Analysis of Motor Insurance Injury Claims and Rehabilitation Outcomes

Batin Cap

Hardy Diakite

Mariam Sibuma Salum

Sean Humphrey

Thana Nassr

Introduction to Programming

Submission date: 2 December 2025

Executive Summary

We use a Typhon extract of 1,000 UK motor bodily injury claims to analyse the relationships between injury patterns & rehabilitation pathways and recovery and overall claim cost. There were 36 variables in the dataset which included demographic profiles, recovery measures, claims process fields and cost breakdowns along with total_claim_amount as the headline outcome. The data preparation tackled considerable missingness in administrative note fields (claims_resolution_exit_notes: 901 missing; claims_management_exit_notes: 896 missing), such as coding them not_provided to retain records, normalising the text fields, and parsing of set-like injury/rehab labels. Two interpretable features were engineered, num_injuries and car_damage_severity_num, which resulted in a total of 1,000 × 38 data points and 0 missing values. Claims costs were highly skewed to the right: mean £21,171 (median £20,639; max £68,163). General_damages predominated the settlement values (mean £19,872), where medical_treatment_costs was relatively low (mean £572), suggesting that experienced recovery and negotiation account for costs as much as direct treatment cost. Injury mix dominated by whiplash were 49.5%, spine/back were 14.1%, and limb were 13.7%. AIS/ISS proxy scoring produced severity bands with minor 76.3%, moderate 21.8%, serious 1.9%. Rehabilitation yielded a large decrease (81.5% recommended vs 59.1% completed). Injury duration provided the best predictive signal; the strongest correlation with total cost ($r = 0.853$) was obtained with injury duration, whereas rental vehicle and home care costs were significant outlier drivers. Operational implications are presented (supported early duration-based triage, rehab utilization enhancement and greater control over ancillary cost pathways), and findings are also associative; they should not be misidentified as causal.

Contents

Executive Summary 2
LIST OF FIGURES 5
Figure 1: Variable Type Summary (Numerical and Categorical count) 5
Figure 2: Missingness profile by variable 5
Figure 3: Severity Band Construction 5
Figure 4: Correlation Matrix 5
Figure 5: Distribution of Total Claim Amount 5
Figure 6: Claim Cost Breakdown (General_damages and medical_treatment_costs) 5
Figure 7: Injury Duration Distribution 5
Figure 8: Medical Attention Delay 5
Figure 9: Top Injury Types 5
Figure 10: Severity Bands 5
Figure 11: Rehab Funnel (Recommended - Completed) 5
Figure 12: Duration by Rehab Status 5
Figure 13: Injury Duration and Total Claim Amount 5
Figure 14: Outlier Drivers (Rental_vehicle_expense and home_care_services_cost) 5
Figure 15: Conceptual Diagram: (Duration - Negotiation/Services - Higher Total Cost) 5
Glossary of Terms (A and Z). 6
1. Introduction 8
Motor insurance and claims analysis. 8
1.2 Problem statement and aims. 8
1.3 Overview of data and analytical techniques. 8
2. Data and Preparation 8
2.1 Data description. 8
2.2 Data cleaning and preprocessing. 9
2.3 Feature engineering. 10
3. Analytical Approach 10
3.1 Descriptive and exploratory analysis 10
3.2 Analysis of severity and rehabilitation. 11
3.3 Predictive modelling / advanced analysis 12
3.4 Limitations of the approach. 13
4. Results 13
4.1 Descriptive statement of claims 13
4.2 Patterns of injury and severity. 18
4.3 Efficacy and outcomes of rehabilitation 20

4.4 Predictive insights.....21
5.1 Summary of main conclusions.....23
Conclusion.....25
APPENDICES.....26
Appendix A: Dataset profile.....26
Appendix B: Missing value (before cleaning).....27
Appendix C: Cleaning and feature engineering summary for Task A.....28
Appendix D: Outlier screening (IQR method, top 10).....29
Appendix E: Severity outputs for Task B.....30
Appendix F: Collaborative Project Management Using GitHub.....31
7.1 Team roles, communication, and coordination.....31
Appendix G: Tables and Figures.....32
REFERENCES.....51

LIST OF FIGURES

Figure 1: Variable Type Summary (Numerical and Categorical count)

Figure 2: Missingness profile by variable

Figure 3: Severity Band Construction

Figure 4: Correlation Matrix

Figure 5: Distribution of Total Claim Amount

Figure 6: Claim Cost Breakdown (General_damages and medical_treatment_costs)

Figure 7: Injury Duration Distribution

Figure 8: Medical Attention Delay

Figure 9: Top Injury Types

Figure 10: Severity Bands

Figure 11: Rehab Funnel (Recommended - Completed)

Figure 12: Duration by Rehab Status

Figure 13: Injury Duration and Total Claim Amount

Figure 14: Outlier Drivers (Rental_vehicle_expense and home_care_services_cost)

Figure 15: Conceptual Diagram: (Duration - Negotiation/Services - Higher Total Cost)

Glossary of Terms (A and Z).

ABI: A UK insurance industry organization providing reports and stats.
Administrative/operational notes: Notes and codes are written by internal claim handlers to explain what happened in a claim.
AIS: A medical scoring system that rates the severity of an injury.
Associative (not causal): A link seen in the data, but it does not prove one causes the other.
Binary variable: Only two values, often 0/1 (No/Yes).
Car damage severity: A label indicating how much damage was done to the car (none, minor, moderate, severe, total loss).
Car_damage_severity_num: The same car damage severity, but numbers instead, hence easier form to analyse.
Category frequency table: A number that displays each category by how many times it appears.
Claims inflation: When claim costs rise over time.
Claim severity: How much a claim costs.
Cleaning / preprocessing: Clean data, format the data, make sure there are no missing values, match fields. It can then be fed into software systems.
Confounding: As in, a hidden factor influences both things compared and is therefore misleading.
Correlation (r): A number to indicate how strongly related two things are.
Credit hire / replacement vehicle costs: Cost of a vehicle for temporary use without the customer's vehicle (in the data: rental_vehicle_expense).
df (DataFrame): The primary Python table to store the data set.
Exploded analysis: Translating list-like fields into multiple rows where you can count up injury or rehab type correctly.
Feature engineering: Adding new useful columns.
FCA: (UK financial regulator) which sets expectations of the insurers for how they deal with claims.
General damages: Money for pain, loss, or inconvenience of life enjoyment (not for medical bills).
GLM: A straightforward and common statistical model found in insurance.
IQR outlier: A value that is flagged as abnormally high/low according to standard rule.
Imputation: Using a normal rule (using median or most common value to fill in missing value).
Injury duration (injury_duration_days): How long it took to recover in days.
Injury labels: Injury types which are found in injury_type_classifications, e.g. whiplash.
ISS: A medical scoring system that totals injury severity by region of the body.
Iss_like: A similar to ISS score based on the injury labels in the dataset (not on a hospital-coded score).
Liability: Whether side accepts blame or contested.
Medical attention delay (medical_attention_delay_days): Days between accident and getting medical treatment.
Missingness: On what part of the dataset is information missing.
Mode: The value most commonly appearing.
Num_injuries: The number of different injury types a claim can cover.
OIC: Online process for UK road traffic injury claims (often in relation to whiplash reforms).
Outlier: A very rare, extreme value, and typically a super high cost.

Predictive modelling: The use of the training data to predict outcomes such as overall claim price.
Rehab recommended/completed: Advice about rehabilitation, and whether rehabilitation was done.
Severity band: A simple group label (minor/moderate/serious) based on iss_like scores.
severity_index: An aggregated score that includes injury severity and items such as duration and vehicle damage to provide an overall approximation of severity.
Selection bias: You are watching data that doesn't fully capture you (only few cases go into rehab for example).
Set-like fields: Columns saved on a text file in form of list/set (for instance, "{'head injury', 'whiplash'}") and have to be converted.
Standardisation (text): Making text consistent (lowercases, space removal) so that all categories match.
Total_claim_amount: The main outcome, the total cost of a claim.
Typhon extract: The exported dataset from the Typhon claims environment/system.

1. Introduction

Motor insurance and claims analysis.

Motor insurance is one of the UK's primary road-traffic risk management tools and is one of the largest personal-lines products, underwriting significant costs for injury and property damage (Adams, 2019; Kester, 2022). Research on UK motor portfolios—from early ratings to updated modelling of claim processes—still informs pricing and portfolio strategies (England, 2022; Guillen et al., 2021). It is evident in the literature that claim frequency and cost are jointly influenced by injury severity, progression in recovery and character of claimant (Davies, 2017; McGlade, 2018). Recent discussions around whiplash reforms, claims inflation and fraud have put motor insurance even more at the centre of UK policy and review conversations (Lewis, 2019; Swaby and Richards, 2024). UK market commentary shows record motor claims costs, translating to increased cost pressure and claims handling performance scrutiny even for ‘typical’ bodily injury portfolios (ABI, 2025; FCA, 2025).

1.2 Problem statement and aims.

However, UK insurers maintain a significant number of operational and medical-cost data, and the relationships between injury patterns, rehabilitation pathways, work disruption and overall claim severity are only partially understood. For example, the literature often addresses pricing, portfolio risk or technology rather than managing bodily injury recovery day-to-day (England, 2022; Johnson, 1971). This report therefore examines motor bodily injury claims data from Typhon to: (i) identify patterns in injury duration, treatment, and rehabilitation; (ii) how these relate to work-related absences and total claim costs; and (iii) conclude with implications for claims handling and rehabilitation initiatives.

1.3 Overview of data and analytical techniques.

Although new policies have been enacted, it still remains the case that motor claims procedures are only efficient when integrated into appropriate rehabilitation programmes and a fair settlement of claims (Lewis, 2019; McGlade, 2018). Insurers must walk a fine line between providing cost discipline and supporting claimants on the right levels, and policymakers must aim to tackle opportunistic claims without restricting access to justice (Adams, 2019; Kester, 2022). This report analyses trends in injury, treatment and recovery across a 1,000 motor bodily injury claims dataset with consideration of time to recovery, severity of injury and the use of resources in line with claimant types. The goal is to create evidence to help make claims management, pricing and rehabilitation decisions, with developed predictive features for triage and reserving.

2. Data and Preparation

2.1 Data description.

The analysis employed a Typhon motor bodily injury extract loaded into df that included 1,000 claims and 36 variables. Variables included demographics (e.g., age of claimant), injury and recovery metrics (e.g., injury_duration_days, medical_attention_delay_days), claims process markers (e.g., liability and closure fields), and cost components (e.g., general_damages, medical_treatment_costs, rental_vehicle_expense), with

total_claim_amount as the headline outcome. This structure mirrors UK regulatory demands for companies to grasp what drives costs/working frictions in motor claims (FCA, 2025).

```
[TaskA] DataFrame.info():
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 36 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   client_id                               1000 non-null   int64
 1   claim_value_category                    1000 non-null   int64
 2   injury_duration_days                    1000 non-null   int64
 3   hospital_visit_required                 1000 non-null   int64
 4   medical_care_sought                     1000 non-null   int64
 5   hospital_admission_required             1000 non-null   int64
 6   currently_unemployed_due_to_injury      1000 non-null   int64
 7   work_absence_duration_days              1000 non-null   int64
 8   work_absence_required                   1000 non-null   int64
 9   rehabilitation_recommended              1000 non-null   int64
 10  rehabilitation_completed                1000 non-null   int64
 11  minor_claimant_at_initial_assessment    1000 non-null   int64
 12  defendant_title_code                    1000 non-null   int64
 13  liability_admission_status              1000 non-null   object
 14  liability_type                          1000 non-null   object
 15  claimant_age_at_incident                1000 non-null   int64
 16  liability_denial_reasons                779 non-null    object
 17  claim_rejection_code                    771 non-null    object
 18  claims_management_exit_notes            104 non-null    object
 19  claims_resolution_exit_notes            99 non-null     object
 20  claims_resolution_closure_code          797 non-null    object
 21  rental_vehicle_expense                  1000 non-null   float64
 22  home_care_services_cost                 1000 non-null   float64
 23  employment_disadvantage_compensation    1000 non-null   float64
 24  career_satisfaction_loss_compensation   1000 non-null   float64
 25  asset_utility_loss_compensation         1000 non-null   float64
 26  medical_treatment_costs                 1000 non-null   float64
 27  general_damages                         1000 non-null   float64
 28  insurance_deductible_amount             1000 non-null   int64
 29  car_damage_severity                     1000 non-null   object
 30  total_claim_amount                      1000 non-null   float64
 31  injury_type_classifications             1000 non-null   object
 32  rehabilitation_program_types            1000 non-null   object
 33  claimant_current_age_years              1000 non-null   int64
 34  medical_attention_delay_days            1000 non-null   int64
 35  emergency_services_attended             1000 non-null   object
dtypes: float64(8), int64(17), object(11)
memory usage: 281.4+ KB
```

Figure 1: Variable Type Summary (Numerical and Categorical count)

2.2 Data cleaning and preprocessing.

Profiling revealed that missingness was primarily concentrated in process-note fields—claims_resolution_exit_notes (901 missing) and claims_management_exit_notes (896 missing)—with smaller gaps in rejection/denial fields. Instead of deleting records, missing administrative text was encoded as not_provided, in order to maintain claim completeness, and indicates no reporting. Text fields were also standardised (strip/lowercase) and set-like strings were parsed to consistently labelled input. In the event of the need for imputation, median/mode analysis was performed in agreement with the most current missing data reporting tradition (Afkanpour et al., 2024; Pham, 2024; Jäger et al., 2021).

```
[TaskA] Missing values per column (before cleaning):
claims_resolution_exit_notes             901
claims_management_exit_notes             896
claim_rejection_code                     229
liability_denial_reasons                 221
claims_resolution_closure_code           203
injury_duration_days                       0
claim_value_category                       0
client_id                                  0
hospital_visit_required                    0
medical_care_sought                        0
currently_unemployed_due_to_injury         0
hospital_admission_required                0
minor_claimant_at_initial_assessment       0
rehabilitation_completed                   0
rehabilitation_recommended                 0
work_absence_required                      0
claimant_age_at_incident                   0
liability_type                             0
liability_admission_status                 0
work_absence_duration_days                 0
defendant_title_code                       0
rental_vehicle_expense                     0
home_care_services_cost                    0
```

Figure 2: Missingness profile by variable

2.3 Feature engineering.

Two interpretable attributes were implemented: num_injuries (number of injury labels per claim) and car_damage_severity_num (ordinal mapping of vehicle damage severity). Following these steps, a working dataset with 38 variables and zero residual missingness was built, allowing for consistency in severity scoring and modelling (Aas, 2024; Wilson et al., 2024).

3. Analytical Approach

3.1 Descriptive and exploratory analysis

The motor insurance recovery extract was profiled with df.info(), df.describe(), counts of missing values and category frequencies (FCA, 2025; Tukey, 1977). Missing operational notes (e.g. claims_resolution_exit_notes and claims_management_exit_notes) were preserved and coded as not_provided (no systematic deletion of claims to acknowledge these fields as administrative) (Little and Rubin, 2019; Pham, 2024; Afkanpour et al., 2024). The text was normalised, set-like injury/rehab data were safely parsed and the engineered variables (num_injuries, car_damage_severity_num) were added to generate 38 variables with no remaining missing values (Jäger et al., 2021).

3.2 Analysis of severity and rehabilitation

Injury and rehabilitation labels were converted into lists and exploded for injury-level comparison. Injury volume was dominated by whiplash (495 claims; 49.5%) and spine/back injury (14.1%); and limb injury (13.7%); but, following recent evidence, whiplash is still high burden, heterogeneous on both symptoms and recovery (He et al., 2024; Särkilahti et al., 2024). In summary, a simple AIS/ISS proxy converted injury labels into the ISS-like score (iss_like) and severity bands: minor 76.3%, moderate 21.8%, serious 1.9% (Ede et al., 2023; van Ditshuizen et al., 2025; Eidenbenz et al., 2025).

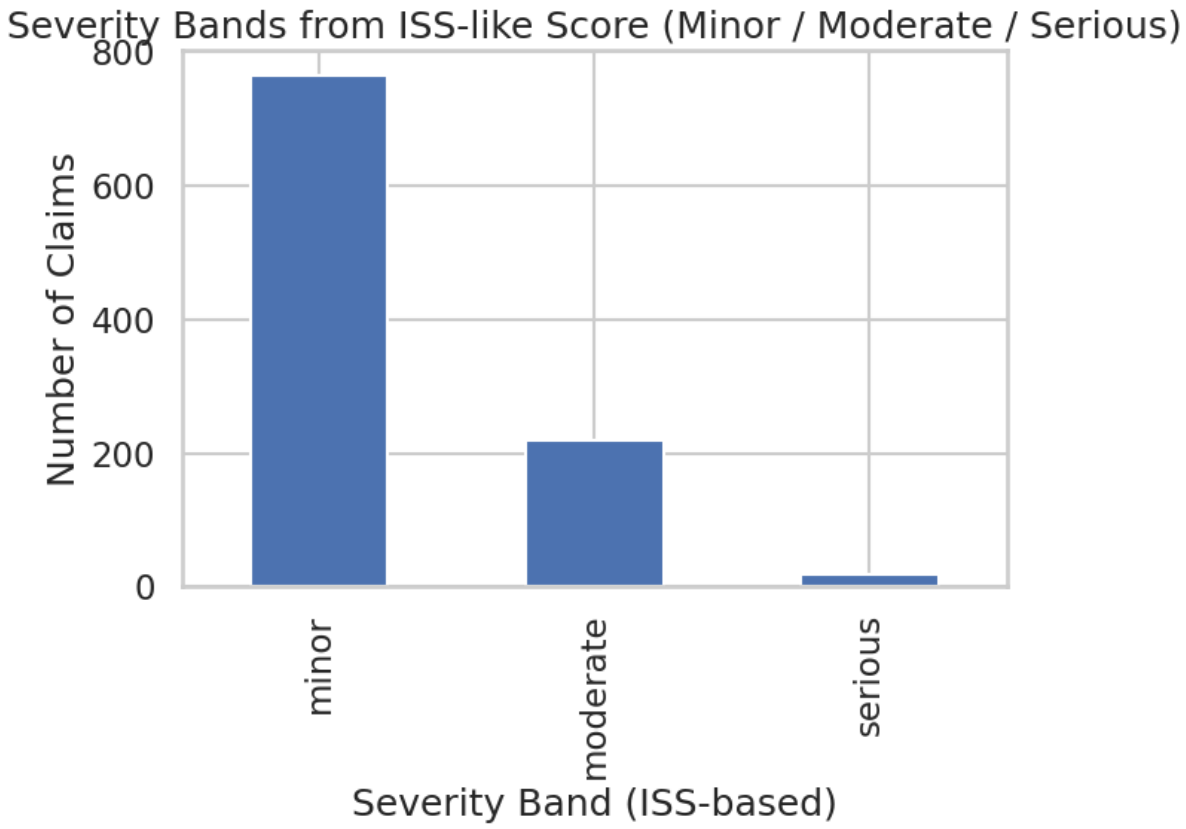


Figure 3: Severity Band Construction

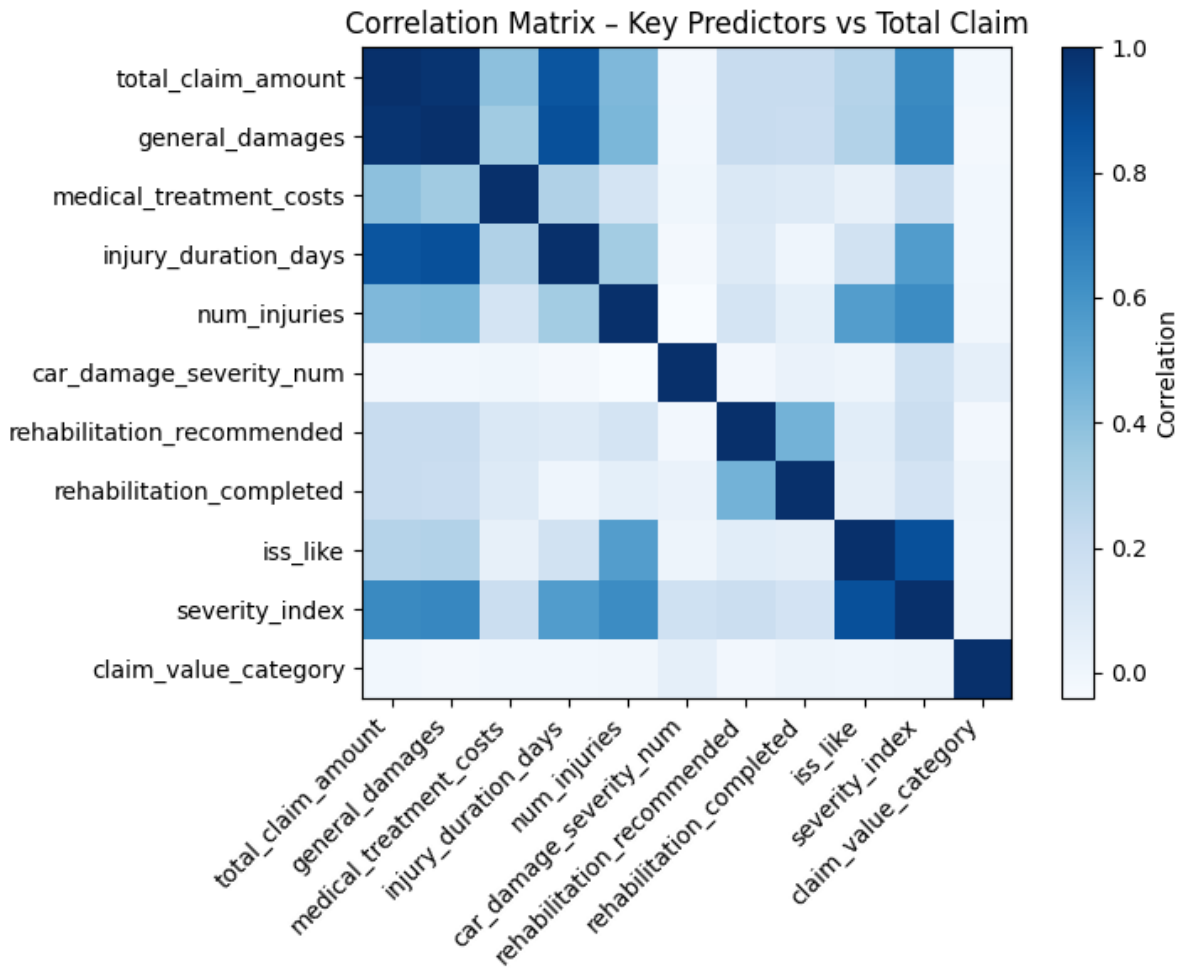


Figure 4: Correlation Matrix

3.3 Predictive modelling / advanced analysis

Correlation screening revealed that injury_duration_days demonstrated the greatest relationship with total_claim_amount (r = 0.853), followed by severity_index (r = 0.640), while iss_like did not (r = 0.276). This is consistent with baseline-first modelling where reasonable models are established prior to more complex learners in regulated insurance models (Wilson et al., 2024; Aas, 2024; Rudin, 2021).

In concrete terms, the model's ladder for claims-cost prediction and triage is: (1) Using linearly based regression (GLM type baseline) to get an interpretable relationship; (2) Employing a Simple Rule Based model (e.g., Duration/Thresholds of Severity) to imitate the way a decision is made on an operating level; then (3) deploying RFs and Gradient Boosting Machine (GBM/Gradient Boosting Regressor) whenever governance allows for non-linearities and interactions (Jaiswal et al., 2024; Holvoet et al., 2023; Rudin, 2021).

This sequencing coincides closely with insurance governance expectations, which state that models impacting claims journeys should be explainable, auditable and monitored. This is especially important when AI/ML is used to influence decisions on behalf of insurance companies. (EIOPA, 2021; ICO, 2020)

3.4 Limitations of the approach

This study is based mainly on observation and finding patterns, so the evidence is about associations. For instance, an observation such as "rehabilitation completion corresponds to higher costs" cannot be treated as rehab leading to more costs as the more severe cases are more likely referred to and are supported by the first time (confounding by indication). In the absence of clearly defined causal designs, findings are meant to be interpreted as evidence for the triage hypotheses rather than causal evidence (Hernán and Robins, 2020; Pearl, 2009).

Secondly, the scores in the severity measures are proxies for data sets: iss_like and severity_index for injury labels represent AIS/ISS-style scores (they are not obtained from systemically coded trauma registry systems), so they are not clinically used, but internally comparison-driven (Eidenbenz et al., 2025; van Ditshuizen et al., 2025).

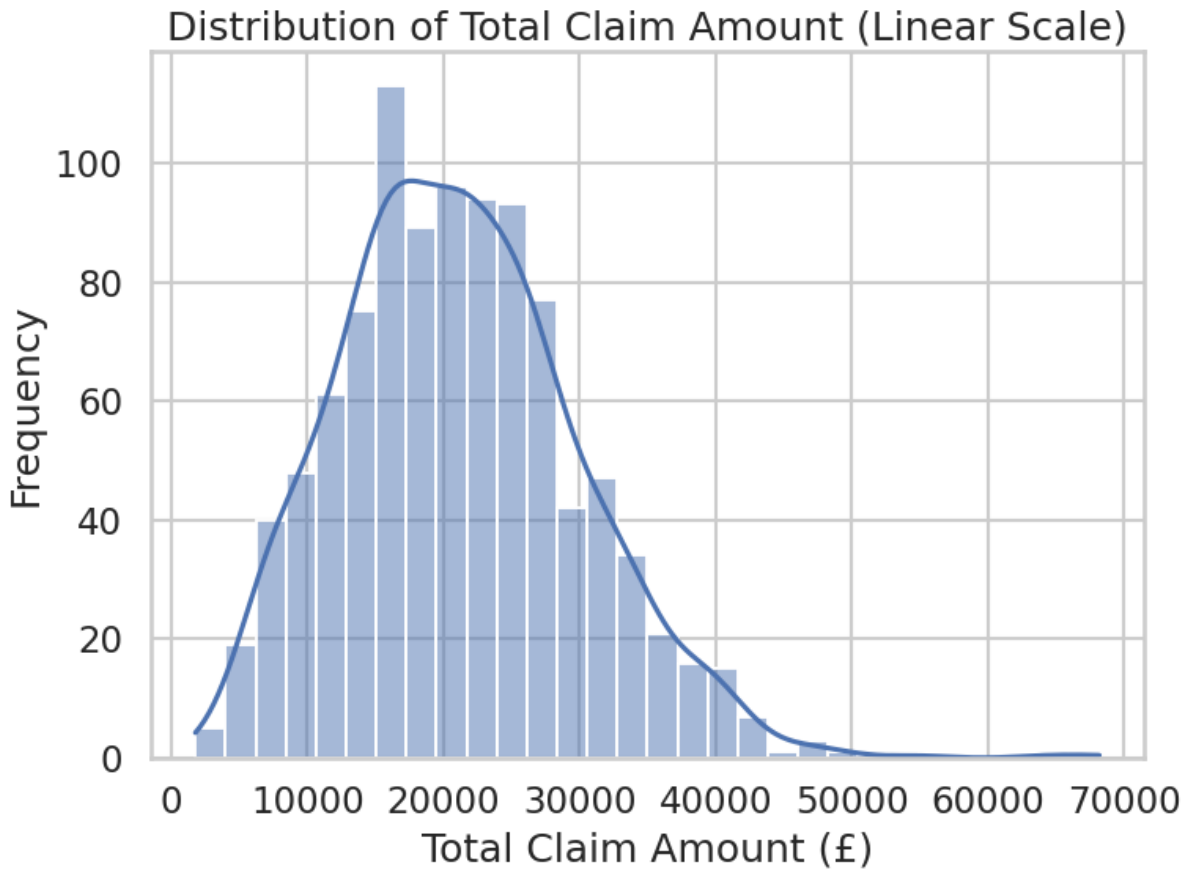
Thirdly, there was high missingness in operational notes and exit fields so not_provided was used for codify documentation which leaves records intact but, in some cases, reduces interpretation of closure rationale and disputes in some cases (Pham, 2024; Afkanpour et al., 2024).

There would be strict governance, transparency and explainability arrangements for operational deployment of models, consistent with guidance on data-driven decision support (ICO, 2020; EIOPA, 2021; EIOPA, 2025).

4. Results

4.1 Descriptive statement of claims

The clean dataset included 1,000 claims with 38 variables. Total claim costs varied greatly: they averaged £21,171 (median £20,639), with a range of £1,792–£68,163. general_damages predominates in the cost structure, with overall cost averages £19,872 and a maximum £56,096; whilst medical_treatment_costs dropped (mean £572 and a maximum £4,466), suggesting that pain, suffering and recovery experience can overshadow direct medical outlay in bodily injury settlement. Recovery was average of a moderate nature; injury_duration_days had an average of 40 days, median 40 and max 150. Length in seeking care was highly skewed: medical_attention_delay_days had a median of 0, but a long tail of 138 days, illustrating the urgency of early engagement and easy claims pathways (FCA, 2025).


Distribution of Total Claim Amount (Linear Scale)
Frequency
0
20
40
60
80
100
0
10000
20000
30000
40000
50000
60000
70000
Total Claim Amount (£)

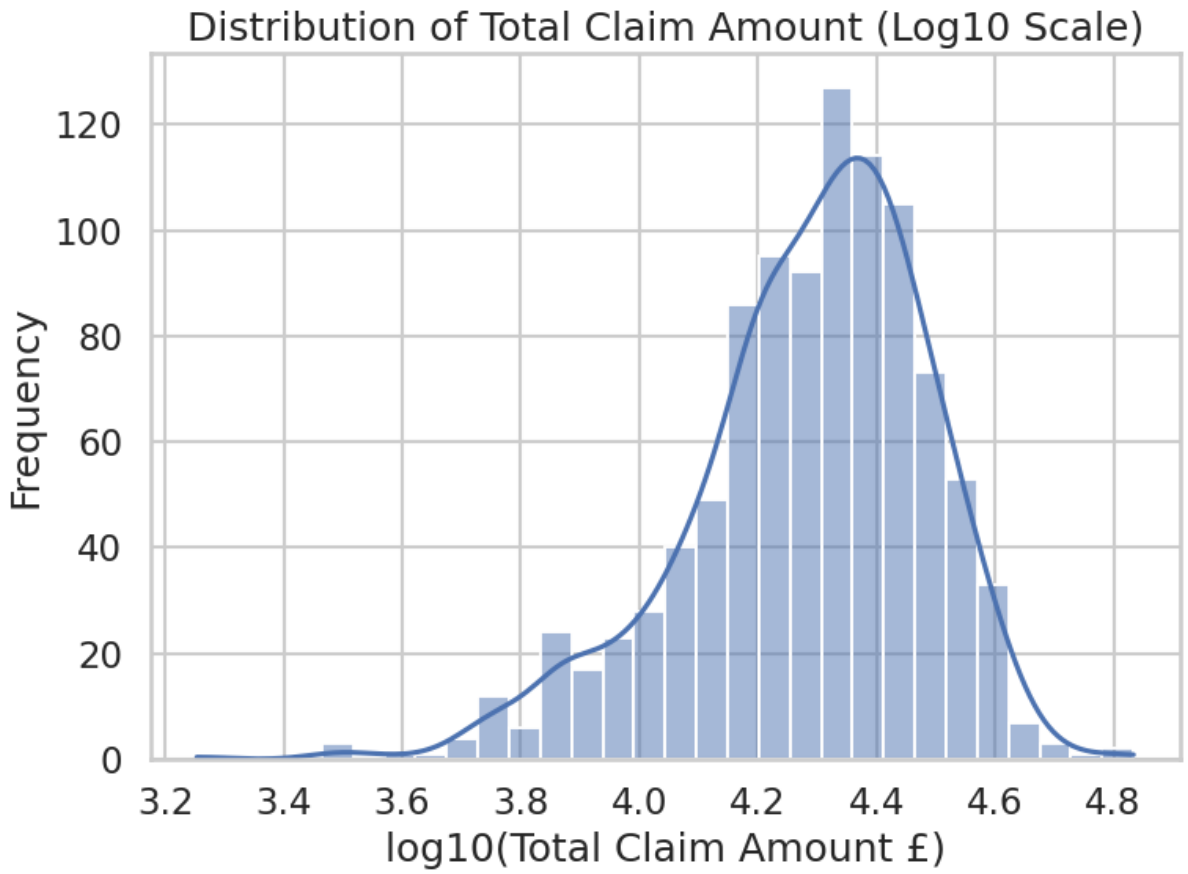
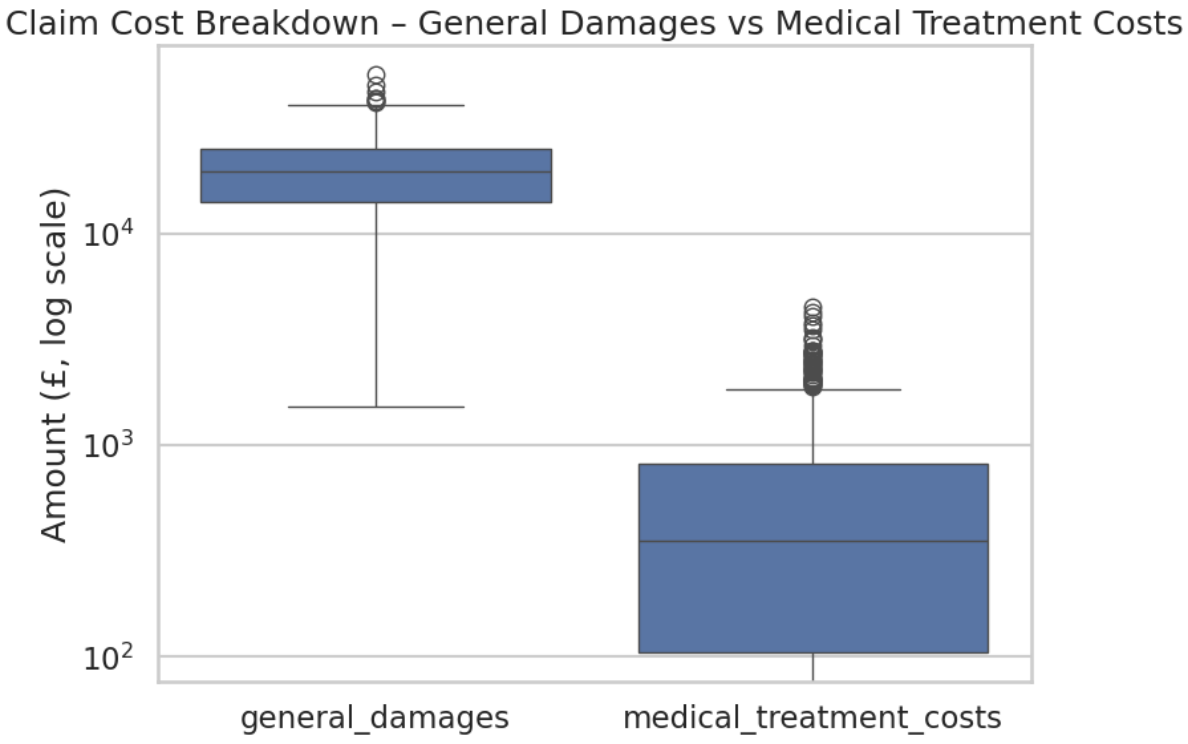


Figure 5: Distribution of Total Claim Amount



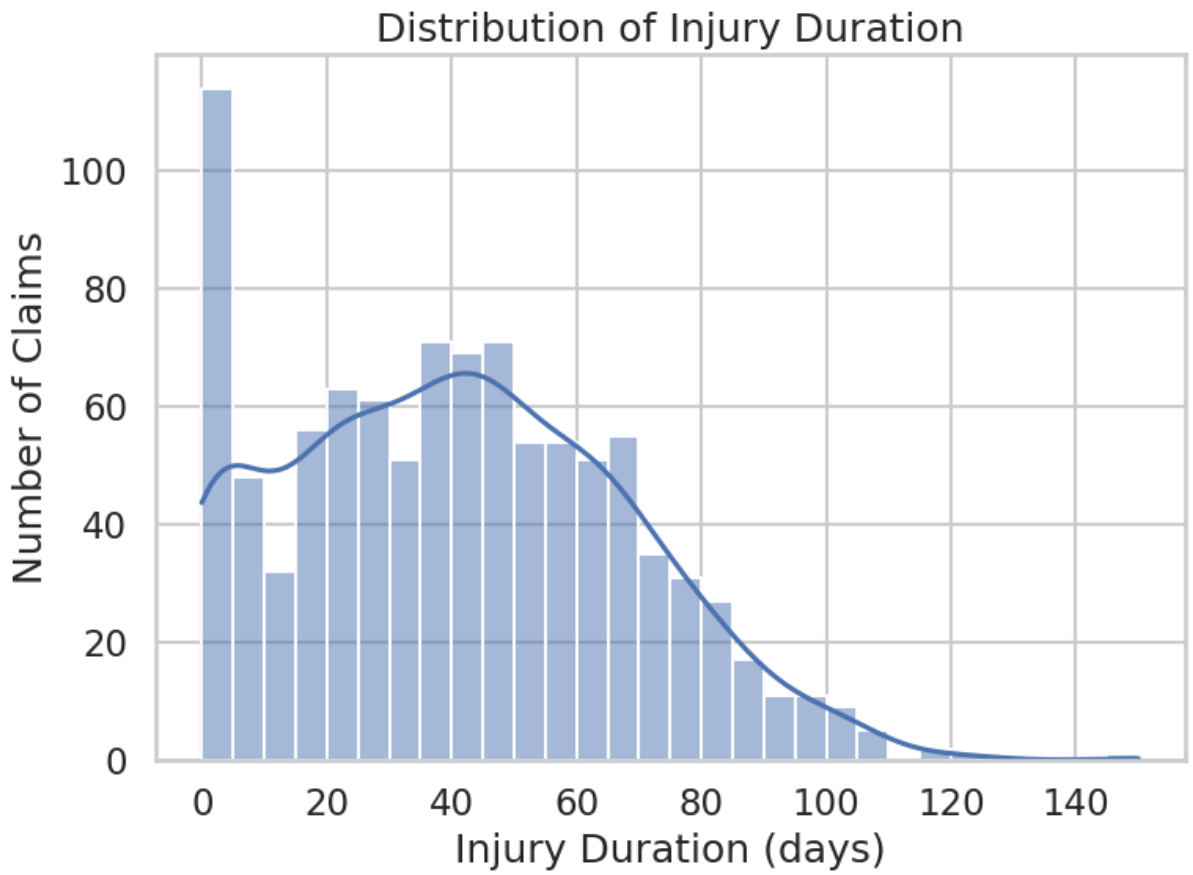


Figure 7: Injury Duration Distribution

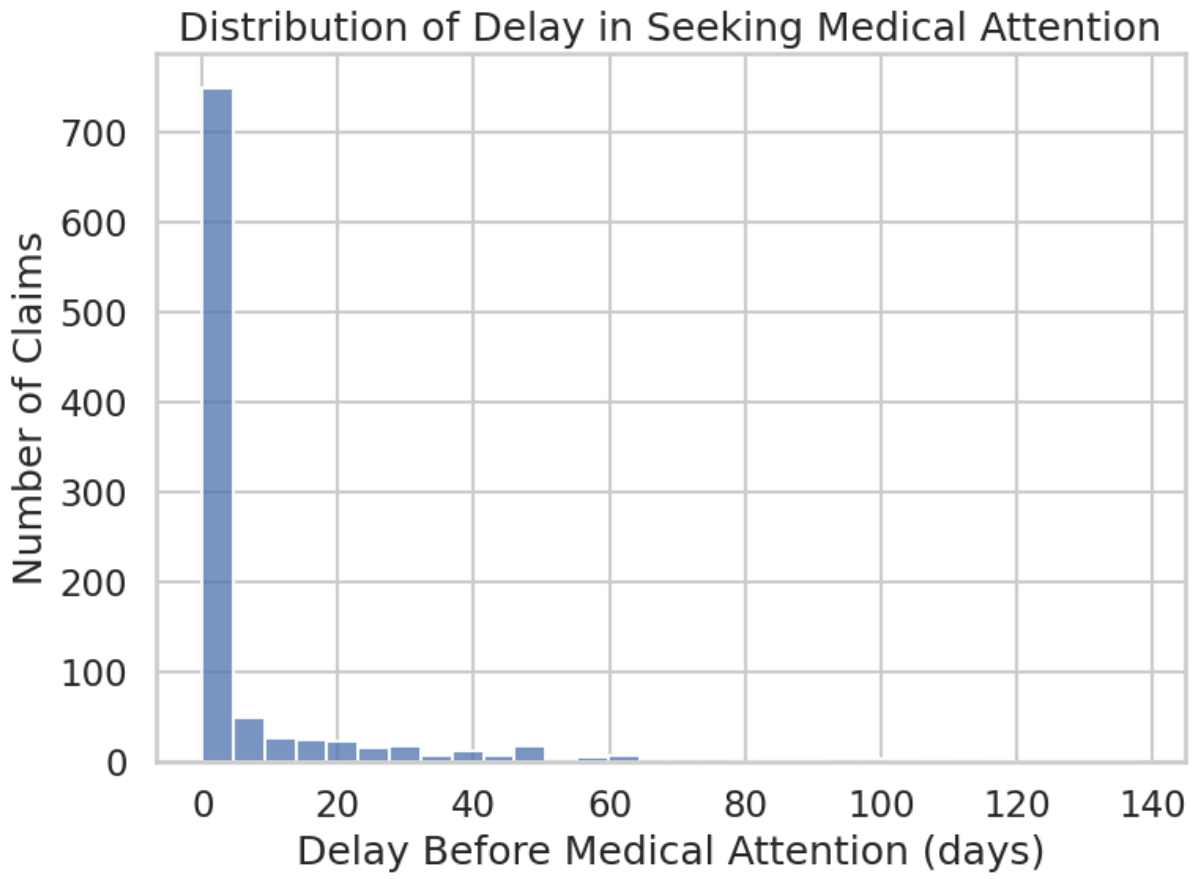


Figure 8: Medical Attention Delay

The spread and upper tail support severity theory in insurance: the costs resulting from bodily injury are right-skewed, with a relatively small number of claims generating disproportional shares of spend —something that is consistent with reports from UK market reports (Aas, 2024; Wilson et al., 2024; ABI, 2025).

4.2 Patterns of injury and severity

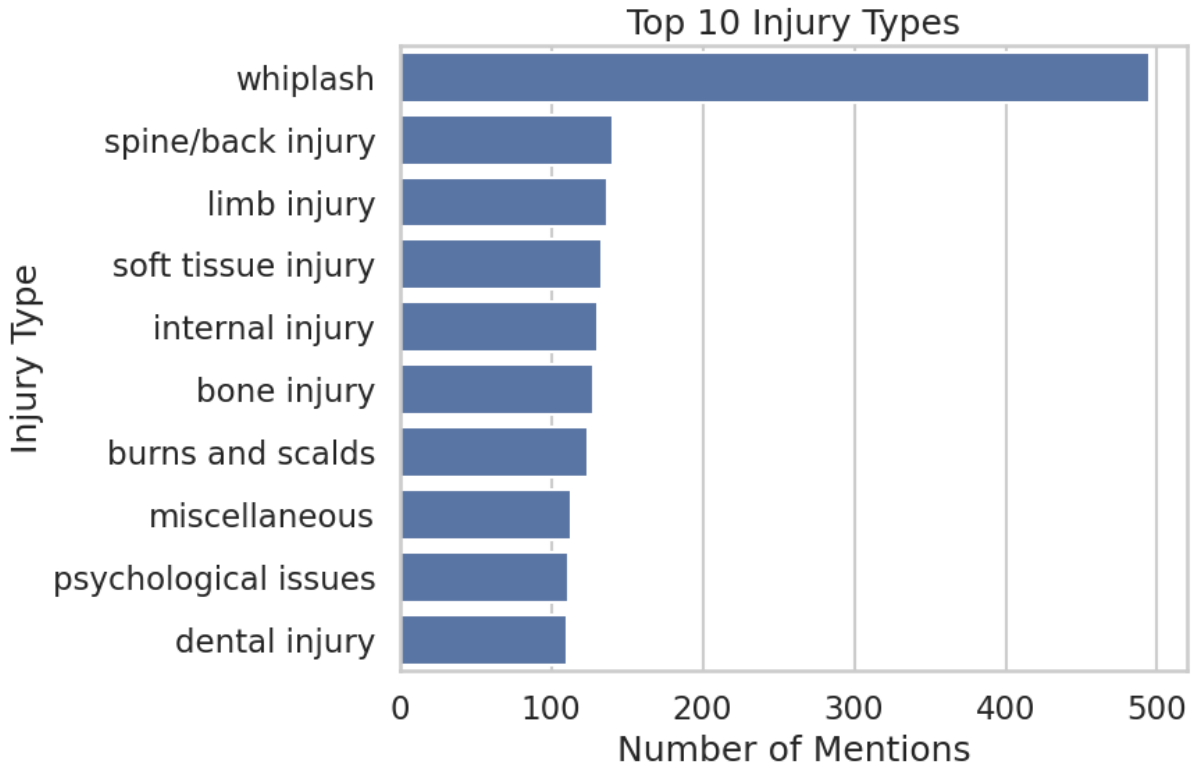
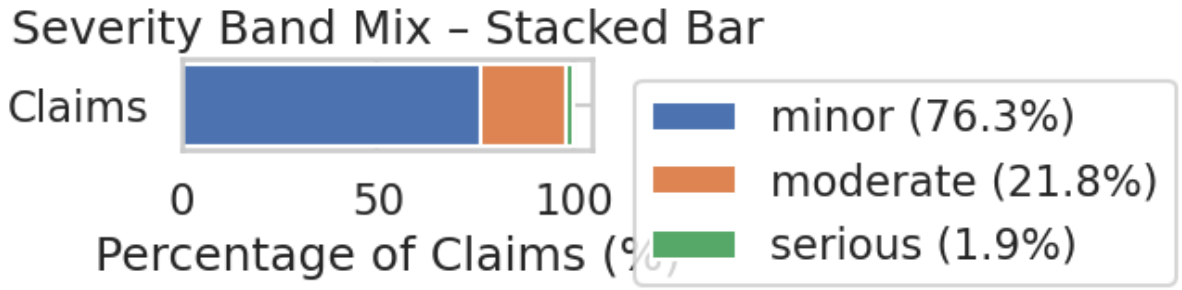


Figure 9: Top Injury Types



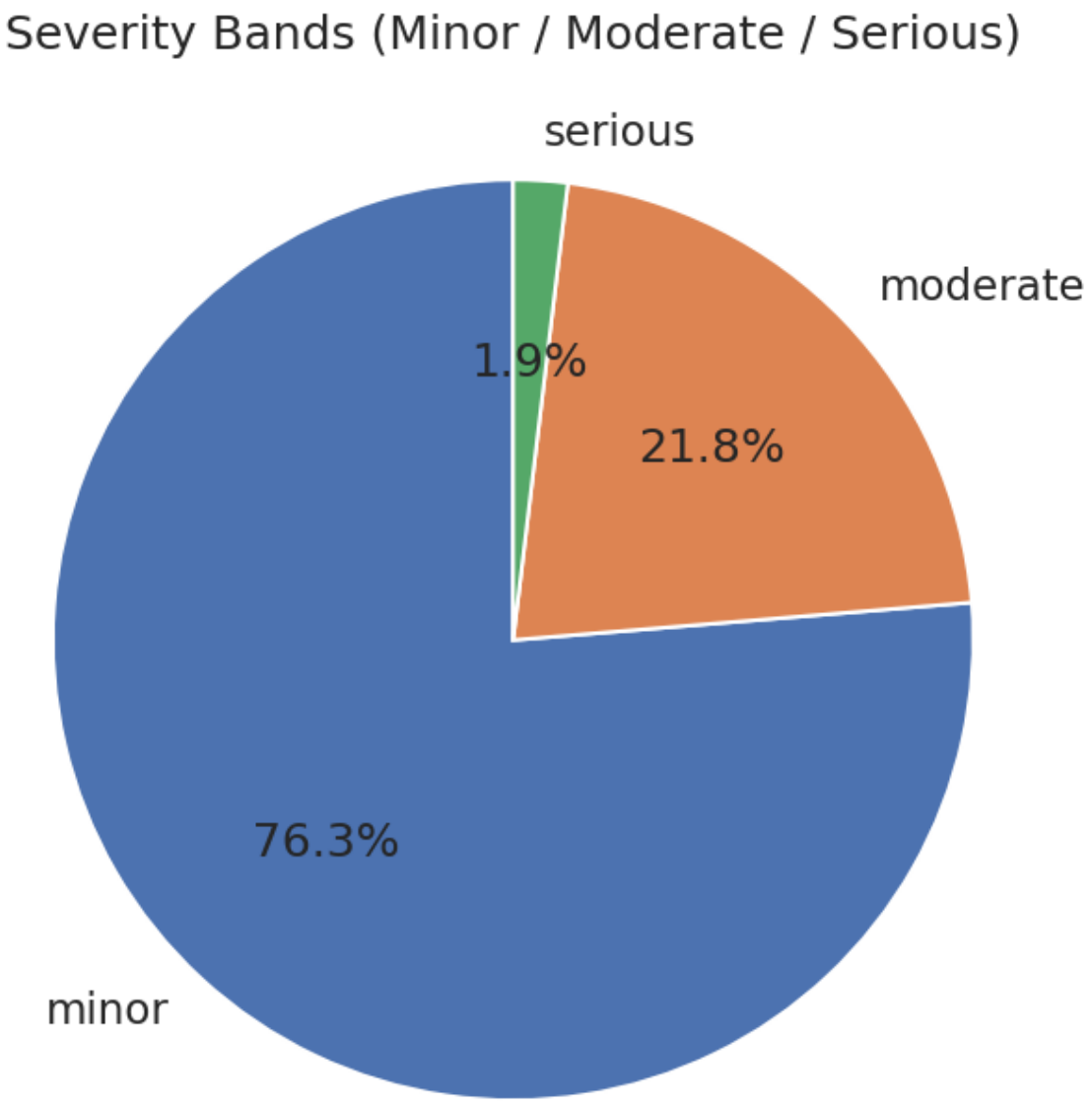


Figure 10: Severity Bands

The most common type of injury is whiplash (49.5%), and following with spine/back, 14.1%, and limb injury, 13.7%. This is consistent with early reviews summarizing that whiplash-associated disorders remain common and variable in terms of burden and recovery modes (He et al., 2024; Särkilahti et al., 2024). The ISS-like scoring resulted in a mean iss_like around 4.83 (max 22) and severity bands: minor 76.3% (763), moderate 21.8% (218), and serious 1.9% (19), consistent with a relatively low severity portfolio with a small high-impact tail (Ede et al., 2023; van Ditshuizen et al., 2025; Eidenbenz et al., 2025).

The prevalence of whiplash further illustrates UK policy design, where low-risk injuries pass through the Official Injury Claim (OIC) route and the whiplash tariff framework shapes evidence requirements and settlement behaviour (OIC, 2025; UK Government, 2025).

4.3 Efficacy and outcomes of rehabilitation

Rehabilitation was often recommended (81.5%) and delivered in 59.1%, leading to a substantial "recommendation-to-completion" disparity. This is important operationally, because protracted duration and fragmented third-party pathways can lengthen settlement times and inflate non-medical costs. This mirrors UK regulatory focus on claims-processing frictions, handovers and customer outcomes (FCA, 2025).

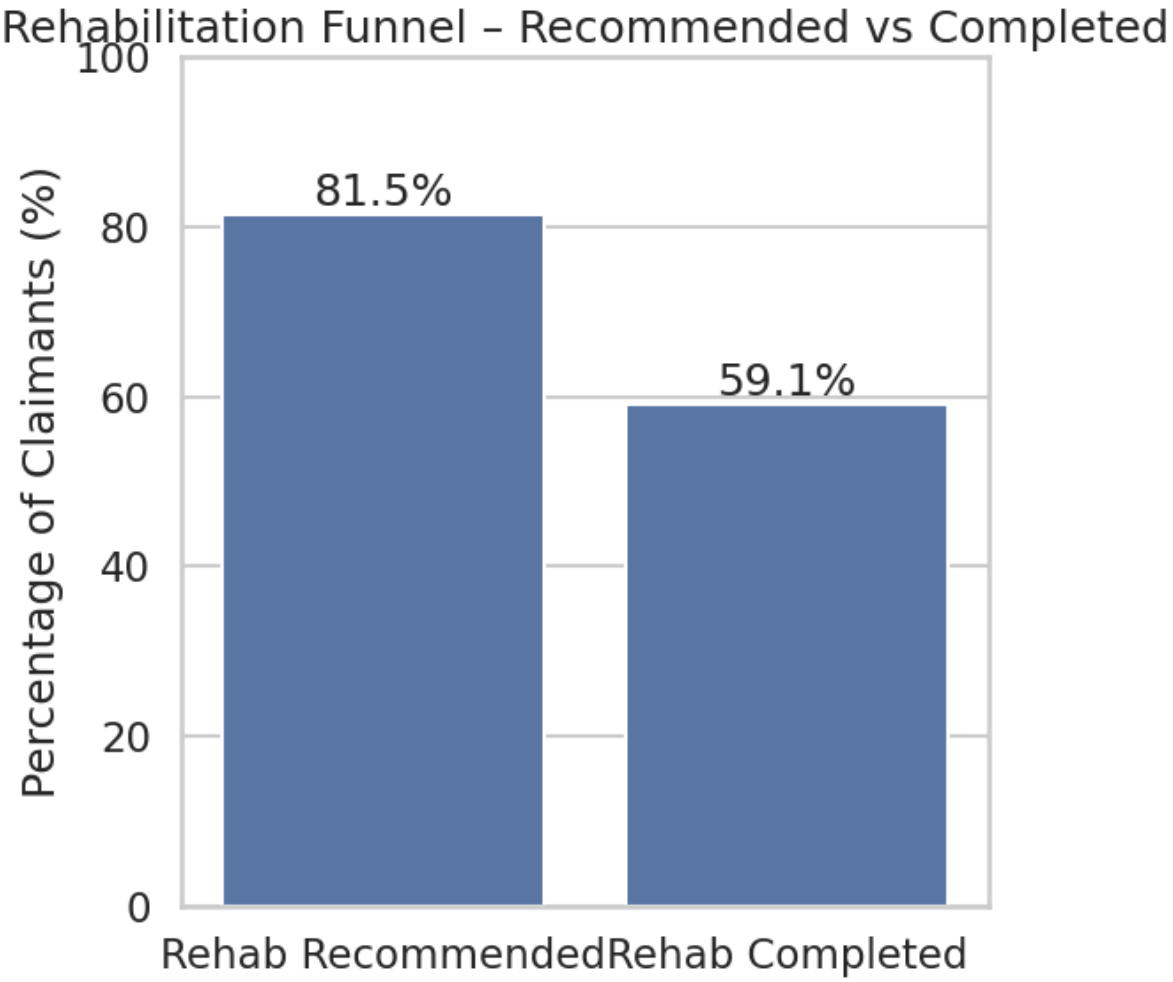


Figure 11: Rehab Funnel (Recommended - Completed)

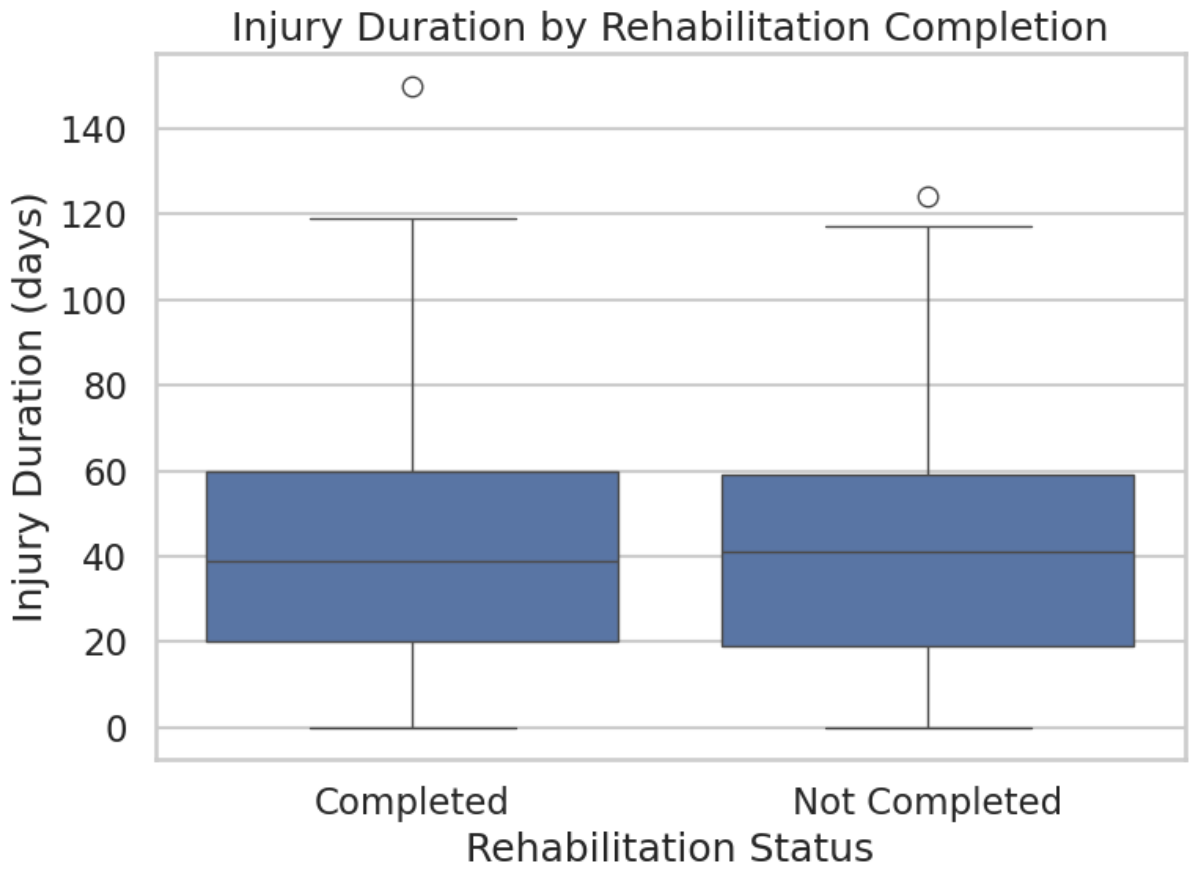


Figure 12: Duration by Rehab Status

4.4 Predictive insights

The strongest statistical signal was duration → cost: injury_duration_days correlated most strongly with total_claim_amount (r = 0.853), followed closely by severity_index (r = 0.640). The anatomical proxy iss_like was weaker (r = 0.276), indicating that recovery and process factors explain more variation in costs than injury labels alone in this data. These further supports initiating with interpretable baselines prior to more complex learners in regulated insurance settings (Rudin, 2021; Wilson et al., 2024; Aas, 2024). Current state of UK market reinforces the importance of solid forecasting in light of claims cost pressure observed (ABI, 2025).

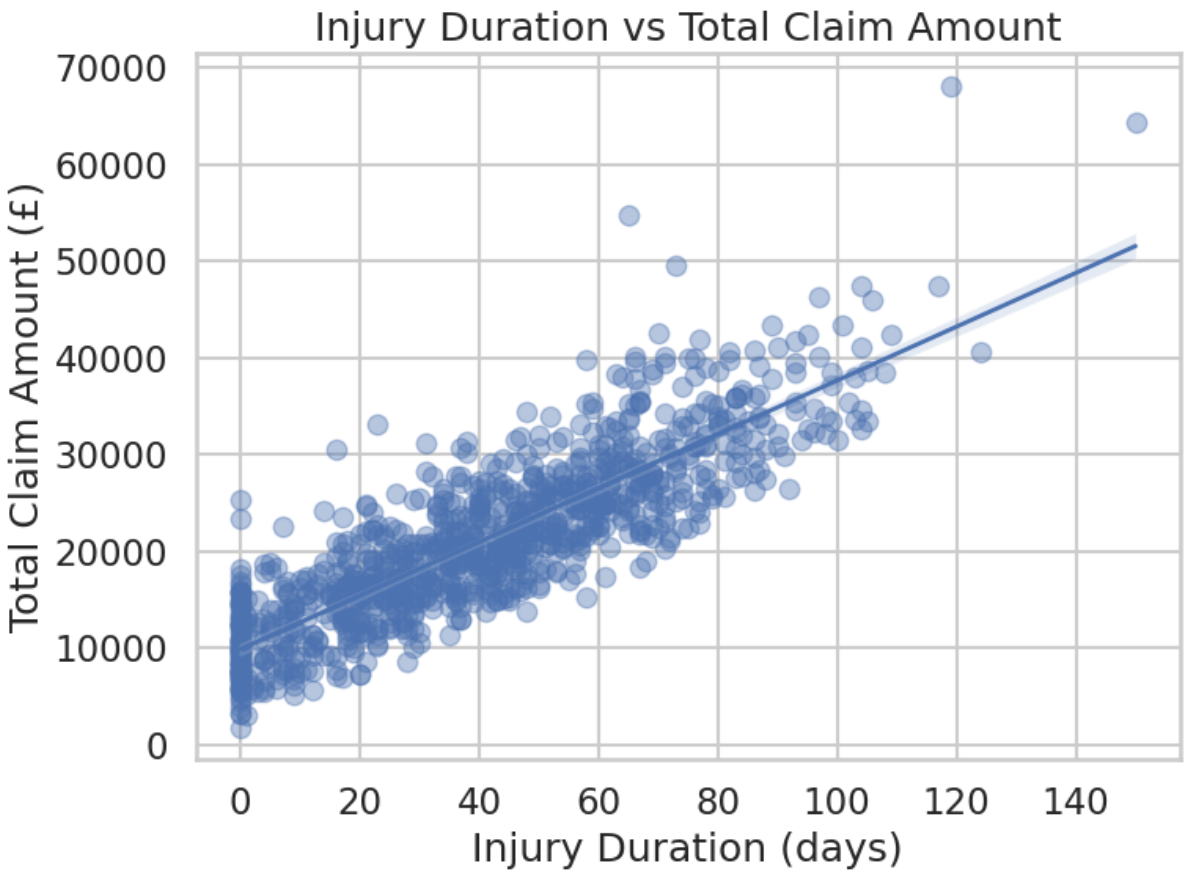


Figure 13: Injury Duration and Total Claim Amount

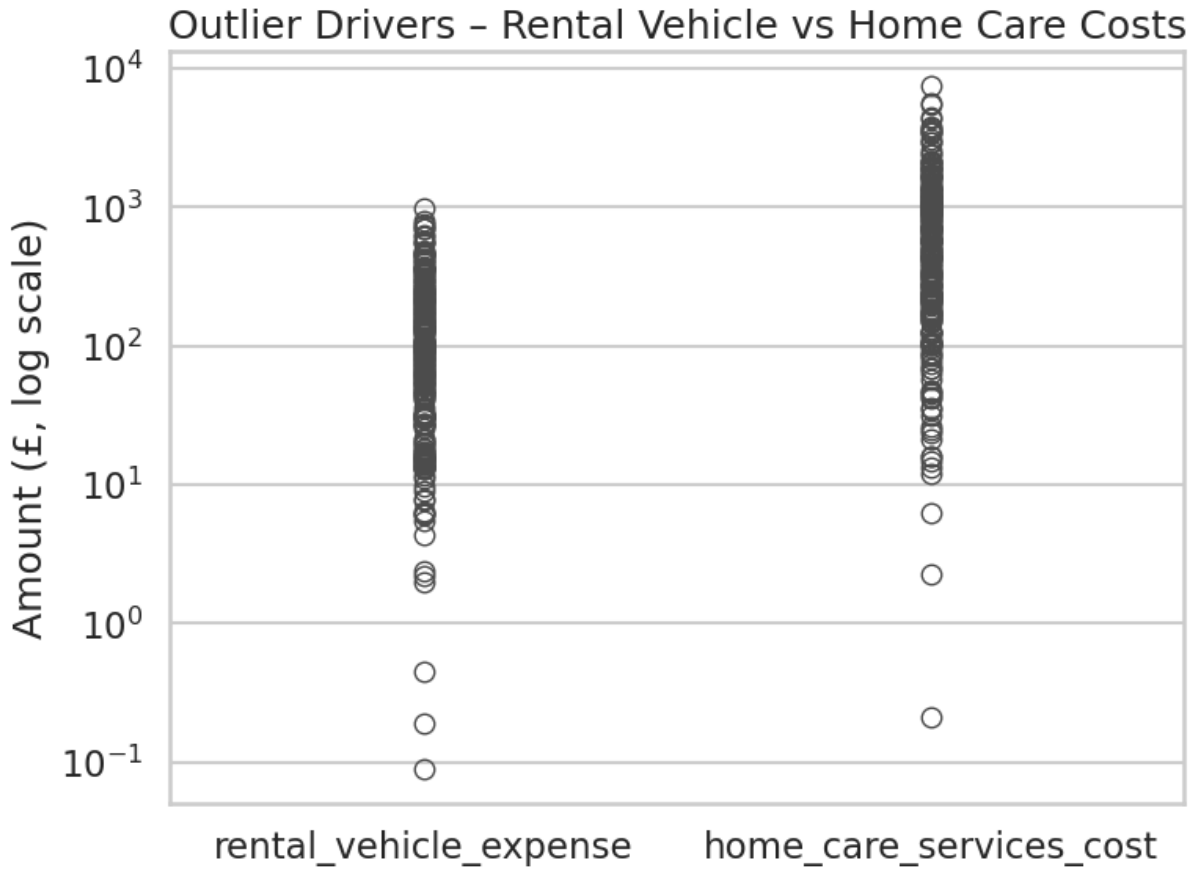


Figure 14: Outlier Drivers (Rental_vehicle_expense and home_care_services_cost)

5.1 Summary of main conclusions.

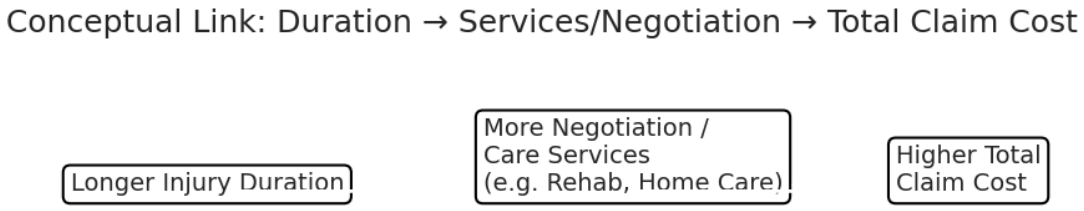


Figure 15: Conceptual Diagram: (Duration - Negotiation/Services - Higher Total Cost)

Altogether, the Typhon extract is similar to the typical UK motor bodily injury portfolio, while most claims are low to moderate in severity, total costs are highly right-skewed, reflecting casualty severity theory and motor-claims cost concentration (Aas, 2024; Wilson et al., 2024). In this dataset, the mean claim cost was £21,171, median was £20,639 and

maximum was £68,163 had general_damages as the cost element (mean £19,872). It also means settlement value is determined more by recovery and negotiation rather than by direct medical costs in most cases. Operational pressure is also evident in cost outliers, particularly rental_vehicle_expense (231 IQR outliers) and home_care_services_cost (229 outliers), suggesting that delays and external service pathways may increase costs in excess of the underlying injury profile (FCA, 2025). These are operational amplifiers; non-injury services e.g., replacement vehicles and support care can go up to absurd levels when the claim stretches long or when multiple middlemen are implicated, and time to resolution is thus a cost driver in and of itself (FCA, 2025). Finally, the pace of recovery is central; injury_duration_days mean was 40 days (max 150) and has the highest relation for claim cost ($r = 0.853$), indicating duration as a practical triage cue. This matches the claims settlement logic: general damages and negotiation track time to recovery and time to resolution, not merely the clinical label of injury (FCA, 2025; Aas, 2024). 5.2 Triggers and patterns. The descriptive analysis, though there are multiple plausible mechanisms that may inform the observed patterns. First, the robust duration–cost relationship likely reflects clusters of longer duration of symptoms, repetitive interventions, work absence and long negotiation, which can carry over both overall damages and ancillary costs. Second, rehabilitation associations need some care: with finished rehab cases you may have selection effects (more complicated claims are referred), not rehab pushing up costs. The causal question is counterfactual — what would have been the costs of the same claims of rehabilitation having been completed versus not being completed — that call for explicit causal designs rather than straightforward group comparisons (Hernán and Robins, 2020; Pearl, 2009). Third, settlement pathways are influenced by the UK policy environment: whiplash reforms, tariffing and the OIC route shapes the required evidence, negotiation behaviour and dispute resolution routes (UK Government, 2025; OIC, 2025). Ultimately, external frictions linked to credit hire and intermediaries can extend timelines and increase costs; this is in line with the dataset's large tail of replacement-vehicle costs and the FCA's focus on quality of claims management (FCA, 2025). 5.3 Limitations. Inferences were hindered by three limitations. The first one was observational results as they are still associative and may have been influenced by unobserved confounding and selection (Hernán and Robins, 2020). Second, severity metrics (iss_like, severity_index) are proxy measures drawn from label mappings, not trauma registry data coded clinically, which supports internal comparison rather than clinical classification (Ede et al., 2023; van Ditshuizen et al., 2025; Eidenbenz et al., 2025). Third, process note and closure were missing from many of the files and not_provided was coded, keeping records but limiting interpretability of closure reasons and escalation pathways (Pham, 2024; Afkanpour et al., 2024). 5.4 Practical implications and guidelines. From a claim's operations perspective, the evidence supports early triage where risk of duration is involved through speedy primary contact, proactive evidence gathering and early health care pathways where initial signs of long recovery are detected. The strong high outlier concentration in rental_vehicle_expense and home_care_services_cost makes the case for stricter control over the journey of a replacement vehicle and faster quantification/liability resolution (FCA, 2025) to prevent unnecessary cost escalation. Given the distance between intervention to compliance (81.5%) and recommendations to completion (59.1%) which relates to rehabilitation strategies, meaningful improvements and action are suggested in respect of claimant engagement, appointment logistics and provider coordination. For analytics and FinTech design purposes the results would inform the establishment of explainable baselines

(linear regression / GLM/ Rule-Based models) as a baseline, using duration, delay, severity proxy, damage indicator for reserving and triage and then proceed to apply them. For advanced approaches (Random Forest/GBM), organizations should strive to produce documentation, monitoring, bias checks, and clear explanation, like the governance model should for automated or data-driven assistance tools (ICO, 2020; EIOPA, 2021; EIOPA, 2025).

Conclusion.

This analysis was conducted on a cleaned Typhon extract comprising 1,000 UK claims of motor bodily injury and uncovered a portfolio of lower-severity harm with significant cost volatility. The average total_claim_amount is £21,171 (Max £68,163), and general_damages (mean £19,872) drives costs, not direct medical expenditure (medical_treatment_costs mean £572). Recovery played an important role: Injury_duration_days averaged 40 days (max 150) and had the greatest association with cost (r = 0.853), suggesting duration is a key operational control point. UK market realities were reflected in the injury mix, with whiplash representing 49.5% of injury claims, highlighting the importance of tariffing, as well as the OIC pathway for evidence and settlement pathways (UK Government, 2025; OIC, 2025). There was high albeit incomplete rehabilitation use (81.5% recommended vs 59.1% completed), suggesting there's potential to reduce preventable delay by improving engagement and provider coordination. The results are compounded by a broader context of increased claims-cost pressure in the UK market with industry commentary of high motor claims costs, and regulators also focusing on claims handling quality and customer outcomes (ABI, 2025; FCA, 2025). The limited design and depth of data will be improved for future work. It requires more robust causal inference (e.g., propensity score methods and causal graphs) to separate the effect of rehabilitation from confounding by severity (Hernán and Robins, 2020; Pearl, 2009). They could further validate severity proxies in richer clinically coded fields to increase validity (Ede et al., 2023; van Ditshuizen et al., 2025). Finally, operational variables which are associated with a UK motor cost inflation—credit-hire duration/rate, repair cycle time, legal representation, number of contracts, dispute timestamps and handoffs—would be incorporated to explain and predict the heavy tail. Any transition to automated triage must remain bottom-up first and be governed with documentation, monitoring and explainability aligned to ICO/EIOPA expectations (ICO, 2020; EIOPA, 2021; EIOPA, 2025).

APPENDICES

Appendix A: Dataset profile

Rows: 1,000 claims, columns: 36 variables.

Data types: 17 integer, 8 float, 11 object (categorical/free-text).

Key values (before cleaning):

total_claim_amount: £1,792.15 to £68,162.57 (mean £21,171.39, median £20,639.37).

general_damages: £1,528.17 to £56,096.19 (mean £19,871.81).

medical_treatment_costs: £0.00 to £4,465.83 (mean £572.20).

injury_duration_days: 0 to 150 (mean 40.67, median 40).

medical_attention_delay_days: 0 to 138 (mean 8.42, median 0).

Appendix B: Missing value (before cleaning)

Column	Missing (count)
Claims_resolution_exit_notes	901
claims_management_exit_notes	896
claim_rejection_code	229
liability_denial_reasons	221
claims_resolution_closure_code	203

After cleaning: operational note fields were coded to not_provided and the data frame finished with 0 missing values.

Appendix C: Cleaning and feature engineering summary for Task A

Object columns normalised (strip + lowercase).

Set-like strings safely parsed into consistent comma-separated labels:

- ✓ injury_type_classifications
- ✓ rehabilitation_program_types

Binary-like columns converted to nullable integer (Int64).

Added engineered features:

- ✓ num_injuries
- ✓ car_damage_severity_num

Final shape: 1,000 × 38 (added the two engineered fields).

Appendix D: Outlier screening (IQR method, top 10)

Column	Outlier count
rental_vehicle_expense	231
home_care_services_cost	229
rehabilitation_recommended	185
medical_attention_delay_days	184
asset_utility_loss_compensation	155
hospital_admission_required	104
medical_care_sought	83
career_satisfaction_loss_compensation	76
num_injuries	54
claim_value_category	53

Appendix E: Severity outputs for Task B

ISS-like score (iss_like) summary

Mean 4.831

Std 4.203

Min 1, Q1 1

Median 4, Q3 8,

Max 22

Composite severity index (severity_index) summary

Mean 12.180

Std 5.456

Min 2.21, Q1 8.2625

Median 11.145, Q3 15.1925

Max 35.49

ISS severity bands

Minor: 763 (76.3%)

Moderate: 218 (21.8%)

Serious: 19 (1.9%)

Correlation matrix highlights

Ijury_duration_days vs total_claim_amount: r = 0.853

Severity_index vs total_claim_amount: r = 0.640

Iss_like vs total_claim_amount: r = 0.276

Data-driven index vs total_claim_amount: r = 0.980

Data-driven index vs injury_duration_days: r = 0.871

Appendix F: Collaborative Project Management Using GitHub

7.1 Team roles, communication, and coordination

The project team consisted of five members, and responsibilities were allocated to reflect individual strengths while maintaining balanced participation. Two members specialised in the data-driven aspects of the work, focusing on data cleaning, exploratory analysis, and the development of severity-scoring and predictive models. Their technical capability ensured the robustness, accuracy, and reproducibility of the analytical outputs. The remaining three members concentrated on research, contextual analysis, and report writing, drawing on strengths in academic communication, critical evaluation, and synthesis of relevant literature. This division of roles allowed the team to integrate quantitative insights with well-evidenced narrative interpretation.

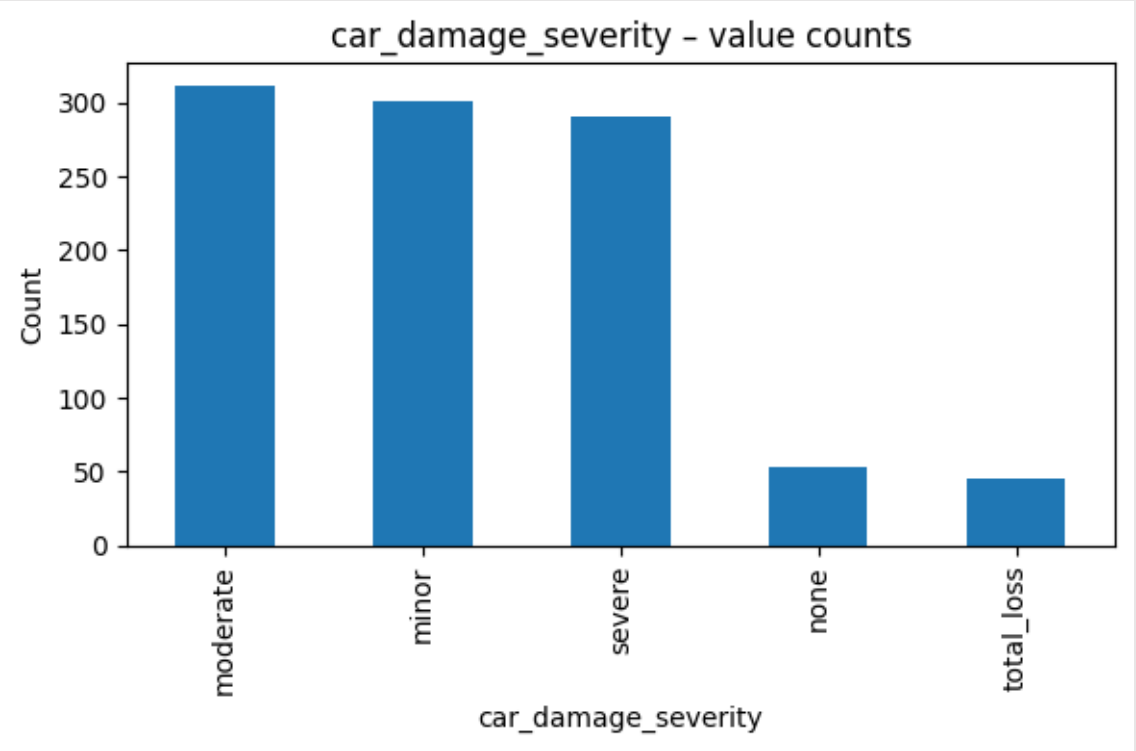
Communication and coordination were achieved through a combination of structured meetings and digital platforms. The group held regular in-person meetings in the university library to discuss progress, allocate tasks, and review drafts collaboratively. Additional group computer lab sessions were used to work synchronously on coding, troubleshoot technical issues, and validate outputs from the analytical models. GitHub served as the central hub for documentation and version control, enabling the team to track contributions, manage commits, and ensure the codebase remained consistent. Pull requests and issue logs helped maintain transparency in the development process.

Alongside these formal structures, the group used WhatsApp as a practical communication channel for sharing documents, arranging meeting times, and coordinating day-to-day tasks. This combination of synchronous meetings, collaborative lab work, and efficient digital communication helped maintain momentum throughout the project.

Overall, the team demonstrated effective coordination by leveraging diverse skill sets, maintaining clear communication channels, and working collaboratively toward a cohesive and evidence-driven final result.

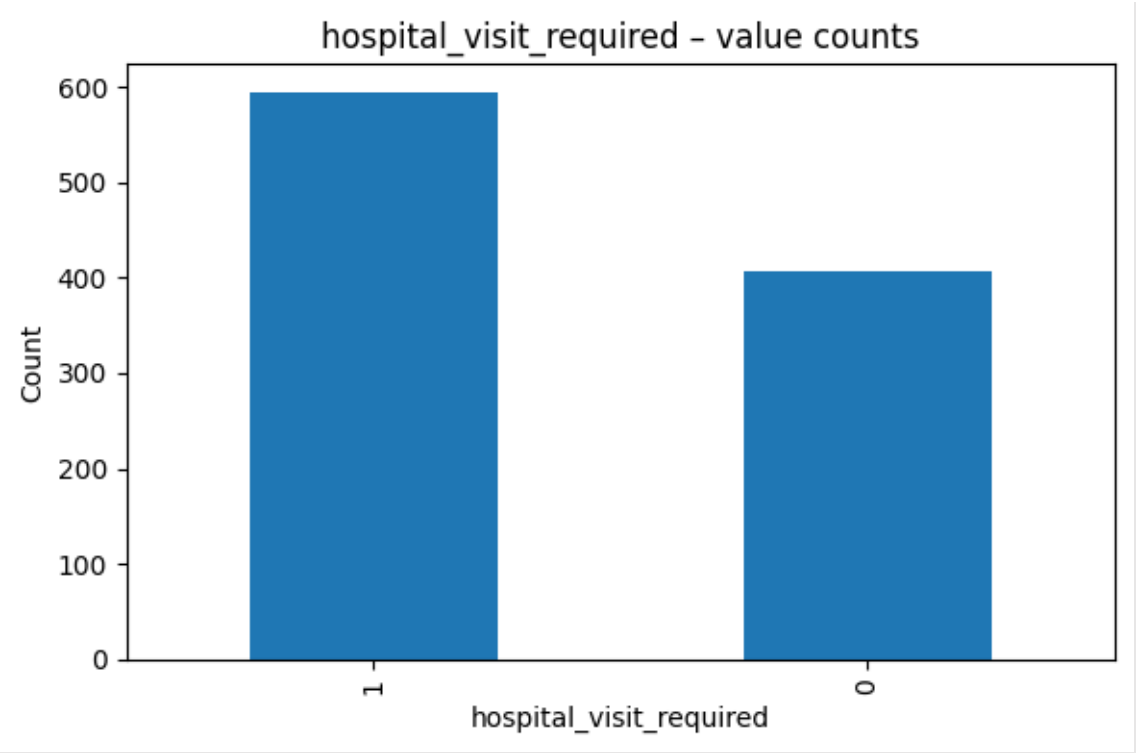
Appendix G: Tables and Figures

moderate	311	31.1
minor	301	30.1
severe	290	29.0
none	53	5.3
total_loss	45	4.5



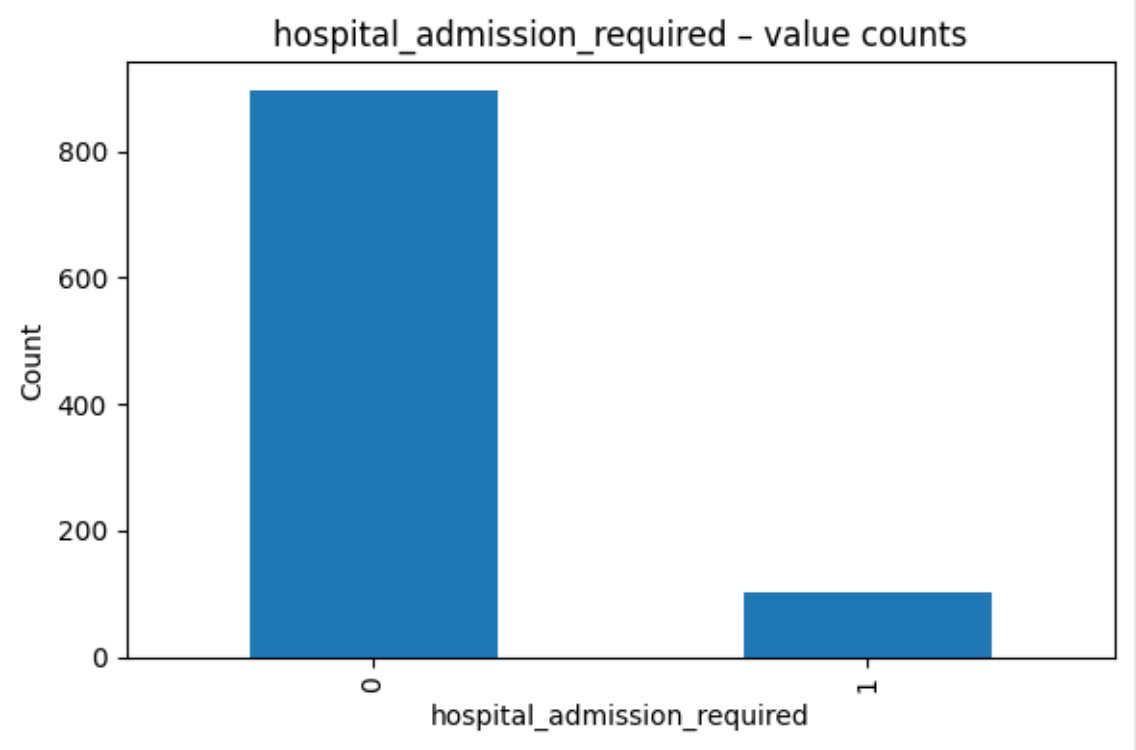
[TaskA] Frequency table for hospital_visit_required:

hospital_visit_required	count	percent
1	594	59.4
0	406	40.6



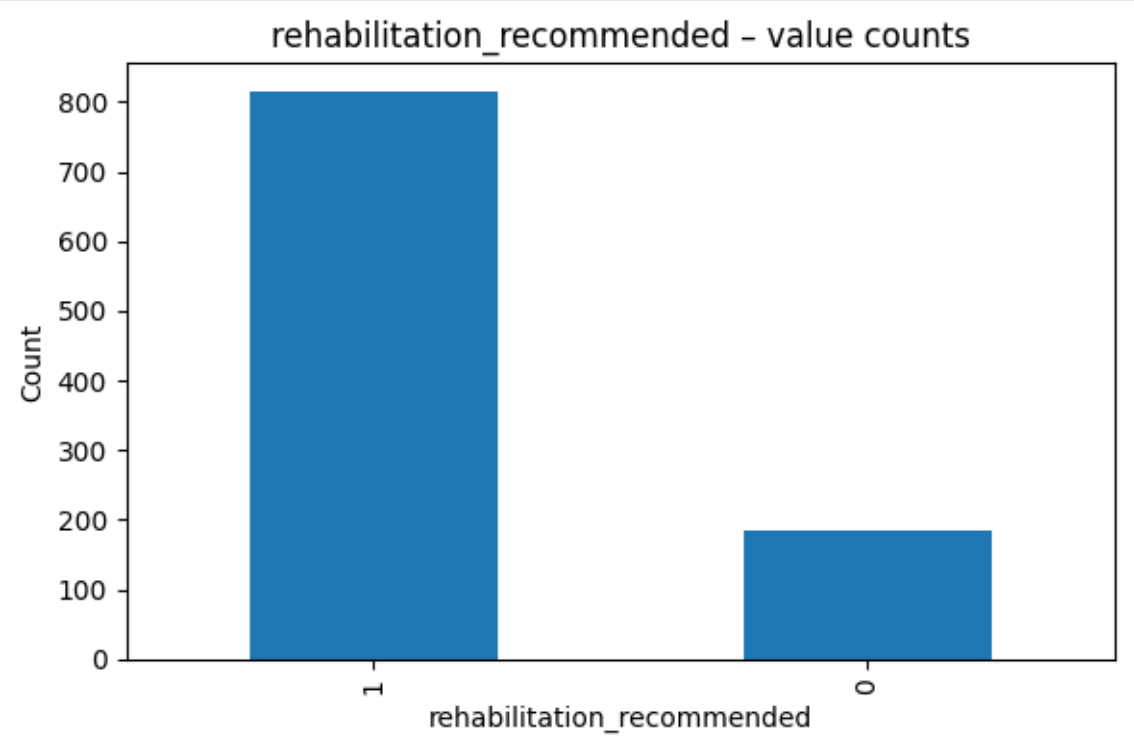
[TaskA] Frequency table for hospital_admission_required:

	count	percent
hospital_admission_required		
0	896	89.6
1	104	10.4



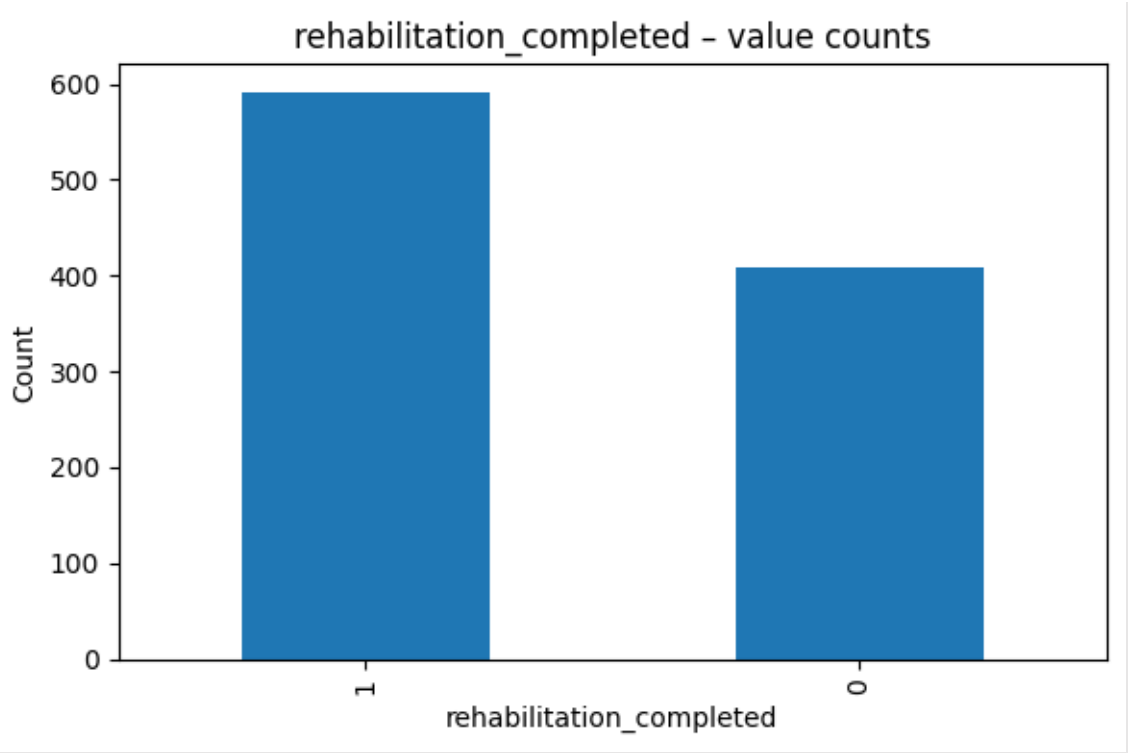
[TaskA] Frequency table for rehabilitation_recommended:

	count	percent
rehabilitation_recommended		
1	815	81.5
0	185	18.5



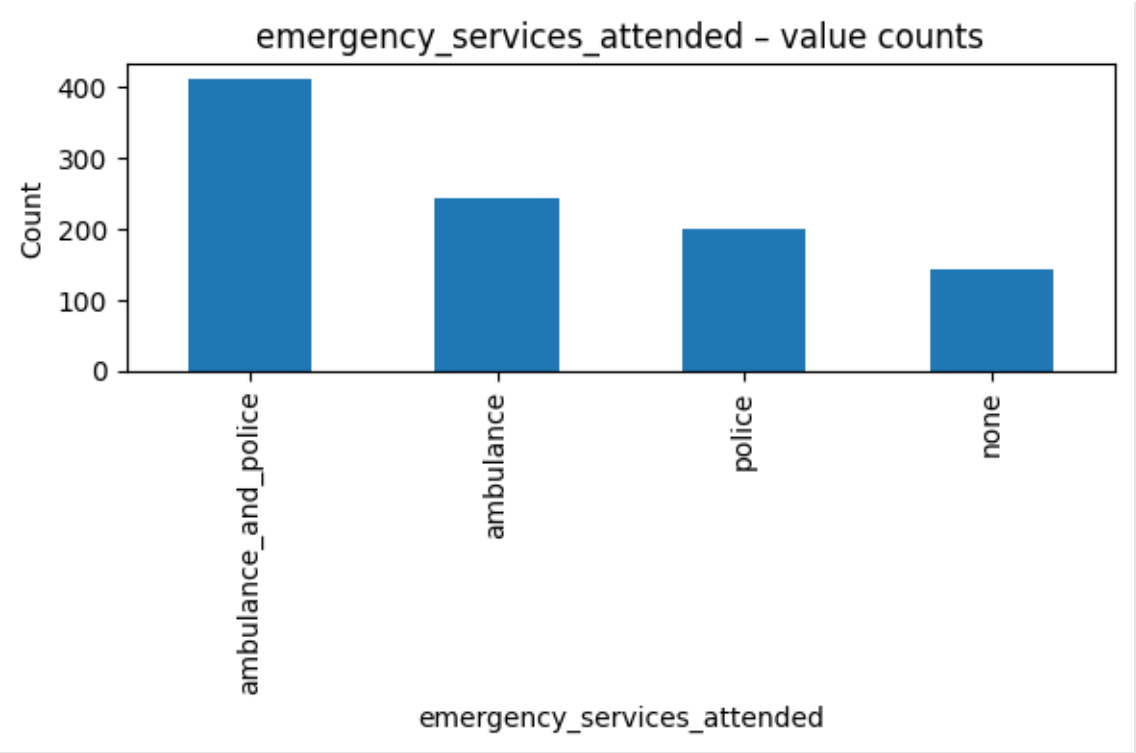
[TaskA] Frequency table for rehabilitation_completed:

	count	percent
rehabilitation_completed		
1	591	59.1
0	409	40.9



[TaskA] Frequency table for emergency_services_attended:

	count	percent
emergency_services_attended		
ambulance_and_police	412	41.2
ambulance	244	24.4
police	200	20.0
none	144	14.4



[TaskA] Summary for total_claim_amount:

count 1000.000000

mean 21171.389070

std 8795.517385

min 1792.150000

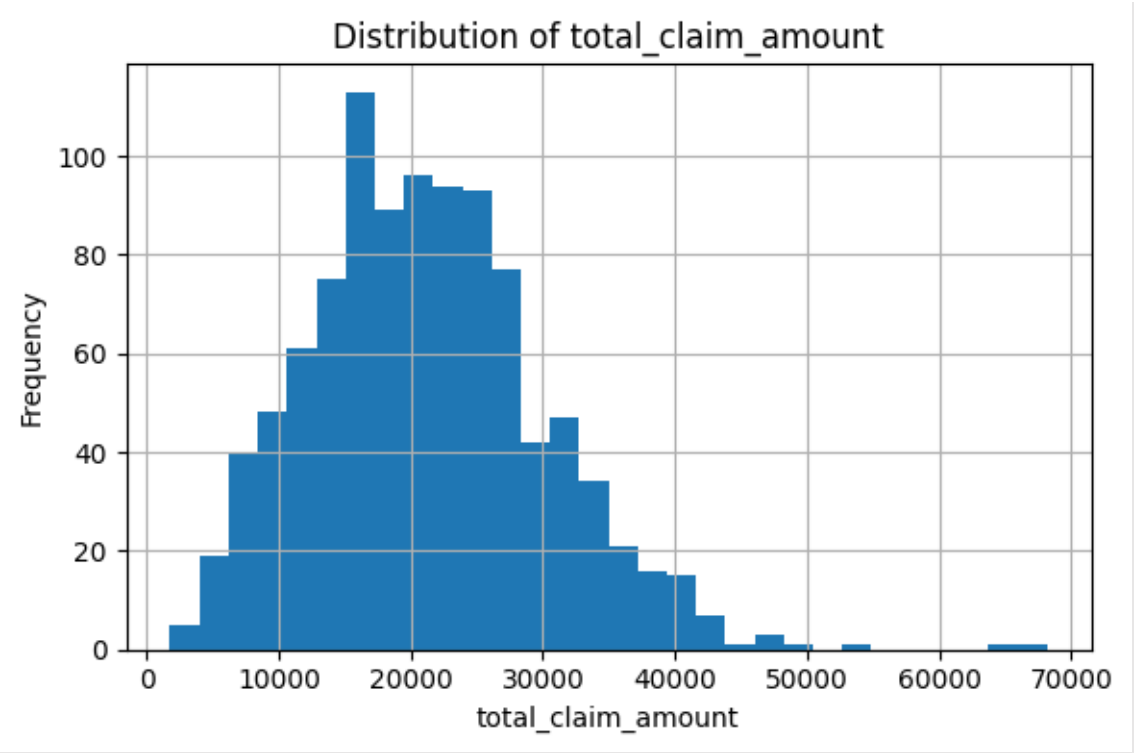
25% 15108.585000

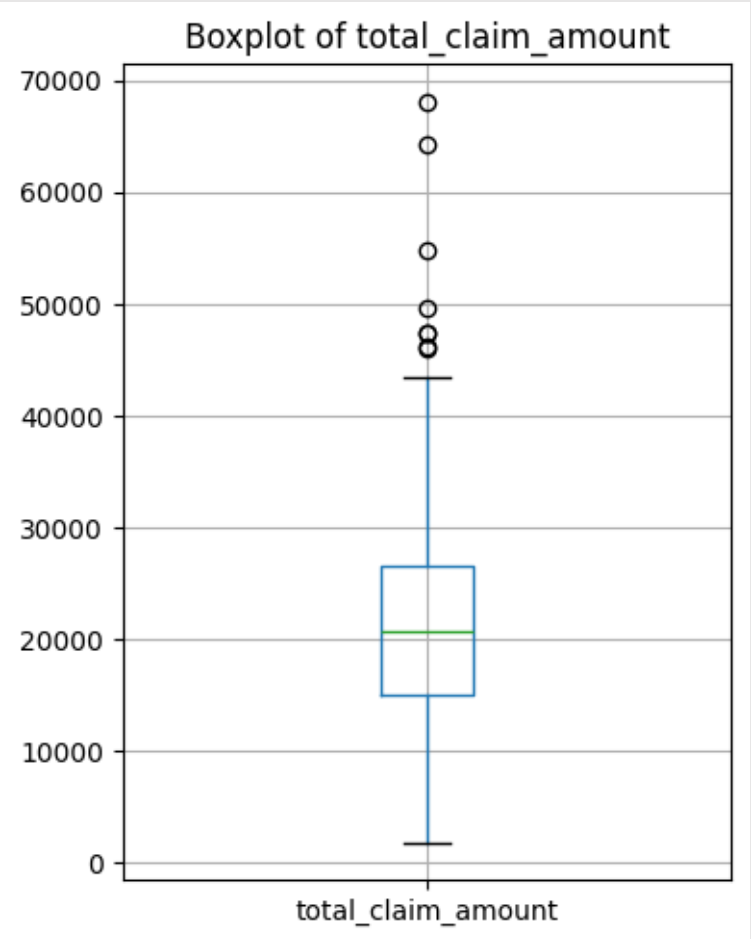
50% 20639.365000

75% 26504.892500

max 68162.570000

Name: total_claim_amount, dtype: float64





[TaskA] Summary for general_damages:

count 1000.000000

mean 19871.809310

std 8094.735036

min 1528.170000

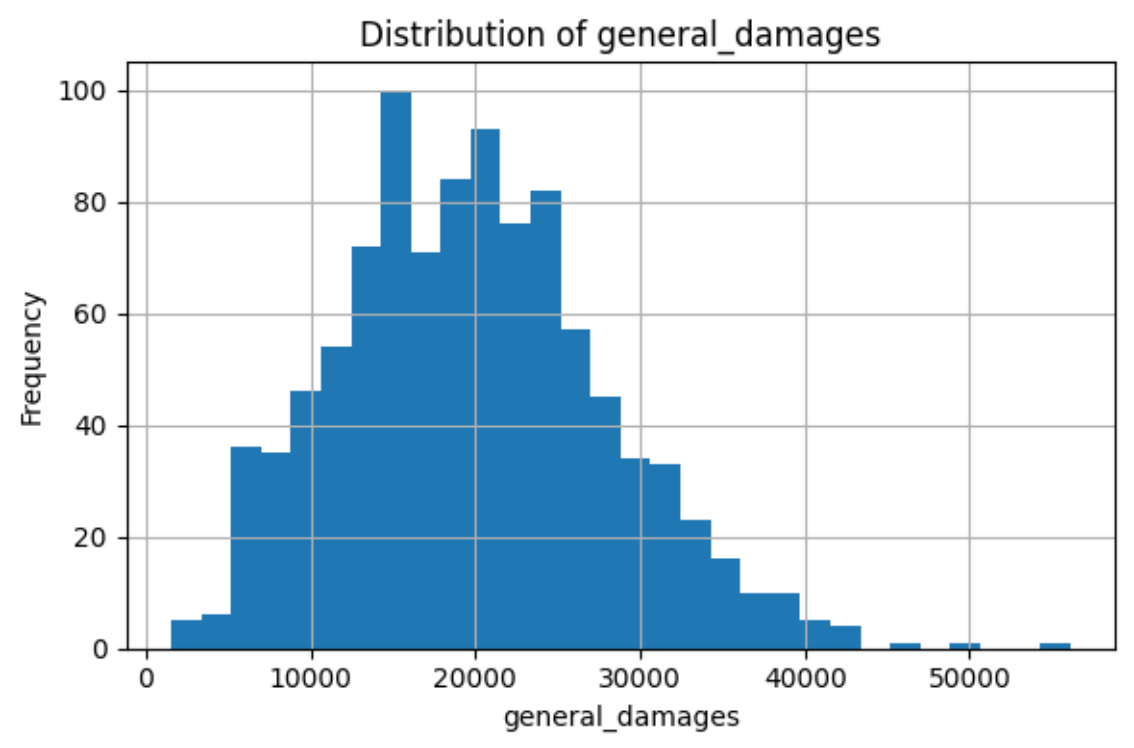
25% 14125.140000

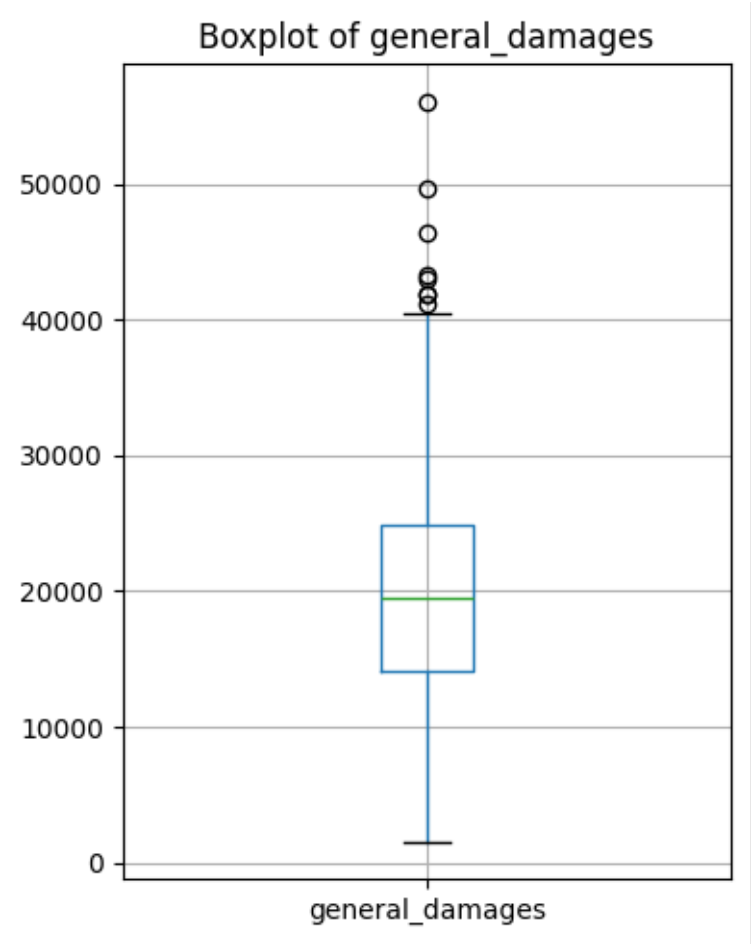
50% 19546.390000

75% 24931.320000

max 56096.190000

Name: general_damages, dtype: float64





[TaskA] Summary for medical_treatment_costs:

count 1000.000000

mean 572.198200

std 653.537592

min 0.000000

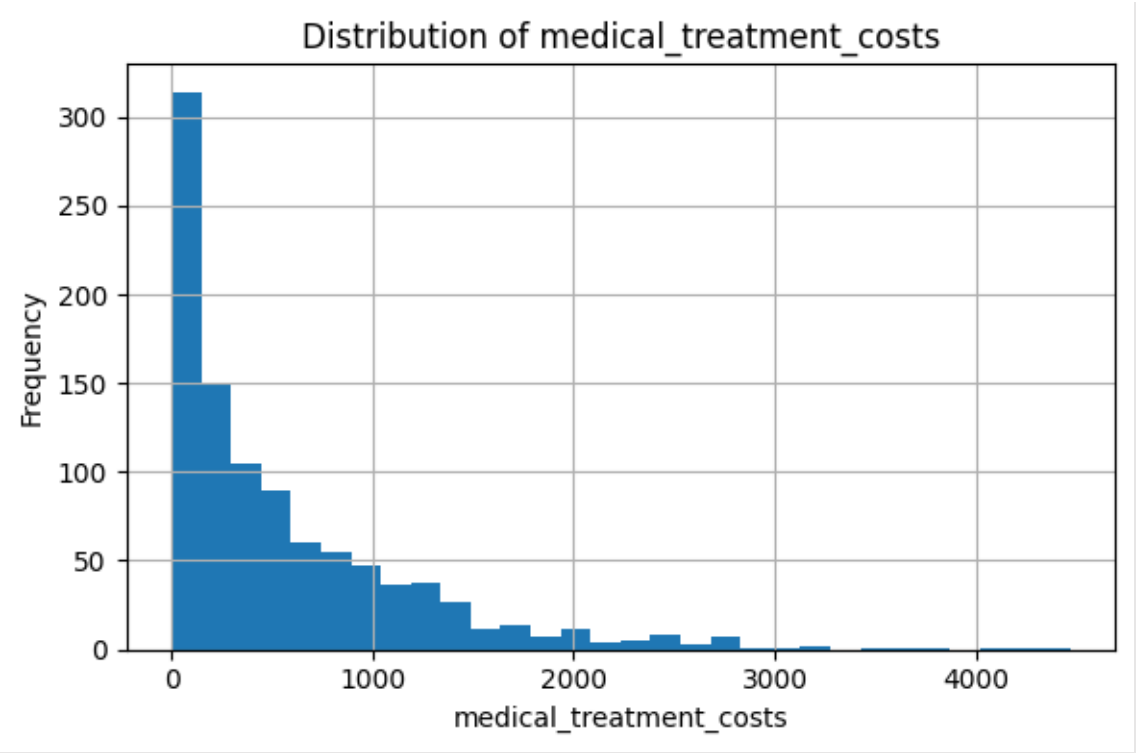
25% 105.305000

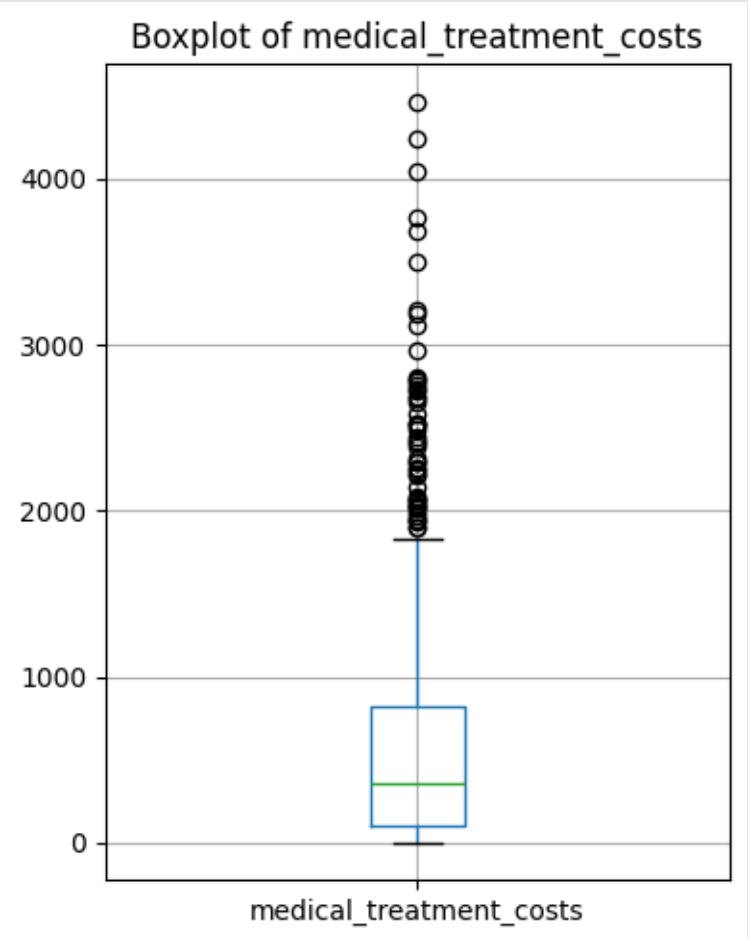
50% 355.760000

75% 821.005000

max 4465.830000

Name: medical_treatment_costs, dtype: float64

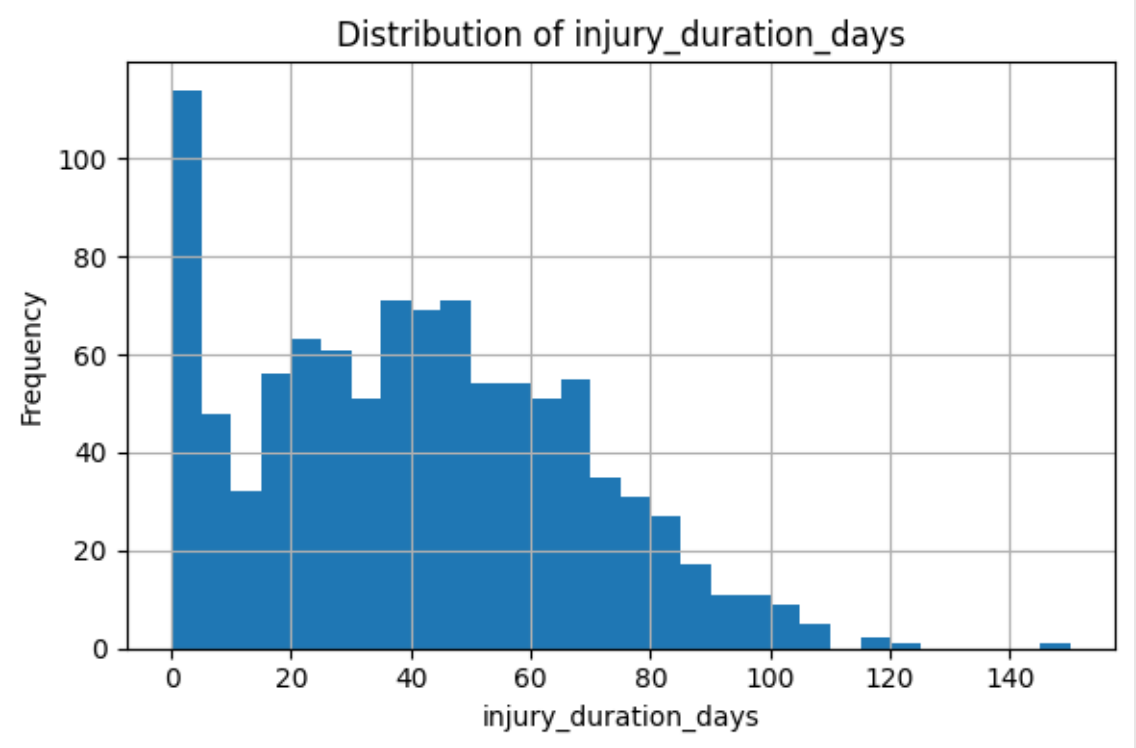


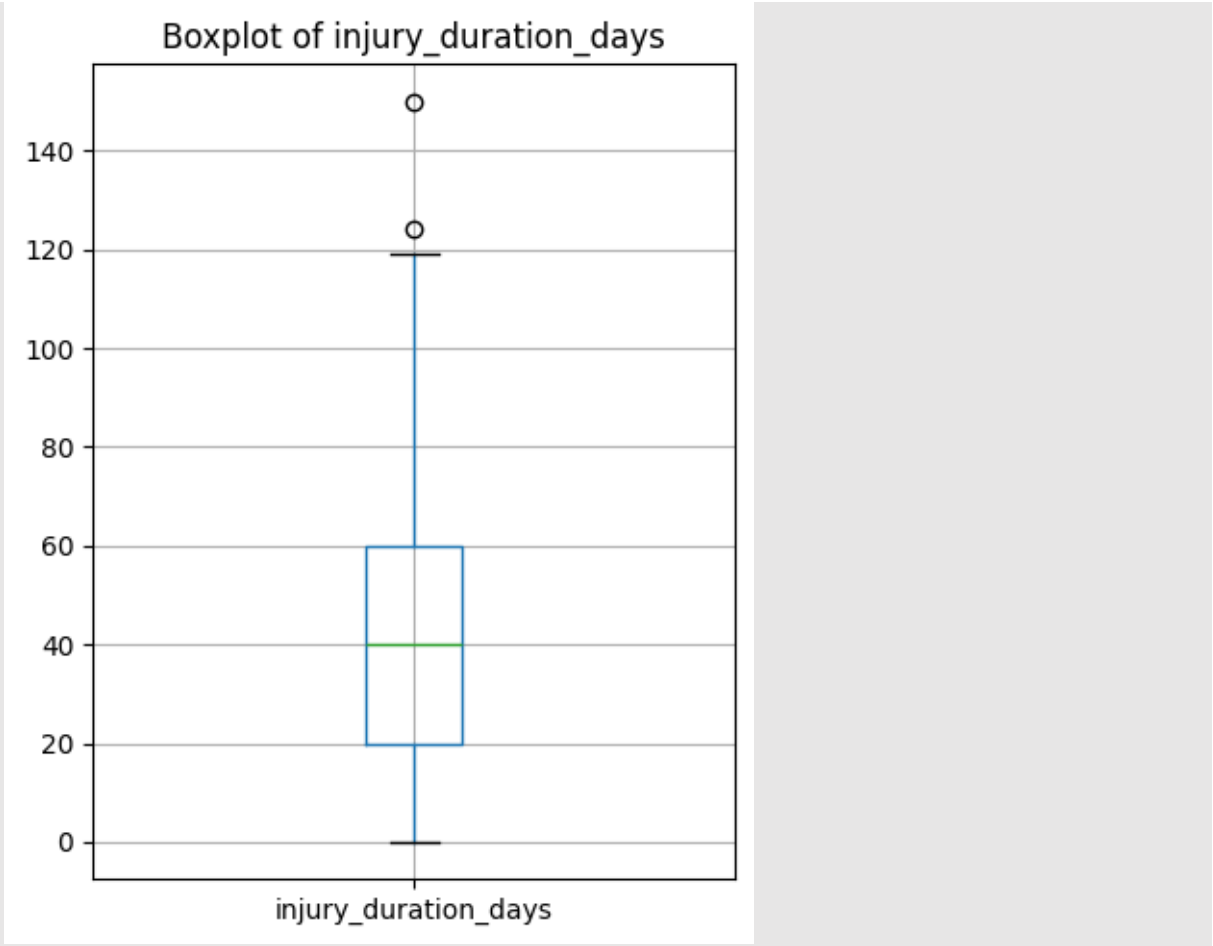


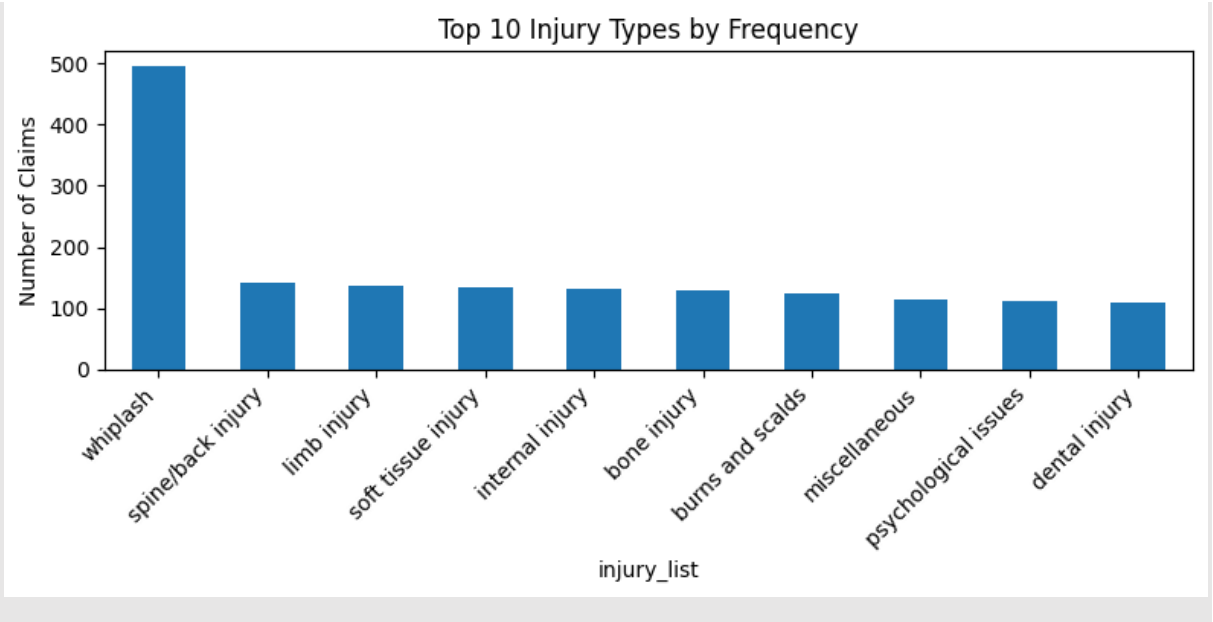
[TaskA] Summary for injury_duration_days:

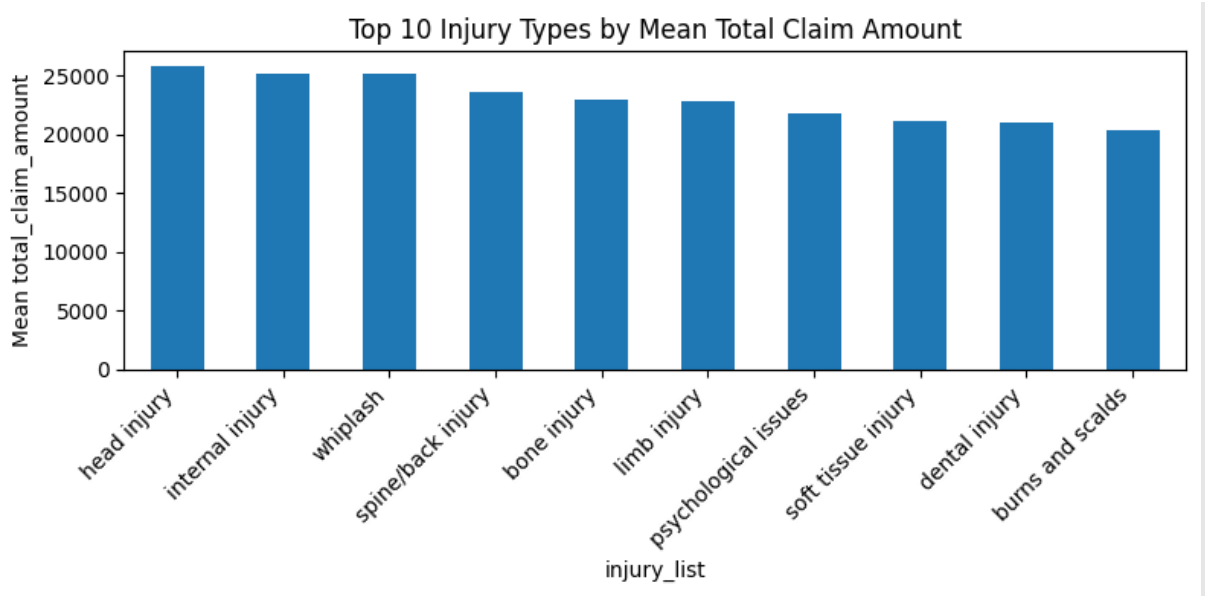
count 1000.000000

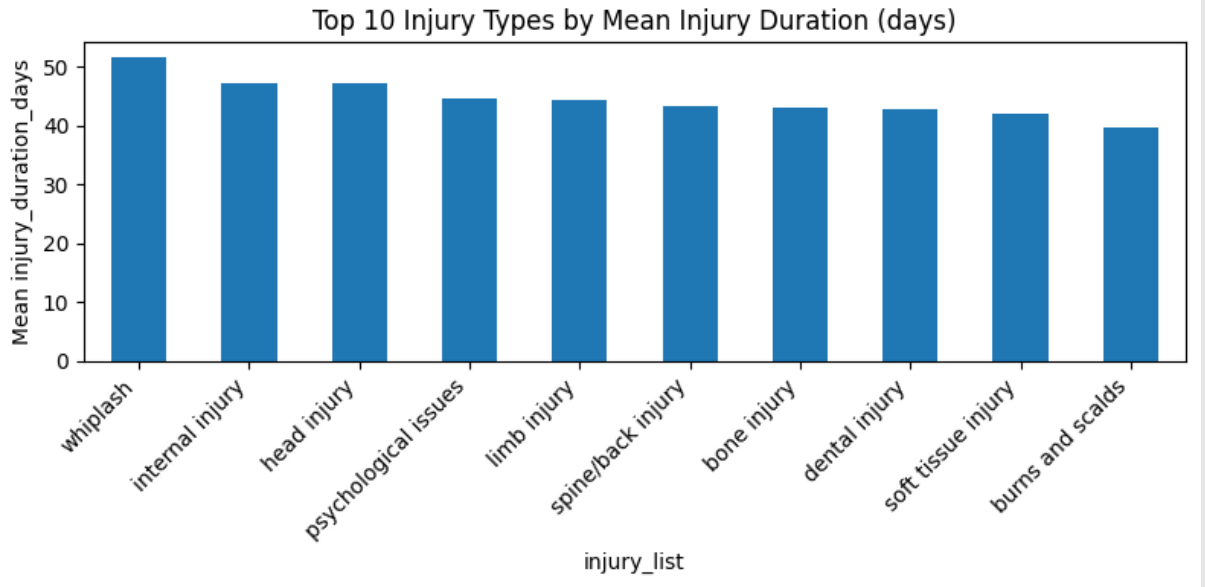
```
mean       40.673000
std        26.987597
min         0.000000
25%        19.750000
50%        40.000000
75%        60.000000
max       150.000000
Name: injury_duration_days, dtype: float64
```

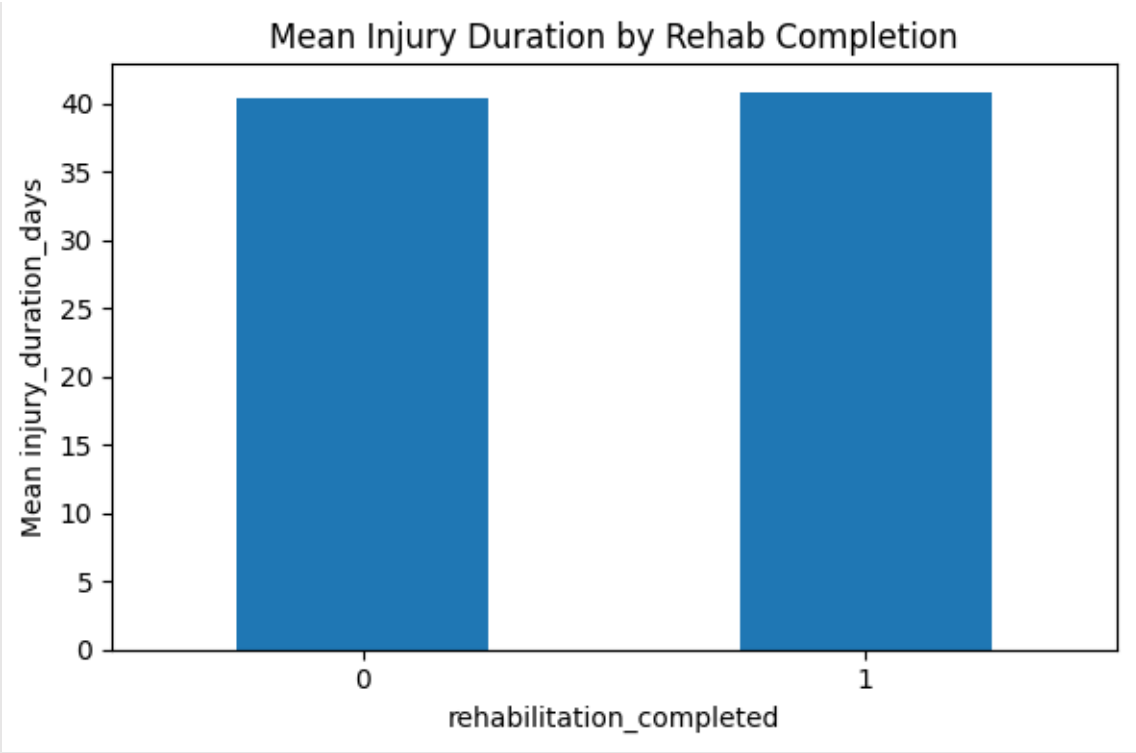


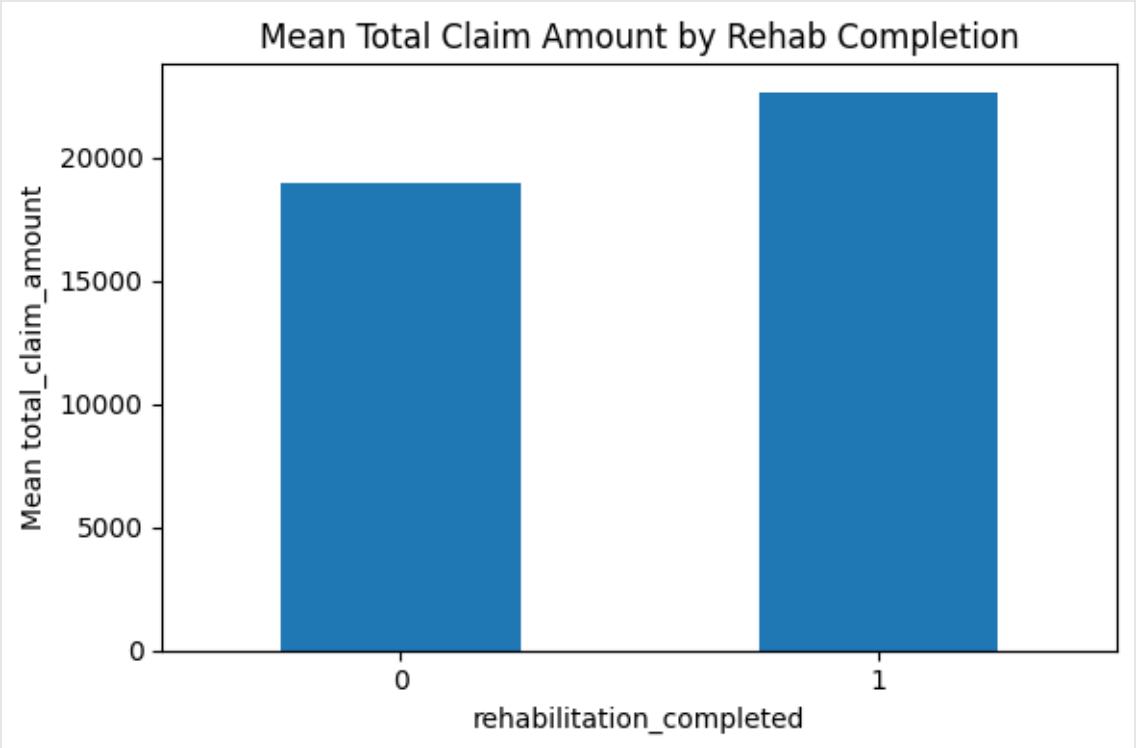

Boxplot of injury_duration_days
140
120
100
80
60
40
20
0
injury_duration_days

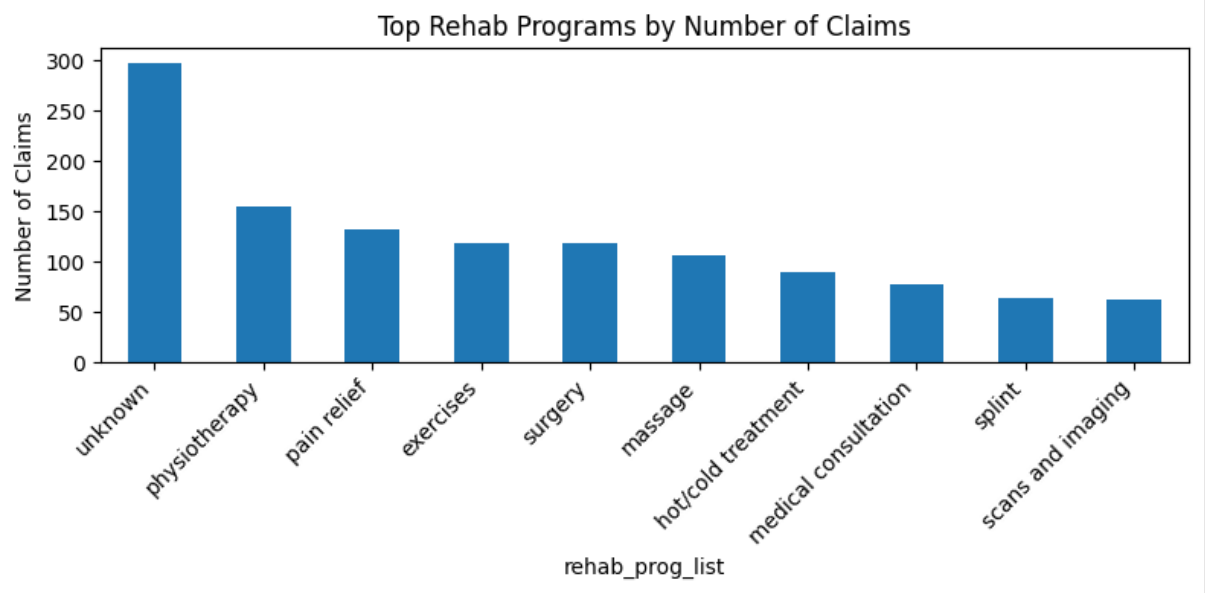

Top 10 Injury Types by Frequency
Number of Claims
500
400
300
200
100
0
whiplash
spine/back injury
limb injury
soft tissue injury
internal injury
bone injury
burns and scalds
miscellaneous
psychological issues
dental injury
injury_list


Top 10 Injury Types by Mean Total Claim Amount
Mean total_claim_amount
25000
20000
15000
10000
5000
0
head injury
internal injury
whiplash
spine/back injury
bone injury
limb injury
psychological issues
soft tissue injury
dental injury
burns and scalds
injury_list


Top 10 Injury Types by Mean Injury Duration (days)
Mean injury_duration_days
50
40
30
20
10
0
whiplash
internal injury
head injury
psychological issues
limb injury
spine/back injury
bone injury
dental injury
soft tissue injury
burns and scalds
injury_list


Mean Injury Duration by Rehab Completion
Mean injury_duration_days
0
5
10
15
20
25
30
35
40
0
1
rehabilitation_completed


Mean Total Claim Amount by Rehab Completion
Mean total_claim_amount
0
5000
10000
15000
20000
0
1
rehabilitation_completed



[TaskA] [TaskB] Rehabilitation Program Analysis completed.

[TaskA] [TaskB] Applying AIS / ISS severity scoring to all claims...

[TaskA] [TaskB] AIS / ISS scoring completed.

[TaskB] ISS-like score summary:

count 1000.00000

mean 4.83100

std 4.20263

min 1.00000

25% 1.00000

50% 4.00000

75% 8.00000

max 22.00000

Name: iss_like, dtype: float64

[TaskB] Severity index summary:

count 1000.00000

mean 12.18028

std 5.45573

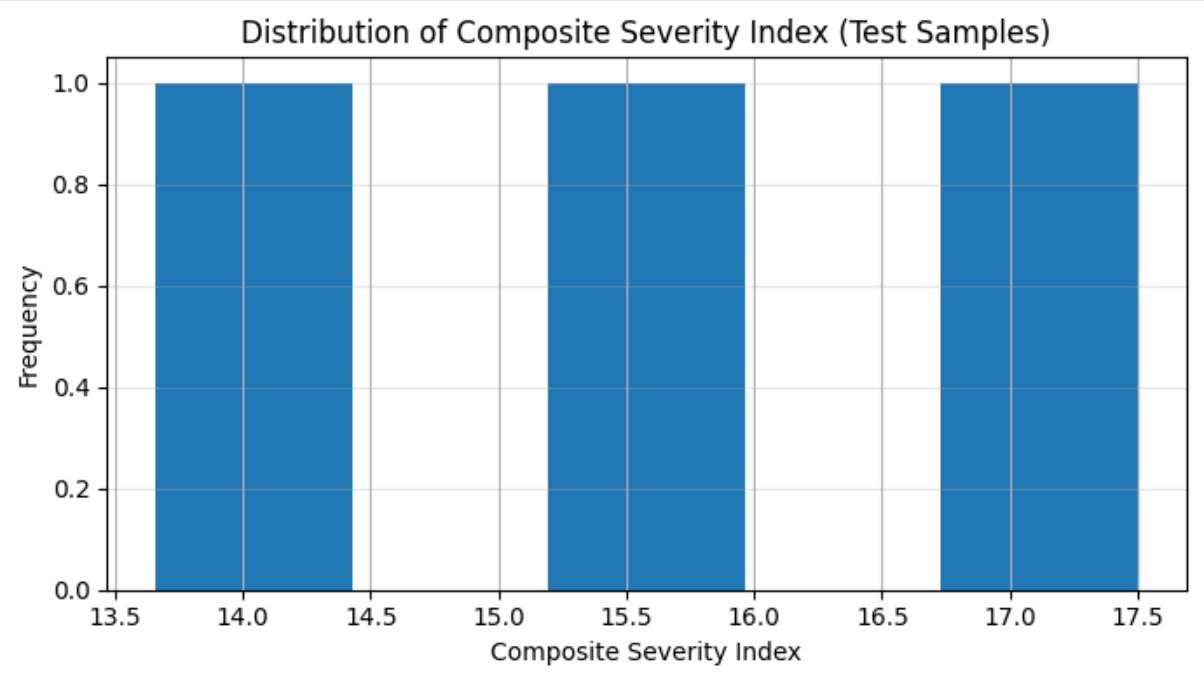
```
min         2.21000
25%         8.26250
50%        11.14500
75%        15.19250
max        35.49000
Name: severity_index, dtype: float64

[TaskB] ISS band counts:
iss_band
minor       763
moderate    218
serious      19
Name: count, dtype: int64
```

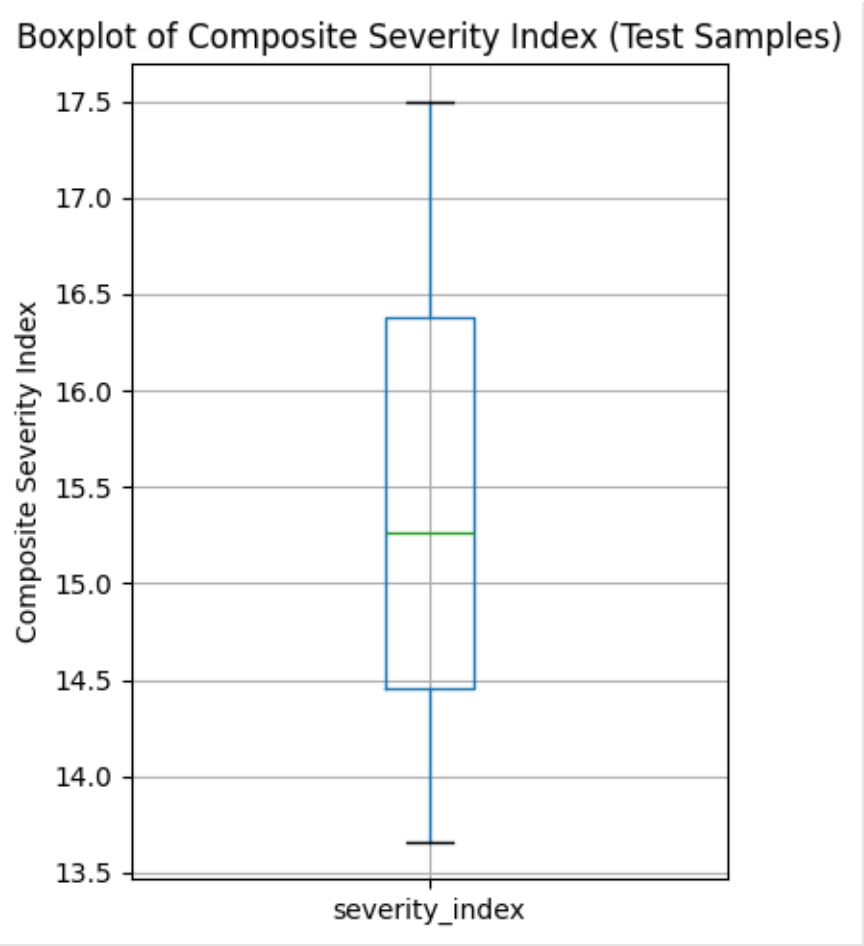
Out:

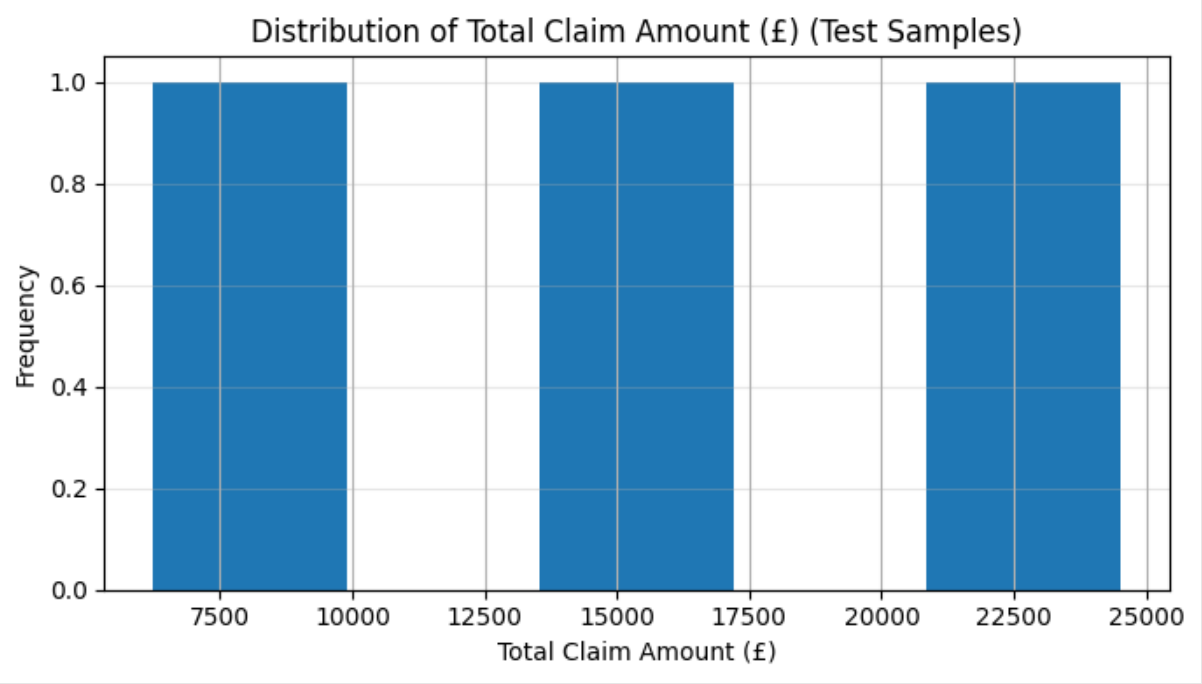
```
[TaskB] Data-driven severity index summary:
count    1000.000000
mean       33.750317
std        14.810488
min         0.000000
25%        23.404830
50%        33.144948
75%        43.045902
max       100.000000
Name: severity_index_data_driven, dtype: float64
```

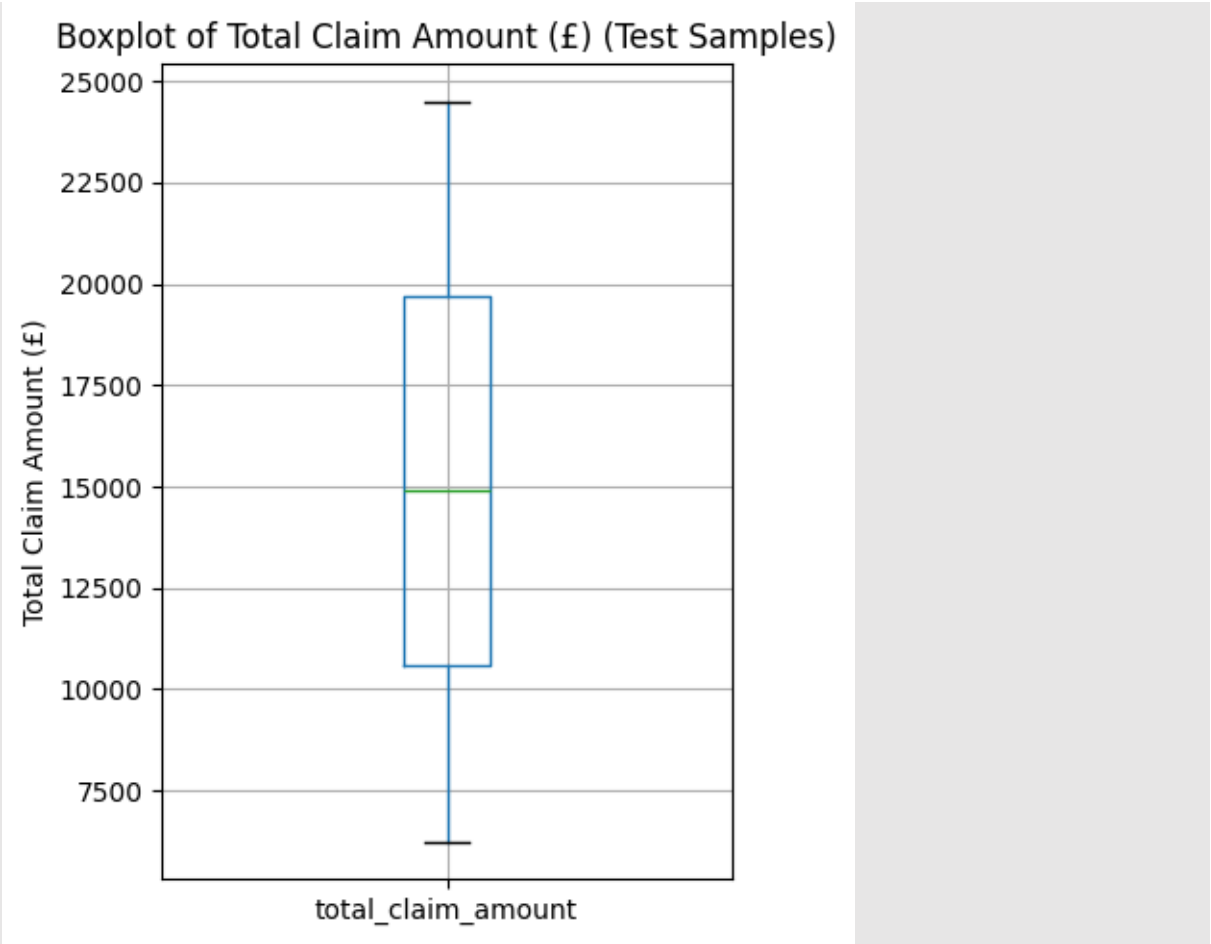
```
CORRELATION MATRIX:
```

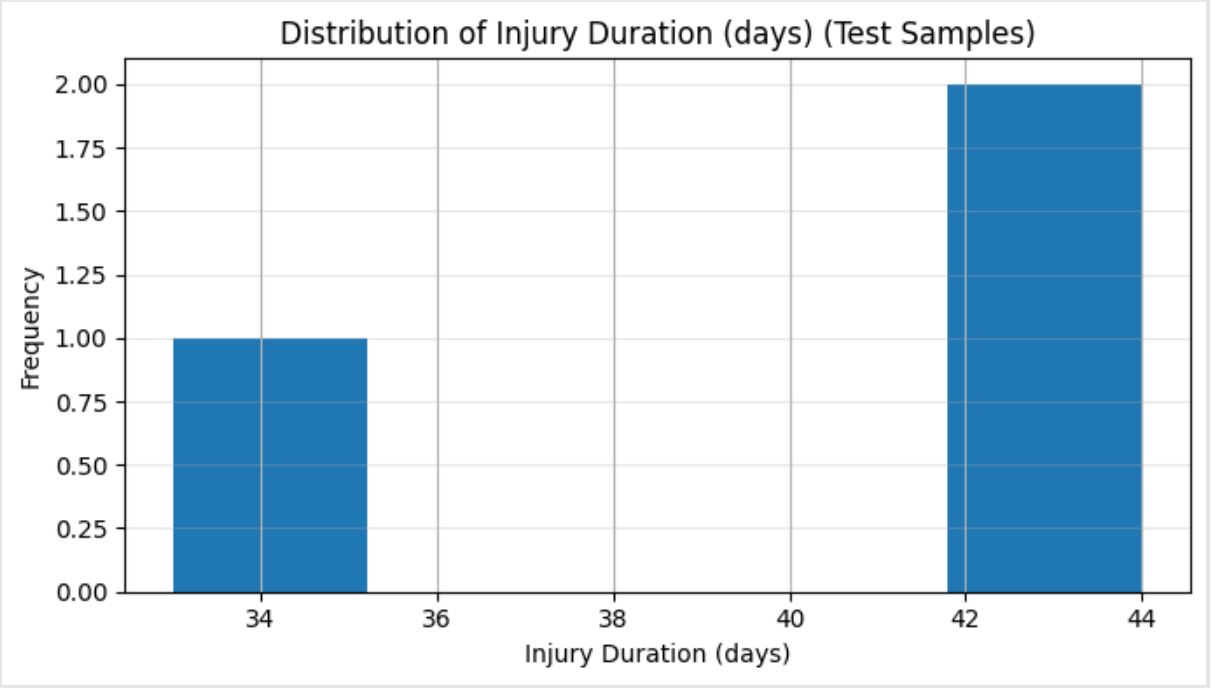
	severity_index	total_claim_amount	injury_duration_days
severity_index	1.000000	0.603848	-0.075502
total_claim_amount	0.603848	1.000000	0.749232
injury_duration_days	-0.075502	0.749232	1.000000

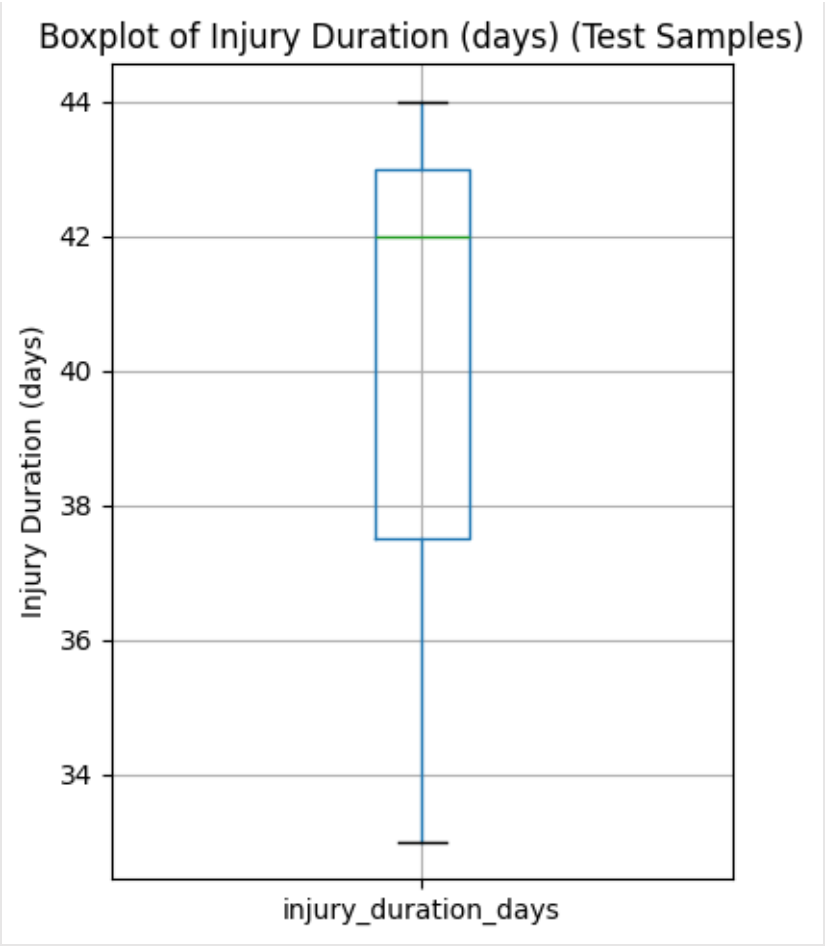



Boxplot of Composite Severity Index (Test Samples)
Composite Severity Index
17.5
17.0
16.5
16.0
15.5
15.0
14.5
14.0
13.5
severity_index


Distribution of Total Claim Amount (£) (Test Samples)
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
7500
10000
12500
15000
17500
20000
22500
25000
Total Claim Amount (£)


Boxplot of Total Claim Amount (£) (Test Samples)
Total Claim Amount (£)
25000
22500
20000
17500
15000
12500
10000
7500
total_claim_amount


Distribution of Injury Duration (days) (Test Samples)
Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
34
36
38
40
42
44
Injury Duration (days)


Boxplot of Injury Duration (days) (Test Samples)
Injury Duration (days)
44
42
40
38
36
34
injury_duration_days

REFERENCES

Aas, Kjersti, Arthur Charpentier, Fei Huang, and Ronald Richman (2024) 'Insurance analytics: prediction, explainability, and fairness', *Annals of Actuarial Science*.

Adams, Mike, Vineet Upreti and Jing Chen (2019) 'Product-market strategy and underwriting performance in UK general insurance', *The European Journal of Finance*, 25(7–8), pp. 735–758.

Afkanpour, Marziyeh, Elham Hosseinzadeh and Hamed Tabesh (2024) 'Identify the most appropriate imputation method for handling missing data: a systematic review', *BMC Medical Research Methodology*.

Association of British Insurers (ABI) (2025) *Motor claims hit record £11.7 billion in 2024*. London: ABI.

Davies, John (2017) '"It's a really grey area": An exploratory case study into the impact of the Jackson Reforms on organised insurance fraud', *International Journal of Law, Crime and Justice*, 49, pp. 20–32.

Ede, Osita, Chisom O. Uzuegbunam, Oke R. Obadaseraye, Kenechi A. Madu, Cajetan U. Nwadinigwe, Chijioke C. Agu, Udo E. Anyaehie and Emmanuel C. Iyidobi (2023) 'Is the New Injury Severity Score (NISS) a better outcome predictor than the Injury Severity Score (ISS)?', *JRSM Open*. Available at: https://journals.sagepub.com/doi/10.1177/22104917231171934 (Accessed: 25 November 2025).

Eidenbenz, David, Tobias Gauss, Tobias Zingg, Vincent Darioli, Cécile Vallot, Pierre-Nicolas Carron, Pierre Bouzat and François-Xavier Ageron (2025) 'Identification of major trauma using the simplified abbreviated injury scale', *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*. Available at: https://sjtrem.biomedcentral.com/articles/10.1186/s13049-025-01320-7 (Accessed: 25 November 2025).

EIOPA (2021) *Artificial Intelligence governance principles: Towards ethical and trustworthy artificial intelligence in the European insurance sector*. Luxembourg: European Insurance and Occupational Pensions Authority. Available at: https://www.eiopa.europa.eu/system/files/2021-06/eiopa-ai-governance-principles-june-2021.pdf (Accessed: 25 November 2025).

EIOPA (2025) *Opinion on Artificial Intelligence governance and risk management*. Luxembourg: European Insurance and Occupational Pensions Authority. Available at: https://www.eiopa.europa.eu/document/download/88342342-a17f-4f88-842f-bf62c93012d6_en?filename=Opinion+on+Artificial+Intelligence+governance+and+risk+management.pdf (Accessed: 25 November 2025).

FCA (2022) 'An agent-based model of motor insurance customer behaviour in the UK with word of mouth', *Journal of Artificial Societies and Social Simulation*, 25(2), article 2.

Financial Conduct Authority (FCA) (2025) *Motor Insurance Claims Analysis: Multi-firm review*. London: FCA. Available at: https://www.fca.org.uk/publication/multi-firm-

reviews/motor-insurance-claims-analysis-multi-firm-review-2025.pdf (Accessed: 25 November 2025).

Guillen, Montserrat, Jens Perch Nielsen and Ana M. Pérez-Marín (2021) 'Near-miss telematics in motor insurance', *Journal of Risk and Insurance*, 88(3), pp. 569–589.

He, Y., Gao, J., Liu, Y. and Qian, J. (2024) 'Global trends and hotspots related to whiplash injury: a bibliometric analysis', [Peer-reviewed journal article]. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11398816/ (Accessed: 25 November 2025).

Hernan, Miguel A. and James M. Robins (2020) *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC.

Holvoeta, Freek, Katrien Antonioa and Roel Henckaerts (2023) 'Neural networks for insurance pricing with frequency and severity data', *arXiv* (preprint). Available at: https://arxiv.org/pdf/2310.12671 (Accessed: 25 November 2025).

Information Commissioner's Office (ICO) (2020) *Explaining decisions made with AI*. Wilmslow: ICO and The Alan Turing Institute.

Jäger, Sebastian, Arndt Allhorn and Felix Bießmann (2021) 'A benchmark for data imputation methods', *Frontiers in Big Data*, 4, 693674. Available at: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.693674/full (Accessed: 25 November 2025).

Jaiswal, Rachana, Shashank Gupta and Aviral Kumar Tiwari (2024) 'Big data and machine learning-based decision support for insurance claim forecasting', *Technological Forecasting and Social Change*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0040162524006279 (Accessed: 25 November 2025).

Kester, Johannes (2022) 'Insuring future automobility: a qualitative discussion of British and Dutch car insurers' responses to connected and automated vehicles', *Research in Transportation Business and Management*, 45(Part C), article 100903.

Lewis, Richard (2019) 'When people matter: finding humanity in tort law', *Journal of Personal Injury Law*, 1, pp. 10–32.

Little, Roderick and Donald Rubin (2019) *Statistical Analysis with Missing Data*. 3rd edn. Hoboken, NJ: Wiley.

McGlade, J. (2018) 'No basis for reforms? An analysis of the data available and proposed changes to the right to whiplash damages and increase the small claims track limit', *Journal of Personal Injury Law*, 3, pp. 63–82.

Official Injury Claim (OIC) (2024) *Guide to Making a Personal Injury Claim* (Version 3.6, August 2024). Available at: https://www.officialinjuryclaim.org.uk/media/2pnf5hcm/guide-to-making-a-personal-injury-claim-v36.pdf (Accessed: 25 November 2025).

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*. 2nd edn. Cambridge: Cambridge University Press.

Pham, Tra My, Nikolaos Pandis and Ian R White (2024) ‘Missing data: Issues, concepts, methods’, [Peer-reviewed journal article]. Available at: https://www.sciencedirect.com/science/article/pii/S1073874624000082 (Accessed: 25 November 2025).

Rudin, Cynthia, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova and Chudi Zhong (2021) ‘Interpretable machine learning: Fundamental principles and 10 grand challenges’, *arXiv* (preprint). Available at: https://arxiv.org/abs/2103.11251 (Accessed: 25 November 2025).

Särkilahti, Niklas, Saara Leino, Jani Takatalo, Eliisa Löyttyniemi and Olli Tenovuo (2024) ‘The symptom profile of people with whiplash-associated disorder: a systematic review’, *Musculoskeletal Science and Practice*. Available at: https://www.sciencedirect.com/science/article/pii/S1360859224002973 (Accessed: 25 November 2025).

Swaby, Gerald and Paul Richards (2024) ‘Ouch! The practicalities of whiplash claims’, *Journal of Personal Injury Law*, 55(1), pp. 1–33.

Tukey, J.W. (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

UK Government (2025) *The Whiplash Injury (Amendment) Regulations 2025* (UK Statutory Instrument). Available at: https://www.legislation.gov.uk/ (Accessed: 25 November 2025).

Van Ditshuizen, Jan C., Menco J. S. Niemeyer, Esther M. M. Van Lieshout, Dennis Den Hartog, Jan-Jaap Visser, Karlijn J. P. van Wessem and Michiel H. J. Verhofstad (2025) ‘Coding traumatic brain injury with the abbreviated injury scale and Injury Severity Score’, *European Radiology*.