Final Project: Data Analysis and Visualization on Insurance Fraud DataSets

Abu, Sayed, Saheli

12/11/2022

Abstract

In this project, we analyzed the datasets named **Claim**, **Demographic**, **Policy**, **Vehicle** and **Fraud** from the datasets of **Insurance Fraud** and implemented the tasks that were asked as the deliverable for the final project. We chose these data sets to analyze because we felt that we can apply the techniques which were learned from the activites throughout the semester and we were finally able to apply those techniques in this project.

Background

For the analysis report, we created 5 distinct type of ggplot visualizations which are mixture of visualizing quantitative and categorical variables, a couple of tables of summary statistics which were obtained using group-wise operations. To do these, we merged the necessary datasets. In addition, the claim dataset was pivoted from wide format to long format to analize the dmage claim amount. The date time variable of our dataset was manipulated by using lubridate functions and one string variable was manipulated by using stringr functions as per the deliverable of this project. We implemented permutation test based on a traditional hypothesis test to see if the average amount of vehicle damage differs by insured gender using an F test. Apart from that, we obtained a parametric and nonparametric bootstrap to estimate standard errors and distribution for the sample median. Finally, we created a dictionary showcasing the variables that were used in our analysis. Apart from the analysis report, we created a dashboard using the flexdashboard package showing few visualizations and summary table as a separate deliverable of this project.

Packages

We used the below packages for our analysis:

```
library(data.table)
library(tidyverse)
library(skimr)
library(ggthemes)
library(broom)
library(boot)
library(dataMeta)
library(stringr)
library(lubridate)
```

Summary of Variables and Exploratory Data Analysis

The dataset depicts the details of the auto insurance claim, like, the fraud details, policy details, vehicle details, claim details and also the customer demographic information etc. But, we used only claim details and the customer demographic information to fulfill the deliverables for this project. The information about attributes are mentioned below:

Claim Information :

• CustomerID : Customer ID

• DateOfIncident : Date of incident

• TypeOfIncident : Type of incident

• TypeOfCollission : Type of Collision - "?" is the missing value

• SeverityOfIncident : Collision severity

• AuthoritiesContacted : Which authorities are contacted

• IncidentState : Incident location (State)

• IncidentCity : Incident location (City)

• IncidentAddress : Incident location (address)

• IncidentTime : time of incident – Hour of the day - the missing value is represented as "-5"

• NumberOfVehicles : Number of vehicles involved

• PropertyDamage : If property damage is there - "?" is the missing value

• BodilyInjuries : Number of bodily injuries

• Witnesses : Number of witnesses - missing value is represented as "MISSINGVALUE"

• PoliceReport : If police report available - "?" is the missing value

• AmountOfTotalClaim : Total claim amount - the missing value is represented as "MISSEDDATA"

• AmountOfInjuryClaim : Claim for injury

• AmountOfPropertyClaim : claim for property damage

• AmountOfVehicleDamage : claim for vehicle damage

The dataset indicates missing data with **"?"**, **"-5"**, **"MISSINGVALUE"**, **"MISSEDDATA"**. We replaced those with **"NA"**.

```
## Rows: 28,836
## Columns: 19
## $ CustomerID            <chr> "Cust10000", "Cust10001", "Cust10002", "Cust1000…
## $ DateOfIncident        <date> 2015-02-03, 2015-02-02, 2015-01-15, 2015-01-19,…
## $ TypeOfIncident        <chr> "Multi-vehicle Collision", "Multi-vehicle Collis…
## $ TypeOfCollission      <chr> "Side Collision", "Side Collision", "Side Collis…
## $ SeverityOfIncident    <chr> "Total Loss", "Total Loss", "Minor Damage", "Min…
## $ AuthoritiesContacted  <chr> "Police", "Police", "Other", "Other", "Fire", "F…
## $ IncidentState         <chr> "State7", "State7", "State8", "State9", "State8"…
## $ IncidentCity          <chr> "City1", "City5", "City6", "City6", "City6", "Ci…
## $ IncidentAddress       <chr> "Location 1311", "Location 1311", "Location 2081…
## $ IncidentTime          <int> 17, 10, 22, 22, 10, 7, 20, 18, 3, 5, 14, 16, 15,…
## $ NumberOfVehicles      <int> 3, 3, 1, 1, 1, 1, 1, 1, 3, 1, 1, 3, 3, 3, 1, 1, …
## $ PropertyDamage        <chr> NA, "YES", "YES", "YES", "NO", "NO", NA, NA, "YE…
## $ BodilyInjuries        <int> 1, 2, 2, 2, 2, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, …
## $ Witnesses             <chr> "0", "1", "3", "3", "1", "2", "2", "2", "0", "2"…
## $ PoliceReport          <chr> NA, "YES", "NO", "NO", "YES", NA, "NO", "NO", NA…
## $ AmountOfTotalClaim    <dbl> 65501, 61382, 66755, 66243, 53544, 53167, 77453,…
## $ AmountOfInjuryClaim   <int> 13417, 15560, 11630, 12003, 8829, 7818, 6476, 57…
## $ AmountOfPropertyClaim <int> 6071, 5919, 11630, 12003, 7234, 8132, 12822, 733…
## $ AmountOfVehicleDamage <int> 46013, 39903, 43495, 42237, 37481, 37217, 58155,…
```

Data summary

Name	claimRawData
Number of rows	28836
Number of columns	19
Key	NULL
_____	

Column type frequency:	
character	11
Date	1
numeric	7
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1.00	8	9	0	28836	0
TypeOfIncident	0	1.00	10	24	0	4	0
TypeOfCollission	5162	0.82	14	15	0	3	0
SeverityOfIncident	0	1.00	10	14	0	4	0
AuthoritiesContacted	0	1.00	4	9	0	5	0
IncidentState	0	1.00	6	6	0	7	0
IncidentCity	0	1.00	5	5	0	7	0
IncidentAddress	0	1.00	13	13	0	1000	0
PropertyDamage	10459	0.64	2	3	0	2	0
Witnesses	46	1.00	1	1	0	4	0
PoliceReport	9805	0.66	2	3	0	2	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
DateOfIncident	0	1	2015-01-01	2015-03-14	2015-01-30	72

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50
IncidentTime	0	1	11.75	6.17	-5	6.00	12.0
NumberOfVehicles	0	1	1.82	0.98	1	1.00	1.0
BodilyInjuries	0	1	0.99	0.78	0	0.00	1.0
AmountOfTotalClaim	50	1	52308.55	25101.17	150	44643.75	58360.0
AmountOfInjuryClaim	0	1	7337.12	4427.64	0	4743.75	7147.0
AmountOfPropertyClaim	0	1	7283.87	4375.84	0	4862.00	7051.0
AmountOfVehicleDamage	0	1	37687.13	17977.05	109	32193.25	42457.5

In Claim data set, arround **35%** data is missing for the variables named **PropertyDamage** and **PoliceReport**.

Demographics Data :

• CustomerID : Customer ID

• InsuredAge : age

• InsuredZipCode : Zip Code

• InsuredGender : Gender - the missing value is represented as “NA”

• InsuredEducationLevel : Education

• InsuredOccupation : Occupation

• InsuredHobbies : Hobbies

• CapitalGains : Capital gains(Financial Status)

• CapitalLoss : capital loss(Financial Status)

• Country : Country

```
## Rows: 28,836
## Columns: 10
## $ CustomerID            <chr> "Cust10000", "Cust10001", "Cust10002", "Cust1000…
## $ InsuredAge            <int> 35, 36, 33, 36, 29, 28, 57, 49, 27, 48, 41, 36, …
## $ InsuredZipCode        <int> 454776, 454776, 603260, 474848, 457942, 457942, …
## $ InsuredGender         <chr> "MALE", "MALE", "MALE", "MALE", "FEMALE", "FEMAL…
## $ InsuredEducationLevel <chr> "JD", "JD", "JD", "JD", "High School", "High Sch…
## $ InsuredOccupation     <chr> "armed-forces", "tech-support", "armed-forces", …
## $ InsuredHobbies        <chr> "movies", "cross-fit", "polo", "polo", "dancing"…
## $ CapitalGains          <int> 56700, 70600, 66400, 47900, 0, 0, 67400, 67400, …
## $ CapitalLoss           <int> -48500, -48500, -63700, -73400, -41500, -41500, …
## $ Country               <chr> "India", "India", "India", "India", "India", "In…
```

Data summary

Name	demographicRawData
Number of rows	28836
Number of columns	10
Key	NULL
_____	
Column type frequency:	
character	6
numeric	4
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1	8	9	0	28836	0
InsuredGender	30	1	4	6	0	2	0
InsuredEducationLevel	0	1	2	11	0	7	0
InsuredOccupation	0	1	5	17	0	14	0
InsuredHobbies	0	1	4	14	0	20	0
Country	0	1	0	5	2	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
InsuredAge	0	1	38.82	8.00	19	33	38	44
InsuredZipCode	0	1	502436.58	72250.87	430104	448603	466691	603
CapitalGains	0	1	23066.57	27637.81	0	0	0	490
CapitalLoss	0	1	-24940.61	27913.21	-111100	-50000	0	0

In the Demographic data set, only the variable named **InsuredGender** has 30 missing values which is only **0.1 %** of overall data.

Policy Information :

• CustomerID : Customer ID

• CustomerLoyaltyPeriod : Duration of customer relationship

• InsurancePolicyNumber : policy number

• DateOfPolicyCoverage : policy commencement date

• InsurancePolicyState : Policy location (State)

• Policy_CombinedSingleLimit : Split Limit and Combined Single Limit

• Policy_Deductible : Deductible amount

• PolicyAnnualPremium : Annual Premium – the missing value is represented as “-1”

• UmbrellaLimit : Umbrella Limit amount

• InsuredRelationship : Realtionship

• TotalCharges : Customer account information (Total). (For this attribute, missing values are denoted as “MISSINGVAL” also)

• DOE : Date of entry as customer

• ElectronicBilling : Customer account information - whether electronic billing

• ContractType : Contract type (For this attribute, missing values are denoted as “NA”)

• PaymentMethod : payment method

The dataset indicates missing data with **“-1”**, **“MISSINGVAL”**. We replaced those with **“NA”**.

```
## Rows: 28,836
## Columns: 10
## $ InsurancePolicyNumber      <int> 110122, 110125, 110126, 110127, 110128, 110…
## $ CustomerLoyaltyPeriod      <int> 328, 256, 228, 256, 137, 27, 212, 235, 447,…
## $ DateOfPolicyCoverage       <date> 2014-10-17, 1990-05-25, 2014-06-06, 2006-1…
## $ InsurancePolicyState       <chr> "State3", "State1", "State1", "State3", "St…
## $ Policy_CombinedSingleLimit <chr> "250/500", "250/500", "500/1000", "250/500"…
## $ Policy_Deductible          <int> 1000, 2000, 1000, 1000, 1000, 500, 500, 500…
## $ PolicyAnnualPremium        <dbl> 1406.91, 1415.74, 1583.91, 1351.10, 1333.35…
## $ UmbrellaLimit              <int> 0, 6000000, 6000000, 0, 0, 0, 0, 4000000, 0…
## $ InsuredRelationship        <chr> "husband", "unmarried", "unmarried", "unmar…
## $ CustomerID                 <chr> "Cust1001", "Cust1004", "Cust1005", "Cust10…
```

Data summary

Name	policyRawData
Number of rows	28836

Number of columns	10
Key	NULL
_____	
Column type frequency:	
character	4
Date	1
numeric	5
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
InsurancePolicyState	0	1	6	6	0	3	0
Policy_CombinedSingleLimit	0	1	7	8	0	9	0
InsuredRelationship	0	1	4	14	0	6	0
CustomerID	0	1	8	9	0	28836	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
DateOfPolicyCoverage	0	1	1990-01-07	2015-02-22	2001-11-27	6779

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25
InsurancePolicyNumber	0	1	129312.52	11114.06	110122	119698.75
CustomerLoyaltyPeriod	0	1	203.07	99.93	1	126.00
Policy_Deductible	0	1	1114.28	546.63	500	622.00
PolicyAnnualPremium	0	1	1255.53	223.01	-1	1122.01
UmbrellaLimit	0	1	983668.03	1969282.04	-1000000	0.00

After analyzing the Policy dataset, we noticed that there is no missing value though it was mentioned in the summary of attributes that the variable named "PolicyAnnualPremium" has the missing value which is represented as "-1".

Vehicle Data:

• CustomerID : Customer ID

• VehicleAttribute : Service signed for

• VehicleAttributeDetails : Value of the vehicle attribute - the missing value is represented as "???"

The dataset indicates missing data with **"???"**. We replaced those with **"NA"**.

```
## Rows: 115,344
## Columns: 3
```

```
## $ CustomerID             <chr> "Cust20179", "Cust21384", "Cust33335", "Cust27…
## $ VehicleAttribute       <chr> "VehicleID", "VehicleModel", "VehicleMake", "V…
## $ VehicleAttributeDetails <chr> "Vehicle8898", "Malibu", "Toyota", "Neon", "Ve…
```

Data summary

Name	vehicleRawData
Number of rows	115344
Number of columns	3
Key	NULL
_____	
Column type frequency:	
character	3
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1	8	9	0	28836	0
VehicleAttribute	0	1	9	12	0	4	0
VehicleAttributeDetails	50	1	2	14	0	28910	0

In the Vehicle data set, only the variable named **VehicleAttributeDetails** has 50 missing values which is only **0.1 %** of overall data.

Fraud Data :

• CustomerID : Customer ID

• ReportedFraud : Fraud or not – Target

```
## Rows: 28,836
## Columns: 2
## $ CustomerID    <chr> "Cust20065", "Cust37589", "Cust24312", "Cust5493", "Cust…
## $ ReportedFraud <chr> "N", "N", "N", "Y", "Y", "N", "Y", "N", "N", "N", "N", "…
```

Data summary

Name	fraudRawData
Number of rows	28836
Number of columns	2
Key	NULL
_____	
Column type frequency:	
character	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1	8	9	0	28836	0
ReportedFraud	0	1	1	1	0	2	0

There is no missing data in the Fraud dataset.

Data Dictionary

Let's create a data dictionary showcasing the variables used in our analysis:

```
##            variable_name            variable_description variable_type
## 1    AmountOfInjuryClaim      The Amount of Injury Claim           int
## 2  AmountOfPropertyClaim    The Amount of Property Claim           int
## 3     AmountOfTotalClaim       The Amount of Total Claim           chr
## 4  AmountOfVehicleDamage    The Amount of Vehicle Damage           int
## 5             CustomerID  Customet Identification Number           chr
## 6         DateOfIncident               Date of Incident          date
## 7          InsuredGender Gender of Auto Insurance Holder           chr
## 8     SeverityOfIncident            Severity of Incident           chr
## 9       TypeOfCollission              Type of Collission           chr
## 10        TypeOfIncident               Type of Incident           chr
```

Data Visualization with Results

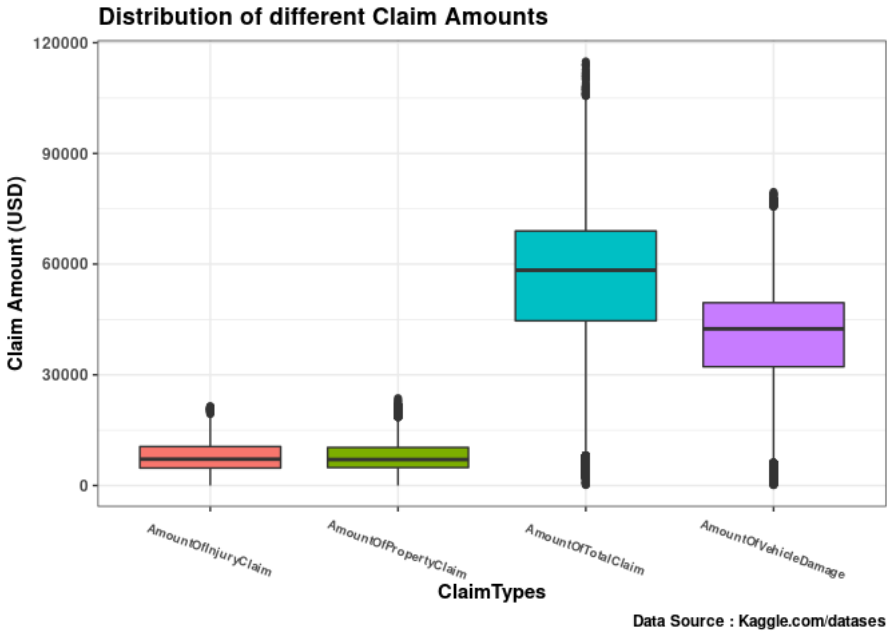
Table of Summary Statistics by merging the necessary tables

```
## # A tibble: 16 × 3
## # Groups:   TypeOfIncident, TypeOfCollission [16]
##    TypeOfIncident           TypeOfCollission     n
##    <chr>                    <chr>            <int>
##  1 Multi-vehicle Collision  Front Collision   3249
##  2 Multi-vehicle Collision  Rear Collision    4311
##  3 Multi-vehicle Collision  Side Collision    4378
##  4 Multi-vehicle Collision  <NA>                28
##  5 Parked Car               Front Collision     17
##  6 Parked Car               Rear Collision      14
##  7 Parked Car               Side Collision      26
##  8 Parked Car               <NA>              2451
##  9 Single Vehicle Collision Front Collision   3969
## 10 Single Vehicle Collision Rear Collision    4212
## 11 Single Vehicle Collision Side Collision    3456
## 12 Single Vehicle Collision <NA>                40
## 13 Vehicle Theft            Front Collision     11
## 14 Vehicle Theft            Rear Collision      24
## 15 Vehicle Theft            Side Collision       7
## 16 Vehicle Theft            <NA>              2643
```

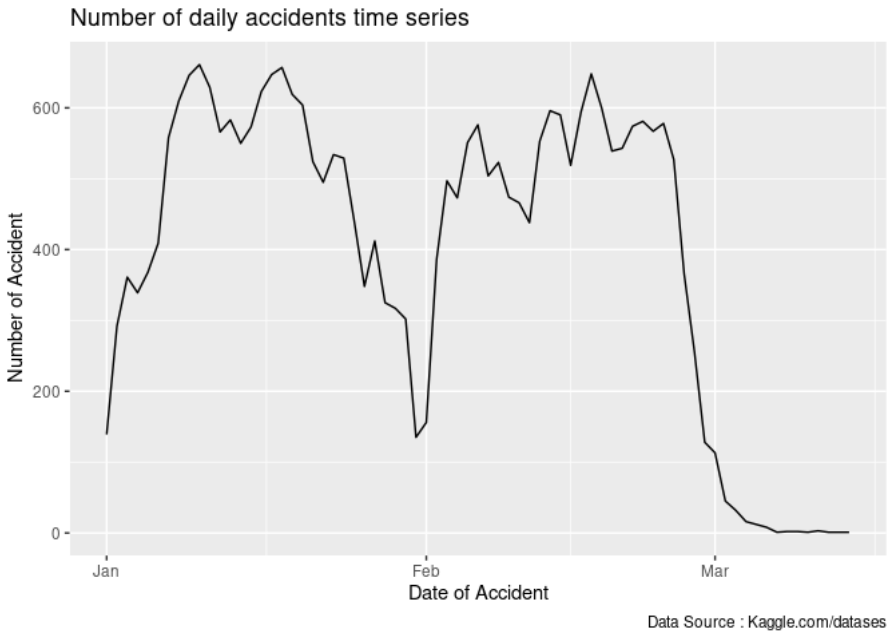
```
## # A tibble: 10 × 6
## # Groups:   Range [5]
##    Range     IsFraud Average Median   Min   Max
##    <chr>     <chr>     <dbl>  <dbl> <dbl> <dbl>
##  1 2001-2005 No         957.    955   931   978
##  2 2006-2010 No         931.    949   863   963
##  3 1996-2000 No         916.    881   849  1021
##  4 1990-1995 No         832.    885   438  1003
##  5 2011-2015 No         409.    448    26   702
##  6 1996-2000 Yes        361.    388   309   394
##  7 2001-2005 Yes        359     361   307   414
##  8 2006-2010 Yes        315.    330   260   347
##  9 1990-1995 Yes        288.    332    88   386
## 10 2011-2015 Yes        220.    238    84   318
```

ggplot Visualizations

Let's analyze Damage claim amounts and to do this let's **pivot** the dataset first to obtain the below distribution of different claim amounts:

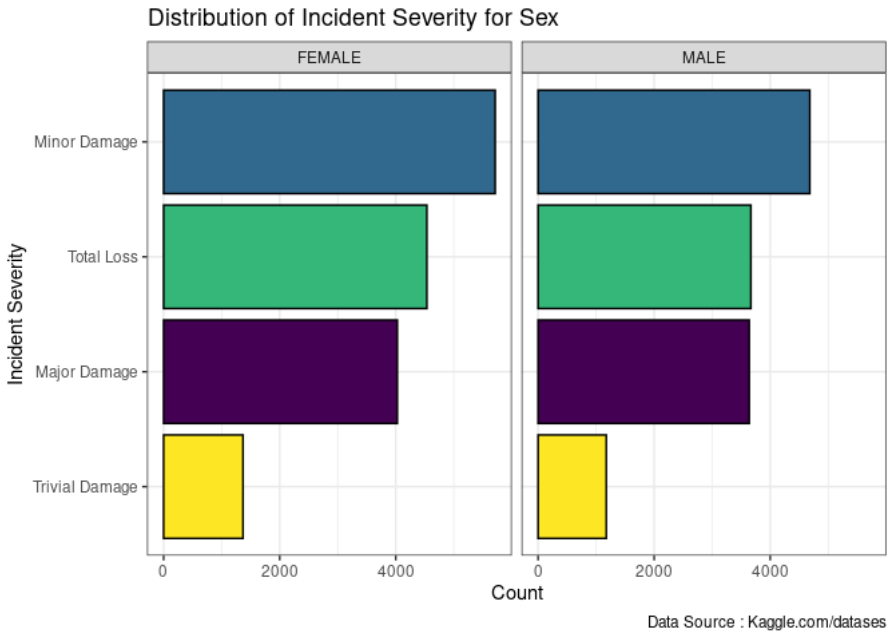


From the above boxplot, we can see that the amount of injury claim and the amount of property claim is similar and low whereas, the amount of total claim and the amount of vehicle claim quite high.

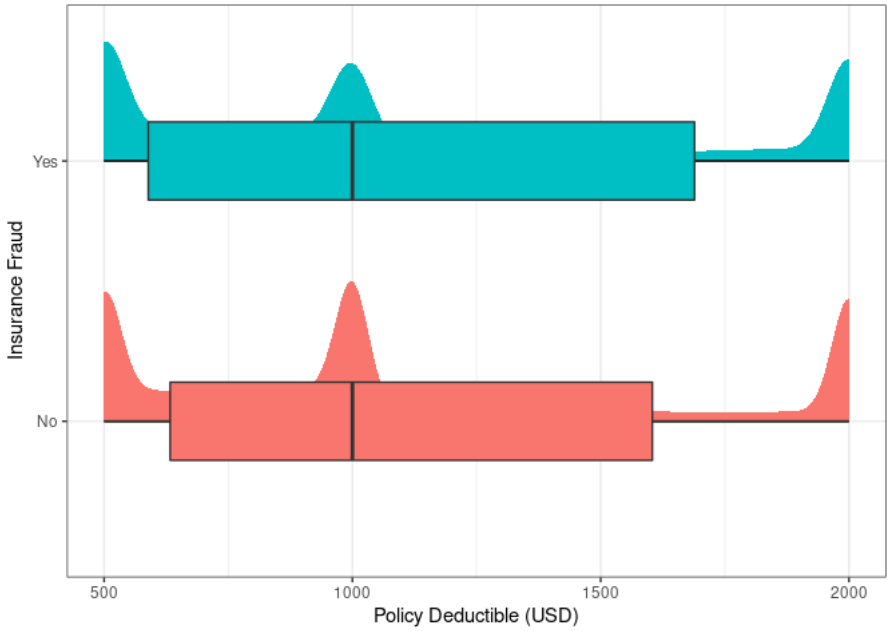


From the time series plot for number of daily accidents, we can see that in the first week of January, third week of January and in the end week of February the number of accidents were at the peak. From starting of the month March, the number of accidents started decreasing.

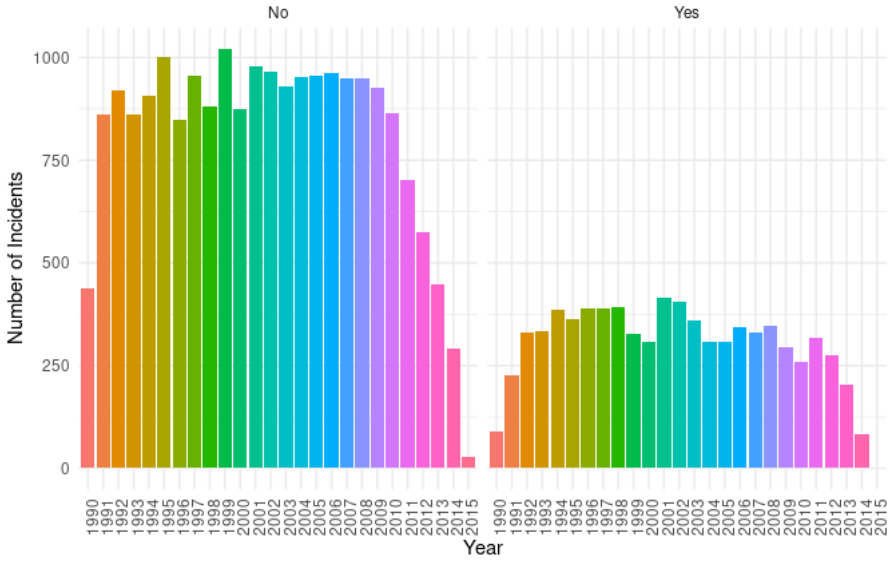
Now, let's merge the claim data and demographic data to see the distribution of incident severity by sex:



From the above side-by-side bar charts we can that the accident rate is higher among female than male for all types of incidents i.e. Trivial, minor, major and total.



From the above plot we can see that for Insurance Fraud “yes”, the policy amount is deducted the most. Whereas, for Insurance Fraud “no”, the policy deductable amount is less.



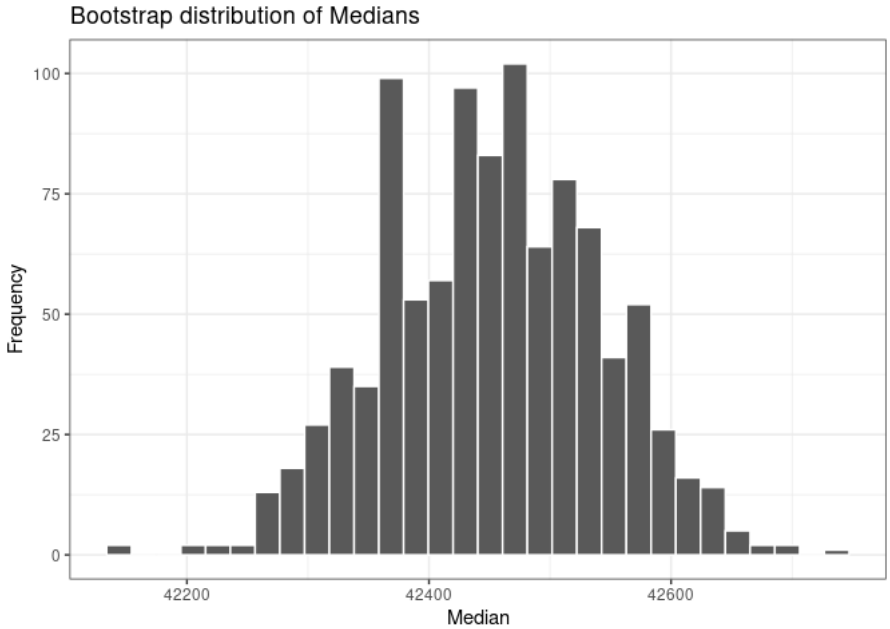
From the above plot, we can see that for IsFraud = Yes i.e. where the policy was claimed without any accident, the number of incidents was high in the year 2001 followed by year 2002. The number of incidents were low for the year 2014 followed by 1990. After year 1990, the number of incidents were increasing till year 1994 after that it was decreased little bit in year 1995. For IsFraud = No, i.e. where policy was claimed for a authentic reason, the number of incidents were high in year 1999 followed by 1995. For year 2001, 2002, 2004 to 2008 the number of incidents were pretty high. Whereas, the number of incidents were at the least in year 2015. Also, we can see that after year 2008, the number of incidents started decreasing.

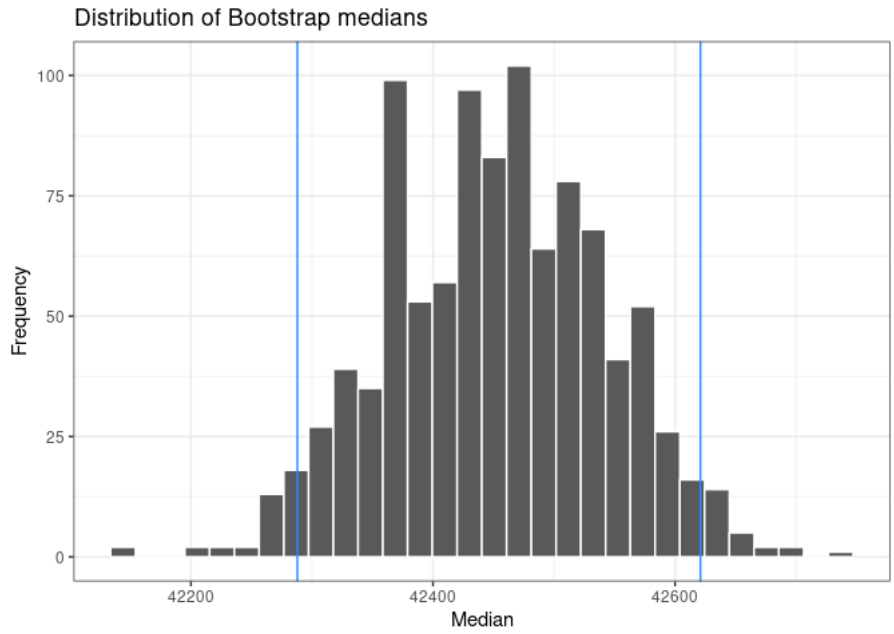
Permutation test based on a traditional hypothesis test

F: 8.0305; P-value: 0.0056667 Q: Do we reject or fail to reject Null Hypothesis? A: Since the p-value of 0.0056667 is smaller than 0.05, we reject Null Hypothesis. We have sufficient evidence at the 5% significance level that the average amount of vehicle damage differs by insured gender (F = 8.0305).

Bootstraping

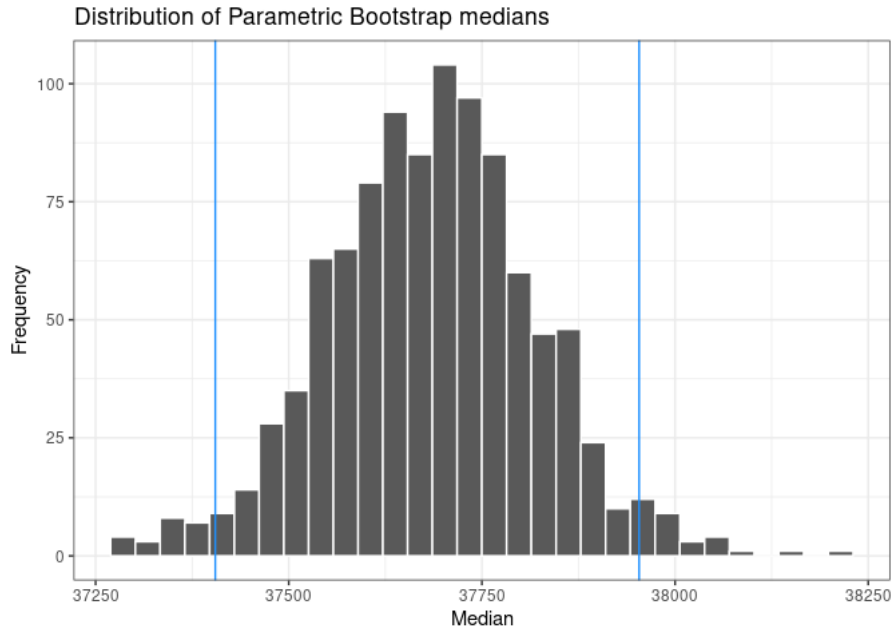
Let's perform **Non-parametric** Bootstrap first:





We are 95% confident that the true median for these distribution is between 4.2288^{4} and 4.2621^{4} .

Now, let's perform the **Parametric** Bootstrap:



We are 95% confident that the true median for these distribution is between 3.7405^{4} and 3.7953^{4}.

Conclusion

Finally, from all the plots, we can conclude that most of the accidents were done by females over males. They have done most of the minor accidents. The most of the accidents took place between January and February and at the starting of month March it got decreased. But surprisingly end of January the accident rate was less. Apart from these observations we have seen that most of the claims are approached for total damage followed by other vehicle damages. Also most of the authentic incidents took place in 1999 followed by 1995 and between year 2004 and 2008 the rate was also pretty high and after year 2008 it started decreasing. Whereas, most of the fraud accidents took place in year 2001 and 2002.

References

1. https://www.kaggle.com/datasets/?search=insurance-fraud
2. https://rdrr.io/cran/dataMeta/man/build_linker.html

3. https://docs.google.com/document/d/1pLFDA0JZUCuGNvpsqRgNhs3KI9DHD8kFtZbvEpd9TdM/edit?usp=sharing
4. https://docs.google.com/document/d/1APSWLqQ_8Wy_mUR3afg9mZbif3YiFCR94uhg8i7D3_g/edit?usp=sharing
5. https://docs.google.com/document/d/1COab_YbF0gDdWz02UDjVCegB9lbkDQgOC3lkvDfroR0/edit?usp=sharing
6. https://docs.google.com/document/d/16UeY3X8I7rJEllTlUm2SWhfKy3WQWiV6gZYZrJzfepk/edit?usp=sharing
7. https://docs.google.com/document/d/1WYPmd750r2ojAiEljpogzqdVG_HPbIK_/edit?usp=sharing&ouid=117678942095329132446&rtpof=true&sd=true
8. https://docs.google.com/document/d/1uFDe5rddt23xqcnuRgg4RSaODQGtg2PwnYmjPRO9ib4/edit?usp=sharing
9. https://docs.google.com/document/d/1xKrR-b9o-GcXzuf3k_aOSr_hJXp4RiJx/edit?usp=sharing&ouid=117678942095329132446&rtpof=true&sd=true
10. https://docs.google.com/document/d/1U-pMTx_HA002arktw9Az6P_r0lP1pseNQ-s6OoyoA88/edit?usp=sharing
11. https://docs.google.com/document/d/17mavgAogTYiZFy8sscdxCdkuQIonhRiAIQhLdlf6cmw/edit?usp=sharing
12. https://docs.google.com/document/d/1CpsVoPgtesTFnwjmyJ3wd1RBg2QaRJcl52q3pGoULU4/edit?usp=sharing
13. https://docs.google.com/document/d/1Xet7QNikLWmsNb3CRSuMX2ve_debbkbq_L17jd_YDWE/edit?usp=sharing
14. https://docs.google.com/document/d/1fqWpnIU4OIgmo4BGwcLNYrdLIBsBPByL/edit?usp=sharing&ouid=117678942095329132446&rtpof=true&sd=true
15. https://docs.google.com/document/d/1K6fARfGLVog9sM3_6EgshjIkZ6jOvKKtQbScWpahA0U/edit
16. https://drive.google.com/file/d/1fAiQTN-KpsWq2J8gUSaYrzZnF8FXbhUF/view?usp=share_link