Homework Assignment 5

BAN 620 - DATA MINING

Group 2

Khush Domadiya – fr9739

Chenghui Tan – sw5299

Mohanasundaram Murugesan – vw4192

Amisha Farhana Shaik – ac1425

ALLIANZ CASE

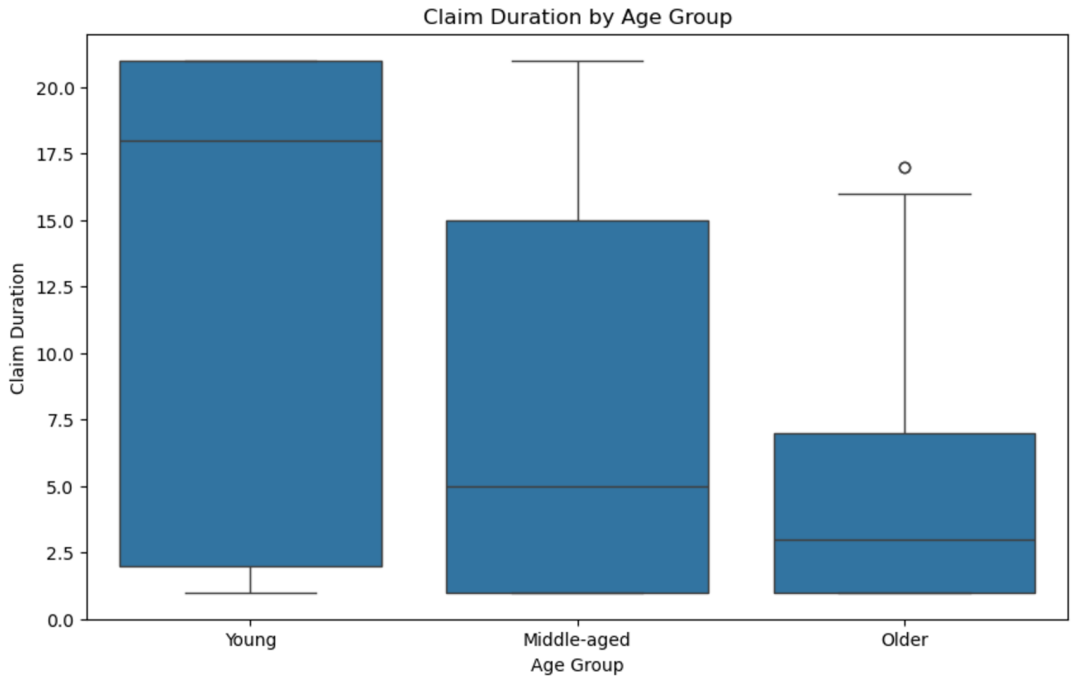
1. Use various visuals to explore the claims data. Are there any patterns among customers and brokers regarding “long” duration claims?

Customer Patterns:

a) **Age Groups:**
From the Duration column we have created another column 'Claim_Category' for long claims with threshold as 10 i.e., if the duration is greater than 10years it is a long claim and hence 'Claim_Category' column would be 1 orelse 0.
And to set this threshold we have looked at the statistics of the Duration column which showed that 75% of the datapoints are above 10.

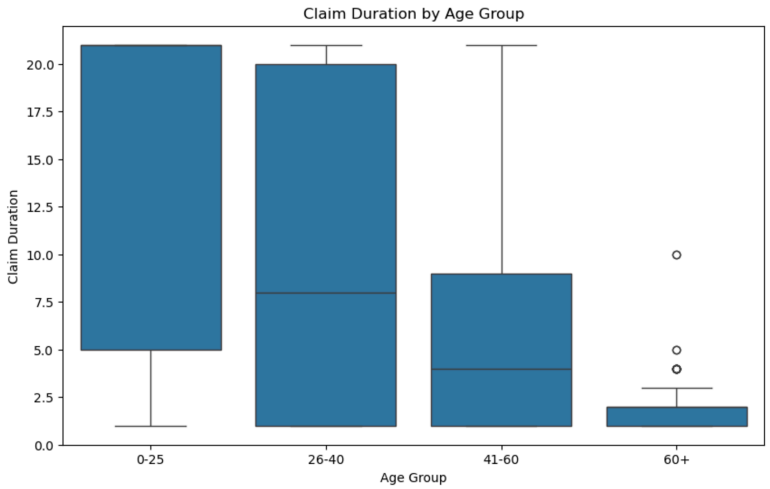
```
count    4938.000000
mean        6.509113
std         6.310578
min         1.000000
25%         1.000000
50%         4.000000
75%        10.000000
max        21.000000
Name: Duration, dtype: float64
```

	Duration	Claim_Category
3	10	1
4	1	0
5	4	0
6	1	0
7	3	0



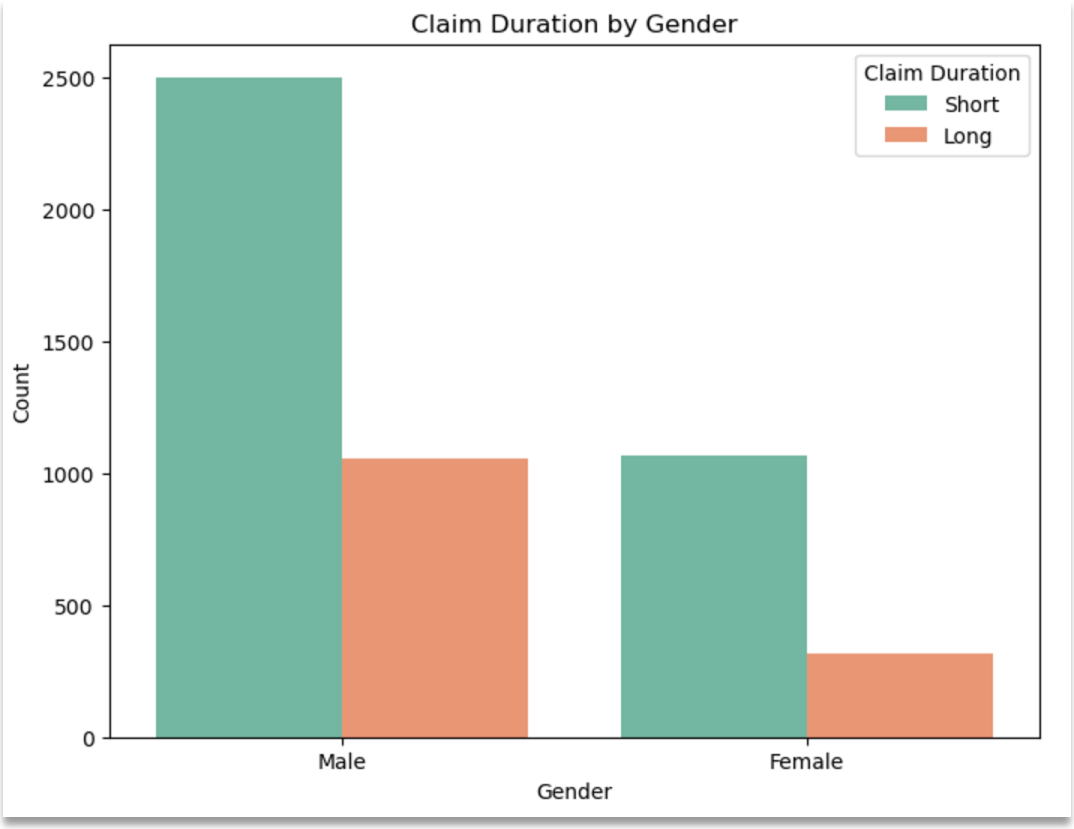
- From the above graph we can see that 50% of the Young population have very high claim durations (i.e., 17.5 and above) than that of more than 75% of the Middle-aged group (i.e., 5 and below).
- Also, 50% of the young population have very high claim durations than the entire population of older age group.

Just to better understand this pattern, we tried 4 different age groups and the main inference is still the same that the young population have higher claim durations than the rest of the age groups.



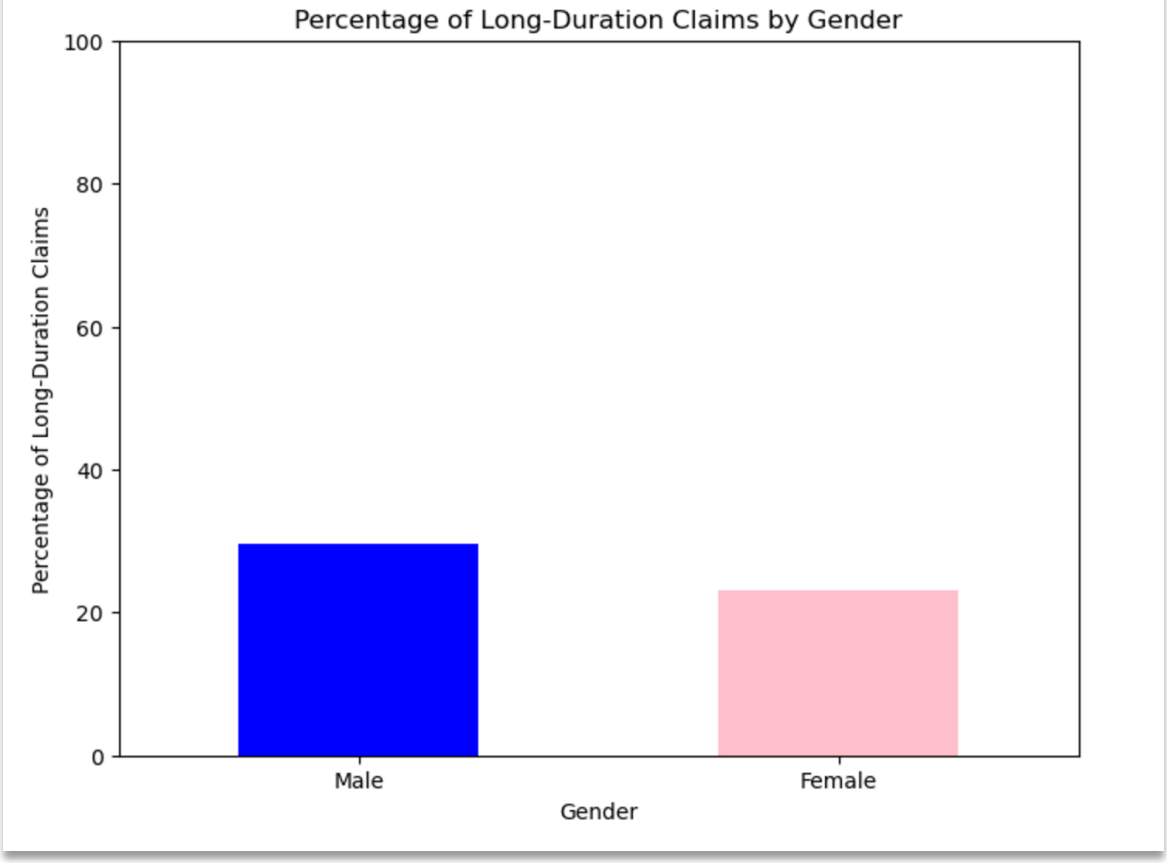
b) Gender:

We have used barcharts to understand the pattern between gender and long duration claims. We have plotted the below graph with the count of long duration claims which shows that male population has had more long duration claims than those of female population.



- However, comparing raw counts can be misleading since the dataset may have unequal numbers of male and female clients. To address this, we have calculated the percentage of long-duration claims for each gender, relative to the total claims in each group, and then plotted a bar chart.
- The percentage of long claims is still higher in male population than that of female. However the below graph is normalized and accounts for the distribution of males and females in the dataset.
- This indicates a potential need for targeted premium adjustments based on gender.

```
Percentage of Long-Duration Claims by Gender:
Sex
1    29.684862
2    23.049133
Name: 1, dtype: float64
```

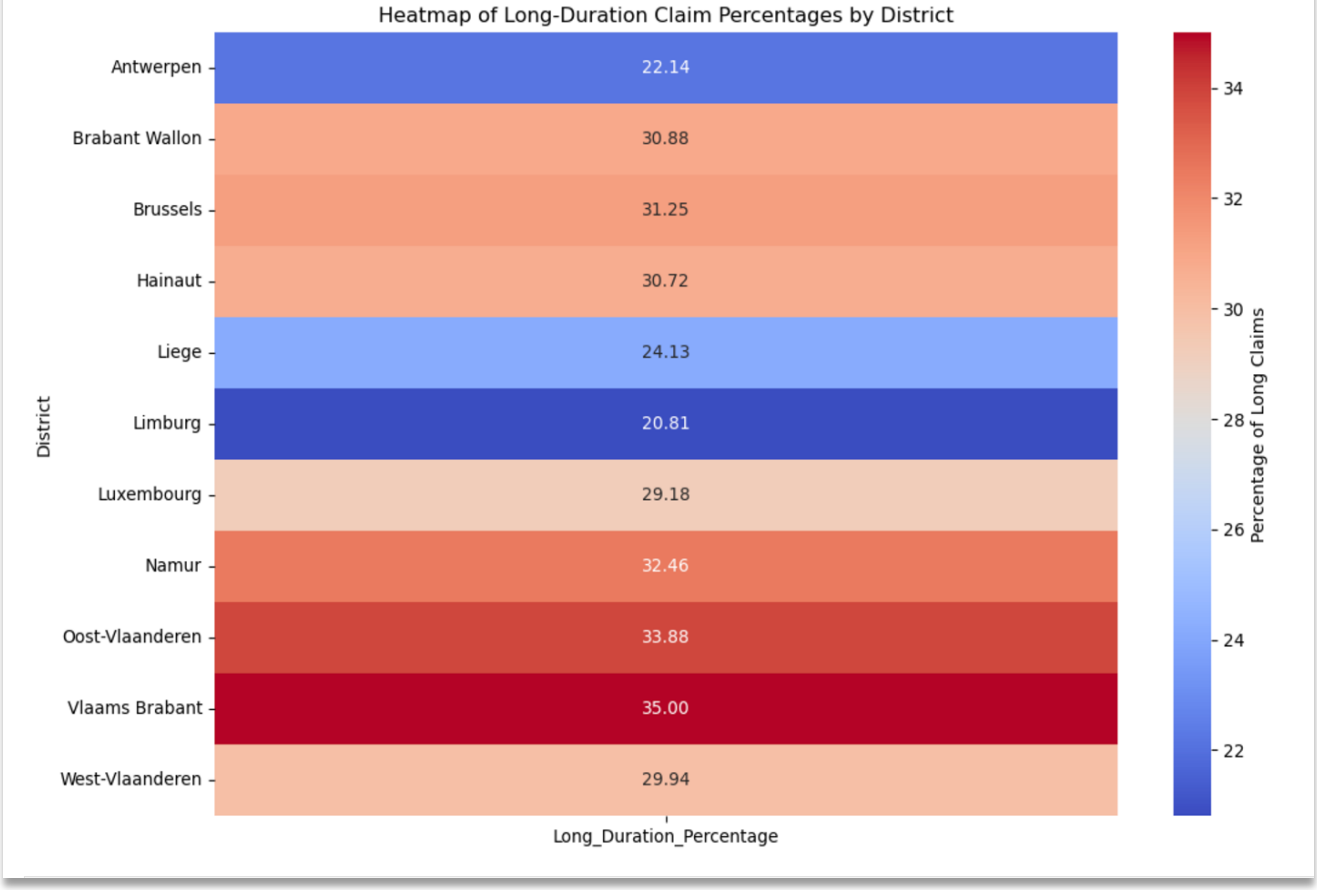


c) Districts:

The below heatmap shows that certain districts have a disproportionately higher percentage of long-duration claims compared to others.

These districts may require further investigation to understand underlying causes (e.g., healthcare access, employment types).

```
Percentage of Long-Duration Claims by District:
District
Antwerpen          22.142857
Brabant Wallon     30.882353
Brussels           31.250000
Hainaut            30.717863
Liege              24.133811
Limburg            20.805369
Luxembourg         29.184549
Namur              32.462687
Oost-Vlaanderen    33.878505
Vlaams Brabant     35.000000
West-Vlaanderen    29.943503
Name: 1, dtype: float64
```



Broker Patterns:

Number of Claims and Long-Duration Percentage were calculated for each broker.

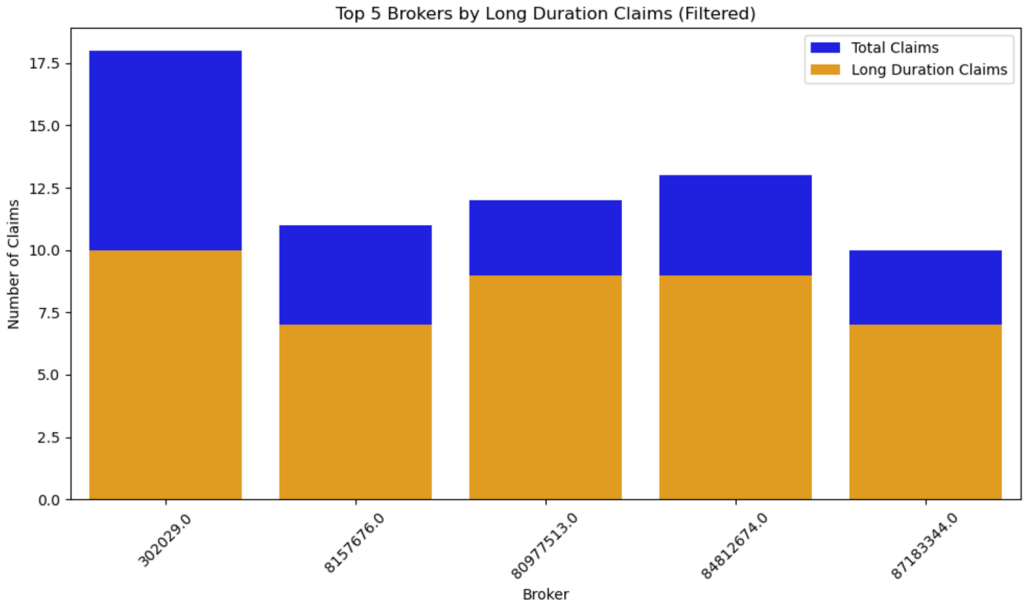
```
Long-Duration Claims Analysis by Broker:
            Long_Duration_Claims_Count  Total_Claims  Long_Duration_Percentage
Broker
51072827.0                         2.0             2                     100.0
35401827.0                         1.0             1                     100.0
36936199.0                         2.0             2                     100.0
81442461.0                         3.0             3                     100.0
36340774.0                         1.0             1                     100.0
...                                ...           ...                       ...
44917920.0                         0.0             1                       0.0
45333378.0                         0.0             1                       0.0
45588280.0                         0.0             2                       0.0
45658495.0                         0.0            16                       0.0
99980375.0                         0.0             1                       0.0

[1323 rows x 3 columns]
```

However, when a broker has very few clients, even one or two long-duration claims can inflate the percentage to 100%, which might not reflect the overall trend. To improve the analysis, we have applied a minimum threshold for the total number of claims.

Filtered brokers with fewer than 10 total claims to exclude those where percentages may not be meaningful. Then sorted them by the percentage of long-duration claims to identify the most impactful ones. Finally retained only the top 5 brokers with significant data for a cleaner bar graph.

Broker	total_claims	long_claims	long_claim_percentage
80977513.0	12	9	75.000000
87183344.0	10	7	70.000000
84812674.0	13	9	69.230769
8157676.0	11	7	63.636364
302029.0	18	10	55.555556
...	...	...	...
9346475.0	22	0	0.000000
36047451.0	11	0	0.000000
15198327.0	12	0	0.000000
45658495.0	16	0	0.000000
86478229.0	10	0	0.000000

[107 rows x 3 columns]



Top Brokers

To Identify the top performing brokers, we have calculated the Total Claims and Long Duration Claims per Broker, which helped identify the top brokers by volume. Then, calculated the number of claims with short durations for each broker. Finally, have ranked Top Brokers by Volume and Performance.

```
Top Brokers by Volume and Short Claim Performance:

Claim_Category  Short_Claims  Long_Claims  Total_Claims  \
Broker
5614684.0                 62            6            68
32704839.0                76           12            88
84457806.0                45            8            53
61603288.0                59           11            70
60246561.0                31            9            40
93602112.0                29           12            41
21206979.0                43           18            61
71731179.0                50           28            78
79779951.0                47           29            76
70075182.0                46           42            88

Claim_Category  Short_Claim_Percentage
Broker
5614684.0                   91.176471
32704839.0                  86.363636
84457806.0                  84.905660
61603288.0                  84.285714
60246561.0                  77.500000
93602112.0                  70.731707
21206979.0                  70.491803
71731179.0                  64.102564
79779951.0                  61.842105
70075182.0                  52.272727
```

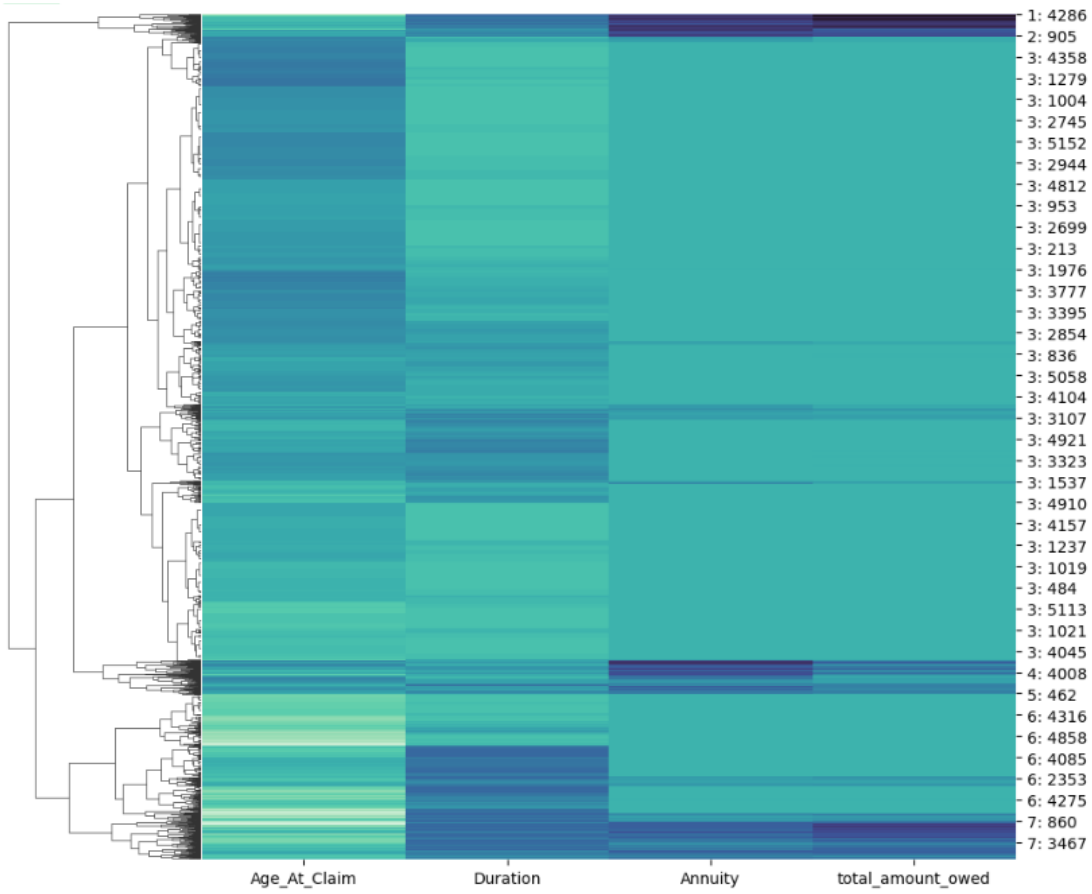
Among the top-performing brokers (by volume), only a few consistently manage to keep the percentage of long-duration claims low. This suggests that specific brokers might have better risk management practices or cater to lower-risk clientele.

b. Segment Allianz's customers using claims data.

Features : Age at claim , duration , Annuity , total amount owed (by allianz)

Method: complete

No of clusters: 7



Cluster 1: very high annuity, very high total amount owed, long duration, average age at claim

Cluster 2: very high annuity, High total amount owed, long duration, average age at claim

Cluster 3: Low annuity, Low total amount owed, Low duration, Average age at claim.

Cluster 4: Very high annuity, Average total amount owed, Average duration, Average age at claim.

Cluster 5: High annuity, Average total amount owed, Average duration, high age at claim.

Cluster 6: low annuity, Low total amount owed, Low age at claim.

Cluster7 : average annuity, high total amount owed, High duration, Average age at claim.

c) What other Machine Learning tools can Allianz employ here?

Machine Learning Tools for Allianz: Forecasting and Risk Assessment

Objective:
Leverage machine learning to forecast disability ratios and assess claim risk to help Allianz optimize resource allocation and manage payouts effectively.

Summary:
The proposed approach leverages LSTM models to forecast future disability ratios (passing Pct_dis1, Pct_dis2, Pct_dis3, Pct_dis4, as input to predict the disability ratio Pct_dis5(disability ratio at year 5). Later using Pct_dis1, Pct_dis2, Pct_dis3, Pct_dis4, Pct_dis5 to predict Pct_dis6. This method can be extended for predicting disability ratios for 21 years (Pct_dis21) and even more, enabling Allianz to estimate its total liability by multiplying these ratios with annuity amounts. Using these forecasts, claims are categorized into high or low risk based on thresholds for annuity (above the 75th percentile) and claim duration (10+ years). To handle the wide range of annuity values, normalization techniques are applied to ensure consistency in risk evaluation. The final risk categorization using classification trees (using grid search or random forest) helps predicting high-risk clients.

Benefits:
This approach empowers Allianz to proactively manage high-risk claims, ensuring efficient allocation of resources. By identifying high-risk clients, Allianz can adjust premium rates to collect more revenue and offset potential losses. The model also provides data-driven insights into claims and risk patterns, offering a scalable framework that can be refined with additional data for greater accuracy and profitability.

d. What will be your final recommendation(s) for the company?

1).Based on district-wise and duration analysis, it was observed that certain district exhibit a lower overall claim duration and annuity compared to others. This indicates a potential difference in customer selection patterns or risk assessment strategies employed by brokers in these regions.

It is recommended that brokers in districts with higher annuity and longer claim durations analyze and adopt the practices of brokers operating in districts with lower annuity and shorter claim durations. Specifically, they should investigate the customer attributes and selection criteria emphasized by brokers in these more profitable districts. By incorporating similar strategies, brokers in high-risk districts can potentially reduce claim rates and improve the company's profitability.

2).The company should conduct an in-depth analysis of customers' employment data to identify patterns associated with risk and profitability. This analysis should categorize customers based on their employment type, salary levels, and associated risk factors.

1. **Employment Type Categorization:** Classify customers' employment types into high-risk, medium-risk, and low-risk categories based on the likelihood of claims (e.g., office jobs versus high-risk physical labour jobs). **study of Belgium:** Metallurgy sector in Belgium is one of the important contributors of Belgium's economy. Most of these industries are present in Antwerpen and nearby districts. These roles are considered high risk while comparing to a traditional employment opportunity. Example a field engineer in these organisations involves both high risk and high salary. Risk classification based on industry and roles taking consideration into salary is important to reduce loss and increase profit.
2. **Salary-Based Premium Adjustment:** Segment customers based on their salary levels (e.g., high salary and low salary). For instance, customers working in high-risk office environments with higher salaries should be charged higher premiums as they present a higher claim risk while contributing to higher loss if not addressed.

This categorization will enable the company to design a dynamic and fair premium structure tailored to customers' profiles.