LA CITY WORKER
COMPENSATION ANALYSIS





DANNIEL WINARTO

JIAYING GU

ZOE HUANG

BAILI LU

RINI MUKHERJEE

Agenda

- Project Overview
- Exploratory Analysis
- Joining Datasets & Cleaning data
- Model Building & Testing
- Insights & Conclusions

Project Overview & Objective

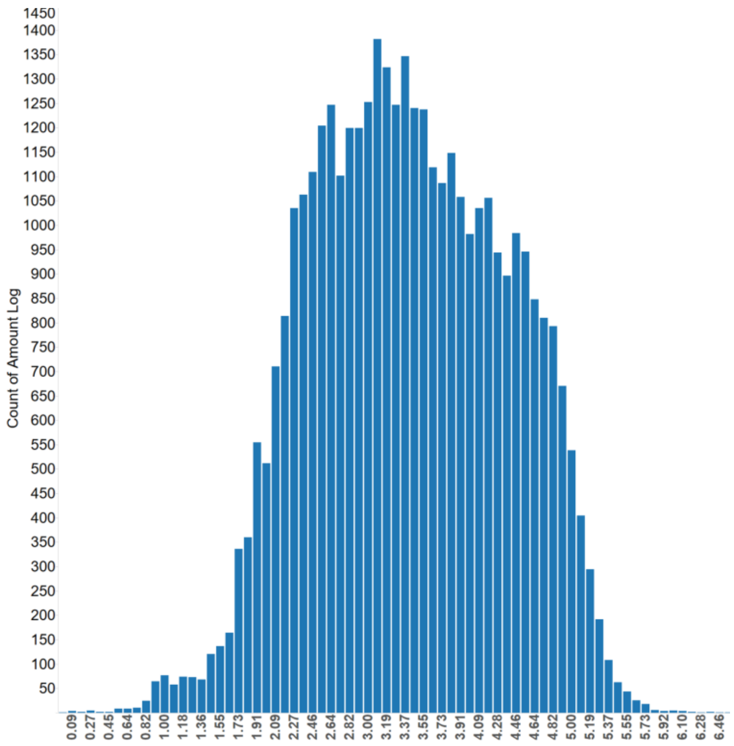
Overview:

1. Dataset contains 3 years record of claimant of LA workers.
2. Entire Dataset contains 39 different Excel files.
3. Response Variable is the claim amount.

Objective:

1. Figure out high risk factors and their patterns.
2. Build a predictive model.

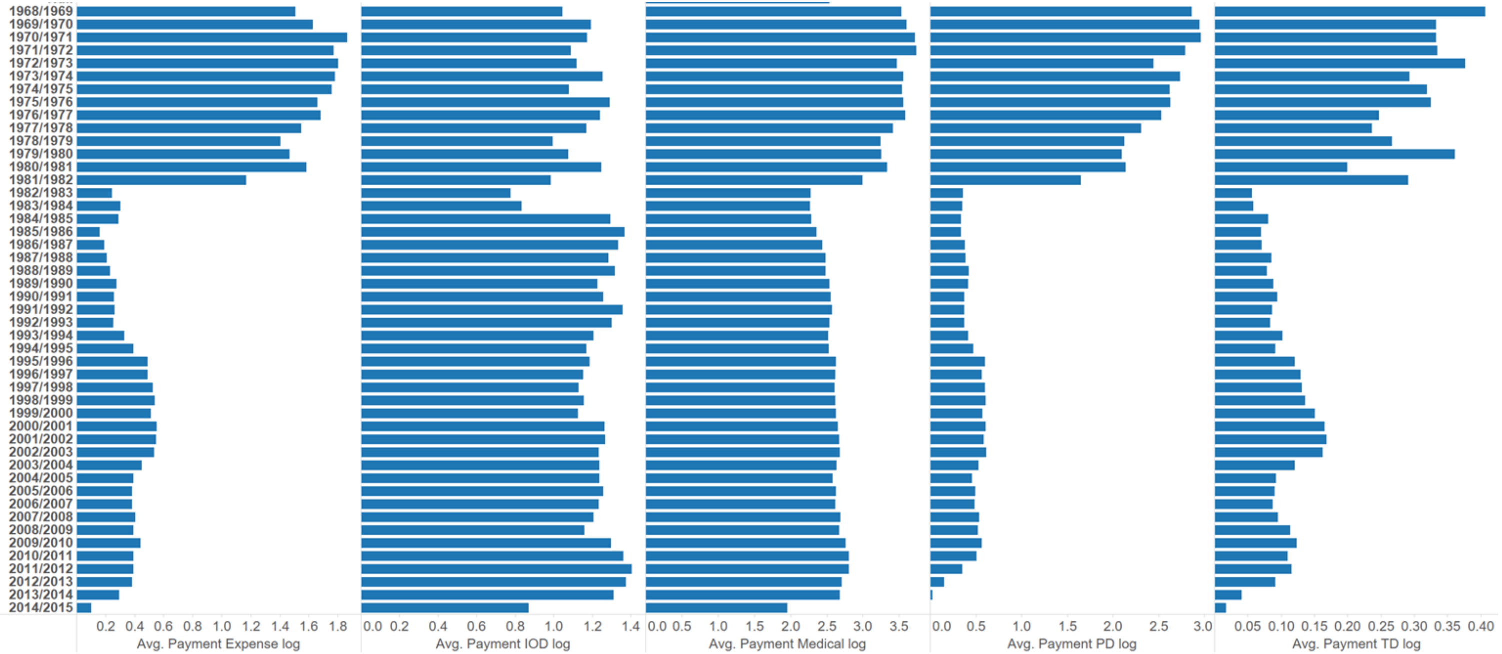
Exploratory Analysis



Histogram of Log Amount

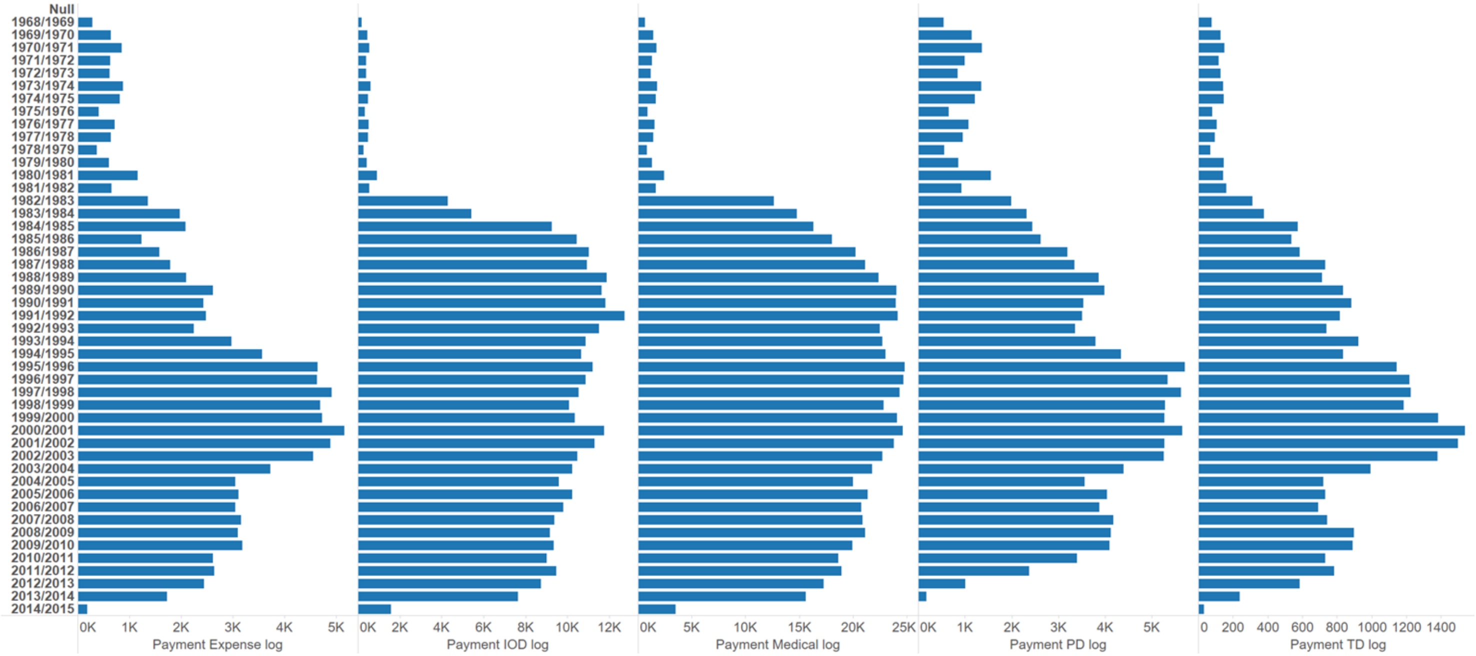
Exploratory Analysis

Average Amount for each Fiscal Year



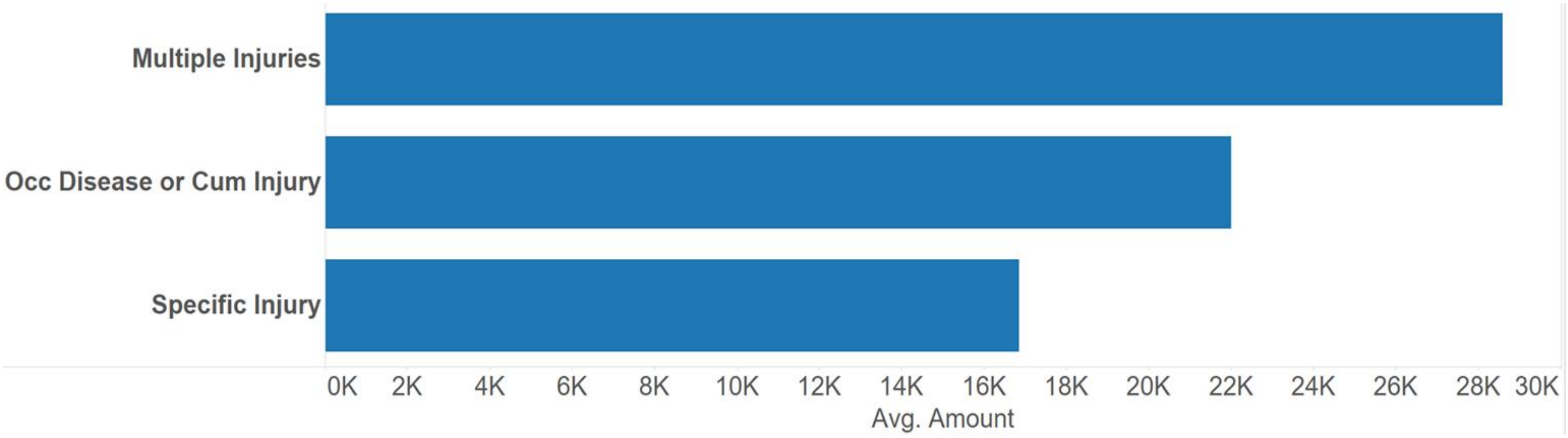
Exploratory Analysis

Total Amount for each Fiscal Year

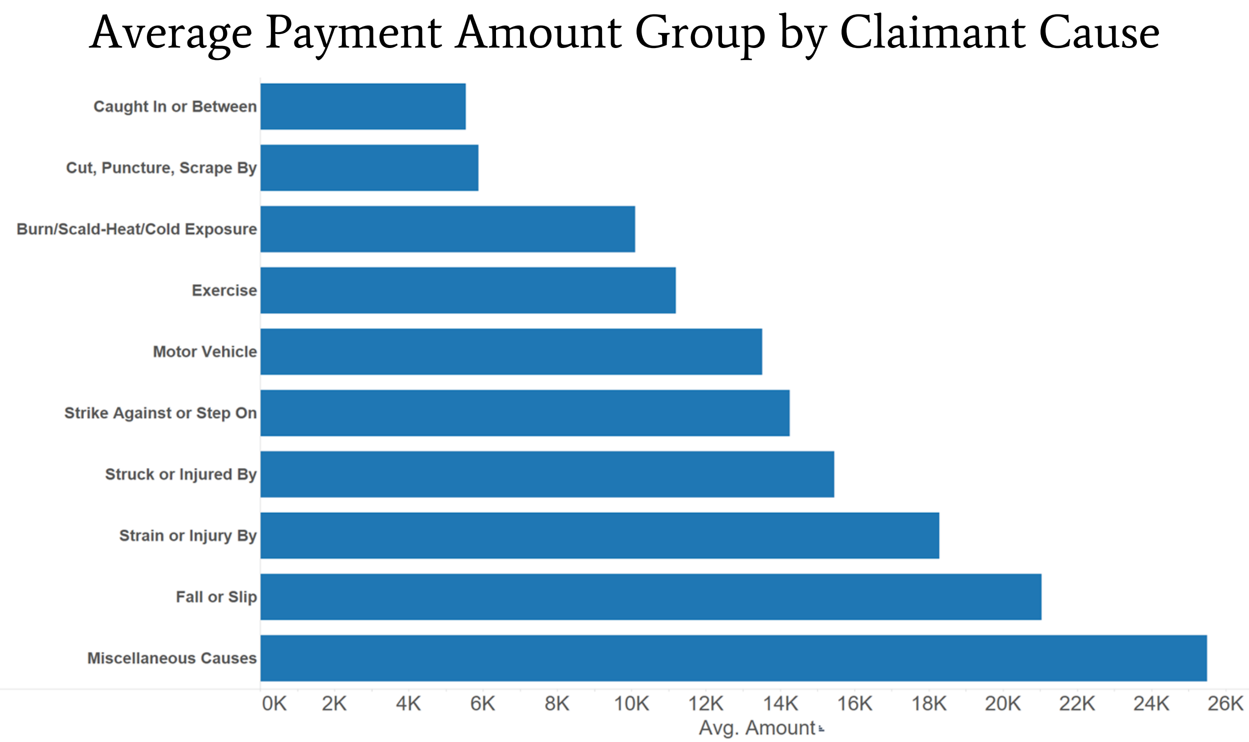


Exploratory Analysis

Average Payment Amount Group by Nature of Injury

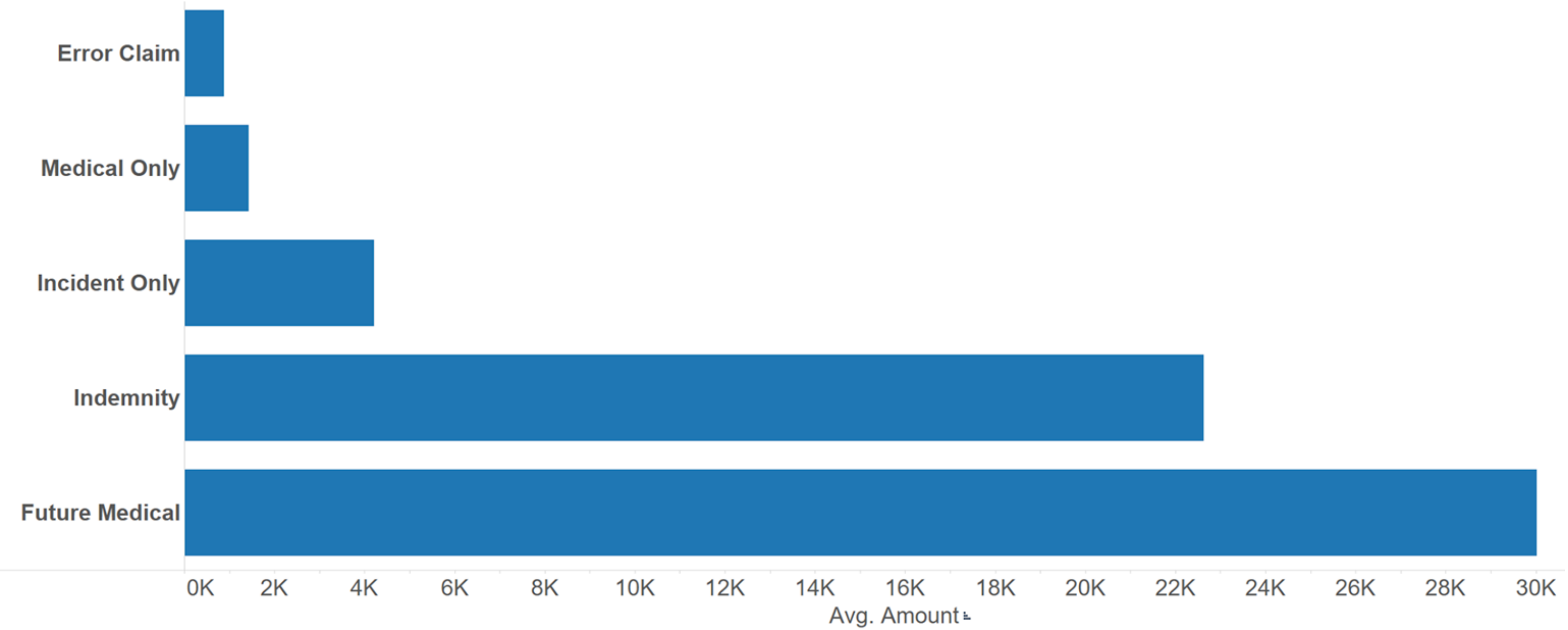


Exploratory Analysis



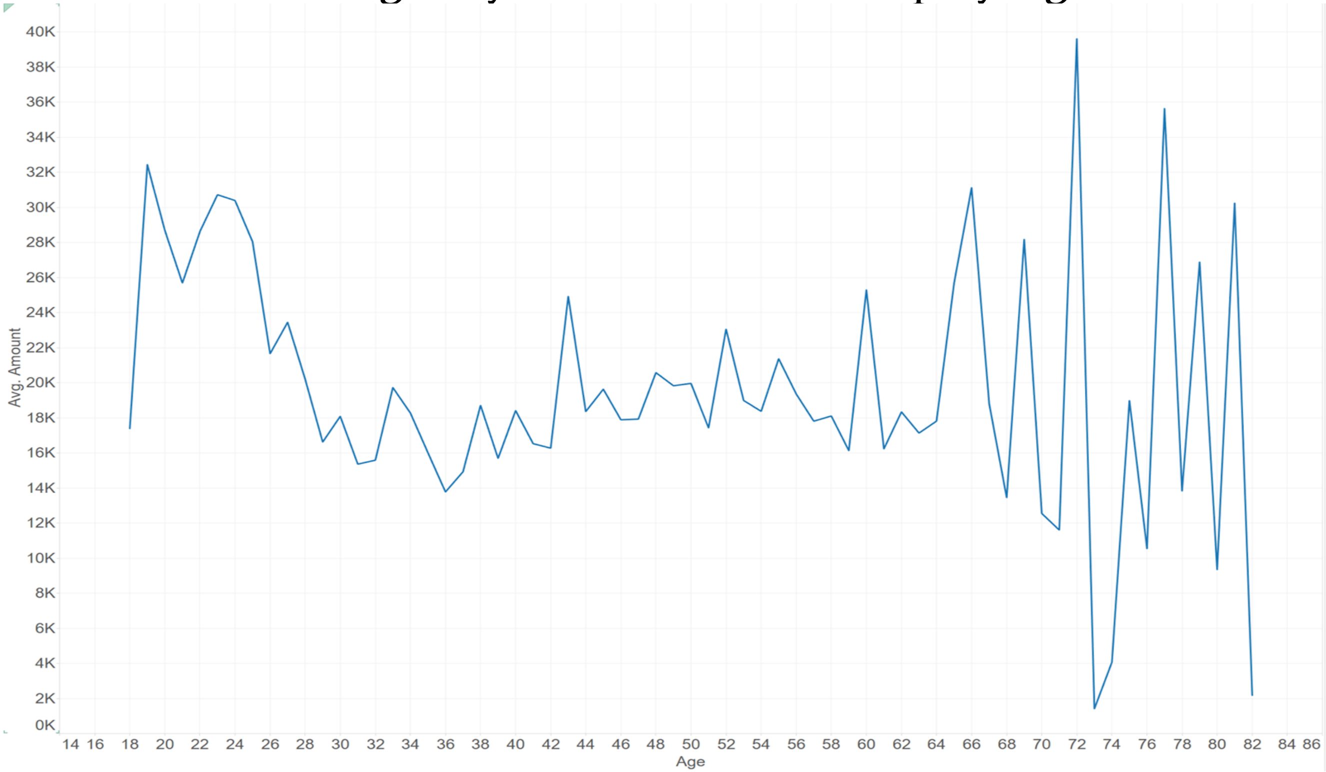
Exploratory Analysis

Average Payment Amount Group by Claimant Type

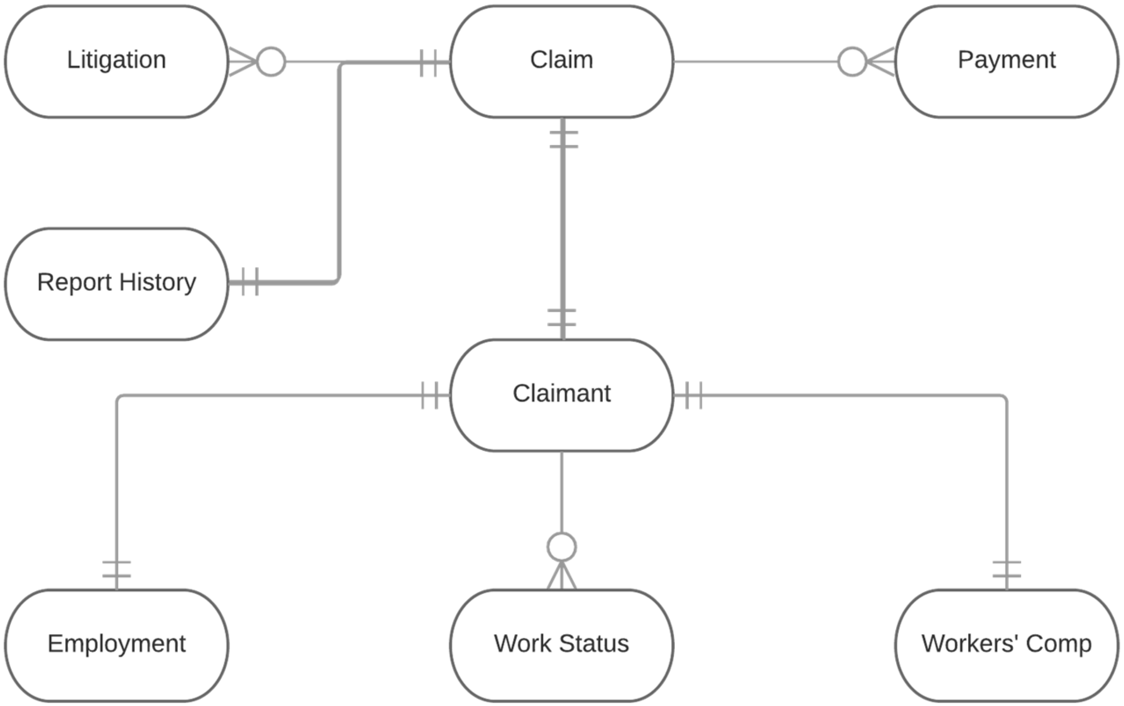


Exploratory Analysis

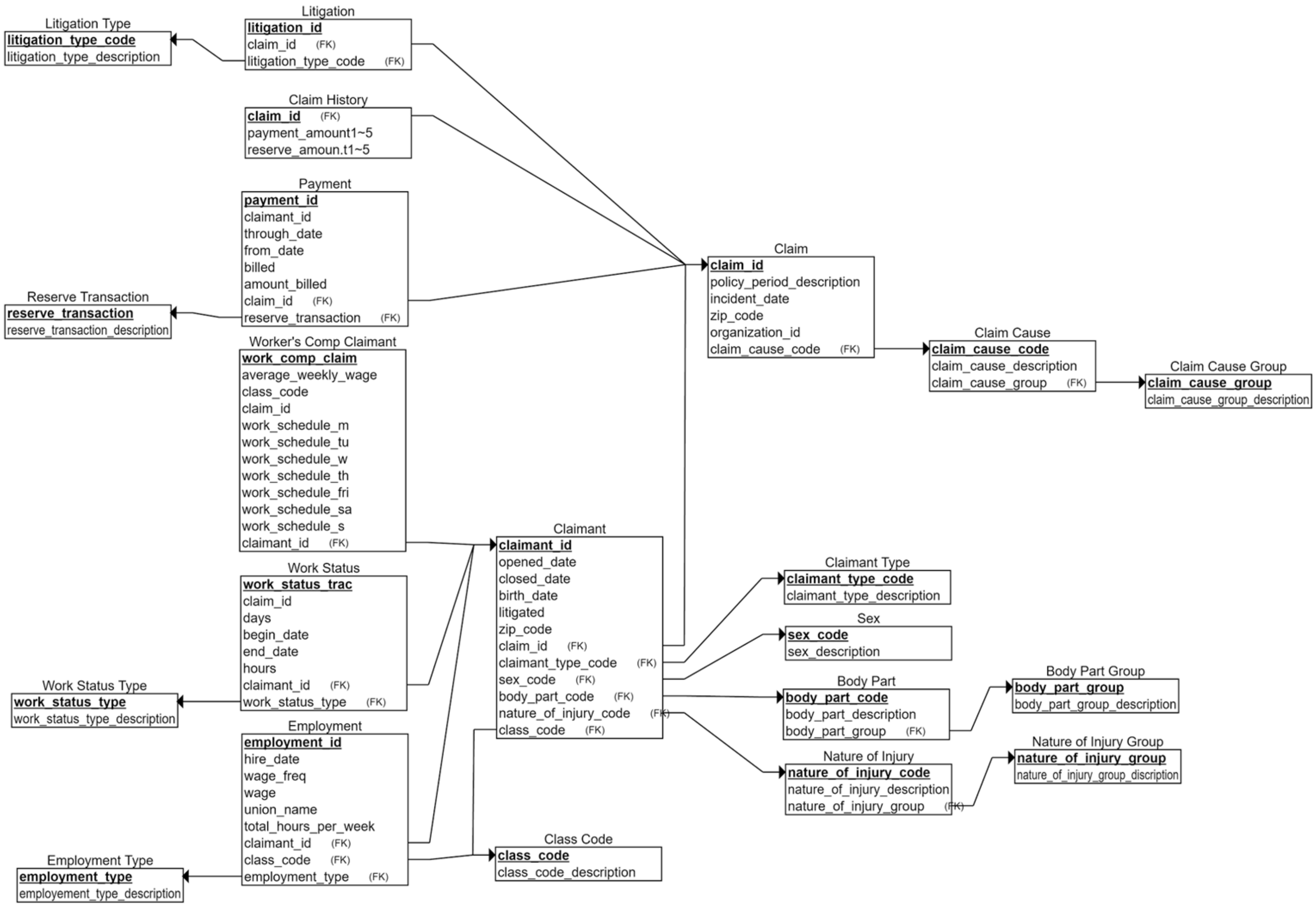
Average Payment Amount Group by Age



Joining Datasets



Joining Datasets



Datasets Joining

– Aggregate datasets to avoid one - many relationship

Payment_id	claim_id	amount_billed	billed	payment_amount
2189111	2	121.5	1	121.5
1874963	2	52.99	1	52.99
3970918	2	14.89	1	14.89
1770126	2	8.8	1	8.8
4495627	2	142.64	1	142.64
183153	3	148.55	1	148.55
498040	3	42.75	1	42.75
5956982	3	138.99	1	138.99
6167279	3	39.29	1	39.29
5957071	3	138.99	1	138.99
5746063	3	6.67	1	6.67
288665	3	6.67	1	6.67
1799068	6	61.38	0	61.38
5219360	6	30.34	1	30.34
2008709	6	61.38	0	61.38
1799085	6	61.38	0	61.38

Payment data is important to our analysis.

We combined the past 3 years payment data together to form a large data set.

For each claim, there might be several payment. But all other datasets we chose can all be uniquely identified by claim_id. So we aggregate payment information by claim_id so it can easily join with all other datasets.

This also reduce the number of the observations from 300K to 30K.

Creating New Variables

Incident Date	Incident Month
	Incident Hour
Birth Date	Age
Incident Date - Hire Date	Hire Year

Data Cleaning

More than 100 variables:

- Delete missing percent > 90%

More than 2400 levels:

- Reduce level

Variable	n.non.miss	n.miss	n.miss.percent	n.unique
claim_id	38712	0	0	38712
claimant_zip_code	38711	1	0	$\boxed{2462}$
organization_id	38711	1	0	1933
occupation_code	38712	0	0	640
org3_code	38332	380	0.98	623
claim_zip_code	31663	7049	18.21	323
org2_code	38711	1	0	176
claim_cause_code	38712	0	0	82
nature_of_injury_code	38712	0	0	54
body_part_code	38712	0	0	51
fiscal_year_desc	38710	2	0.01	48
org1_code	38711	1	0	45
incident_hour	38712	0	0	24
incident_month	38712	0	0	12
claim_cause_group	38712	0	0	10
employment_type	10909	27803	$\boxed{71.82}$	8
body_part_group_code	38712	0	0	6
claimant_type_code	38712	0	0	5
claimant_status_code	38712	0	0	3
sex_code	38712	0	0	3
work_schedule_fri	38711	1	0	3

Data Cleaning

Categorical Variables Level Reduction methodology:

```
# 0%        20%        40%        60%        80%          100%
#   1.000    370.474   1393.350   4972.160   23140.756 3747884.280
```

1. Zip Code

   Splitted based on the quantile of claim amount, into 5 groups

2. Incident Hour

   Splitted every 4 hours into 5 groups

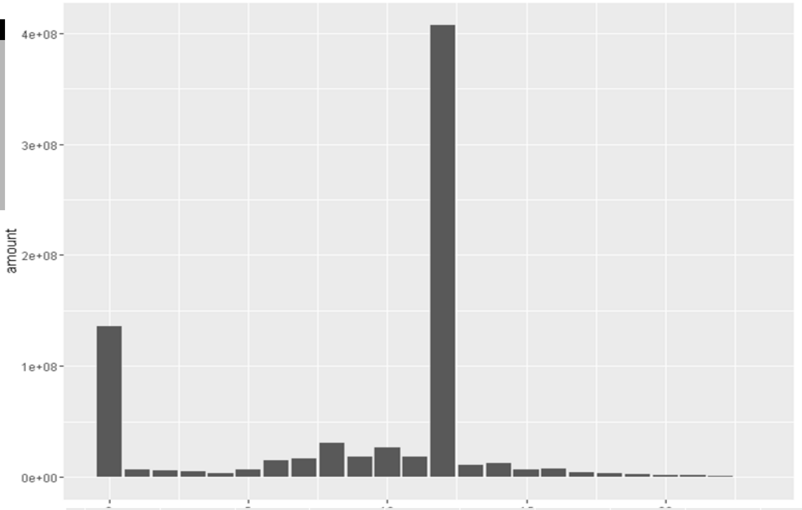
   “Midnight - 5AM”, “6AM - 11AM”, “Noon - 5PM”, 7PM - Midnight”

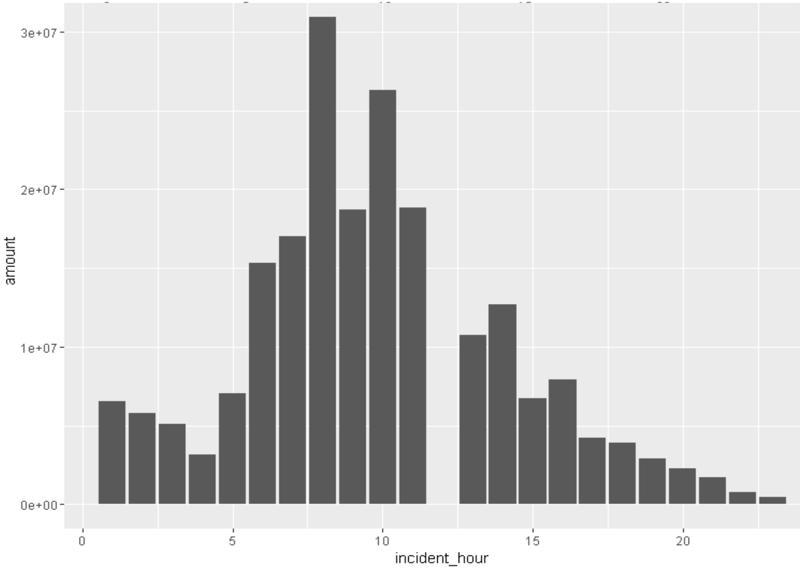
3. Occupation group

   Splitted based on the quantile claim amount, into 5 groups

4. Organization Code group

   Splitted based on the quantile claim amount, into 5 groups





Data Cleaning

Class Level Information		
Class	**Levels**	**Values**
litigated	2	0 1
nature_of_injury_gro	3	Multiple Injuries Occ Disease or Cum Injury Specific Injury
X.fiscalYearGroup	6	2010/2011 2011/2012 2012/2013 2013/2014 2014/2015 pre2010
zipGroup	5	z1 z2 z3 z4 z5
claimant_type	5	Error Claim Future Medical Incident Only Indemnity Medical Only

Building Model

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	72184	4511.48143	15309.6	<.0001
Error	25583	7538.86375	0.29468		
Corrected Total	25599	79723			

Root MSE	0.54285
Dependent Mean	7.26792
R-Square	0.9054
Adj R-Sq	0.9054
AIC	-5660.53445
AICC	-5660.50771
SBC	-31124

We split the dataset:
80 % training
20% testing

We integrate the levels of categorical variables

(litigated, nature of injury, fiscal year, zip code and claimant type)

The predictive model is significant

The Adjusted R-Squared is 0.9054

Building Model

- All the levels of litigated, nature of injury, fiscal year and zip code group are significant

- Only part of the levels of claimant type are significant

Parameter Estimates

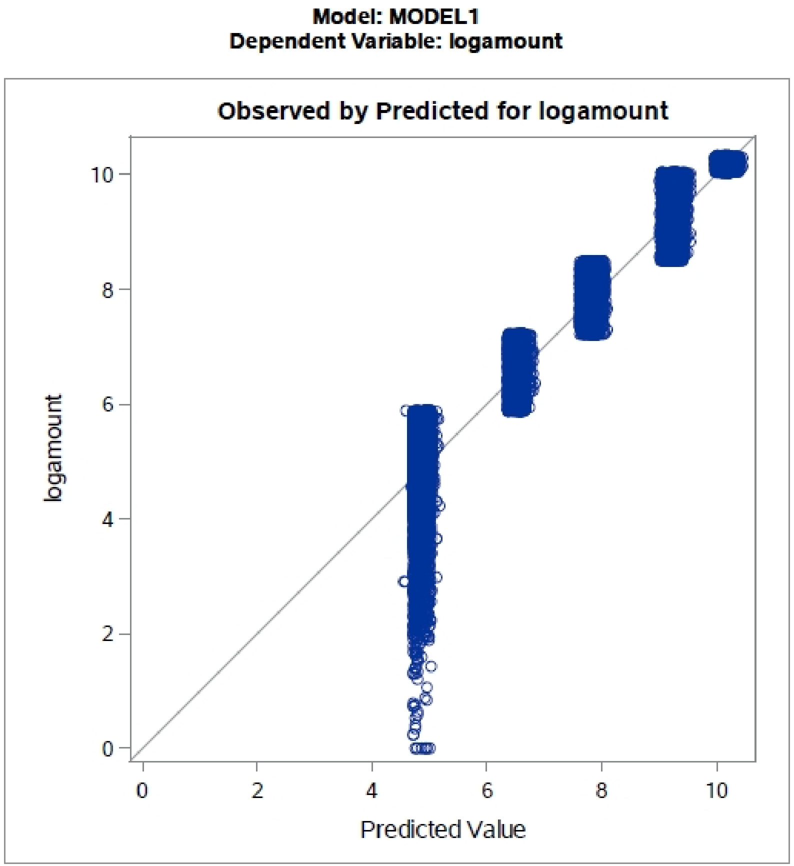
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	10.088614	0.021091	478.33	<.0001
litigated	0	1	-0.119873	0.009601	-12.49	<.0001
litigated	1	0	0	.	.	.
nature_of_injury_gro	Multiple Injuries	1	0.034685	0.009370	3.70	0.0002
nature_of_injury_gro	Occ Disease or Cum Injury	1	0.011677	0.011924	0.98	0.3275
nature_of_injury_gro	Specific Injury	0	0	.	.	.
X.fiscalYearGroup	2010/2011	1	-0.034207	0.014676	-2.33	0.0198
X.fiscalYearGroup	2011/2012	1	0.165375	0.011728	14.10	<.0001
X.fiscalYearGroup	2012/2013	1	0.212269	0.011692	18.16	<.0001
X.fiscalYearGroup	2013/2014	1	0.202276	0.012059	16.77	<.0001
X.fiscalYearGroup	2014/2015	1	0.178403	0.018673	9.55	<.0001
X.fiscalYearGroup	pre2010	0	0	.	.	.
zipGroup	z1	1	-5.217934	0.019214	-271.57	<.0001
zipGroup	z2	1	-3.579806	0.018955	-188.86	<.0001
zipGroup	z3	1	-2.314464	0.018704	-123.74	<.0001
zipGroup	z4	1	-0.922768	0.018473	-49.95	<.0001
zipGroup	z5	0	0	.	.	.
claimant_type	Error Claim	1	-0.188751	0.171867	-1.10	0.2721
claimant_type	Future Medical	1	0.120457	0.012606	9.56	<.0001
claimant_type	Incident Only	1	-0.120265	0.313582	-0.38	0.7013

Parameter Estimates

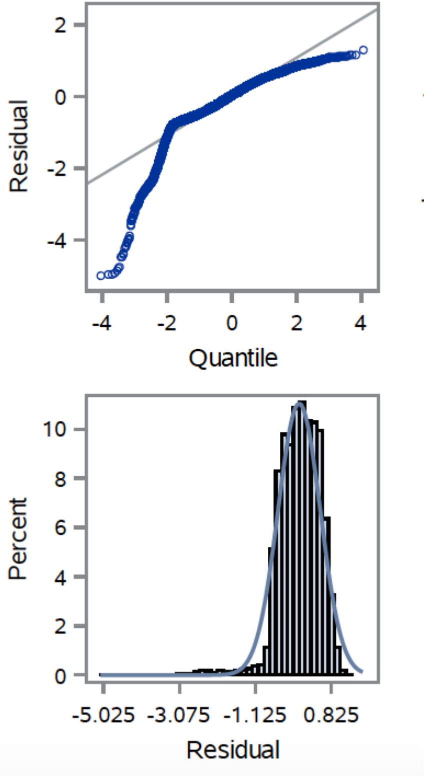
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
claimant_type	Indemnity	1	0.039629	0.008925	4.44	<.0001
claimant_type	Medical Only	0	0	.	.	.

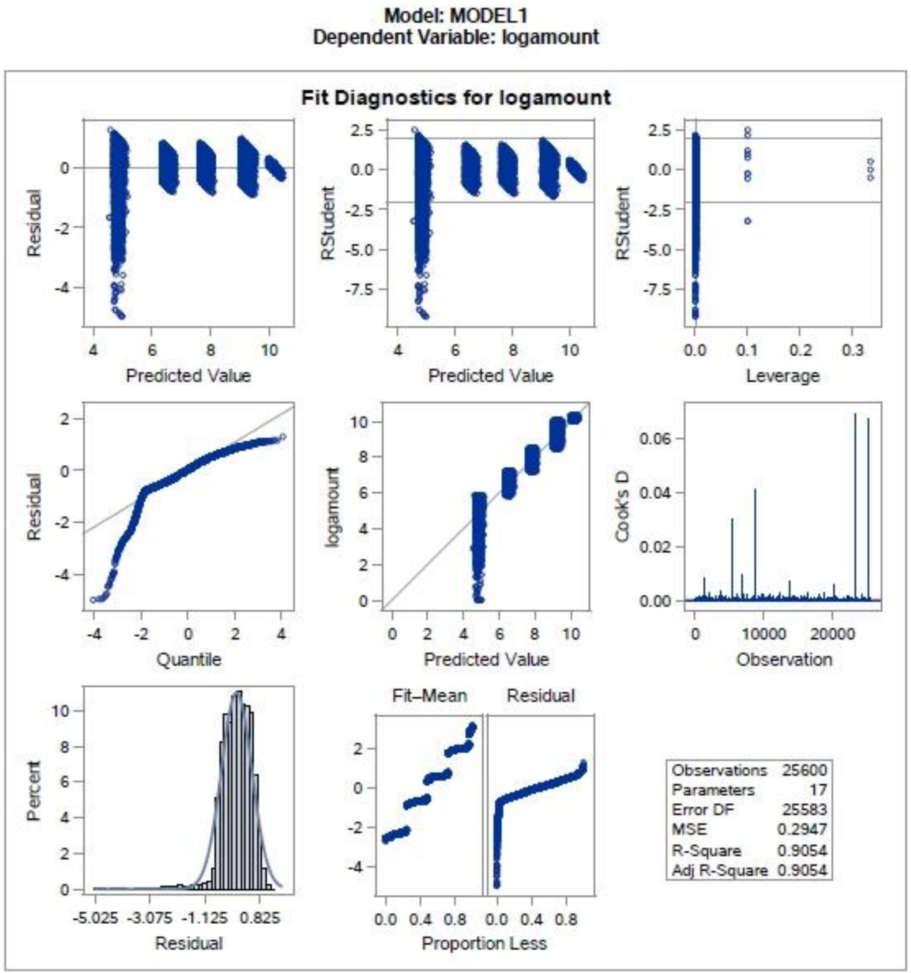
Building Model

Since the trend of many variables is seriously right-skewed, we do the log transformation for the dependent variable and fit the linear model



Building Model





Building Model

Log.Amount =
10.089 +
C1 x Litigated[0/1] +
C2 x Nature of Injury Group +
C3 x Fiscal Year Group +
C4 x Zip Group+
C5 x Claimant Type

Where:
C1 = [-.120, 0]
C2 = [.035,.012, 0]
C3 = [-.034, .165, .212, .202, .178, 0]
C4 = [-5.218, -3.580, -2.314, -.923, 0]
C5 = [-.189, .120, -.120]

Parameter Estimates

Parameter		DF	Estimate
Intercept		1	10.088614
litigated	0	1	-0.119873
litigated	1	0	0
nature_of_injury_gro	Multiple Injuries	1	0.034685
nature_of_injury_gro	Occ Disease or Cum Injury	1	0.011677
nature_of_injury_gro	Specific Injury	0	0
X.fiscalYearGroup	2010/2011	1	-0.034207
X.fiscalYearGroup	2011/2012	1	0.165375
X.fiscalYearGroup	2012/2013	1	0.212269
X.fiscalYearGroup	2013/2014	1	0.202276
X.fiscalYearGroup	2014/2015	1	0.178403
X.fiscalYearGroup	pre2010	0	0
zipGroup	z1	1	-5.217934
zipGroup	z2	1	-3.579806
zipGroup	z3	1	-2.314464
zipGroup	z4	1	-0.922768
zipGroup	z5	0	0
claimant_type	Error Claim	1	-0.188751
claimant_type	Future Medical	1	0.120457
claimant_type	Incident Only	1	-0.120265

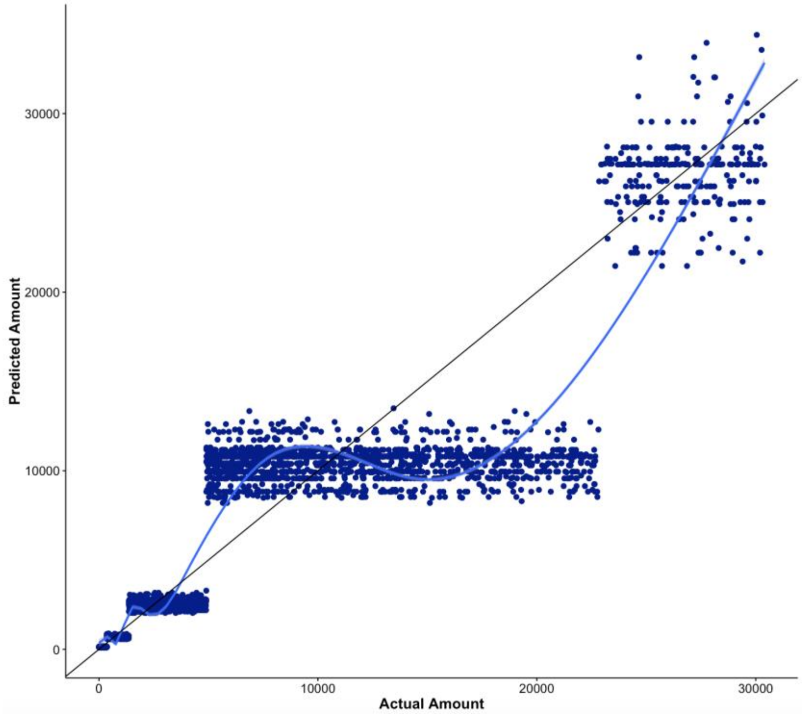
Parameter Estimates

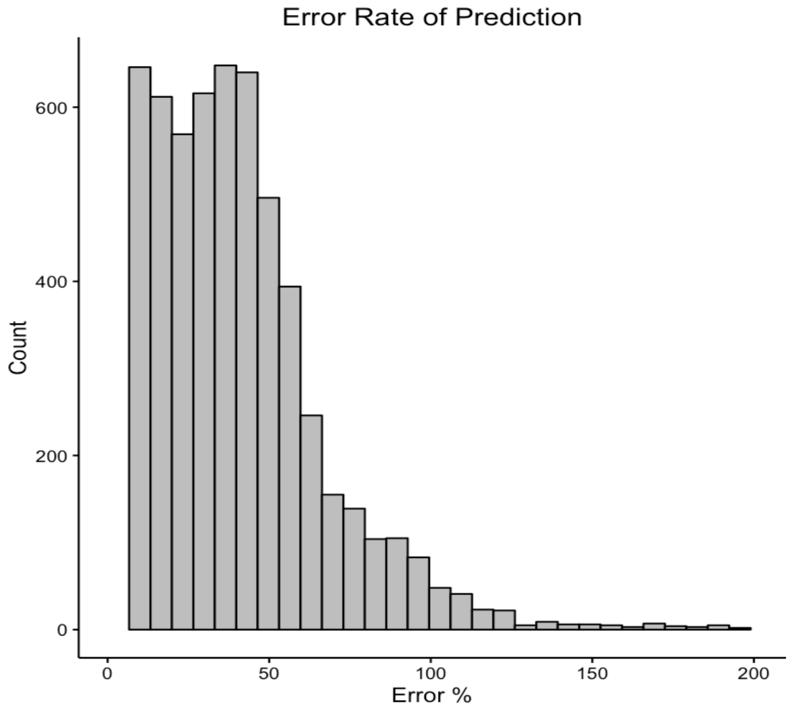
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
claimant_type	Indemnity	1	0.039629	0.008925	4.44	<.0001
claimant_type	Medical Only	0	0	.	.	.

Model Prediction

> THE MODEL WELL PREDICT THE TESTING DATA

> Most of our predicted results are off within 50% of the actual data, meaning that our model can capture the true values.





Business Insight

List of Importance (stepwise selection) :

Zip Group
Fiscal Year Group
Litigated
Claimant Type
Nature of Injury Group

USC Marshall

Business Insight - Zip Code

Zip Group is the most significant variable, we found out by applying stepwise regression methodology.

This finding is intuitively make sense, since some areas may have condition(eg. Road quality, rainfall intensity, average income) that increase chance of the worker getting injured (Based on the assumption that generally people tend to live close where they work).

Example of most freq in each zip group:

Z1 = 91350 -> Santa Clarita
Z2 = 91342 -> Sylmar
Z3 = 93065 -> Simi Valley
Z4 = 93551 -> Palmdale
Z5 = 91709 -> Chino hills

	claimant_zip_code (int)	count (int)
1	91350	120
2	91709	98
3	93065	95

zipGroup	z1	1	-5.217934
zipGroup	z2	1	-3.579806
zipGroup	z3	1	-2.314464
zipGroup	z4	1	-0.922768
zipGroup	z5	0	0

We should pay attention more on this area to reduce the claimant frequency in the future

Business Insight - Fiscal Year

Claims that happened in fiscal year 2012/2013 have the highest amount.

All those that happened before 2011 are lower than after 2011.

After 2013, the average amount per claim decreased. 2013-2014 fiscal year Is higher than 2014-2015 fiscal year.

X.fiscalYearGroup	2010/2011	1	-0.034207	0.014676	-2.33	0.0198
X.fiscalYearGroup	2011/2012	1	0.165375	0.011728	14.10	<.0001
X.fiscalYearGroup	2012/2013	1	0.212269	0.011692	18.16	<.0001
X.fiscalYearGroup	2013/2014	1	0.202276	0.012059	16.77	<.0001
X.fiscalYearGroup	2014/2015	1	0.178403	0.018673	9.55	<.0001
X.fiscalYearGroup	pre2010	0	0	.	.	.

USC Marshall

Business Insight - Litigated

Claims that are litigated have higher amount than those that are not litigated.

Meaning that in order to cut cost, reduce the number of litigated claims would help.

litigated	0	1	-0.119873	0.009601	-12.49	<.0001
litigated	1	0	0	.	.	.

USC Marshall

Business Insight - Claimant Type

Error claims have the lowest amount, and future medical claims have the highest.

Incident only claims are lower than medical only.

In order to cut cost, special attention should be applied to future medical claims.

claimant_type	Indemnity	1	0.039629	0.008925	4.44	<.0001
claimant_type	Medical Only	0	0	.	.	.
claimant_type	Error Claim	1	-0.188751	0.171867	-1.10	0.2721
claimant_type	Future Medical	1	0.120457	0.012606	9.56	<.0001
claimant_type	Incident Only	1	-0.120265	0.313582	-0.38	0.7013

Business Insight - Nature of Injury

Multiple injuries claims have the highest average amount,
occ disease or cum injury comes next and the lowest is specific injury.

Cost could be reduced by paying attention to claims of multiple injuries.

nature_of_injury_gro Multiple Injuries	1	0.034685	0.009370	3.70	0.0002
nature_of_injury_gro Occ Disease or Cum Injury	1	0.011677	0.011924	0.98	0.3275
nature_of_injury_gro Specific Injury	0	0	.	.	.

USC Marshall

Suggestions & Takeaways

1. Since our model suggests that Zip Group, Fiscal Year Group ,Litigated ,Claimant Type, Nature of Injury Group are risky factors for claim amount. LA City could cut cost by focusing on these.
2. There are other variables that we considered to put into our model, but because of existence of missing values or wrong values, we excluded them from our model. If these variables are maintained better in the future, they can be contributed to the model. For example, employment type.
3. If LA City would like to obtain more accurate predictive results, we will recommend to gather more numerical data in the future.

USC Marshall

THANK YOU
&
QUESTIONS ARE WELCOME

DANNIEL
WINARTO

JIAYING
GU

ZOE
HUANG

BAILI
LU

RINI
MUKHERJEE