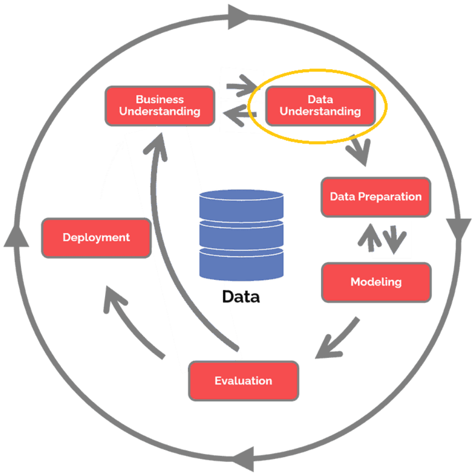
CRISP-DM Data Understanding and Preparation

Jesús David García

September 7,2023

The purpose of this document is to implement the CRISP-DM methodology [1] for a hypothetical "Chain Ladder Method Project for Provision Estimation" for the company *Chubb Limited*. More specifically, it will follow IBM´s steps for the second and third phase of the CRISP-DM process, which is denominated *Data Understanding and Preparation*. The company´s information was obtained from its Annual Report [2] and its webpage [3]. This project might include hypothetical information in order to simulate the business case (for example, in the section "current solution"). For code reference, please visit our research notebook at this link. Data can be downloaded at this other link.



Data Understanding

This phase seeks to examine the general properties of the data, explore some of its characteristics, revise its quality, and report on the results. More specifically, it will go into the following details:

Data Description Data Exploration Data Quality

Data Description

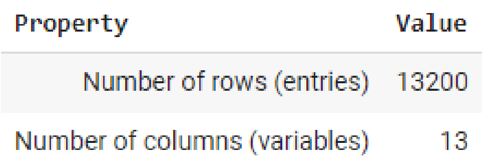
This dataset has the historic claims on workers compensation insurance for several years. From the company's database we had access to the following variables:

- **GRCODE**: NAIC (National Association of Insurance Commissioners) company code, a unique identifier for each insurance company or group.
- **GRNAME**: NAIC company name, the name of the insurance company or group corresponding to the NAIC code.
- **AccidentYear**: The year in which accidents or claims occurred, ranging from 1988 to 1997.
- **DevelopmentYear**: The year in which the claim is developed or reported, with a range of up to 10 years since the accident year.

- **DevelopmentLag**: The development lag, likely calculated as (AY-1987 + DY-1987 - 1), representing the time period between when the accident occurred and when the losses were reported or settled.
- **IncurLoss**: Incurred losses and allocated expenses reported at the end of the specified year, containing financial data related to costs incurred by the insurer due to claims and expenses.
- **CumPaidLoss**: Cumulative paid losses and allocated expenses at the end of the specified year, representing the total amount paid out by the insurer for claims and expenses up to that year.
- **BulkLoss**: Bulk and IBNR (Incurred But Not Reported) reserves on net losses and defense and cost containment expenses reported at the end of the year, potentially containing data related to reserves set aside for future claims and expenses.
- **PostedReserve97**: Posted reserves in the year 1997, taken from the Underwriting and Investment Exhibit – Part 2A, including net losses unpaid and unpaid loss adjustment expenses, likely representing a specific reserve amount for the year 1997.
- **EarnedPremDIR**: Premiums earned at the incurral year - direct and assumed, potentially containing data related to premiums earned by the insurer for policies issued in the specified year.
- **EarnedPremCeded**: Premiums earned at the incurral year - ceded, representing premiums earned but then ceded to reinsurance companies.
- **EarnedPremNet**: Premiums earned at the incurral year - net, likely representing the net premiums earned after accounting for both direct and ceded premiums.
- **Single**: A binary indicator where 1 indicates a single insurance entity, and 0 indicates a group insurer. This column could be used to classify insurance companies as either standalone entities or part of a group.

We will proceed further describing the variables by answering some initial questions:

0.0.1 How big is the dataset?

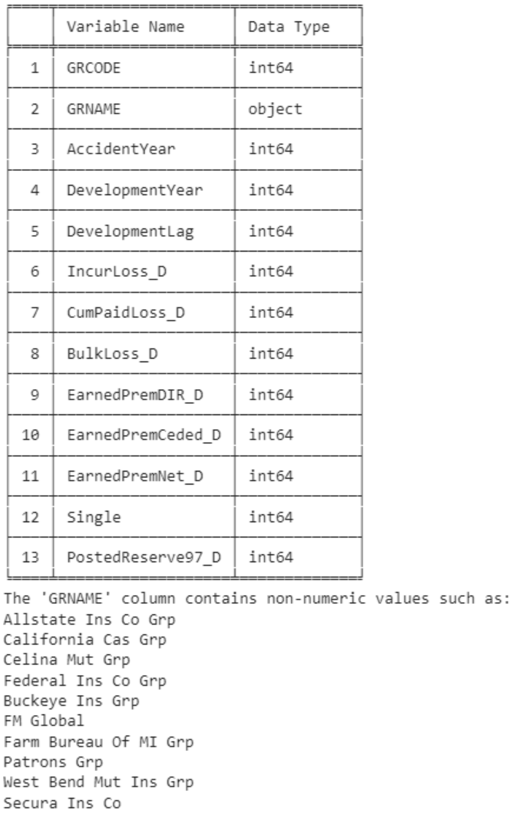


Property	Value
Number of rows (entries)	13200
Number of columns (variables)	13

We have a total of 13200 entries among 13 variables. The dataset size is 13 x 13200.

0.0.2 What are the data types of the dataset?

As expected, all variables except for GRNAME are integers. Checking further into "GRNAME", which appeared as "object", we found it only contains non-numeric values. Printing some of its values, we can see they effectively correspond to companies names:

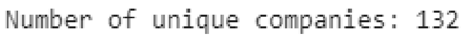


	Variable Name	Data Type
1	GRCODE	int64
2	GRNAME	object
3	AccidentYear	int64
4	DevelopmentYear	int64
5	DevelopmentLag	int64
6	IncurLoss_D	int64
7	CumPaidLoss_D	int64
8	BulkLoss_D	int64
9	EarnedPremDIR_D	int64
10	EarnedPremCeded_D	int64
11	EarnedPremNet_D	int64
12	Single	int64
13	PostedReserve97_D	int64

```
The 'GRNAME' column contains non-numeric values such as:
Allstate Ins Co Grp
California Cas Grp
Celina Mut Grp
Federal Ins Co Grp
Buckeye Ins Grp
FM Global
Farm Bureau Of MI Grp
Patrons Grp
West Bend Mut Ins Grp
Secura Ins Co
```

Data Exploration

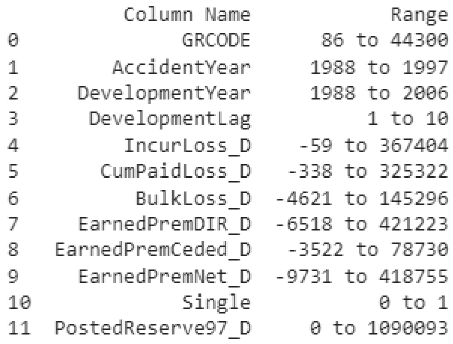
0.0.3 What are the data types of the dataset?



```
Number of unique companies: 132
```

There 132 companies, which means we will have 132 run-off triangles for our model and analysis. This should be enough data for a robust analysis.

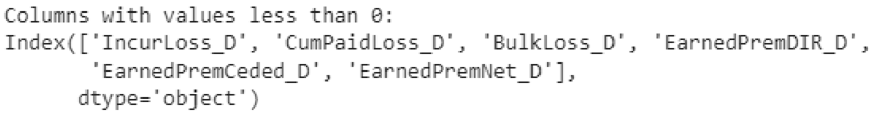
0.0.4 What is the range for the numeric variables?



```
          Column Name            Range
0              GRCODE      86 to 44300
1        AccidentYear     1988 to 1997
2     DevelopmentYear     1988 to 2006
3      DevelopmentLag          1 to 10
4         IncurLoss_D    -59 to 367404
5       CumPaidLoss_D   -338 to 325322
6          BulkLoss_D  -4621 to 145296
7     EarnedPremDIR_D  -6518 to 421223
8   EarnedPremCeded_D   -3522 to 78730
9     EarnedPremNet_D  -9731 to 418755
10             Single           0 to 1
11  PostedReserve97_D     0 to 1090093
```

Surprisingly, some variables have negative and positive values. As a first thought, this shouldn´t be the case as these columns are losses, and should be expressed as an absolute term. Are these negative values an income for the company? If that is the case, how does this income work? Earned Premiums also have negative values, which also seems weird as it is supposed to be the income for the company.

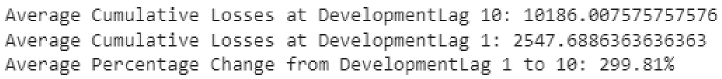
0.0.5 Are there negative values in the numeric columns? (Double Check)



```
Columns with values less than 0:
Index(['IncurLoss_D', 'CumPaidLoss_D', 'BulkLoss_D', 'EarnedPremDIR_D',
       'EarnedPremCeded_D', 'EarnedPremNet_D'],
      dtype='object')
```

Just as a double check, we indeed confirm there are several columns with values less than 0.

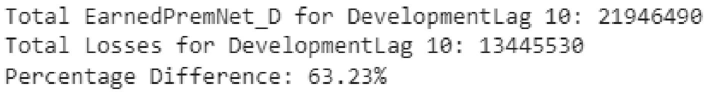
0.0.6 How much does Cumulative Losses change from Year of the accident to ten years later (development year == 10)?



```
Average Cumulative Losses at DevelopmentLag 10: 10186.007575757576
Average Cumulative Losses at DevelopmentLag 1: 2547.6886363636363
Average Percentage Change from DevelopmentLag 1 to 10: 299.81%
```

The average difference from Year 1 to Year 10 on the claims is 299.81%. This reinforces the importance of an accurate model to predict provisions values for the company to have an adequate and realistic reserve for future claims. Doing a broad claim, an insurer can expect the ultimate expense for a year to be about 4 times the expenses for the claims reported on the first year.

0.0.7 What is the difference between Cumulative Losses and Net Premiums Earned (i.e., how much is earned on average)?

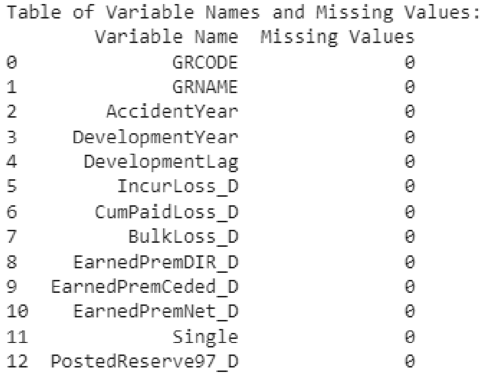


```
Total EarnedPremNet_D for DevelopmentLag 10: 21946490
Total Losses for DevelopmentLag 10: 13445530
Percentage Difference: 63.23%
```

The insurance industry is quite profitable. The difference between the earned net premiums and the cumulative losses at year 10 after an accident is 63.23

Data Quality

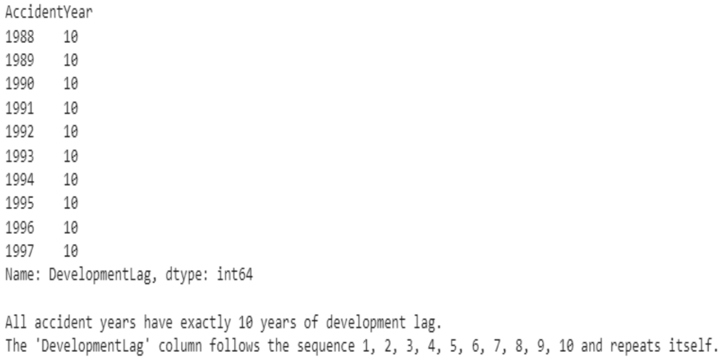
0.0.8 Are there missing values in any column?



```
Table of Variable Names and Missing Values:
        Variable Name  Missing Values
0              GRCODE               0
1              GRNAME               0
2        AccidentYear               0
3     DevelopmentYear               0
4      DevelopmentLag               0
5         IncurLoss_D               0
6       CumPaidLoss_D               0
7          BulkLoss_D               0
8     EarnedPremDIR_D               0
9   EarnedPremCeded_D               0
10    EarnedPremNet_D               0
11             Single               0
12  PostedReserve97_D               0
```

There are no missing values in the dataset, so NA values won't be a problem for the analysis.

0.0.9 Is it the case that every accident year has 10 years of development lag?



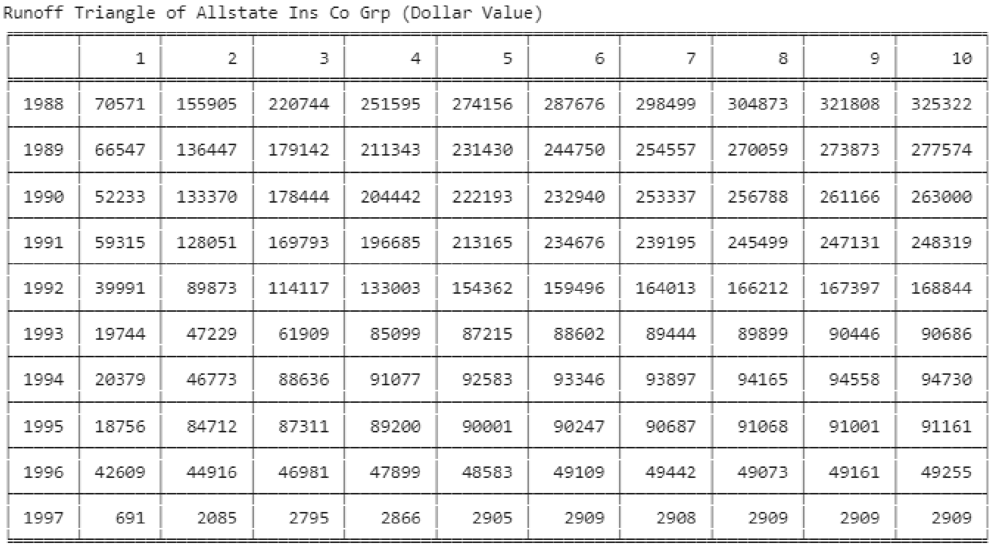
```
AccidentYear
1988    10
1989    10
1990    10
1991    10
1992    10
1993    10
1994    10
1995    10
1996    10
1997    10
Name: DevelopmentLag, dtype: int64

All accident years have exactly 10 years of development lag.
The 'DevelopmentLag' column follows the sequence 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and repeats itself.
```

All years have 10 development lags, so the run-off triangles will be complete and won't have missing data problems.

Data Understanding

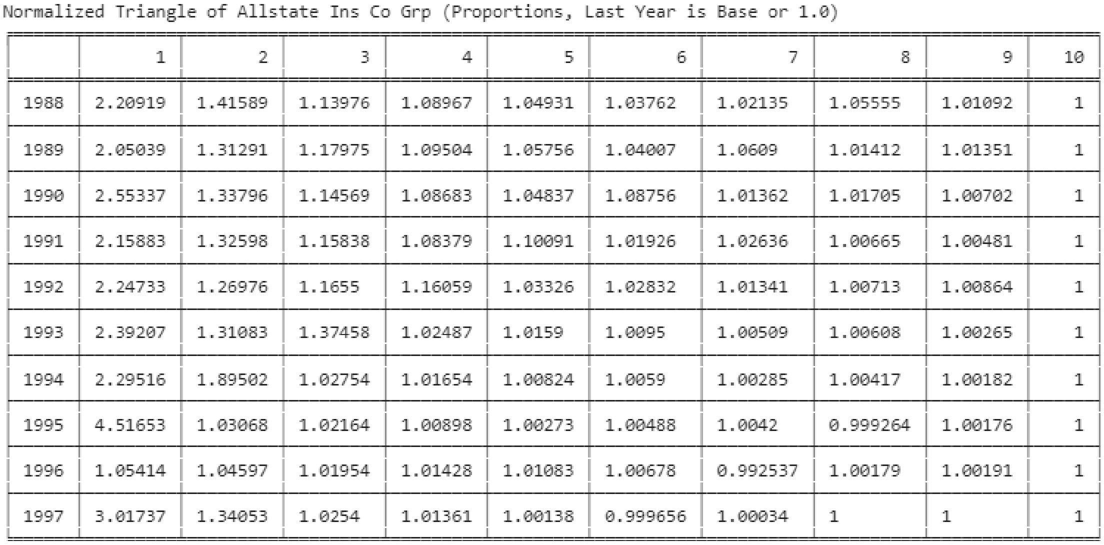
Building the Run Off Triangles for 1 Company

We will build a run-off triangle for the first company in the list, to check the format of the triangles:



Runoff Triangle of Allstate Ins Co Grp (Dollar Value)

	1	2	3	4	5	6	7	8	9	10
1988	70571	155905	220744	251595	274156	287676	298499	304873	321808	325322
1989	66547	136447	179142	211343	231430	244750	254557	270059	273873	277574
1990	52233	133370	178444	204442	222193	232940	253337	256788	261166	263000
1991	59315	128051	169793	196685	213165	234676	239195	245499	247131	248319
1992	39991	89873	114117	133003	154362	159496	164013	166212	167397	168844
1993	19744	47229	61909	85099	87215	88602	89444	89899	90446	90686
1994	20379	46773	88636	91077	92583	93346	93897	94165	94558	94730
1995	18756	84712	87311	89200	90001	90247	90687	91068	91001	91161
1996	42609	44916	46981	47899	48583	49109	49442	49073	49161	49255
1997	691	2085	2795	2866	2905	2909	2908	2909	2909	2909



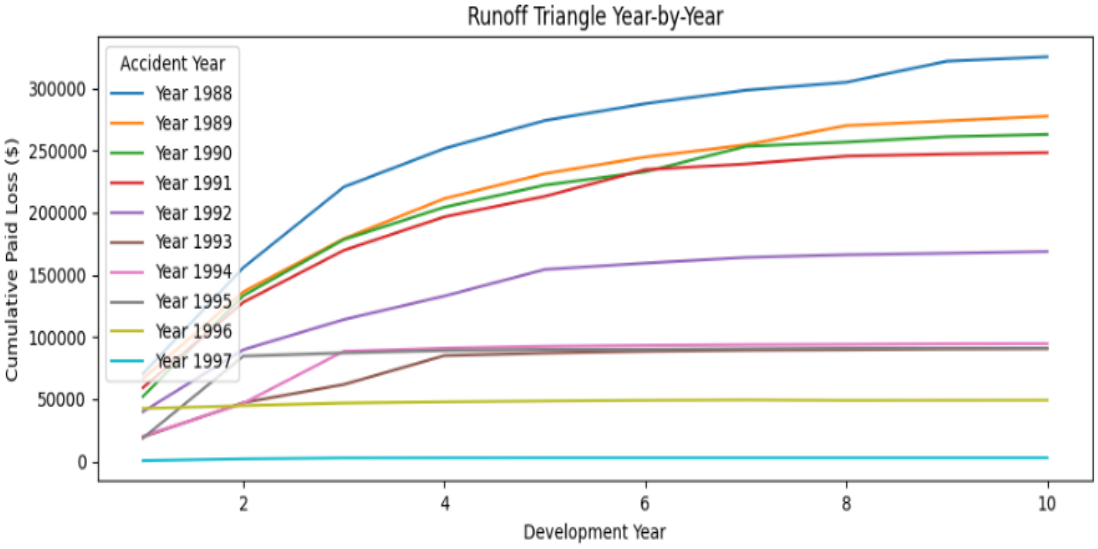
Normalized Triangle of Allstate Ins Co Grp (Proportions, Last Year is Base or 1.0)

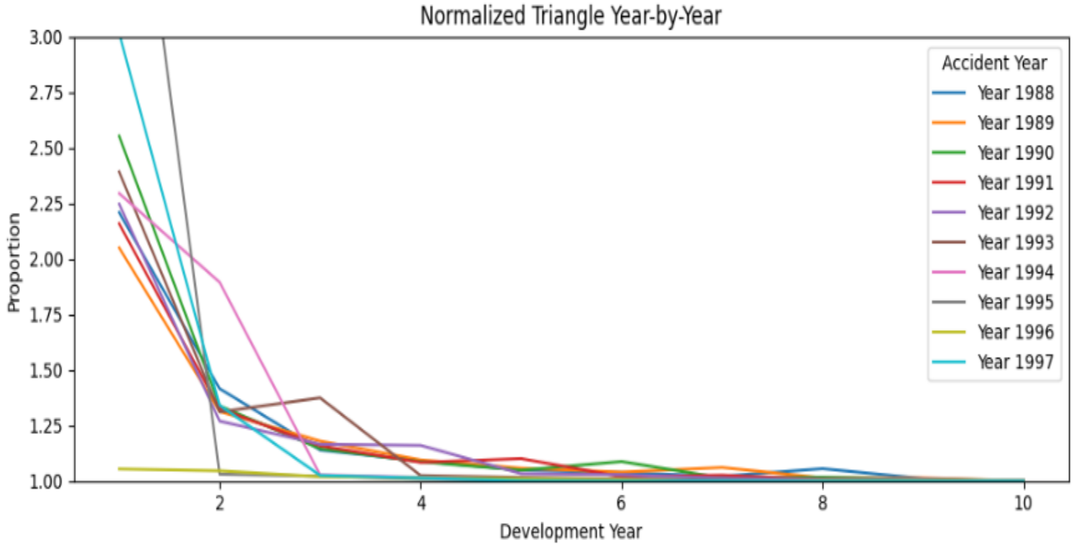
	1	2	3	4	5	6	7	8	9	10
1988	2.20919	1.41589	1.13976	1.08967	1.04931	1.03762	1.02135	1.05555	1.01092	1
1989	2.05039	1.31291	1.17975	1.09504	1.05756	1.04007	1.0609	1.01412	1.01351	1
1990	2.55337	1.33796	1.14569	1.08683	1.04837	1.08756	1.01362	1.01705	1.00702	1
1991	2.15883	1.32598	1.15838	1.08379	1.10091	1.01926	1.02636	1.00665	1.00481	1
1992	2.24733	1.26976	1.1655	1.16059	1.03326	1.02832	1.01341	1.00713	1.00864	1
1993	2.39207	1.31083	1.37458	1.02487	1.0159	1.0095	1.00509	1.00608	1.00265	1
1994	2.29516	1.89502	1.02754	1.01654	1.00824	1.0059	1.00285	1.00417	1.00182	1
1995	4.51653	1.03068	1.02164	1.00898	1.00273	1.00488	1.0042	0.999264	1.00176	1
1996	1.05414	1.04597	1.01954	1.01428	1.01083	1.00678	0.992537	1.00179	1.00191	1
1997	3.01737	1.34053	1.0254	1.01361	1.00138	0.999656	1.00034	1	1	1

The first run-off table was built using the Cumulative Losses for the Company Allstate Ins Co Grp. For example, losses for the accident year 1988 were $70.571 as reported for that year but amounted a total of $325.322 after 10 years. That shows that many claims for the accident year 1988 were reported several years after the incident occurred.

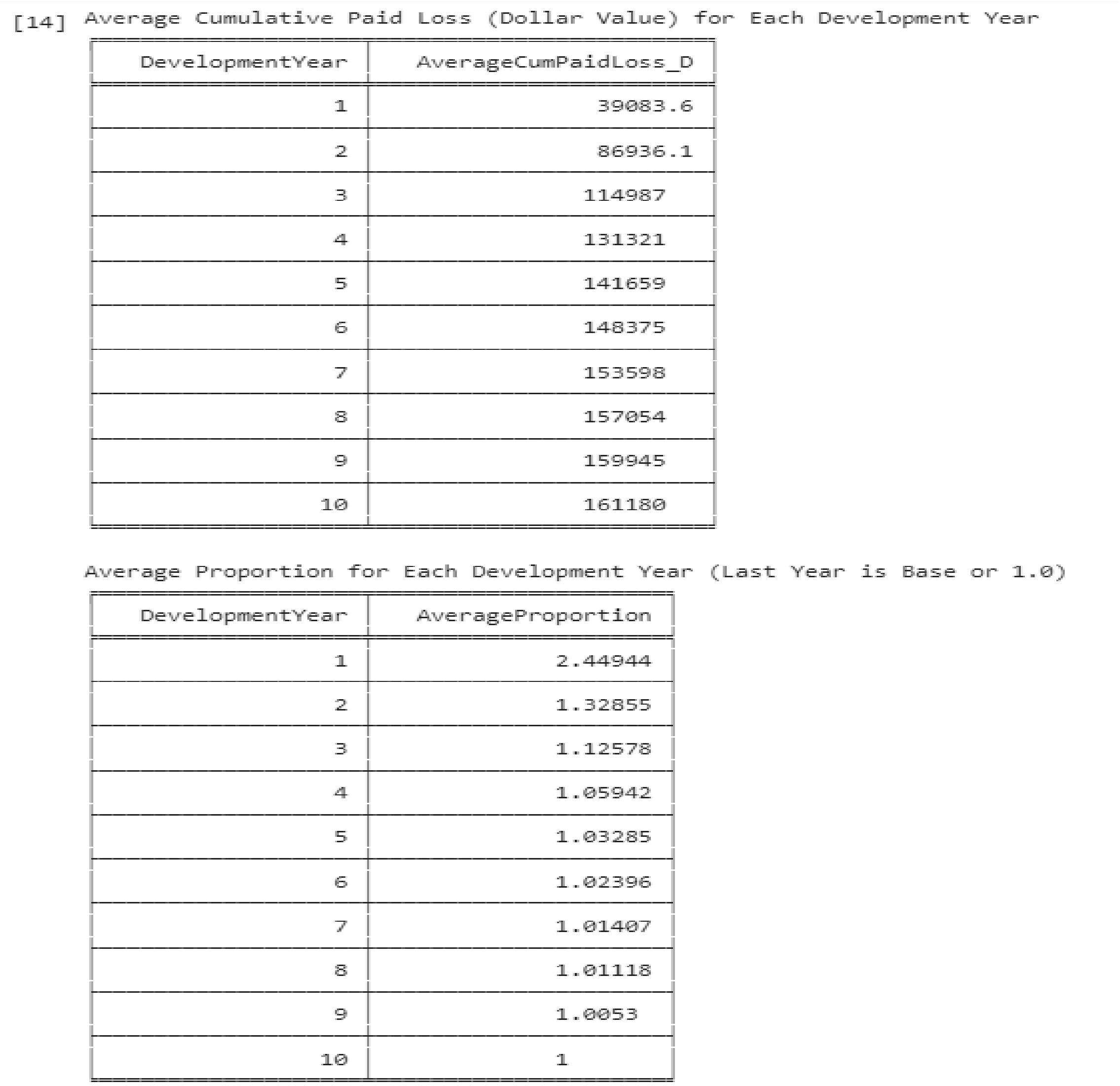
The second run-off table uses the dollar value of the last year (year 10) as the base (of 1.0), and calculates the proportion of cumulative losses of one year compared to the next one. To be more clear, the 2.20919 in the year 1 of 1988 means that the cumulative losses reported on the development year 1989 (the second year) were 2.20919 times the ones reported on 1988, the previous year.

Just to validate this, the cumulative losses as of 1988 were $70.571, and as for the next year, 1989, the value was $155.905. The value of $155.905 is effectively 2.20919 times the previous value of $70.571. If we had this information and were standing in the first year of an accident year, we know the next year we can expect cumulative losses to increase by 2.20919 times, and with this information we can make the appropiate provision.



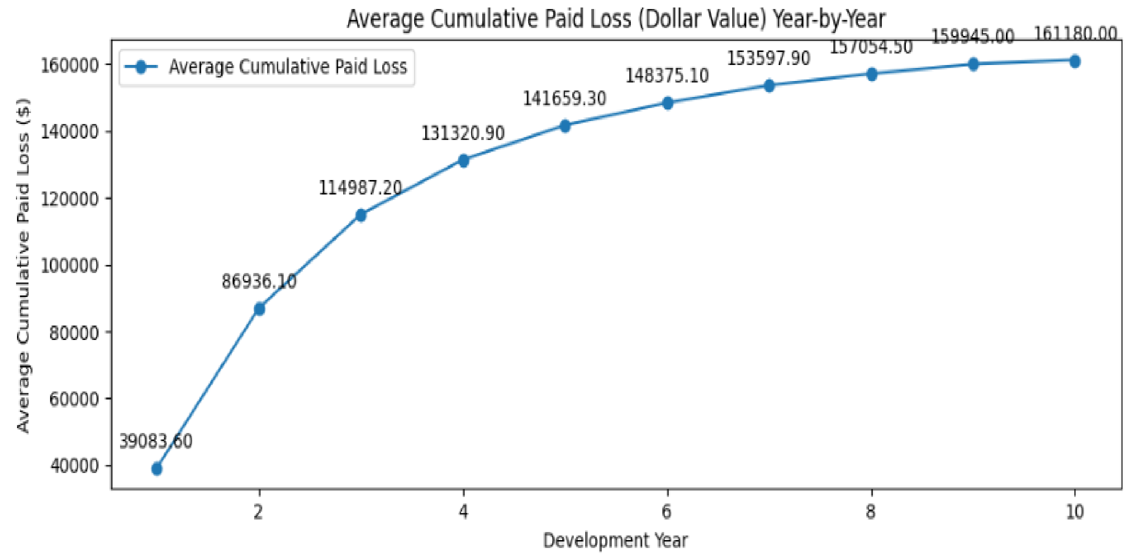


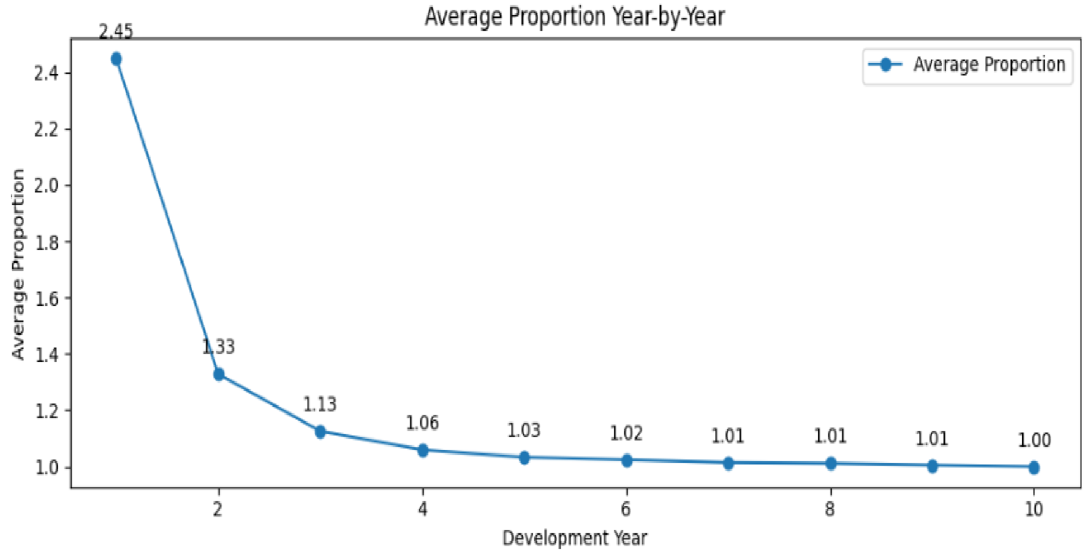
We can see the biggest growth on cumulative losses tend to be in the first 2 or 3 years (which makes sense, as the majority of the claims of an accident year will be reported as soon as possible and won't have such a big delay after a few years). However, the value of the claims keeps growing over time, so it is indeed necessary to have an adequate provision system.



```
[14] Average Cumulative Paid Loss (Dollar Value) for Each Development Year
     ╒════════════════════╤═══════════════════════╕
     │    DevelopmentYear │   AverageCumPaidLoss_D │
     ╞════════════════════╪═══════════════════════╡
     │                  1 │                39083.6 │
     ├────────────────────┼───────────────────────┤
     │                  2 │                86936.1 │
     ├────────────────────┼───────────────────────┤
     │                  3 │               114987   │
     ├────────────────────┼───────────────────────┤
     │                  4 │               131321   │
     ├────────────────────┼───────────────────────┤
     │                  5 │               141659   │
     ├────────────────────┼───────────────────────┤
     │                  6 │               148375   │
     ├────────────────────┼───────────────────────┤
     │                  7 │               153598   │
     ├────────────────────┼───────────────────────┤
     │                  8 │               157054   │
     ├────────────────────┼───────────────────────┤
     │                  9 │               159945   │
     ├────────────────────┼───────────────────────┤
     │                 10 │               161180   │
     ╘════════════════════╧═══════════════════════╛

     Average Proportion for Each Development Year (Last Year is Base or 1.0)
     ╒════════════════════╤═════════════════════╕
     │    DevelopmentYear │   AverageProportion │
     ╞════════════════════╪═════════════════════╡
     │                  1 │             2.44944 │
     ├────────────────────┼─────────────────────┤
     │                  2 │             1.32855 │
     ├────────────────────┼─────────────────────┤
     │                  3 │             1.12578 │
     ├────────────────────┼─────────────────────┤
     │                  4 │             1.05942 │
     ├────────────────────┼─────────────────────┤
     │                  5 │             1.03285 │
     ├────────────────────┼─────────────────────┤
     │                  6 │             1.02396 │
     ├────────────────────┼─────────────────────┤
     │                  7 │             1.01407 │
     ├────────────────────┼─────────────────────┤
     │                  8 │             1.01118 │
     ├────────────────────┼─────────────────────┤
     │                  9 │             1.0053  │
     ├────────────────────┼─────────────────────┤
     │                 10 │             1       │
     ╘════════════════════╧═════════════════════╛
```

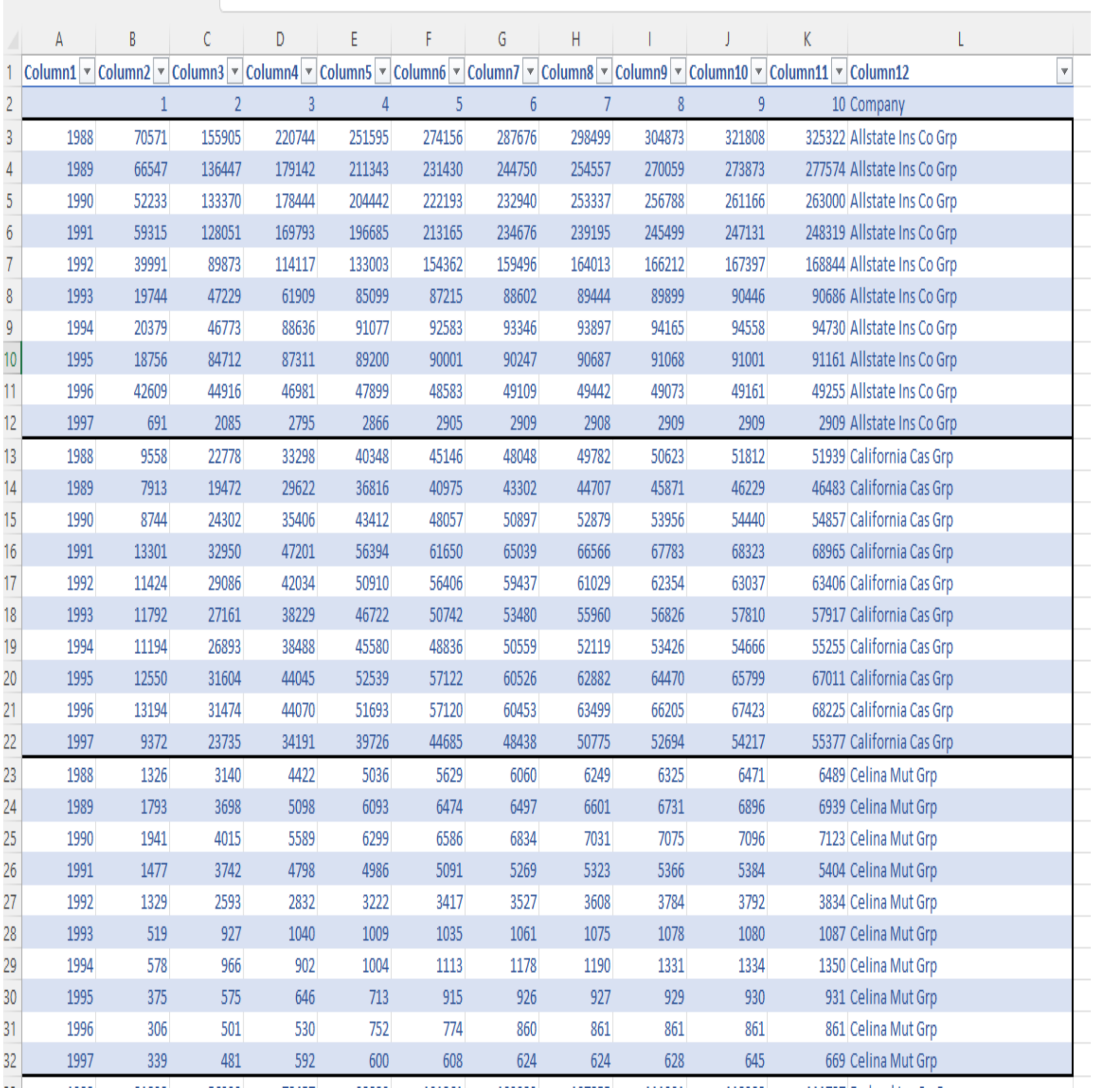
These tables show the average values of how the cumulative losses change through time. Let's build graphs to analyze this:





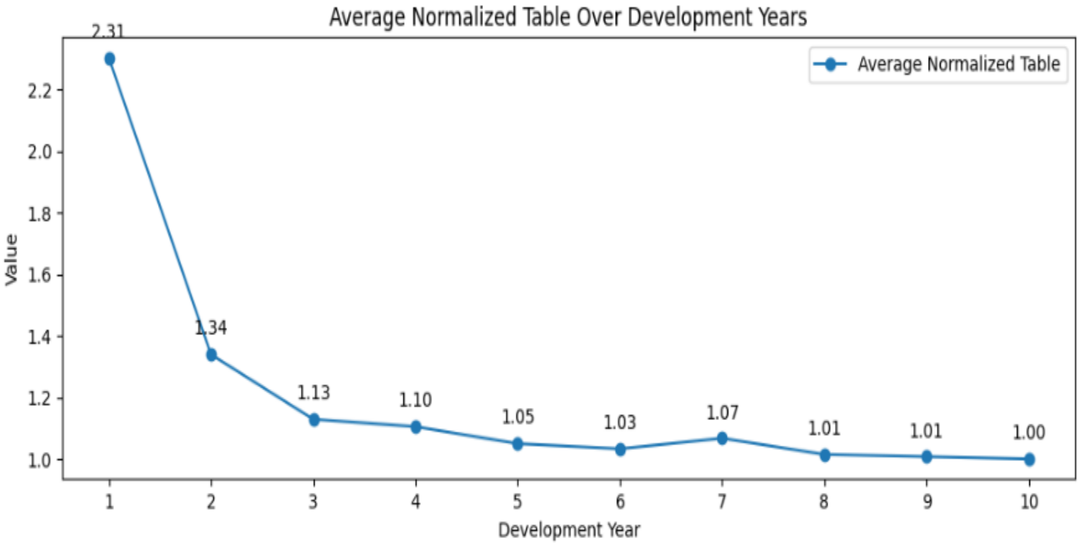
These two graphs show the average change in the cumulative losses for the company Allstate Ins Co Grp. For instance, the claims reported for the second year after an accident are, on average, 2.45 times the ones reported on the first year. Taking this into consideration, if for a year the company has losses for 100 dollars, it can expect cumulative losses to sum 245 dollars for the next year (meaning an additional 145 dollars were reported on the second year).

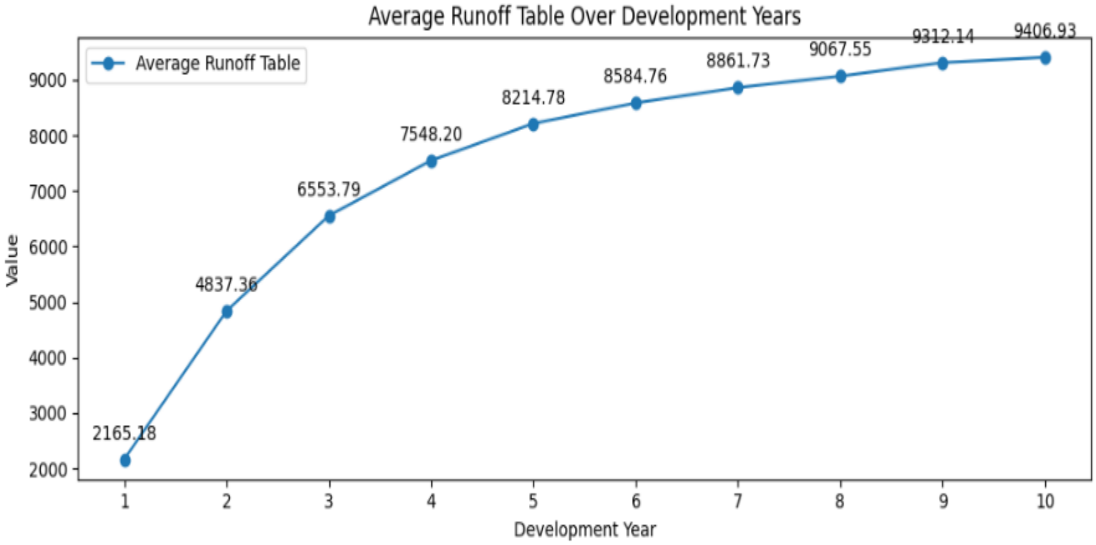
Let´s apply the previous logic for all companies. This code will create two csv files that have the run-off tables (in dollar amounts and normalized) for ALL companies:



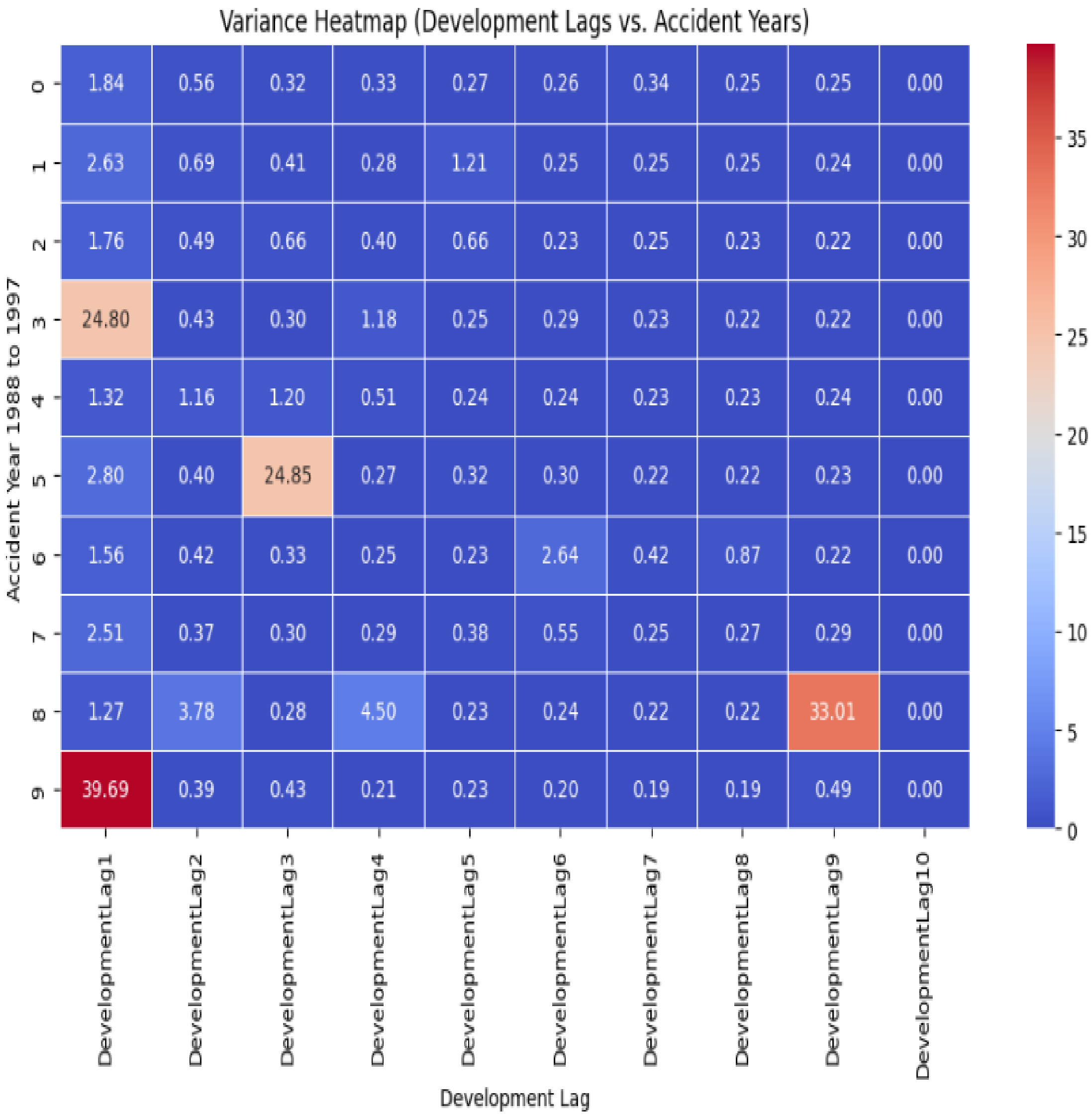
	A	B	C	D	E	F	G	H	I	J	K	L
1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12
2		1	2	3	4	5	6	7	8	9	10	Company
3	1988	70571	155905	220744	251595	274156	287676	298499	304873	321808	325322	Allstate Ins Co Grp
4	1989	66547	136447	179142	211343	231430	244750	254557	270059	273873	277574	Allstate Ins Co Grp
5	1990	52233	133370	178444	204442	222193	232940	253337	256788	261166	263000	Allstate Ins Co Grp
6	1991	59315	128051	169793	196685	213165	234676	239195	245499	247131	248319	Allstate Ins Co Grp
7	1992	39991	89873	114117	133003	154362	159496	164013	166212	167397	168844	Allstate Ins Co Grp
8	1993	19744	47229	61909	85099	87215	88602	89444	89899	90446	90686	Allstate Ins Co Grp
9	1994	20379	46773	88636	91077	92583	93346	93897	94165	94558	94730	Allstate Ins Co Grp
10	1995	18756	84712	87311	89200	90001	90247	90687	91068	91001	91161	Allstate Ins Co Grp
11	1996	42609	44916	46981	47899	48583	49109	49442	49073	49161	49255	Allstate Ins Co Grp
12	1997	691	2085	2795	2866	2905	2909	2908	2909	2909	2909	Allstate Ins Co Grp
13	1988	9558	22778	33298	40348	45146	48048	49782	50623	51812	51939	California Cas Grp
14	1989	7913	19472	29622	36816	40975	43302	44707	45871	46229	46483	California Cas Grp
15	1990	8744	24302	35406	43412	48057	50897	52879	53956	54440	54857	California Cas Grp
16	1991	13301	32950	47201	56394	61650	65039	66566	67783	68323	68965	California Cas Grp
17	1992	11424	29086	42034	50910	56406	59437	61029	62354	63037	63406	California Cas Grp
18	1993	11792	27161	38229	46722	50742	53480	55960	56826	57810	57917	California Cas Grp
19	1994	11194	26893	38488	45580	48836	50559	52119	53426	54666	55255	California Cas Grp
20	1995	12550	31604	44045	52539	57122	60526	62882	64470	65799	67011	California Cas Grp
21	1996	13194	31474	44070	51693	57120	60453	63499	66205	67423	68225	California Cas Grp
22	1997	9372	23735	34191	39726	44685	48438	50775	52694	54217	55377	California Cas Grp
23	1988	1326	3140	4422	5036	5629	6060	6249	6325	6471	6489	Celina Mut Grp
24	1989	1793	3698	5098	6093	6474	6497	6601	6731	6896	6939	Celina Mut Grp
25	1990	1941	4015	5589	6299	6586	6834	7031	7075	7096	7123	Celina Mut Grp
26	1991	1477	3742	4798	4986	5091	5269	5323	5366	5384	5404	Celina Mut Grp
27	1992	1329	2593	2832	3222	3417	3527	3608	3784	3792	3834	Celina Mut Grp
28	1993	519	927	1040	1009	1035	1061	1075	1078	1080	1087	Celina Mut Grp
29	1994	578	966	902	1004	1113	1178	1190	1331	1334	1350	Celina Mut Grp
30	1995	375	575	646	713	915	926	927	929	930	931	Celina Mut Grp
31	1996	306	501	530	752	774	860	861	861	861	861	Celina Mut Grp
32	1997	339	481	592	600	608	624	624	628	645	669	Celina Mut Grp

This is a snapshot of the .csv files we can build, which has 132 run-off triangles (1 for every company). This csv file can be obtained running the code in the colab document. Doing the same averages and graphs as previously done, we can obtain an average run-off triangle for ALL companies:

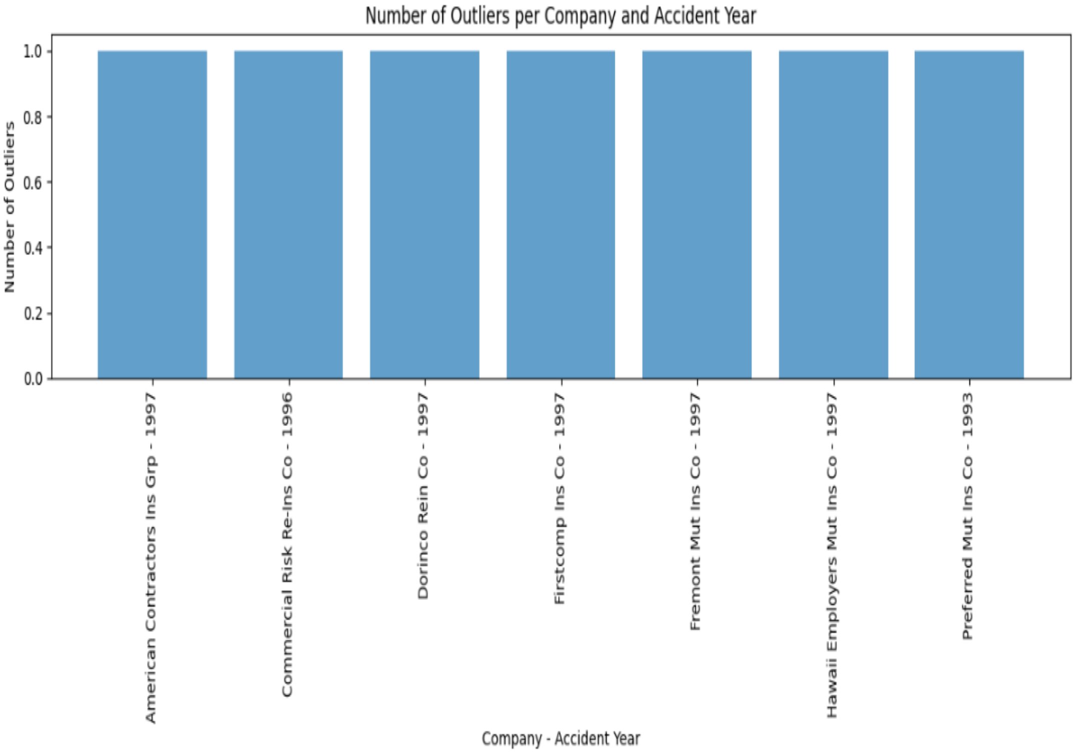




This is a snapshot of the .csv files we can build, which has 132 run-off triangles (1 for every company). This csv file can be obtained running the code in the colab document. Besides the average, it is important to know if we have outliers, that is, years that have uncommon claims and may be affecting the analysis. For this, we can take a look at the graph on the following page. From it we can see that in the development years 1,3, and 9 there are uncommon behaviours from one year to another.



We should take a further look at which companies in what years have this behaviour:



For further analysis, we should exclude this companies for this years as they had abnormal behaviours in the claims.

References

[1] Chubb Limited. (2023). Form 10-K: Annual report pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934 for the fiscal year ended December 31, 2022. United States Securities and Exchange Commission. https://investors.chubb.com/financials/sec-filings/default.aspx

[2] IBM. (n.d.). Business Understanding - IBM SPSS Modeler. IBM SPSS Modeler Documentation. https://www.ibm.com/docs/en/spss-modeler/saas?topic=guide-business-understanding

[3] Chubb. (2023, August 24). Chubb Insurance. Chubb. https://about.chubb.com/

[4] StateRequirement. (n.d.). What Is Property and Casualty Insurance? StateRequirement. https://t.ly/2U01V

[5] The Org. (n.d.). Chubb. The Org. https://theorg.com/org/chubb