BUSINESS ANALYTICS AND MODELLING PRESENTATION

Analysis on Insurance data

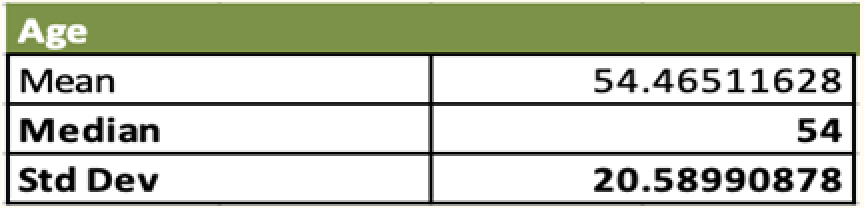
Submitted to: Matthew Horrigan

Aim and Scope of the Project

The project aims to guide through the entire exploratory data analysis (EDA) process, extracting valuable insights from the provided data to aid the organization and stakeholders in their decision-making processes.

Descriptive Analysis of Original Dataset

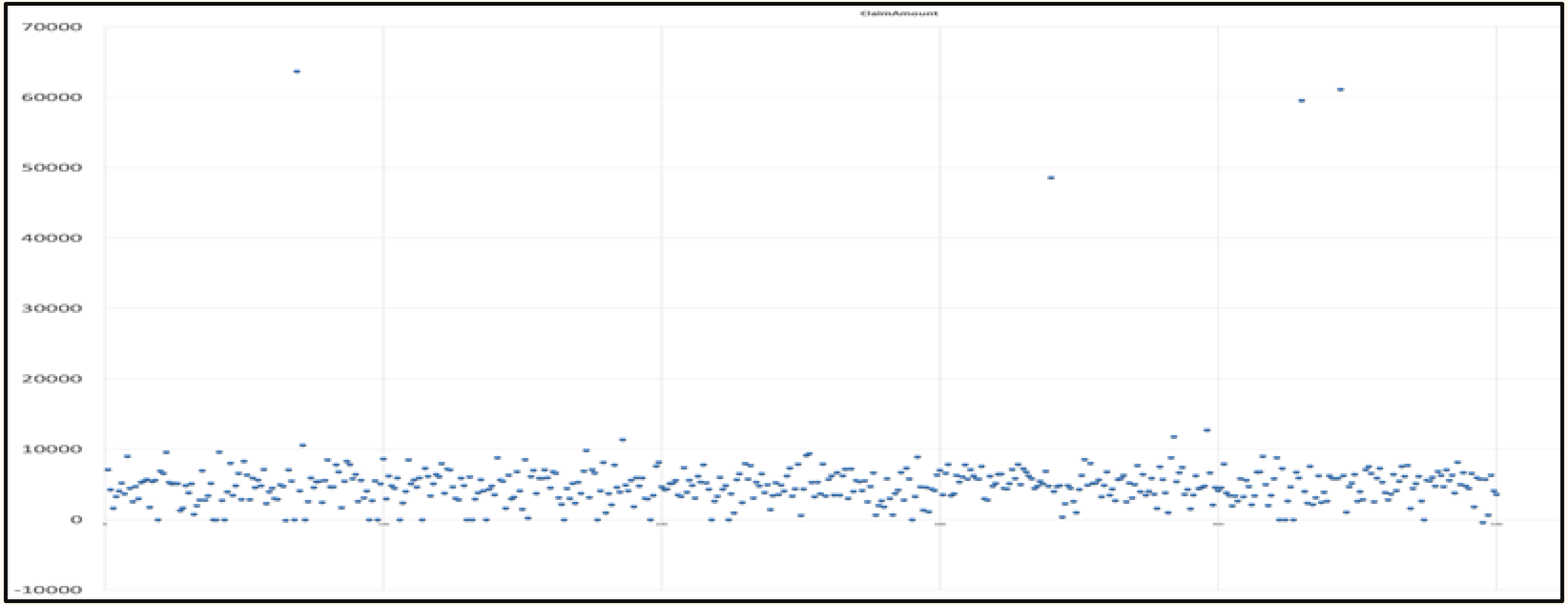
- Age Insights
  - Average age in the dataset is approximately 54.47 years, with a median age of 54 years.
  - The standard deviation for age is approximately 20.59 years, reflecting a broad age range within the dataset.
  - The Minimum age provided in the dataset is 18 and the maximum age is 97 years



Age	
Mean	54.46511628
Median	54
Std Dev	20.58990878

Descriptive Analysis of Original Dataset

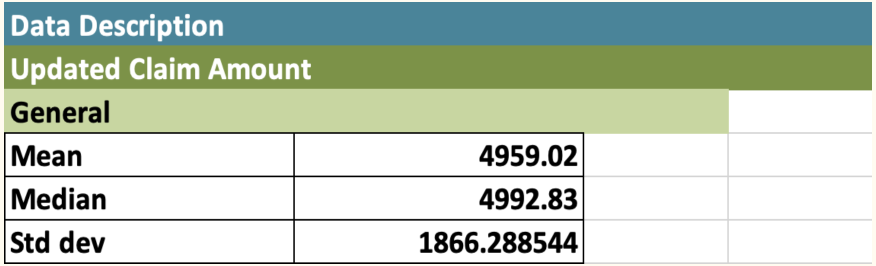
Descriptive Statistics for Claim Amount: By looking into the scatter plot of Claim amount vs the claim ID, we can conclude that the data set has some outliers which would impact the descriptive statistics for the column Claim Amount. So it was important to identify these datapoints and exclude them from analysis to get more reliable results. One can observe that majority of the data points lie in between 900 and 10,000. So for preliminary analysis, we have considered these as the class boundary beyond which data would be considered as an outlier.



Descriptive Analysis of Original Dataset

- The section represents Claim Amounts, with a mean value of approximately €5,420.99 and a median of $4,992.83.
- The data has a standard deviation of around €1,866.29, indicating moderate variation in claim amounts.
- Control Limits for Claim Amount
  - Calculated Control Limits (UCL and LCL) for Claim Amount:
    - Upper Class Limit (UCL): €8,725.41
    - Lower Class Limit (LCL): €1,260.25
- These control limits serve as benchmarks for monitoring and maintaining acceptable levels of claim amounts.

Limits=Median ± 2x Standard Deviation

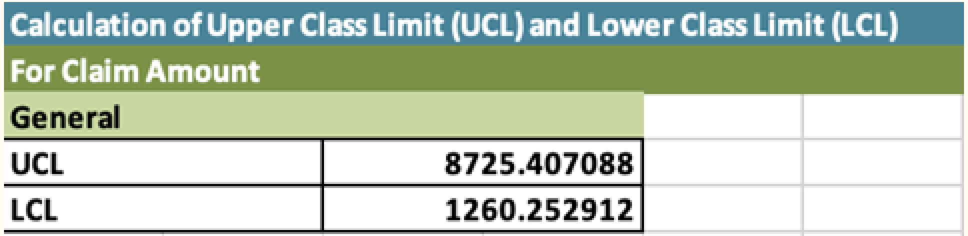


Data Description

Updated Claim Amount

General

Mean	**4959.02**
Median	**4992.83**
Std dev	**1866.288544**



Calculation of Upper Class Limit (UCL) and Lower Class Limit (LCL)

For Claim Amount

General

UCL	**8725.407088**
LCL	**1260.252912**

Issues with the Original Dataset

- **Missing Data**: Missing data can affect the reliability and accuracy of the data model which includes visual representation of the dataset as well, making it challenging to understand the data and use the dataset for analysis and forecasting. Visualization can be misleading if missing values are not clearly indicated.
- **Imputation Uncertainty**: When missing data is imputed (replaced with estimated values), there is inherent uncertainty in the imputation process. The choice of imputation method can affect the results, and the uncertainty should be considered in the interpretation of the findings.
- Missing data can lead to a loss of information, reducing the effective sample size for analysis. This can impact the statistical power of tests and make it challenging to detect significant relationships or patterns in the data.
- Reduced Precision: Missing data can reduce the precision of estimates and increase the standard errors of statistical analyses, which can affect the reliability of results.

Impact of clean data

- Clean data is data that is free from errors, inconsistencies, and any form of corruption, making it reliable and accurate. The impact of having clean data is substantial and can positively affect various aspects of an organization's operations.

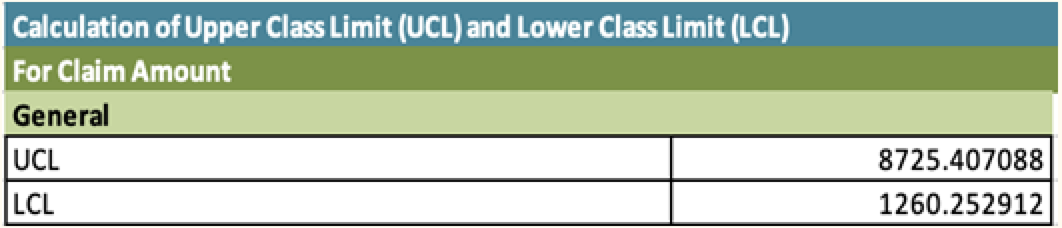
- Clean data is a fundamental asset for organizations. It supports better decision-making, enhances analytics, reduces costs, and has a positive impact on customer satisfaction, compliance, and innovation. Investing in data quality is an investment in the success and sustainability of the organization.

Advantages of clean data

- **Accuracy**: Clean data is error-free, ensuring reliable and trustworthy information.
- **Informed Decisions**: It empowers better decision-making with accurate insights.
- **Efficient Operations**: Reduces time and costs associated with data cleaning.
- **Customer Satisfaction**: Improves customer service and personalized experiences.
- **Compliance**: Ensures adherence to regulatory standards, avoiding legal complications.
- **Innovation**: Fosters data-driven strategies, opening doors to new opportunities.

Data PreProcessing

- The Technique Followed here is : **Data Trimming**
- In this approach, any data points that fall outside the specified range i.e; UCL and LCL are simply omitted from the final dataset used for analysis.
- Calculation of UCL and LCL using Limits=Median ± 2x Standard Deviation

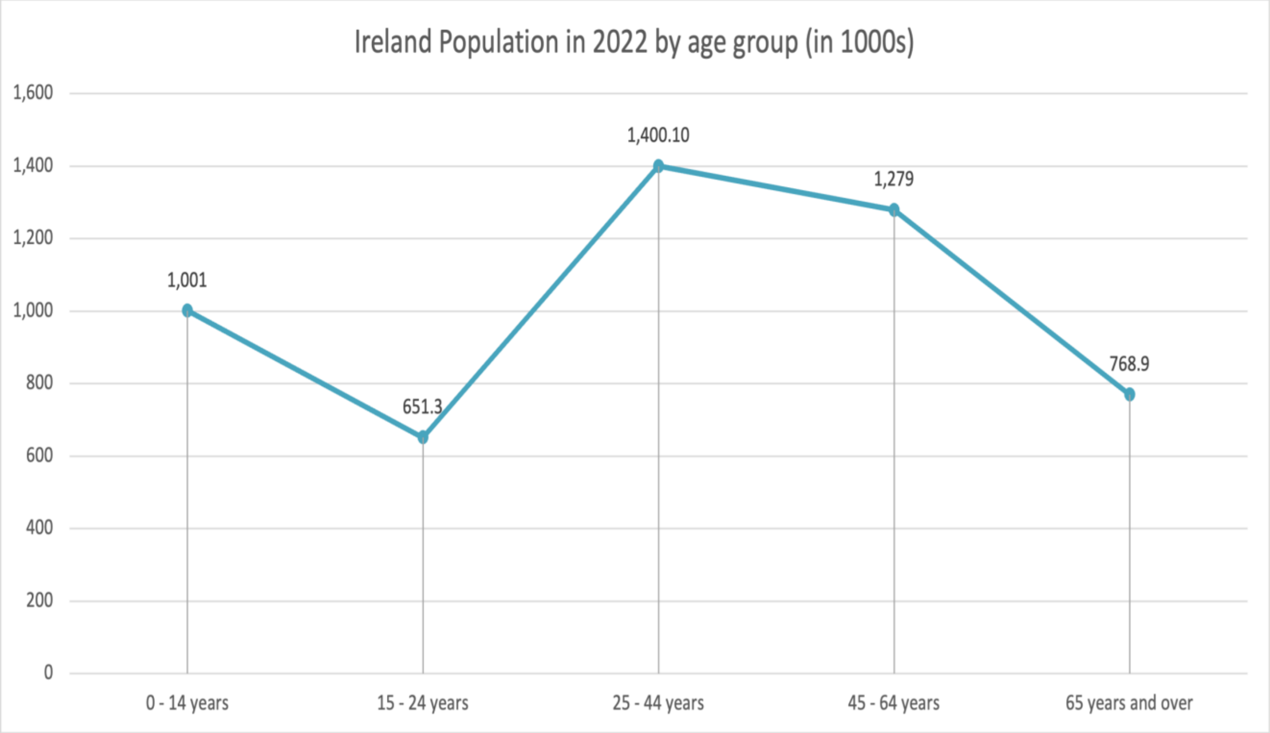


Calculation of Upper Class Limit (UCL) and Lower Class Limit (LCL)	
For Claim Amount	
General	
UCL	8725.407088
LCL	1260.252912

- Since there was no strong link between age and other factors (like gender, disease, or claim amount), missing age values weren't filled in to avoid skewing the analysis towards a specific age range.

- Instead of individual ages, a new "Age Class" column was created for grouping ages. Missing claim amounts were filled by using average claim amounts for specific age groups within different disease categories, presented in the pivot table for further analysis.

Population Analysis of Ireland 2022

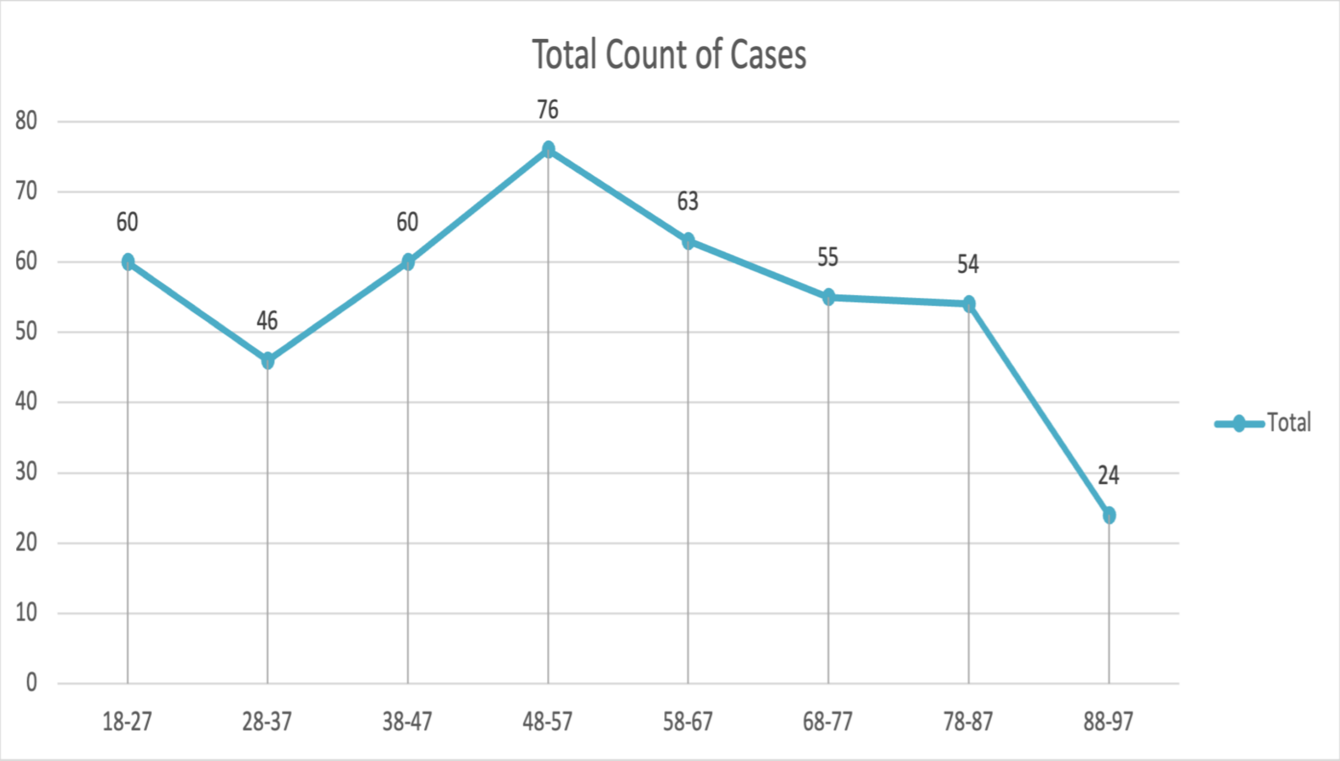


Data Source: Statista :: Population of the Republic of Ireland in 2022, by age group (in 1,000s)

After doing a basic research about population distribution across various age groups, it is evident that the age group of 0-14 exhibits a high population, followed by a noticeable decrease in the 15-24 age group. However, there is an increase in population from ages 25-44, which then decreases for the other two age groups beyond that.

Data Analysis

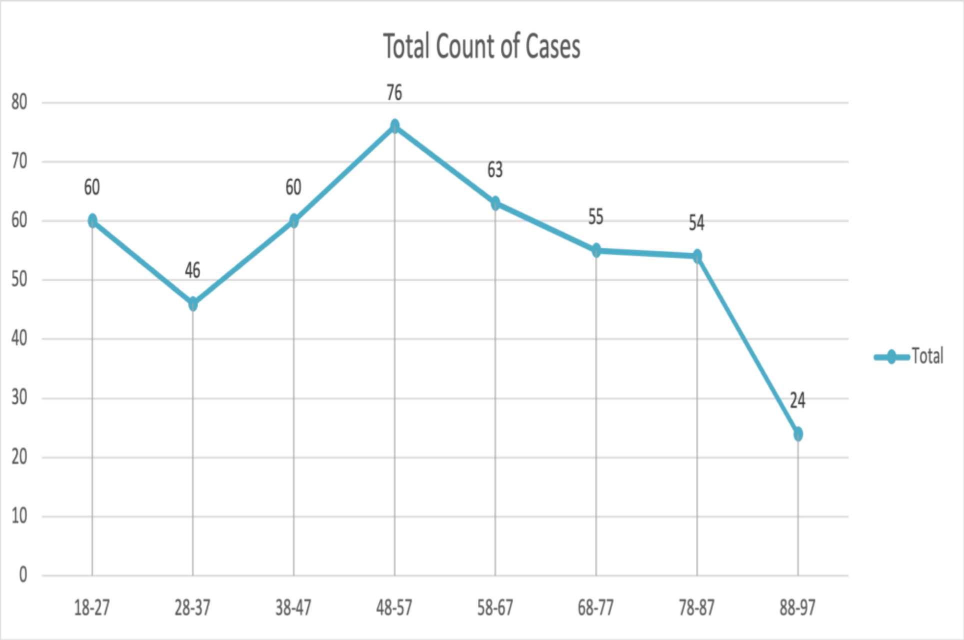
Similar pattern has been observed in the provided Insurance data set as well where people having age around 38 to 67 has the maximum number of cases as justified by the population distribution of Ireland.



Data Analysis

The line chart illustrates a predominant concentration of cases within the age range of 38 to 67 across various diseases, except for Cardiology, where cases decrease with age. This divergence in Cardiology could indicate differing lifestyle impacts on disease occurrence as individuals grow older.

For all the other disease category, as the chart indicates, there's an up trend for number of cases till the age of 57 and then a down trend for age beyond 57. Considering this pattern, we have clustered the dataset into two divisions Group 01 (ages 18 to 57) and Group 02 (ages 58 to 97) for all diseases except Cardiology. This grouping aims to explore differences in disease incidence across broader age ranges, although it's crucial to consider that multiple factors beyond age may influence these patterns.

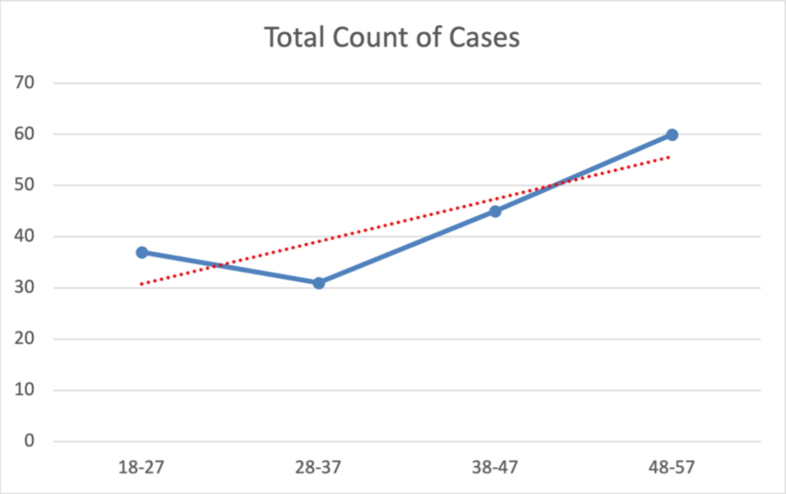


Relation between Age and Disease Count with respect to population distribution by Age Groups

Data Cluster using Age Groups : Age Group 01 :18 to 57 : For All Diseases except Cardiology

- The data shows that as people in age group 01 (18-57) get older, the chances of developing the specified diseases increase. This information can be used to decide on insurance premiums and suggests the need for a health check before providing insurance.

- However, it's essential to remember that while age is strongly linked to disease risk, other health factors and individual circumstances should also be considered for a complete understanding of an individual's health profile before determining insurance terms.



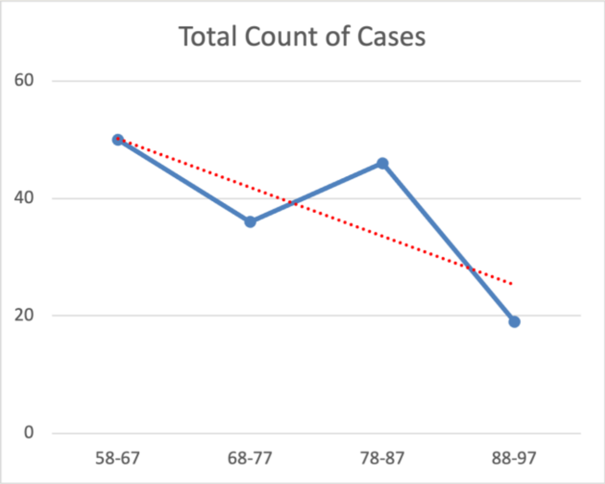
Data Cluster using Age Groups :: Age Group 01 : 18 to 57 :: For All Diseases except Cardiology

Correlation Coefficient	0.853586083

Data Cluster using Age Groups :: Age Group 02 :58 to 97: For All Diseases except Cardiology

- In age group 02(58-97), there's a reverse trend: as people grow older, the number of cases decreases, suggesting a negative correlation between age and disease count. This information could also guide the formulation of insurance premium amounts.

- Similar to age group 01(18-57), while age seems to be inversely related to disease count in this older age category, a comprehensive assessment considering various health aspects is crucial for setting insurance premiums.



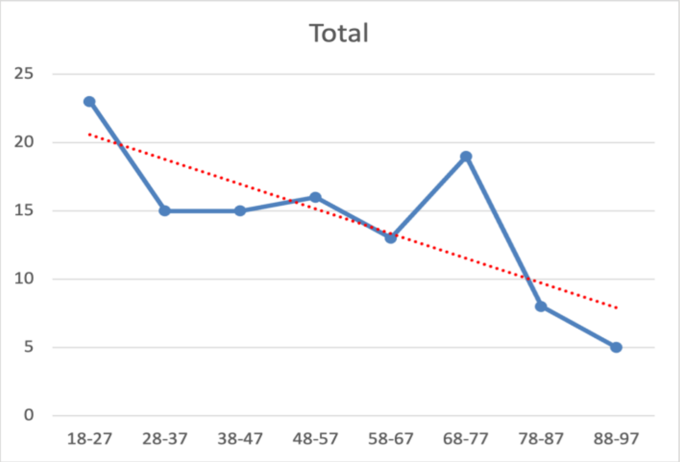
Data Cluster using Age Groups :: Age Group 02 : 58 to 97:: For All Diseases except Cardiology

Correlation Coefficient	-0.775498

Data Analysis of Disease Category: Cardiology with respect to age group

- In Cardiology cases, there's a clear pattern: as age rises, the number of cases decreases, supported by a correlation coefficient of -0.77. This relationship could inform insurance premium calculations and highlight the need for medical checks and a person's fitness assessment before granting insurance.

- The strong negative correlation between age and Cardiology cases indicates that older age is associated with a reduced number of cases. This information can guide insurance premium calculations and underscores the importance of medical examinations and evaluating an individual's physical health before approving an insurance product.

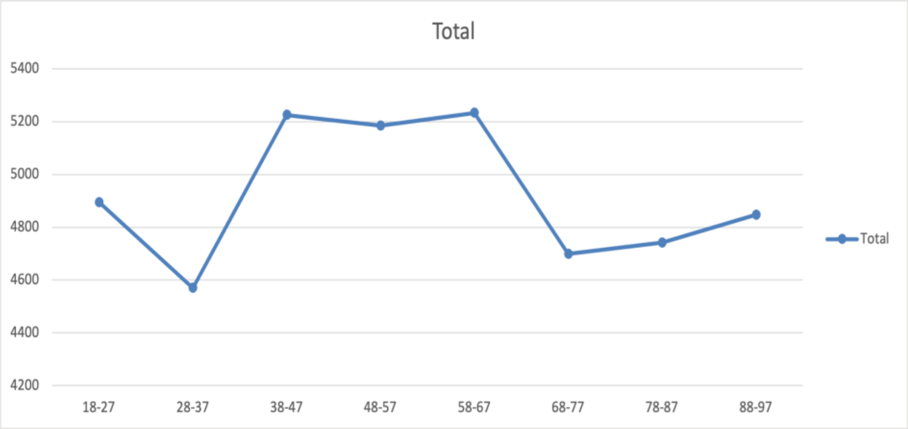


Data Analysis of Disease Category: Cardiology with respect to age group

Correlation Coefficient	-0.774100953

Relation between Age Class and Average Disease Claim Amount

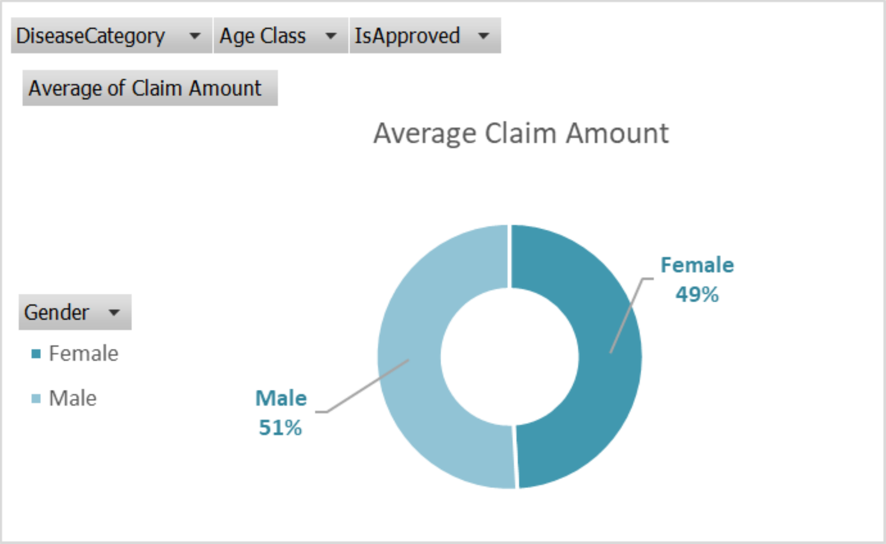
The Analysis explores the relationship between Age class and average claim amounts for various diseases. While individual disease categories show limited variation in claim amounts, combining all diseases reveals a substantial change. This change suggests that between ages 38 to 67, the average claim amount per person exceeds that of the rest of the population.



Relation between Age Class and Average Disease Claim Amount

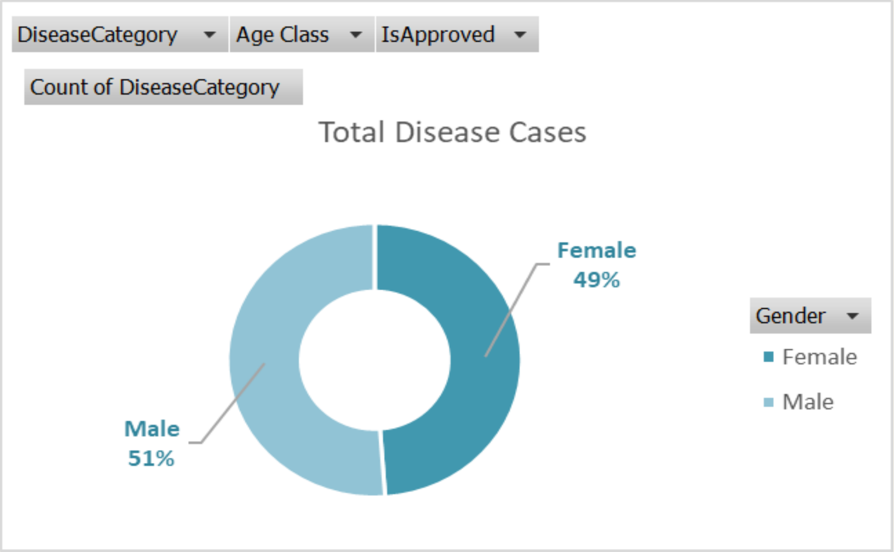
Relation between Gender Vs Average Disease Claim Amount

The pie chart displays how average claim amounts are distributed based on gender. It indicates that there is little to no difference in the average claim amounts between genders.



Relation between Gender Vs Count of Cases

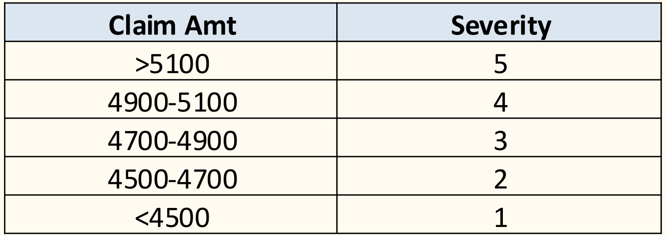
The pie chart represents the total number of disease cases identified among different genders. Similar to the pie chart displaying claim amounts by gender, it also demonstrates a relatively balanced distribution of disease count between the two genders.



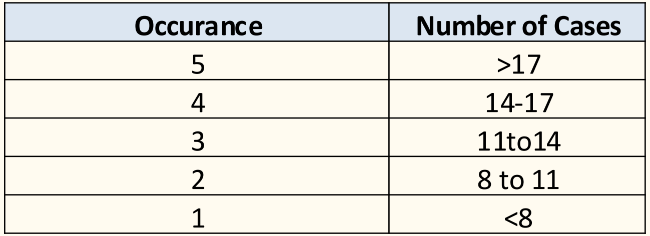
Risk Detection and Mitigation of Risk due to realization of higher claim values

- The risk identification process using RPN (Risk Priority Number) involves recognizing potential risk events, such as higher claim amounts, and determining the frequency of these events occurring—measured by the number of qualifying cases. It also assesses how easily these events can be detected; however, in this scenario, it assumes all cases are promptly detected by medical professionals in a hospital.

- Following the initial risk identification and frequency assessment, a severity number and an occurrence number are assigned to each event and its frequency. These values are utilized to calculate the RPN (Risk Priority Number), using the formula $\mathrm{RPN} = \mathrm{S} \times \mathrm{O} \times \mathrm{D}$, where "D" (detection) is constant and considered as 1. This RPN helps to pinpoint which events pose a higher risk and potential profit loss.



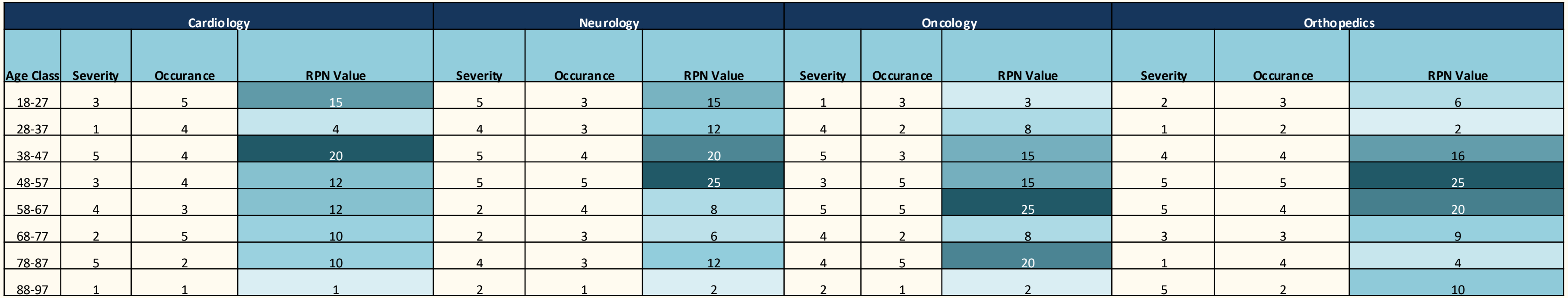
Claim Amt	Severity
>5100	5
4900-5100	4
4700-4900	3
4500-4700	2
<4500	1



Occurance	Number of Cases
5	>17
4	14-17
3	11to14
2	8 to 11
1	<8

Risk Detection and Mitigation of Risk due to realization of higher claim values

From the table, higher risk events can be easily identified. For example, let's consider cardiology. For this disease, the highest RPN value is 20 which is for the age class of 38 to 47, so whenever a person of age 38 to 47 plans to buy insurance against cardiological diseases, an extra premium amount should be levied from the person to adjust the risk of realizing the claim amount.



	Cardiology			Neurology			Oncology			Orthopedics		
Age Class	Severity	Occurance	RPN Value	Severity	Occurance	RPN Value	Severity	Occurance	RPN Value	Severity	Occurance	RPN Value
18-27	3	5	15	5	3	15	1	3	3	2	3	6
28-37	1	4	4	4	3	12	4	2	8	1	2	2
38-47	5	4	20	5	4	20	5	3	15	4	4	16
48-57	3	4	12	5	5	25	3	5	15	5	5	25
58-67	4	3	12	2	4	8	5	5	25	5	4	20
68-77	2	5	10	2	3	6	4	2	8	3	3	9
78-87	5	2	10	4	3	12	4	5	20	1	4	4
88-97	1	1	1	2	1	2	2	1	2	5	2	10

Using the above data, three products are suggested as demonstrated below. Standard Product which would not include high risk categorized disease coverage, Premium Product which would include all the diseases and a third product which provides high risk categorized disease insurance coverage after specific add on are bought and medical verification has been completed.

Risk Detection and Mitigation of Risk due to realization of higher claim values

Age Class	Product: Standard				Product: Standard Plus				Product: Premium			
	CRDLG	NRLG	ONLG	ORTH	CRDLG	NRLG	ONLG	ORTH	CRDLG	NRLG	ONLG	ORTH
18-27	N	O	Y	Y	O	Y	Y	Y	Y	Y	Y	Y
28-37	Y	N	O	Y	Y	O	Y	Y	Y	Y	Y	Y
38-47	N	O	Y	Y	O	Y	Y	Y	Y	Y	Y	Y
48-57	Y	O	Y	N	Y	Y	Y	O	Y	Y	Y	Y
58-67	Y	Y	N	O	Y	Y	O	Y	Y	Y	Y	Y
68-77	N	Y	Y	O	O	Y	Y	Y	Y	Y	Y	Y
78-87	Y	O	N	Y	Y	Y	O	Y	Y	Y	Y	Y
88-97	Y	Y	O	N	Y	Y	Y	O	Y	Y	Y	Y

Legends

Where	Represents
Y	Yes
O	Option Add On
N	Not Available

Using the above data, three products are suggested as demonstrated below. Standard Product which would not include high risk categorized disease coverage, Premium Product which would include all the diseases and a third product which provides high risk categorized disease insurance coverage after specific add on are bought and medical verification has been completed.

Conclusions

- The correlation between the first age group and second age group follows a linear trend. This could be attributed to the smaller population size within the older age groups, as indicated in the table illustrating Ireland's population demographics.
- In the field of Cardiology, a noticeable trend is the reduction in case count with increasing age, supported by a correlation coefficient of -0.77. This relationship is valuable for estimating insurance premiums and establishing guidelines for necessary medical evaluations and assessing an individual's physical fitness before approving an insurance product.
- The association between gender and both the average claim amount and total disease cases remains consistent. This suggests that there is no notable variation in the average claim amount concerning gender. Similarly, the distribution of disease count is relatively even across both genders.
- Using the RPN values, determination of higher risk consumers can be identified and accordingly product can be suggested including requirement of medical verification to be done before providing the insurance product.