to effectively support decision making such as mobile and self-serve analytical applications.

3) Recruit, retain and effectively leverage analytical talent such as data scientists as they are a critical element of the shift to an Analytics 3.0 world. They have the skills to extract and structure the complex, high volume data sets in use by organizations. However, they need to work closely with IT and traditional quantitative analysts to develop insights for the business.

4) Insure that analytics groups are aligned with the business and focusing on critical business questions and decisions.