Assessment 1 - Part A (10%)

Due Sunday, 15 March 2020 11:59PM AWST

The file *Insurance Fraud Claims.csv* contains data relating to 500 insurance policy holders who have made a claim. Amongst the data collected were the holder's marital status, personal income, type of injury sustained from the accident, the amount claimed and received, and whether the claim was flagged as fraudulent by the insurance company.

First, copy the code below to a R script. Enter your student ID into the command **set.seed(.)** and run the whole code. The code will create a sub-sample that is unique to you.

```
#You may need to change/include the path of your working directory
dat <- read.csv("Insurance Fraud Claims.csv")  #Import the dataset into R Studio.

set.seed(Enter your student ID here)

#Randomly select 400 rows
selected.rows <- sample(1:nrow(dat),size=400,replace=FALSE)
mydata <- dat[selected.rows,]  #Your sub-sample of 400 observations

dim(mydata)  #check the dimension of your sub-sample

#Write to a csv file. You will need to submit this file along with your solution
write.csv(mydata,"mydata.csv")
```

Complete the following tasks in R using the sub-sample that you have just created.

(1) Complete the table below for the three categorical features, **Marital Status**, **Injury Type** and **Fraud Flag**. [3 marks]

Marital Status		Injury Type		Fraud Flag	
Category	**N (%)**	**Category**	**N (%)**	**Category**	**N (%)**
Single	1.50	Back	22.75	No	65.25
Married	7.25	Broken limbs	31.50	Yes	34.75
Divorced	2.25	Soft Tissue	6.75	Missing	0
Missing	89.00	Serious	33.00		
		Missing	6.00		

(2) Complete the table below for the three continuous features, **Income of Policy Holder Status**, **Claimed Amount** and **Claimed Amount Received**. [6 marks]

Feature	Missing N (%)	Minimum	Mean	Median	Max	IQR	Skewness
Income of Policy Holder Status	67	16730	39480	39438	71284	13304.75	0.291639
Claimed Amount	0	-99999	15441	5682	270200	9284.75	2.280109
Claimed Amount Received	3.75	0	13680	3847	295303	8254	4.825968

(3) Based on statistics in the summary tables of (1) and (2), can you identify any potential/obvious issue (if any) with your dataset? If so, what are they? [3 marks]

Solution

The issues I found based on table 1 and 2 are as follow:

- 89% of material status values are missing which make it inconclusive to categories the users based on the material status of the user
- In the Income of Policy Holder to 67% of the values are not available such high level of missing values make mean, median, IQR and skewness value less relevant the subjected data.
- The minimum value in the claim amount is -99999 which is incorrect as the claim amount must always contain positive values

(4) Suppose that you wish to find the median claimed amount received for each injury type. However, you have noticed that the data for **Claimed Amount Received** for a few of the

individuals are missing. Examine these individuals closely and outline whether you would (1) remove these individuals from the analysis or (2) replace/impute the values. Justify your answer.

[3 marks]

Solution

Based on the data set only 15 values contain NA for the claim amount received, so in my understanding instead of replacing the missing values to the median a better approach would be skip those records

	Median
Back	3016.05
Broken Limb	3933.0
Serious	4653.5
Soft Tissue	3556.5

Injury Type	Median
Back	3016.05
Broken Limb	3933.0
Serious	4653.5
Soft Tissue	3536.5

Submission Instructions:

Your submission must include the following:

- This document with your answers
- A copy of your R code
- The dataset containing your sub-sample

The three files must be submitted through **Blackboard as three separate files.**

Note that no marks will be given if the results you have provided cannot be confirmed by your code.

Academic Misconduct

Edith Cowan University regards academic misconduct of any form as unacceptable. Academic misconduct, which includes but is not limited to, plagiarism; unauthorised collaboration; cheating in examinations; theft of other student's work; collusion; inadequate and incorrect referencing; will be dealt with in accordance with the ECU Rule 40 Academic Misconduct (including Plagiarism) Policy.

Ensure that you are familiar with the Academic Misconduct Rules.

Assignment Extensions

Applications for extensions must be completed using the ECU Application for Extension form, which can be accessed online.

Before applying for an extension, please check out the ECU Guidelines for Extensions which details circumstances that can and cannot be used to gain an extension. For example, normal work commitments, family commitments and extra-curricular activities are not accepted as grounds for granting you an extension of time because you are expected to plan ahead for your assessment due dates.

Please submit applications for extensions via email to both your tutor and the Unit Coordinator.

Where the assignment is submitted no more than 7 days late, the penalty shall, for each day that it is late, be 5% of the maximum assessment available for the assignment. Where the assignment is more than 7 days late, a mark of zero shall be awarded.