

# Strategic Insights for Insurance 4 You

## **1. Background and Context**

Insurance-4-You (I4U) operates in a highly competitive market where accurate risk assessment and targeted product offerings are critical for profitability. Currently, the organization is reviewing its actuarial models for health insurance billing and preparing to launch new marketing campaigns for its auto and travel divisions. The core business problem is to transition from assumption-based decision-making to a data-driven strategy.

I was tasked with analysing four distinct datasets—client health records, travel mode preferences, historical safety data, and regional police reports. The objective was to identify statistical patterns that could optimize billing accuracy, refine marketing targeting for families, and improve regional risk zoning. This report details the data wrangling, statistical validation, and strategic insights derived from these analyses to support I4U's objectives.

## **2. Analytical Approach**

My analysis was conducted entirely within RStudio, utilising the tidyverse suite for its robust capabilities in data manipulation (dplyr) and visualisation (ggplot2), alongside specific libraries such as skimr and moments for statistical testing.

### Data Preparation and Cleaning

The foundation of the analysis was ensuring data integrity across all four sources.

- **Health Data:** Upon importing insurance.csv, I inspected the structure using str() and summary(). I calculated descriptive statistics (mean, median, IQR) to understand the central tendency of Body Mass Index (BMI).
- **Seatbelt Data:** I identified 209 missing values exclusively in the seatbelt column of seatbelt.csv. To prevent data loss, I imputed these values with 0 and rounded all numerical columns to two decimal places for consistency.

```
# Determine the sum of missing values
sum(is.na(seatbelt))
sum(is.na(seatbelt$seatbelt))

# Replace NA with 0.
seatbelt[is.na(seatbelt)] = 0
head(seatbelt)
sum(is.na(seatbelt$seatbelt))
```

- **Police Data:** To eliminate bias and reduce dimensionality, I removed 12 administrative columns (e.g., MDC, lat, long) from police.csv, reducing the dataset from ~51,000 to ~19,000 relevant observations. I also employed string manipulation (str\_to\_title) to standardize inconsistent text formatting.

```
# Drop unnecessary columns.
police_df <- select(police_new, -c('X', 'idNum', 'date', 'MDC', 'preRace',
                                   'race', 'lat', 'long', 'policePrecinct',
                                   'citationIssued', 'personSearch',
                                   'vehicleSearch'))

colnames(police_df)
dim(police_df)

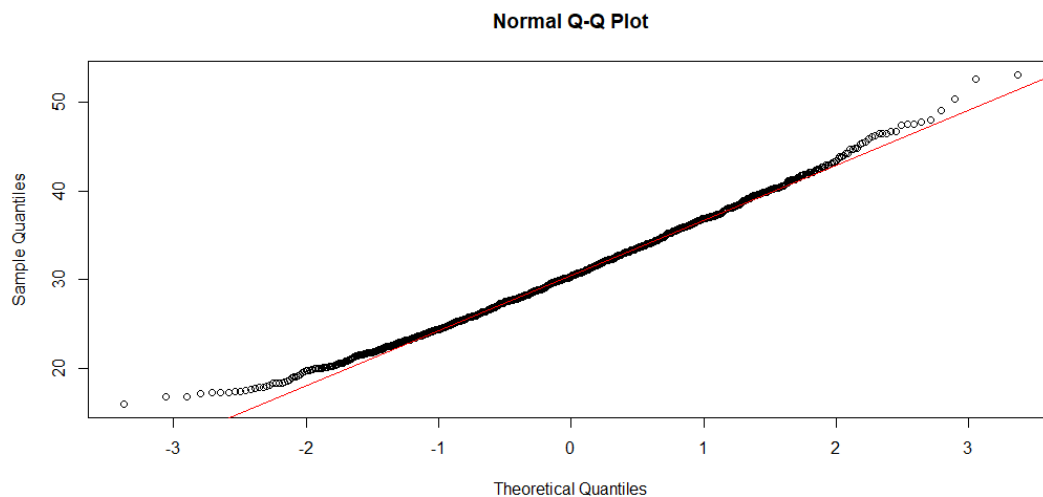
# Rename column names with first letter to uppercase.
names(police_df) <- str_to_title(names(police_df))
```

- **Travel Data:** I renamed cryptic headers (e.g., vcost to vehicle\_cost) to ensure readability for stakeholders and used mutate() to create new variables, such as calculating the total cost of travel by summing vehicle and general costs.

## Statistical Methodology

I applied rigorous statistical testing to validate assumptions before running correlations.

- **Normality Testing:** For the BMI analysis, I moved beyond simple averages. I generated Q-Q plots and performed the Shapiro-Wilk test, which returned a p-value < 0.05. This indicated that while BMI is visually symmetric, it is not perfectly normal.



```
shapiro-wilk normality test
data: (health$bmi)
W = 0.99389, p-value = 2.605e-05
```

- Correlation Testing: Crucially, I found that the age variable was not normally distributed. Consequently, I selected Spearman's rank correlation (rather than Pearson's) to accurately test the relationship between age, BMI, and insurance charges.
- Aggregations: For the Travel and Police datasets, I used `group_by()` and `summarise()` functions to create subsets. This allowed me to isolate specific demographics, such as families with more than two dependents, to analyse their specific behaviours distinct from the general population.

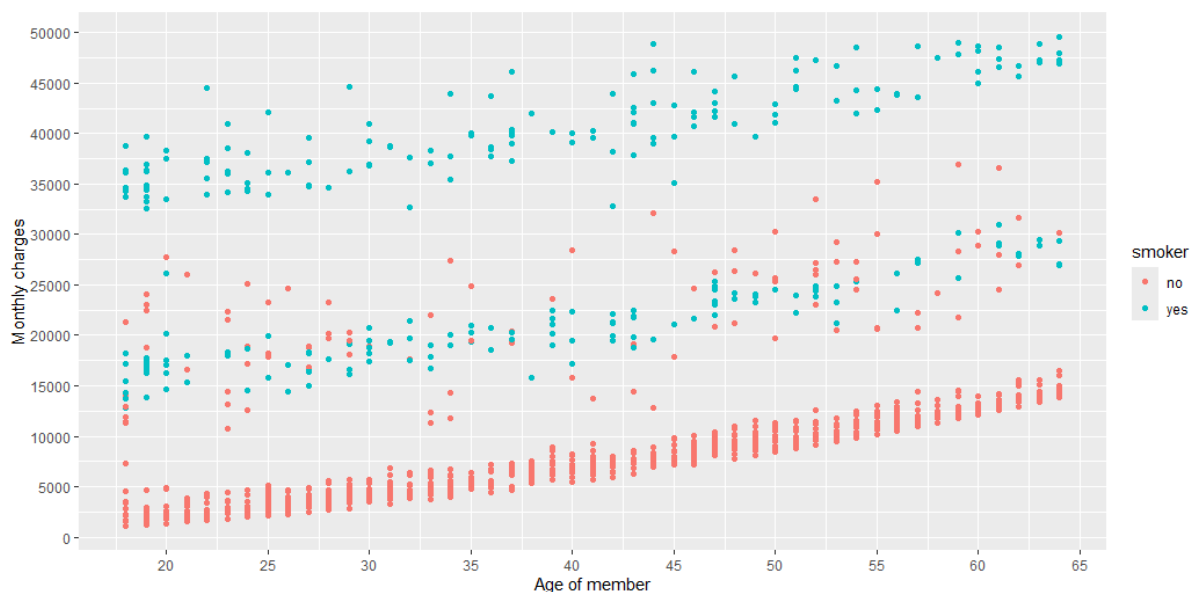
### 3. Visualisation and Insights

The visual analysis provided critical context to the raw numbers, revealing non-linear patterns that simple statistics missed.

#### Health and Billing Insights

I generated scatterplots to visualize the relationship between BMI and Charges.

- The "Smoker Gap": The visualization revealed a distinct "forked" distribution. While the statistical correlation between BMI and charges was weak ( $r \approx 0.02$ ), the plot showed that high costs are driven by the *interaction* of BMI and smoking. Non-smokers remained in a low-cost band regardless of their weight, whereas smokers exhibited exponential cost increases as their BMI rose.



## Travel Behaviour Insights

By subsetting the travel mode data for families, I visualized travel preferences using frequency counts.

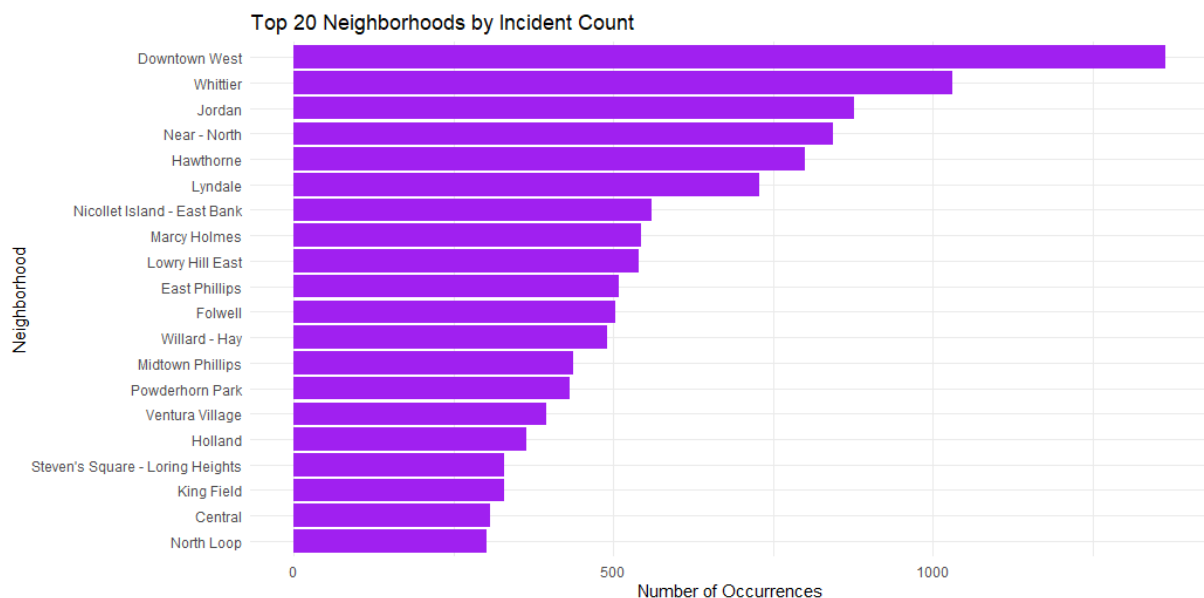
- **Mode Hierarchy:** The data revealed a clear preference hierarchy for families: Car (37), Train (28), Air (24), and Bus (7).
- **Cost Sensitivity:** Further analysis explained this behaviour. I compared the mean general costs across modes, revealing that Cars (\$98.63) were significantly cheaper than Buses (\$129.46) or Trains (\$142.81). This insight confirms that I4U's family clients are highly price-sensitive and prioritize private transport.

```
> arrange(mean_costs, desc(mean_GC))
  vehicle mean_VC mean_GC
1   train 50.69792 142.8125
2    bus 32.38542 129.4583
3    car 11.98958  98.6250
```

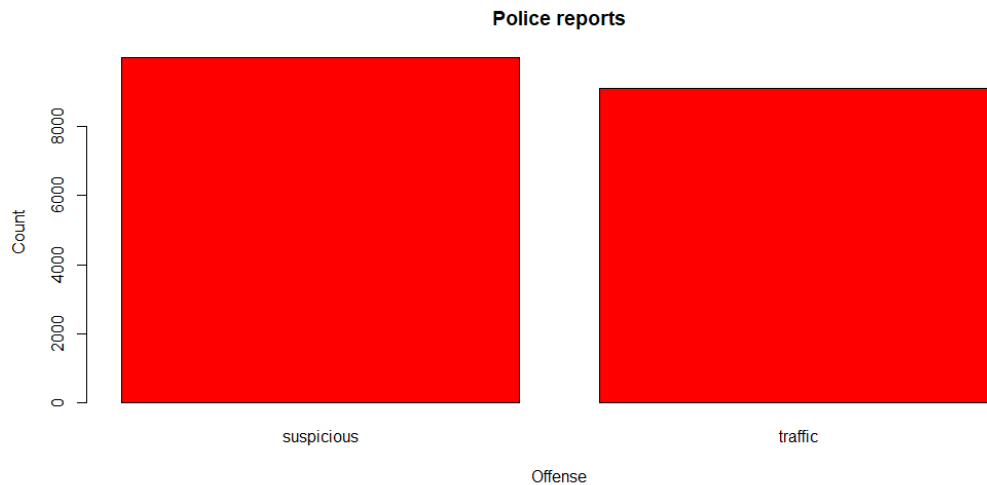
## Regional Risk Profiling

Bar charts generated from the police.csv data highlighted specific high-activity zones.

- **Hotspots:** Neighbourhoods like "Downtown West" and "Whittier" appeared as significant outliers for police stops.

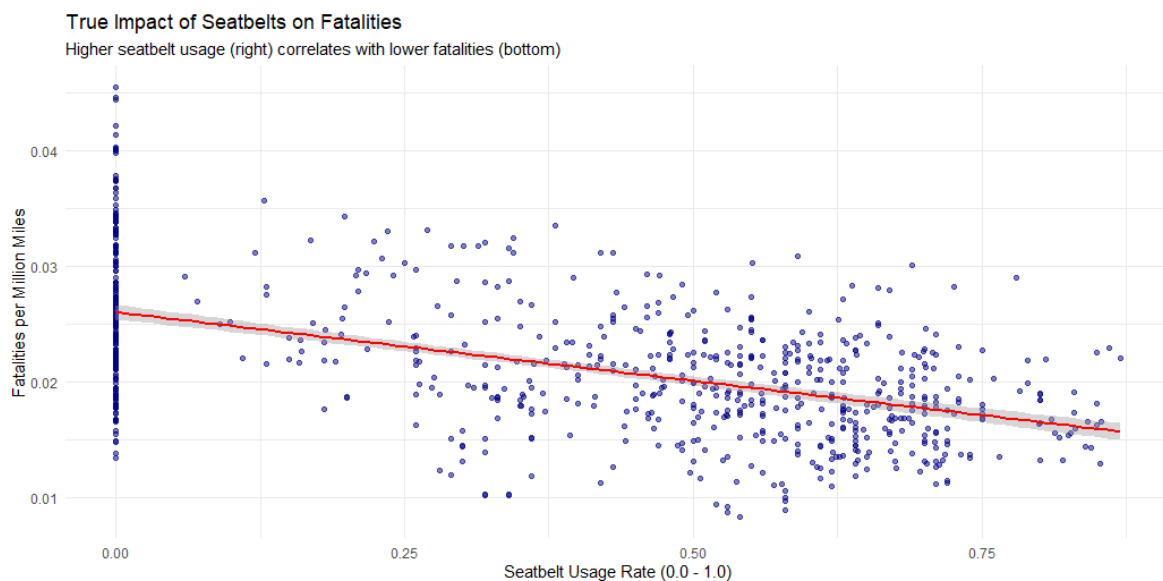


- **Incident Types:** By visualizing the Problem column, I identified that "Suspicious" stops (9,984) slightly outnumbered "Traffic" stops (9,100). This suggests that risk in these areas is often behavioural rather than purely vehicular, which impacts how we should assess comprehensive coverage in those zones.



## Safety Trends

Boxplots of the seatbelt dataset revealed outliers in mileage data, likely representing large states like California or Texas. Despite these outliers, the aggregate analysis (using `apply()`) showed a clear historical trend where increased seatbelt usage correlated with lower fatality rates, validating the safety features I4U promotes.



## 4. Patterns and Predictions

Based on the synthesized data, I have identified three key patterns that define the strategic direction for I4U.

### Prediction 1: The "Interaction Effect" in Health Pricing

The data contradicts the hypothesis that BMI alone is a strong predictor of cost. The pattern clearly shows that Smoking Status is the primary multiplier of risk.

- Prediction: If I4U adopts a billing model weighted heavily on BMI without accounting for smoking status, we predict the company will lose profitability. You risk overcharging obese non-smokers (who are actually lower cost/risk), causing them to leave for competitors. Simultaneously, you risk undercharging thin smokers (who are high cost/risk).
- Strategy: Implement a tiered-risk model where high premiums are triggered only by the combination of Smoking + High BMI.

### Prediction 2: Family Mobility is Driven by Value

The travel data shows a robust pattern: families systematically avoid the most expensive options (Train/Bus) in favour of Cars.

- Prediction: Marketing campaigns that focus on "luxury" or "convenience" will fail with this demographic.
- Strategy: The new Auto Insurance campaign must emphasize "Cost-Efficiency" and "Family Safety." The data proves families are already choosing cars to save money; our product should reinforce that value proposition.

### Prediction 3: Geographic Risk Concentration

Risk is not evenly distributed across the city; it is concentrated in specific nodes.

- Prediction: A flat-rate premium for home or auto insurance across the city will expose I4U to higher claims ratios in neighbourhoods like Downtown West.
- Strategy: I4U should utilize the "Suspicious" vs. "Traffic" ratios calculated in the police analysis to refine zoning. Neighbourhoods with high "Suspicious" counts warrant higher comprehensive deductibles due to increased risk of theft or vandalism, whereas "Traffic" heavy zones warrant higher collision premiums.