* **986 rows and 13 columns**: a moderate‐sized dataset ideal for our tree‐based and linear models.
* Non‐null counts show every record is present—no rows were dropped or truncated.
* Observation:
* 1. Binary flags flagged as outliers (e.g. Any Transplants, Any Chronic Diseases, Known Allergies, History Of Cancer In Family) because their IQR is zero (all values 0 or 1), every ‘1’ is marked as an outlier. We will exclude binary columns from further outlier handling and focus on continuous variables.
* 2. Weight: 16 points (1.62%) lie beyond [37 kg, 117 kg].
* 3. BMI: 22 points (2.23%) outside [12.34, 41.81].
* 4. NumberOfMajorSurgeries: 16 points (1.62%) at the extreme value 3.
* 5. PremiumPrice: 6 points (0.61%) beyond [₹10,500, ₹38,500].
* Insights:
* 1. BMI Distribution: About 2.2% of records beyond [12.3, 41.8] in BMI appear as extreme underweight/obese. Removing them yields a more symmetric, bell-shaped distribution, which can help regression models that assume normality.
* 2. PremiumPrice Skew: Very high premiums (≥₹38,500) are rare (0.6% of data). Clipping these reduces the heavy right tail, lowering variance and making error metrics (e.g. RMSE) less sensitive to a few extreme cases.
* 3. Weight Tail Effects: The top ~1.6% of weights above 117 kg distort the histogram’s tail. Excluding them produces a cleaner, more centralized weight distribution—beneficial for models that implicitly assume homoscedasticity.
* 4. The rare “3 major surgeries” cases (~1.6% of data) vanish if we blindly drop IQR outliers. These high-risk individuals are actually important for underwriting.
* By visualizing before/after, we can decide to cap or transform these continuous outliers rather than drop them entirely, preserving most data while reducing undue influence from extremes.
* Insights:
* 1. The age distribution is fairly uniform between 18–66, with slight peaks around the early 30s and late 40s.
* 2. Roughly 18 % of individuals report a chronic disease and 82 % do not. This imbalance (1 in 5 have chronic conditions) suggests the model will need to handle a minority‐class flag; it’s an important risk signal but won’t dominate the dataset.
* 3. Only about 5–6 % of applicants have had a transplant. This very rare event is still clinically important but must be treated carefully (e.g. as a high-impact binary indicator) rather than dropped or overly down-weighted.
* 4. Premiums are right-skewed: the bulk lie between ₹15 000–₹30 000, with a long tail up to ₹40 000. A log transformation or robust model (e.g. tree-based) will help mitigate that skew and reduce the influence of a few very high premiums.
* Insights:
* 1. Premium by Chronic Disease:
* Policyholders with a chronic disease (-- right box) have a higher median premium (₹28 000) than those without (₹23 000).
* Their entire IQR is shifted up by about ₹4 000–₹5 000, and the top whisker extends to ₹40 000 vs. ₹39 000 for non-chronic.
* Chronic conditions are a clear cost driver—worth treating as a prime risk flag.
* 2. Premium by Diabetes:
* Those with diabetes pay a slightly higher median (₹25 000) than non-diabetics (₹23 000).
* The diabetic group also shows more high-end outliers (premiums near ₹40 000), suggesting greater variability in their risk profiles.
* Diabetes has a moderate but meaningful upward effect on premiums, and introduces extra tail risk.
* 2. Premium vs. Age:
* There’s a strong positive trend: younger adults (18–30) cluster around ₹15 000–₹25 000, while premiums steadily rise into the ₹25 000–₹35 000 range for ages 40–60.
* A handful of seniors (60+) hit the top end (₹35 000–₹40 000).
* Age is our single strongest continuous predictor—older age almost always means higher cost.
* 3. Premium vs. BMI:
* Here we see a weak, noisy upward tilt: very low BMI (< 18) and very high BMI (> 40) sometimes co-occur with higher premiums, but the cloud is broad.
* Most premiums sit between ₹20 000–₹30 000 across BMI 20–35.
* BMI by itself isn’t as powerful as Age or chronic-disease flags, but extreme BMI values do tend to be pricier.
* Insights:
* 1. Orange points (those with chronic diseases) are mostly found in the upper‐right of the PremiumPrice plots—confirming that chronic‐disease holders tend to pay higher premiums for a given age or BMI.
* Insights:
* 1. Age (r = 0.70): By far the single strongest linear relationship. Older applicants consistently pay more, so Age should be front-and-center in any model.
* 2. AnyTransplants (r = 0.29): Transplant recipients incur substantially higher premiums. Even though it’s a rare flag, its presence signals major risk.
* 3. NumberOfMajorSurgeries (r = 0.26): Each additional surgery lifts premiums—surgeries serve as a proxy for underlying health complexity.
* 4. Diabetes (r = 0.08) and Family Cancer History (r = 0.08) show minimal direct correlation in isolation—these may interact with other features or be non-linear.
* 5. KnownAllergies and Height are essentially uncorrelated with price (|r|<0.05).
* 6. Age is your powerhouse predictor. Among health flags, transplants, surgery count, and chronic-disease status are the next most important
* Insights:
* 1. T-Tests on PremiumPrice:
* - Significant premium differences for all major health flags except KnownAllergies (p ≈ 0.71).
* - Diabetes, BP problems, transplants, chronic diseases, and family cancer history each show p < 0.05, confirming they raise mean premiums.
* - No effect from KnownAllergies, so that flag may add little predictive value in isolation.
* 2. Premiums vary meaningfully by the number of major surgeries (ANOVA F = 26.14, p ≈ 2.9 × 10⁻¹⁶), so surgery count should be treated as an ordinal risk factor rather than a simple binary flag.
* 3. Chi-Square Associations among Flags:
* - Diabetes is significantly associated with BP problems (p ≈ 0.00008), chronic diseases (p ≈ 0.0064), and allergies (p ≈ 0.0148).
* - KnownAllergies and HistoryOfCancerInFamily are also linked (p ≈ 0.00046).
* - Most other pairs (e.g., transplants vs. chronic diseases, BP vs. allergies) show no association, indicating they capture distinct risk dimensions.
* 4. There is no evidence of an association between having a chronic disease and a family history of cancer (χ² = 0.02, p ≈ 0.89), indicating these two risk flags capture distinct aspects of health risk and can be modeled independently
* 5. **\*\*Quantify Marginal Impact\*\*** in concrete terms—for example, the model tells us that, all else equal, each extra year of age adds about ₹323 to the premium.
* 6. An R² of 0.54, OLS **\*\*assesses overall fit\*\***, showing that these five predictors together explain roughly 54% of the variation in premiums. This combination of confounding control, precise effect sizes, and a global goodness-of-fit measure gives us a solid baseline before exploring non-linear models or interaction effects.
* **\*\*Random Forest\*\*** is the best overall choice here:
* \* It achieves the **\*\*highest test R² (0.8625)\*\***—so it explains the most variance on unseen data.
* \* It has the **\*\*lowest RMSE (₹2,421)\*\***, meaning its predictions are, on average, closest to the true premiums.
* \* Its cross-validated R² (0.7208) is very close to the Gradient Boost’s (0.7351), but the RF model generalizes better at test time.
* By contrast:
* \* **\*\*Decision Tree\*\*** is weakest (CV R² ≈ 0.50, test R² ≈ 0.79, RMSE ≈ ₹3,013).
* \* **\*\*Gradient Boosting\*\*** has a slightly higher CV R² (0.735) but under-performs RF on the test set (test R² = 0.8549, RMSE ≈ ₹2,487).
* So for a balance of stability (CV) and real-world accuracy (test), Random Forest is your top performer.

After Hyperparameter tuning

Here’s how your **\*\*tuned Decision Tree\*\*** stacks up against the **\*\*linear baseline\*\*** and the **\*\*untuned tree models\*\***:

| Model                              | RMSE (₹) | MAE (₹) | R²     | CV R² |

| ---------------------------------- | -------- | ------- | ------ | ----- |

| **\*\*Linear Regression (original y)\*\*** | 3,494.4  | 2,586.2 | 0.7136 | 0.617 |

| **\*\*Decision Tree (default)\*\***        | 3,013.4  | 1,333.3 | 0.7870 | 0.500 |

| **\*\*Decision Tree (tuned)\*\***          | 2,755.1  | 1,757.3 | 0.8220 | 0.668 |

| **\*\*Random Forest (untuned)\*\***        | 2,421.4  | 1,439.5 | 0.8625 | 0.721 |

| **\*\*Gradient Boosting (untuned)\*\***    | 2,487.1  | 1,675.1 | 0.8549 | 0.735 |

**### 🔍 Key Takeaways**

1. **\*\*Baseline Linear Regression\*\***

   \* RMSE ≈ ₹3.5 K, MAE ≈ ₹2.6 K, R² ≈ 0.71

   \* Solid interpretability but relatively high error.

2. **\*\*Default Decision Tree\*\***

   \* Improved over linear: RMSE down to \~₹3.0 K, R² up to 0.79, MAE \~₹1.3 K—but severely overfits (CV R² \~0.50 vs test R² \~0.79).

3. **\*\*Tuned Decision Tree\*\***

   \* RMSE drops further to \~₹2.75 K and test R² climbs to \~0.82.

   \* CV R² improves to \~0.67, meaning you’ve reduced overfitting.

   \* **\*\*Trade-off:\*\*** MAE rose to ₹1.76 K (versus default’s ₹1.33 K), suggesting the tree now makes fewer large errors (lower RMSE) but more consistent mid-range errors (higher MAE).

4. **\*\*Random Forest & Gradient Boosting\*\***

   \* Still outperform even the tuned tree:

     \* **\*\*RF:\*\*** RMSE \~₹2.42 K, R² \~0.86, CV R² \~0.72

     \* **\*\*GBM:\*\*** RMSE \~₹2.49 K, R² \~0.85, CV R² \~0.74

---