A Stochastic Investigation Into Swedish Insurance Data

Group: 26

Tom Sijbers
1709208

Martin Georgiev
1681826

October 2023

Contents

1 Introduction

This report encompasses a stochastic investigation into insurance data provided by the Swedish insurance company Wasa. The investigation aims to leverage a variety of data analysis and simulation methods with the goal of gaining a deeper understanding of the trends and relations within the data set.

Included in the data set, is information about insurance claims received by Wasa between 1994 and 1998. For each claim, the following information was stored.

* The driver age
* The geographic zone
* The vehicle age
* the claim cost

The investigation uses this data set to examine the existence of correlations between the claim cost and the other factors. Specifically, does the age of the driver or vehicle influence the expanse of the claim? Or is there any discrepancies between the geographic zone of a claim and its cost?

These questions are investigated through careful analysis of the data in chapter 2 Data Analysis. Chapter 3, then focuses on using a stochastic simulation to simulate the number of claims and the total claim cost.

2 Data Analysis

2.1 Chapter Introduction

This chapter focuses on the data analysis performed during this assignment. It outlines the techniques used to gather useful summary statistics and insightful plots of the data.

Section 2.2 focuses specifically on plots and figures showing summary statistics about the data set based on driver age, vehicle age or geographic zone. This is followed by section 2.3 which focuses on plotting and analyzing claim cost distributions and how they differ based on driver age, vehicle age or geographic zone.

2.2 Descriptive Statistics

2.2.1 Cleaning the Data

Initial plotting of the data was hindered by an abundance of entries with a claim cost of 0 SEK.

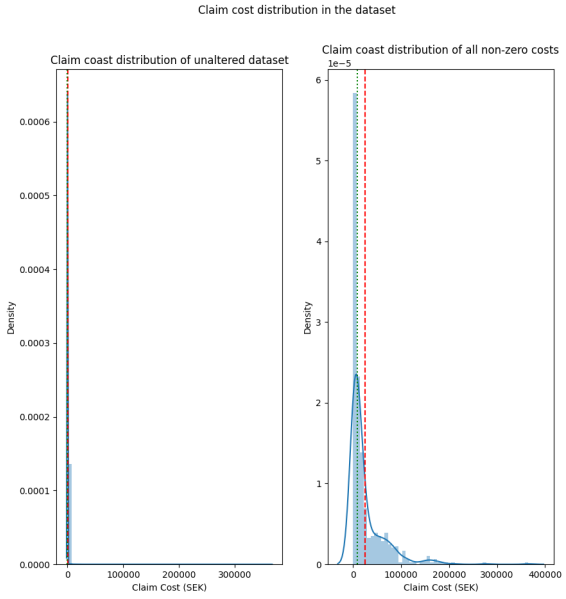

Figure 1: Cost distribution of unaltered vs altered data set

Figure 1 shows two graphs plotting the claim cost distribution of the data

set. On the left side the complete unaltered data set is used, but when inspecting the graph it quickly becomes apparent that there is an issue with the original data preventing a proper plotting of the cost distribution. This issue is due to only 670 entries of the original data 64548 entries having a cost value above 0 SEK. This means over 98% of the data entries have no cost and as such, the remaining data that provides a proper insight into the cost distribution is hidden by an abundance of faulty data. To correct this issue, a new data set was made containing only the values of the original data set where the claim cost is greater than 0 SEK. On the right side of figure 1, this data is plotted as a cost distribution.

2.2.2 Driver Age

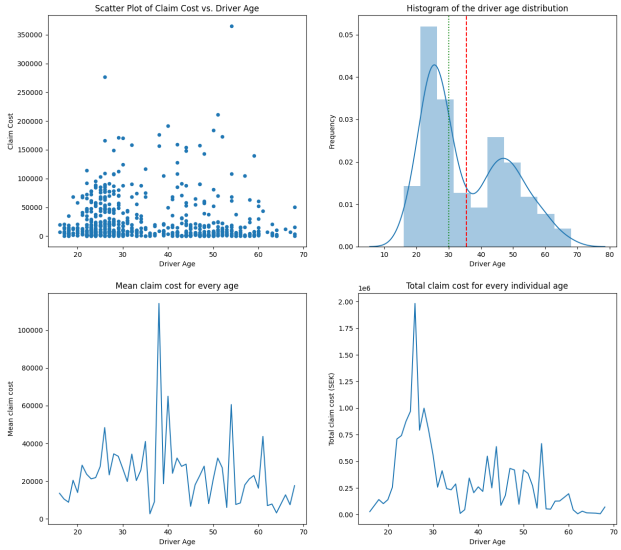

Figure 2: Driver Age summary statistic plots

Figure 2 consists of four different plots. The first plots is a scatter plot that for every entry in the data set plots the driver age against the claim cost. Next, the second plot shows a histogram showing the distribution of driver ages in the data set. Thereafter, the third plot is a line plot that shows the mean claim cost for every individual driver age. Lastly, the fourth plot shows another line plot, this time showing the total claim cost for every individual age. In this section of the report, these plots will be analyzed and explained.

The scatter plot in figure 1 provides insight into clusters and outliers in the data set. For example, most claims are clustered to have a cost between 0 and 50000 SEK. The majority of claims are also made by younger people below the age of 35. Then there is a large cluster of claims made by 45-60 year old drivers. Lastly, it clearly shows claims above 50000 SEK are much less likely, but when they do occur they often occur among younger drivers.

Subsequently, figure 1 contains a histogram showing the distribution of driver ages. Along the histogram there is a green dotted line showing the median driver age is 30 and a red line showing the mean driver age is 35. Noticeably, the graph reveals over half the claims are made by drivers under the age of 30. This could show more experienced drivers are less likely to have accidents and therefore file fewer insurance claims.

Following this, figure 1 contains a line plot. To create this plot, the mean claim cost for every individual driver age in the data set was calculated, this graph was then created to showcase how the mean cost differentiates for different ages. However, there seems to be no clear relationship linking the amount of the claim cost to the drivers age. Throughout the graph, mean costs seem to vary among younger ages as much as they do for older ages. The only noticeable spike is at around age 40 where damages spike to over 100000 SEK. Analysis of the scatter plot of data points around this age reveals that there are fewer low cost claims at this age which results in the mean cost being much higher. Yet, since this is a unique spike along the graph, more data is needed to replicate these results before any conclusions can be drawn.

Lastly, figure 1 contains another line plot. Unlike the previous line plot, this one shows the total claim cost of all different claims added together for every individual age in the data set. It provides a better overview of how much it costs to insure different ages. Noticeably, the line plot resembles the histogram showing the distribution of driver ages. Specifically, the graph reaches its maximum between the ages of 20 and 30, this is explained by helf the total claims being filed by people in this age group. Older drivers have a lower total claim cost since they have fewer accidents.

2.2.3 Vehicle Age

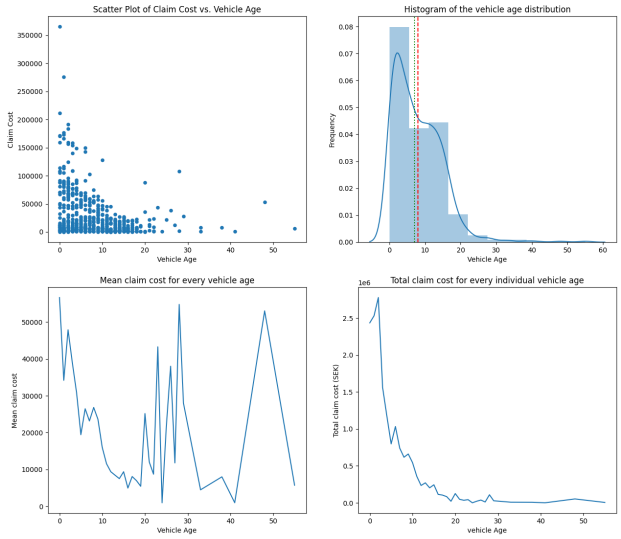

Figure 3: Vehicle Age summary statistic plots

Figure 3 shows the exact same plots as figure 2, except this time using vehicle ages instead of driver ages. So first, a scatter plot graphing the vehicle age against the claim cost for every data entry. Followed by a histogram showing the distribution of vehicle ages in the data set. Lastly, it contains two line plots, the first showing the mean claim cost for every individual vehicle age and the second showing the total claim cost for every individual vehicle age.

The scatter plot in figure 3 is very revealing. Almost all data points are clustered together before the vehicle age of 20. Furthermore, there is a clear pattern showing a larger number of claims and more expensive claims the younger a vehicle is. This shows a clear relation between the cost of a claim and the age of the vehicle where younger vehicles are likely to have higher claim costs.

The histogram showing the distribution of vehicle ages could reveal a reason for this relationship. The green dotted line shows the median vehicle age in the data set us around 18 while the red dotted line shows the mean age is around 19. Combining this with the fact that claims for cars above the age of 30 are almost non-existent, the data could be interpreted to suggest that there are fewer older vehicles being driven and therefore there are less incidents with them.

The line plot showing the mean claim cost for every individual vehicle age is less reliable this time around. It shows the mean claim cost for vehicles around 5 years old to be equal to those around 30 and 50 years old. However, when inspecting the scatter plot above it reveals that there is only around 1 claim made per vehicle age for vehicles older than 30 years old. As a result, the mean cost doesn't tell us much since one data point could easily be an outlier and therefore skew the results.

Lastly, the final line plot shows a line plot of the total claim cost for every individual vehicle age. It again closely resembles the vehicle age distribution shown by the histogram because more claims result in a higher total cost. Yet, it does enforce that younger vehicles are much more expensive in terms of total claim costs.

2.2.4 Geographic Zones

	1	2	3	4	5	6	7
count	173	162	118	190	9	17	1
mean	32022.91	29698.56	21378.2	19866.47	11637.67	16943.82	650
std	42861.39	38827.03	42188.59	32167.32	15067.48	21643.1	NaN
min	77	35	221	16	540	1600	650
25%	4904	3723.5	2842.5	2068.5	2260	4740	650
50%	14500	11816.5	8000	6392.5	4209	7990	650
75%	47841	43688	20062.25	20523.25	14675	13148	650
max	276298	191462	365347	183579	46541	68000	650

Figure 4: Claim Cost Distribution in Different Driver Age Groups

Figure 4 gives us the summary statistics of the Claim Cost in different Regions. In terms of sample sizes in every region we see that regions 1-4 have much larger sample size than regions 5-7. This suggests that regions 1-4 have more extensive data sets to draw conclusions from than 5-7. For 7 specifically there is practically no way to characterise the region with only a single sample.
It is also apparent that regions 1-4 have a much higher mean claim cost than 5-7. Furthermore the standard deviation greatly varies between regions and implies that regions 1-4 have higher variability in Claim Cost compared to regions 5-7. Both of these observations can be used to suggest marketing or increased interest for the company to provide more insurances to those regions.
The higher variability of the data sets for regions 1-4 can also be observed in the min, max, and different quartile values.

2.3 Distribution Fitting

2.3.1 Driver Age

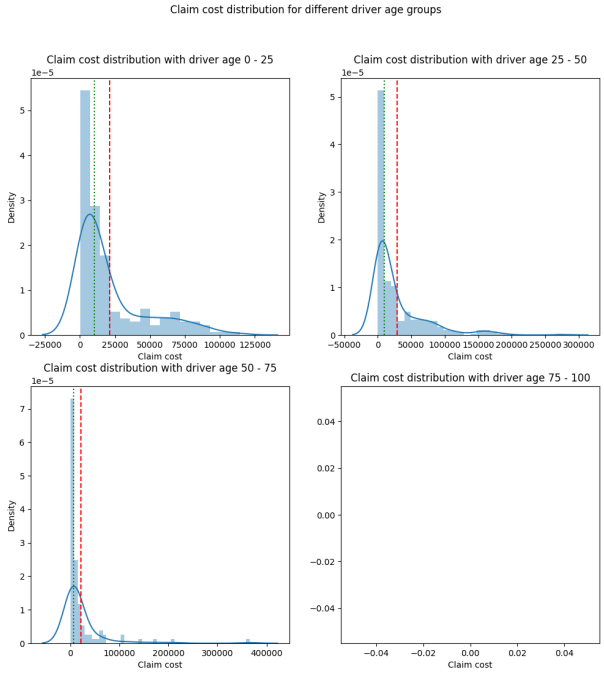

Figure 5: Claim Cost Distribution in Different Driver Age Groups

The figure above shows the claim cost distribution for 4 different age groups. This was done by splitting the original data set into 4 smaller data sets based on the age of the driver in each claim. The claim cost distribution was then plotted as a histogram for each age group. The first histogram shows the distribution for drivers between the age of 0 and 25, the seconds shows the distribution for drivers between the age of 25 and 50, the third shows the distribution for drivers between the age of 50 and 75, and finally, the fourth shows the distribution for drivers between the age of 75 and 100.

Upon inspecting the graphs its noticeable that the 4th histogram, showing the claim cost distribution for drivers between the ages of 75 and 100, is empty. This is because there are no claims made by drivers over the age of 75.

Now for the remaining three graphs, each graph contains a red and green dotted line. the green line represents the median value in this distribution while

the red line represents the mean value. These values can help us get a better understanding of the differences in each data set. The following table shows the mean and median value for each age group.

Driver Age Group	**Median Claim Cost**	**Mean Claim Cost**
0-25	10245.5 SEK	21289.816 SEK
25-50	9349 SEK	28564.392 SEK
50-75	6540 SEK	21762.934 SEK
75-100	NaN	NaN

The mean values for each driver age group still falls within the 20000 SEK range, which means each group has a similar mean claim cost. The median values for age group 0 to 25 and 25 to 50 are also quite similar with age group 50 to 75 being a few thousand SEK lower. This tells us that in terms of average cost, the age groups do not differ to much. However, age group 0-25 seems to have a max claim cost around 125000 SEK which is lower than the max claim cost for age groups 25 to 50 and 50 to 75 which both equal around 200000 SEK. This shows that older drivers seem to have more outliers with high claim costs.

Additionally, the density curve for the two histograms at the top reaches much higher. This represents the much higher number of claims filled by younger drivers. Overall, this tells us that age group 0-25 tends to have the most claims but does not have as many extreme claim cost values. Age group 25 to 50 has fewer claims filled but they have a higher maximum claim cost. While age group 50 to 75 has the least amount of claims filled but their claims do also have a higher maximum value.

2.3.2 Vehicle Age

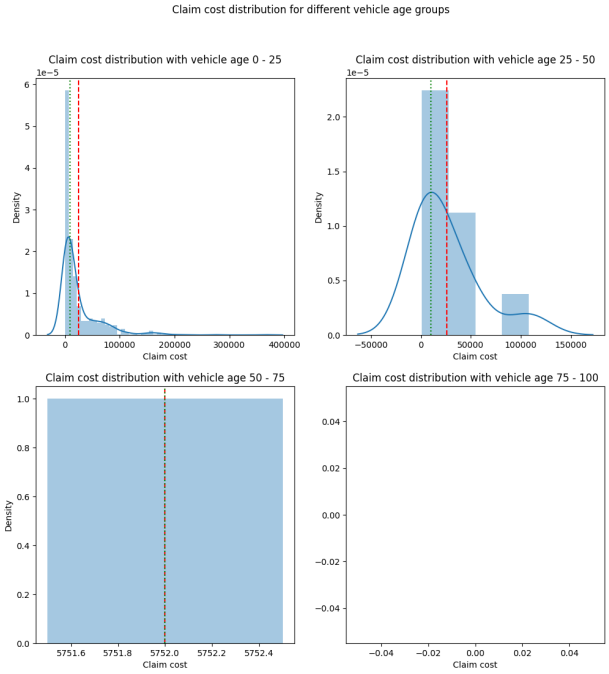

Figure 6: Claim Cost Distribution in Different Vehicle Age Groups

The figure above shows 4 histograms. The first shows the claim cost distribution for vehicles between 0 and 25 years old, the second does the same for vehicles between the age of 25 and 50, the third shows vehicles between the age of 50 and 75, and the last shows vehicles between the age of 75 and 100.

Inspecting the table reveals that there is only one claim for vehicles above the age of 50 and no claims at all for vehicles over the age of 75. As a result the graphs are either empty or contain only a single value.

Again, each graph has a red and green line representing the mean and median respectively, below is a table showing the mean and median distribution values for each vehicle age group.

Vehicle Age Group	Median Claim Cost	Mean Claim Cost
0-25	9030 SEK	25459.402 SEK
25-50	9911 SEK	25832.2 SEK
50-75	5752 SEK	5752 SEK
75-100	NaN	NaN

The most obvious result produced by this table is that the mean and median values of age groups 0 to 25 and age groups 25 to 50 are almost identical being less than 1000 SEK apart. This is surprising since the histogram shows lots of variation in claim cost amount for age group 0 to 25 and much less for age group 25 to 50. This means that cars in age group 0 to 25 have a much more varied claim cost value while age group 25 to 50 claims tend to have similar cost values.

The mean and median for age group 50 to 75 does not tell us much as there is only one claim made in this age group meaning there is very little this tells us about the cost distribution for cars in this age range. This one value might be an outlier that over or under represents the cost of claims made for these vehicles and so using it for results would not be advisable.

2.3.3 Geographic Zones

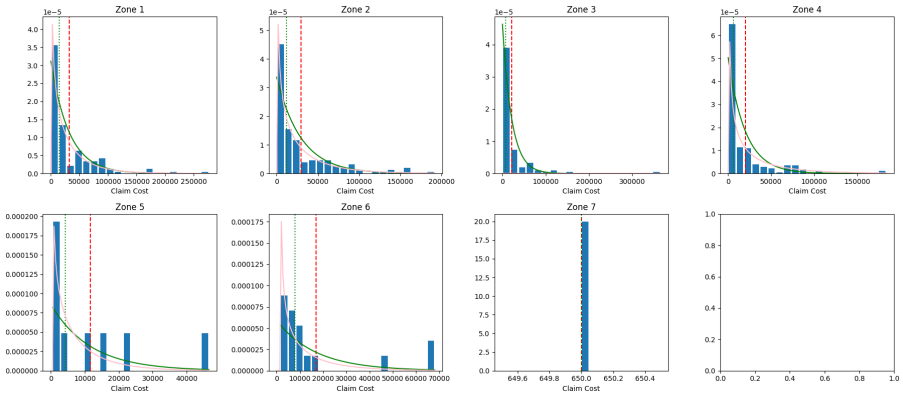

Figure 7: Claim Cost Distributions for Different Geographic Zones with Fitted Distributions (Gamma = Pink, Exponential = Green)

Having plotted the distributions of the Claim Cost samples for the different regions we noticed that the distributions are similar to the Exponential and Gamma distributions. Next we created fits for the Exponential and Gamma distribution over the sample, plotted them in the Figure 7, and tested them using the Kolmogorov-Smirnov test.

Region	Exponential Dist p-value	Gamma Distribution p-value
1	0.00013387635803192507	0.3903800433343818
2	$2.2200580403244777e-05$	0.4941361256911242
3	$3.366800955333299e-07$	$1.248e-320$
4	$4.017397734773498e-11$	0.0026481529240668412
5	0.5207082706722153	0.8919754969479589
6	0.2638250923457027	0.14645690161454827
7	0.7357588823428847	NaN

From the results we can conclude that the Gamma Distribution is a more suitable fit in regions 1, 2, 4, and 5 (higher p-value), while the fitted Exponential Distribution is better for regions 3, 6 7. What is strange is that Region 3 appears to have similar distribution to regions 1, 2, 4, yet has such a vast difference in the fitted Gamma Distribution. This is likely due to the single outlier with claim cost above 300 000 SEK.

3 Stochastic Simulation

3.1 Chapter Introduction

Chapter 3 covers the stochastic simulation element of this investigation. It begins by providing a detailed description of the simulation model used and how it was constructed in section 3.2. Following this, section 3.3 provides information about the accuracy of the model and the number of runs it performed. Section 3.4 presents the simulation results and summary statistics. Overall, the goal of this chapter is to give the reader a deeper understanding of the stochastic simulation used and how the results of this investigation where achieved.

3.2 Model Description

Our model simulates the accumulation of damage costs in a single year as a compound Poisson process. The accumulation of damage costs in a single year consists of the arrival of a claim and the addition of the claim's damage cost to the cumulative yearly damage costs.
The claim interarrival times are simulated using a Poisson distribution with a lambda equal to the average arrivals per day which is estimated using the number of claims in the 5 year period (1994-1998).

```
lam = n_claims / (365 * 5)
...
interarrival_time = np.random.exponential(1 / lam)
```

Then, we randomly choose the region from which the claim arrived according to an estimation for the distribution of the claim regions:

```
claim_region_distribution = [len(df[df['zone'] == 1])/n_claims,
                             len(df[df['zone'] == 2])/n_claims,
                             len(df[df['zone'] == 3])/n_claims,
                             len(df[df['zone'] == 4])/n_claims,
                             len(df[df['zone'] == 5])/n_claims,
                             len(df[df['zone'] == 6])/n_claims,
                             len(df[df['zone'] == 7])/n_claims]
...
claim_region = np.random.choice(a=claim_region_elements, size=1
, p=claim_region_distribution)
```

The amount of that claim is simulated using the appropriate fitted distribution from the Data Analysis of the Claim Costs in the different zones.

```
claim_amount_distribution =
[stats.gamma(a = fitShape1, loc = fitLoc1, scale = fitScale1),
 stats.gamma(a = fitShape2, loc = fitLoc2, scale = fitScale2),
 stats.expon(np.mean(df[df['zone'] == 3]['claimCost'])),
 stats.gamma(a = fitShape4, loc = fitLoc4, scale = fitScale4),
 stats.gamma(a = fitShape5, loc = fitLoc5, scale = fitScale5),
 stats.expon(np.mean(df[df['zone'] == 6]['claimCost'])),
 stats.expon(np.mean(df[df['zone'] == 7]['claimCost']))]
...
claim_amount = claim_amount_distribution[claim_region[0] - 1].rvs()
```

This is repeated while the time elapsed i.e sum of interarrival times is less than a year.

```
while time_elapsed < 365:  # 1 year in days
        # Sample interarrival time
        interarrival_time = np.random.exponential(1 / lam)
        time_elapsed += interarrival_time
        if time_elapsed <= 365:
        ....
            claims += 1
            claim_cost += claim_amount
```

3.3 Accuracy and Number of runs

The simulation model above is run 10000 times. This number of run was chosen to balance the results for the distribution and summary statistics and the time the simulation takes to run. With 10000 runs we get precise graphs of the distributions and the values of the mean and std while keeping run-time approximately around 130 seconds.

3.4 Simulation Results

From the figure below, we notice that compound Poisson processes, the cumulative damage cost and number of claims, follow a distribution similar to the Normal Distribution with the respective means and std.

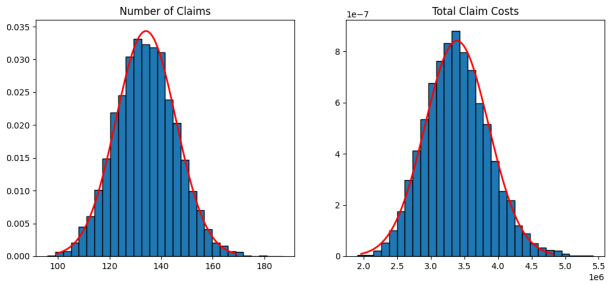

Figure 8: Claim Cost Distribution in Different Vehicle Age Groups

	Number of Claims	**Cumulative Claim Cost**
Mean	134.1112	3389280.437236682
Std	11.618787998754431	474435.2936364452

Overall we get a mean of 134 claims filed per year totaling a cumulative claim cost of 3389280.4 SEK. The number of claims has a standard deviation of 12 claims while the cumulative claim cost has a standard deviation of 474435.3

SEK. Overall, this normal distribution suggests there is a reliable way to simulate the expected number of claims a Wasa will receive in a year, furthermore, simulating the expected cost of these claims allows Wasa to better prepare and manage their financial budget.

4 Conclusions and Recommendations

This investigation into Wasa insurance data revealed some intriguing trends and relations between the different factors included in each claim. Specifically, data analysis into the data set revealed a disproportionate amount of claims made by younger cars and drivers. However, driver age didn't seem to affect the average claim cost much, only the number of claims. Vehicle age on the other hand seems to have a larger correlation, with younger cars almost exclusively having the most expensive claim costs. While we can speculate about the cause of these trends, there is no way to conclusively justify them from numeric data alone.

Investigating claim distributions between the different geographic zones also proved insightful. Analysis revealed a large difference in the number of claims between different regions. Zone's 1 to 4 had noticeably more claims filed in them while zones 5 to 7 had a total of 27 claims between them. This contrast in the number of claims filed between these zones could be due to the nature of these zones. Zones 1, 2 and 3 are from within large cities, suburbs and lesser towns, each of these have a more urban environment meaning more vehicles and therefore more accidents. Zone 4 covers small towns and countryside while zone 5 and 6 cover small towns and countryside in the northern side of Sweden. The large difference here is that northern Sweden is the least populated area of Sweden, therefore zone 5 and 6 have fewer entries than zone 4. Lastly zone 7 covers Sweden's largest island, which still has a smaller population and therefore less insurance claims are filled.

While investigating this data set has produced compelling results, it is important to note that these these results are drawn from a limited data set in a rather short time span. Therefore, further investigations need to be conducted where this data is combined with newer insurance claim data to test whether these results can be replicated and corroborated. Until such research is performed, it is difficult to claim with certainty that the results of this study are proof of real trends rather than premature conclusions made off limited data.

5 Appendix

5.1 Driver & Vehicle Age Data Analysis Code

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

#Variable stores the excel file name
excel_file = "assignment2data.xlsx"

#Excel file is read and turned into a dataframe
df = pd.read_excel(excel_file)

df_no_zeros = df[(df['claimCost'] > 0)]

#This figure shows the claim cost distribution before and after
    cleaning the data
#It is important so we canm highlight why certain data was removed
    in our graphing

fig, ax_1 = plt.subplots(nrows=1, ncols=2, figsize=(10, 10))

#Show a plot of the claimcost distribution in the unaltered dataset
sns.distplot(df['claimCost'], ax = ax_1[0])
ax_1[0].set_title('Claim coast distribution of unaltered dataset')
ax_1[0].set_xlabel('Claim Cost (SEK)')
ax_1[0].axvline(np.mean(df['claimCost']), ls='--', c='r', label='
    Mean')
ax_1[0].axvline(np.median(df['claimCost']), ls=':', c='g', label='
    Median')

#Show a plot of the distribution if 0 costs are taken out.
sns.distplot(df_no_zeros['claimCost'], ax=ax_1[1])
ax_1[1].set_title('Claim coast distribution of all non-zero costs')
ax_1[1].set_xlabel('Claim Cost (SEK)')
ax_1[1].axvline(np.mean(df_no_zeros['claimCost']), ls='--', c='r',
    label='Mean')
ax_1[1].axvline(np.median(df_no_zeros['claimCost']), ls=':', c='g',
     label='Median')

fig.suptitle('Claim cost distribution in the dataset')

# plt.hist(df_no_zeros['claimCost'])
# plt.show()

#
#
#Driver Age Plots
#
#

#Create different datasets for different age groups
#Driver Age group 0-25
df_0_to_25 = df_no_zeros[df_no_zeros['driverAge'] <= 25]
```

```
#Driver Age group 25-50
df_25_to_50 = df_no_zeros[(df_no_zeros['driverAge'] > 25) & (
    df_no_zeros['driverAge'] <= 50)]

#Driver Age group 50-75
df_50_to_75 = df_no_zeros[(df_no_zeros['driverAge'] > 50) & (
    df_no_zeros['driverAge'] <= 75)]

#Driver Age group 75-100
df_75_to_100 = df_no_zeros[(df_no_zeros['driverAge'] > 75) & (
    df_no_zeros['driverAge'] <= 100)]


#Create a figure to show the claim cost distribution in these
    driver age groups
fig, ax_2 = plt.subplots(nrows=2, ncols=2, figsize=(12, 12),
    squeeze=False)

#Plot the df_0_to_25 dataset as a histogram
sns.distplot(df_0_to_25['claimCost'], ax = ax_2[0, 0])
ax_2[0, 0].set_title('Claim cost distribution with driver age 0 -
    25')
ax_2[0, 0].set_xlabel('Claim cost')
ax_2[0, 0].axvline(np.mean(df_0_to_25['claimCost']), ls='--', c='r
    ', label = 'Mean')
ax_2[0, 0].axvline(np.median(df_0_to_25['claimCost']), ls=':', c='g
    ', label='Median')

#Plot the df_25_to_50 dataset as a histogram
sns.distplot(df_25_to_50['claimCost'], ax = ax_2[0, 1])
ax_2[0, 1].set_title('Claim cost distribution with driver age 25 -
    50')
ax_2[0, 1].set_xlabel('Claim cost')
ax_2[0, 1].axvline(np.mean(df_25_to_50['claimCost']), ls='--', c='r
    ', label='Mean')
ax_2[0, 1].axvline(np.median(df_25_to_50['claimCost']), ls=':', c='
    g', label='Median')

#Plot the df_50_to_75 dataset as a histogram
sns.distplot(df_50_to_75['claimCost'], ax = ax_2[1, 0])
ax_2[1, 0].set_title('Claim cost distribution with driver age 50 -
    75')
ax_2[1, 0].set_xlabel('Claim cost')
ax_2[1, 0].axvline(np.mean(df_50_to_75['claimCost']), ls='--', c='r
    ', label='Mean')
ax_2[1, 0].axvline(np.median(df_50_to_75['claimCost']), ls=':', c='
    g', label='Median')

#Plot the df_75_to_100 dataset as a histogram
sns.distplot(df_75_to_100['claimCost'], ax = ax_2[1, 1])
ax_2[1, 1].set_title('Claim cost distribution with driver age 75 -
    100')
ax_2[1, 1].set_xlabel('Claim cost')
ax_2[1, 1].axvline(np.mean(df_75_to_100['claimCost']), ls='--', c='
    r', label='Mean')
ax_2[1, 1].axvline(np.median(df_75_to_100['claimCost']), ls=':', c
    ='g', label='Median')
```

```
fig.suptitle('Claim cost distribution for different driver age
    groups')

#
#Data Analysis Plots
#

#Create a figure for all the plots
fig, ax_3 = plt.subplots(nrows=2, ncols=2, figsize=(15, 15))

#Scatter plot of Driver Age vs Claim Cost
df_no_zeros.plot(kind='scatter', x='driverAge', y='claimCost', ax=
    ax_3[0, 0])
ax_3[0, 0].set_title('Scatter Plot of Claim Cost vs. Driver Age')
ax_3[0, 0].set_xlabel('Driver Age')
ax_3[0, 0].set_ylabel('Claim Cost')

#Plot a histogram of the driver age distribution in the claims
sns.distplot(df_no_zeros['driverAge'], bins=10, ax = ax_3[0, 1])
ax_3[0, 1].set_title('Histogram of the driver age distribution')
ax_3[0, 1].set_xlabel('Driver Age')
ax_3[0, 1].set_ylabel('Frequency')
ax_3[0, 1].axvline(np.mean(df_no_zeros['driverAge']), ls='--', c='r
    ', label='Mean')
ax_3[0, 1].axvline(np.median(df_no_zeros['driverAge']), ls=':', c='
    g', label='Median')

#Get the mean claimCost for every age and store it
mean_claimCost_by_age = df_no_zeros.groupby('driverAge')['claimCost
    '].mean()

#Create a line plot showing the mean cost for every age
mean_claimCost_by_age.plot(ax=ax_3[1, 0])
ax_3[1, 0].set_title('Mean claim cost for every age')
ax_3[1, 0].set_xlabel('Driver Age')
ax_3[1, 0].set_ylabel('Mean claim cost')

#Get the total claimCost for every driver age and store it
total_claimCost_by_age = df_no_zeros.groupby('driverAge')['
    claimCost'].sum()

#Create a line plot showing the total cost for every driver age
total_claimCost_by_age.plot(ax=ax_3[1, 1])
ax_3[1, 1].set_title('Total claim cost for every individual age')
ax_3[1, 1].set_xlabel('Driver Age')
ax_3[1, 1].set_ylabel('Total claim cost (SEK)')



#
#
#Vehicle Age plots
#
#

#Create different datasets for different vehicle age groups
```

```
#Vehicle Age group 0-25
df_car_0_to_25 = df_no_zeros[df_no_zeros['vehicleAge'] <= 25]

#Vehicle Age group 25-50
df_car_25_to_50 = df_no_zeros[(df_no_zeros['vehicleAge'] > 25) & (
    df_no_zeros['vehicleAge'] <= 50)]

#Vehicle Age group 50-75
df_car_50_to_75 = df_no_zeros[(df_no_zeros['vehicleAge'] > 50) & (
    df_no_zeros['vehicleAge'] <= 75)]

#Vehicle Age group 75-100
df_car_75_to_100 = df_no_zeros[(df_no_zeros['vehicleAge'] > 75) & (
    df_no_zeros['vehicleAge'] <= 100)]

#Create a figure to show the claim cost distribution in  vehicle
    age groups
fig, ax_4 = plt.subplots(nrows=2, ncols=2, figsize=(12, 12),
    squeeze=False)

#Plot the df_car_0_to_25 dataset as a histogram
sns.distplot(df_car_0_to_25['claimCost'], ax = ax_4[0, 0])
ax_4[0, 0].set_title('Claim cost distribution with vehicle age 0 -
    25')
ax_4[0, 0].set_xlabel('Claim cost')
ax_4[0, 0].axvline(np.mean(df_car_0_to_25['claimCost']), ls='--', c
    ='r', label='Mean')
ax_4[0, 0].axvline(np.median(df_car_0_to_25['claimCost']), ls=':',
    c='g', label='Median')

#Plot the df_car_25_to_50 dataset as a histogram
sns.distplot(df_car_25_to_50['claimCost'], ax = ax_4[0, 1])
ax_4[0, 1].set_title('Claim cost distribution with vehicle age 25 -
     50')
ax_4[0, 1].set_xlabel('Claim cost')
ax_4[0, 1].axvline(np.mean(df_car_25_to_50['claimCost']), ls='--',
    c='r', label='Mean')
ax_4[0, 1].axvline(np.median(df_car_25_to_50['claimCost']), ls=':',
     c='g', label='Median')

#Plot the df_car_50_to_75 dataset as a histogram
sns.distplot(df_car_50_to_75['claimCost'], ax = ax_4[1, 0])
ax_4[1, 0].set_title('Claim cost distribution with vehicle age 50 -
     75')
ax_4[1, 0].set_xlabel('Claim cost')
ax_4[1, 0].axvline(np.mean(df_car_50_to_75['claimCost']), ls='--',
    c='r', label='Mean')
ax_4[1, 0].axvline(np.median(df_car_50_to_75['claimCost']), ls=':',
     c='g', label='Median')

#Plot the df_car_75_to_100 dataset as a histogram
sns.distplot(df_car_75_to_100['claimCost'], ax = ax_4[1, 1])
ax_4[1, 1].set_title('Claim cost distribution with vehicle age 75 -
     100')
ax_4[1, 1].set_xlabel('Claim cost')
ax_4[1, 1].axvline(np.mean(df_car_75_to_100['claimCost']), ls='--',
     c='r', label='Mean')
```

```
ax_4[1, 1].axvline(np.median(df_car_75_to_100['claimCost']), ls
    =':', c='g', label='Median')

fig.suptitle('Claim cost distribution for different vehicle age
    groups')

#
#Data Analysis Plots
#

#Create a figure for all the different plots
fig, ax_5 = plt.subplots(nrows=2, ncols=2, figsize=(15, 15))

#Create a scatter plot of vehicle Age vs Claim Cost
df_no_zeros.plot(kind='scatter', x='vehicleAge', y='claimCost', ax=
    ax_5[0, 0])
ax_5[0, 0].set_title('Scatter Plot of Claim Cost vs. Vehicle Age')
ax_5[0, 0].set_xlabel('Vehicle Age')
ax_5[0, 0].set_ylabel('Claim Cost')


#Get a histogram of the vehicle age distribution in the claims
sns.distplot(df_no_zeros['vehicleAge'], bins=10, ax = ax_5[0, 1])
ax_5[0, 1].set_title('Histogram of the vehicle age distribution')
ax_5[0, 1].set_xlabel('Vehicle Age')
ax_5[0, 1].set_ylabel('Frequency')
ax_5[0, 1].axvline(np.mean(df_no_zeros['vehicleAge']), ls='--', c='
    r', label='Mean')
ax_5[0, 1].axvline(np.median(df_no_zeros['vehicleAge']), ls=':', c
    ='g', label='Median')

#Get the mean claimCost for every vehicle age and store it
mean_claimCost_by_vehicleAge = df_no_zeros.groupby('vehicleAge')['
    claimCost'].mean()

#Create a plot showing the mean cost for every vehicle age
mean_claimCost_by_vehicleAge.plot(ax=ax_5[1, 0])
ax_5[1, 0].set_title('Mean claim cost for every vehicle age')
ax_5[1, 0].set_xlabel('Vehicle Age')
ax_5[1, 0].set_ylabel('Mean claim cost')

#Get the total claimCost for every vehicle age and store it
total_claimCost_by_vehicle_age = df_no_zeros.groupby('vehicleAge')
    ['claimCost'].sum()

#Create a line plot showing the total cost for every vehicle age
total_claimCost_by_vehicle_age.plot(ax=ax_5[1, 1])
ax_5[1, 1].set_title('Total claim cost for every individual vehicle
     age')
ax_5[1, 1].set_xlabel('vehicle Age')
ax_5[1, 1].set_ylabel('Total claim cost (SEK)')


plt.show()

#This code printed some summary statistics needed for the report
```

```
# print('!!!!')
# print(np.mean(df_car_0_to_25['claimCost']))
# print(np.mean(df_car_25_to_50['claimCost']))
# print(np.mean(df_car_50_to_75['claimCost']))
# print(np.mean(df_car_75_to_100['claimCost']))
# print('!!!!')

# print('!!!!')
# print(np.median(df_car_0_to_25['claimCost']))
# print(np.median(df_car_25_to_50['claimCost']))
# print(np.median(df_car_50_to_75['claimCost']))
# print(np.median(df_car_75_to_100['claimCost']))
# print('!!!!')
```

5.2 Geographic Zone Data Analysis Code

```
df_origin = pd.read_excel("assignment2data.xlsx")
#print(df_origin.info())
df = df_origin[df_origin["claimCost"] > 0]
df = df[['zone', 'claimCost']]
print(df.info())


#
#
# DATA ANALYSIS SECTION
#
# Summary Statistics

summary_stats = df.groupby('zone')['claimCost'].describe()
print(summary_stats)

# Create subplots
fig, axes = plt.subplots(nrows=2, ncols=4, figsize=(16, 8))

# Flatten the axes array to access each subplot
axes = axes.ravel()

# Loop through zones and create histograms
# for the claim cost distributions
for i in range(0,7):
    ax = axes[i]
    dataPoints = df[df['zone'] == i+1]['claimCost']
    ax.hist(dataPoints, bins=20, rwidth=0.8, density=True)
    ax.axvline(np.mean(dataPoints), ls='--', c='r', label="Mean")
    ax.axvline(np.median(dataPoints), ls=':', c='g', label="Median
    ")
    ax.set_title(f'Zone {i+1}')
    ax.set_xlabel('Claim Cost')

    # Fit an Exponential distribution
    # using the mean value of claim costs
    # as an estimator for the scale
    sampleMean = np.mean(df[df['zone'] == i+1]['claimCost'])
    fitExpDist = stats.expon(scale = sampleMean)
    xs = np.arange(min(dataPoints), max(dataPoints), 0.1)
```

```
    ys = fitExpDist.pdf(xs)
    ax.plot(xs, ys, color="green")

    # Fit a Gamma distribution
    if (i+1) != 7:
        xs = np.linspace(-1, max(dataPoints), 100)
        fitShape, fitLoc, fitScale = stats.gamma.fit(dataPoints)
        fitGamDist = stats.gamma(a = fitShape, loc = fitLoc, scale
    = fitScale)
        ys = fitGamDist.pdf(xs)
        ax.plot(xs, ys, color="pink")


    # Run the Kolmogorov Smirnov Test
    print("KS Test Region Expo" + str(i+1) + " :" + str(stats.
    kstest(dataPoints, fitExpDist.cdf)))
    print("KS Test Region Gamma" + str(i+1) + " :" + str(stats.
    kstest(dataPoints, fitGamDist.cdf)))

fig.tight_layout()
plt.show()
```

5.3 Stochastic Simulation Code

```
#
#
# SIMULATION ZONE
#
#
import numpy as np
import matplotlib.pyplot as plt

# Define parameters
n_claims = len(df)
lam = n_claims / (5*365)       #average arrival rate per day
                        # total number of claims / the period of
    claim sampling  January 1994- December 1998
                        # 670/5*356
                        # or 64548
print(n_claims)
#expo = stats.expon(scale = 1/lam)

claim_region_elements = [1, 2, 3, 4, 5, 6, 7]
# estimate the region distribution
# using the historical sample
claim_region_distribution = [len(df[df['zone'] == 1])/n_claims,
                             len(df[df['zone'] == 2])/n_claims,
                             len(df[df['zone'] == 3])/n_claims,
                             len(df[df['zone'] == 4])/n_claims,
                             len(df[df['zone'] == 5])/n_claims,
                             len(df[df['zone'] == 6])/n_claims,
                             len(df[df['zone'] == 7])/n_claims]  #
    Replace with your region distribution

fitShape1, fitLoc1, fitScale1 = stats.gamma.fit(df[df['zone'] ==
    1]['claimCost'])
```

```
fitShape2, fitLoc2, fitScale2 = stats.gamma.fit(df[df['zone'] ==
    2]['claimCost'])
fitShape4, fitLoc4, fitScale4 = stats.gamma.fit(df[df['zone'] ==
    4]['claimCost'])
fitShape5, fitLoc5, fitScale5 = stats.gamma.fit(df[df['zone'] ==
    5]['claimCost'])


claim_amount_distribution = [stats.gamma(a = fitShape1, loc =
    fitLoc1, scale = fitScale1),
                             stats.gamma(a = fitShape2, loc =
    fitLoc2, scale = fitScale2),
                             stats.expon(np.mean(df[df['zone'] ==
    3]['claimCost'])),
                             stats.gamma(a = fitShape4, loc =
    fitLoc4, scale = fitScale4),
                             stats.gamma(a = fitShape5, loc =
    fitLoc5, scale = fitScale5),
                             stats.expon(np.mean(df[df['zone'] ==
    6]['claimCost'])),
                             stats.expon(np.mean(df[df['zone'] ==
    7]['claimCost']))]  # Replace with your amount distribution



# Simulation setup
num_simulations = 10000
total_claims = []
total_claim_costs = []

for _ in range(num_simulations):
    time_elapsed = 0
    claims = 0
    claim_cost = 0

    while time_elapsed < 365:  # 1 year in days
        # Sample interarrival time
        interarrival_time = np.random.exponential(1 / lam)
        time_elapsed += interarrival_time
        if time_elapsed <= 365:
            # Sample claim region and amount
            claim_region = np.random.choice(a=claim_region_elements
    , size=1, p=claim_region_distribution)

            claim_amount = claim_amount_distribution[claim_region
    [0] - 1].rvs()

            # Update claims and claim cost
            claims += 1
            claim_cost += claim_amount

    total_claims.append(claims)
    total_claim_costs.append(claim_cost)

# Calculate and print statistics
mean_claims = np.mean(total_claims)
std_claims = np.std(total_claims)
```

```
mean_claim_costs = np.mean(total_claim_costs)
std_claim_costs = np.std(total_claim_costs)

# Plot histograms
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
plt.hist(total_claims, bins=30, density=True, edgecolor='k')
# Plot the normal distribution over it
mu = mean_claims
sigma = std_claims
x = np.linspace(mu - 3 * sigma, mu + 3 * sigma, 100)
pdf = stats.norm.pdf(x, mu, sigma)
plt.plot(x, pdf, 'r-', lw=2)  # 'r-' indicates a red solid line
plt.title('Number of Claims')
plt.subplot(1, 2, 2)
plt.hist(total_claim_costs, bins=30, density=True, edgecolor='k')
# Plot the normal distribution over it
mu = mean_claim_costs
sigma = std_claim_costs
x = np.linspace(mu - 3 * sigma, mu + 3 * sigma, 100)
pdf = stats.norm.pdf(x, mu, sigma)
#
plt.plot(x, pdf, 'r-', lw=2)  # 'r-' indicates a red solid line
plt.title('Total Claim Costs')
plt.show()

# Interpret the results and report your findings
print(f"Mean Number of Claims: {mean_claims}")
print(f"Standard Deviation of Number of Claims: {std_claims}")
print(f"Mean Total Claim Costs: {mean_claim_costs}")
print(f"Standard Deviation of Total Claim Costs: {std_claim_costs
    }")
```