Home Assignment - Atidot Data Analyst

Leshem Ben Hur

This document outlines the comprehensive workflow, methodology and technical steps taken to complete each part.
All scripts and appendices are marked as footnotes. To run the full script please find attached Readme.md.

First, I reviewed the dataset to understand the meaning and context behind each column. For the columns I was not familiar with, I conducted additional research to clarify what they represent (for example: what “riders” are in insurance policy).

Data Understanding & Cleaning[1]

Using Claude Code, I generated a script to run sanity checks on the dataset.
My main approach was to profile the data to better understand the dataset and identify data quality issues.

Key evaluations performed:

1. Shape, dtypes and categorial distribution
2. Missing values (and possible justifications)
3. Out of range values (for date logic and numeric values)
4. Consistency checks (ex. if policy_end_date exists, churned must beTRUE)
5. Duplicates (Full row duplicates, policy_id, customer_id)
6. Numerical outliers

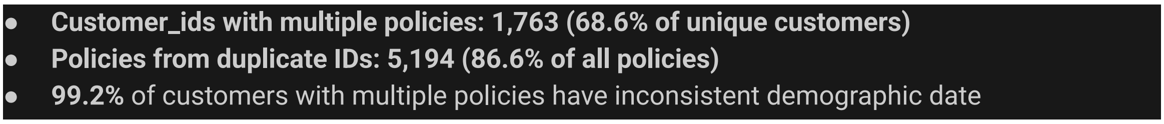
The full output and analysis results are in sanity_checks_report.

In practice, the treatment of missing or duplicated data depends on the specific business question and information needs. In this analysis, a general approach is applied to optimize the dataset.

The main data quality issues identified are:

1. **Inconsistent customer data** for the same customer ID.

The duplications are not a problem as one customer might have several insurances or might add policies over time.
The same customer_id appears in multiple records with different values for demographic attributes such as gender and country.



- **Customer_ids with multiple policies: 1,763 (68.6% of unique customers)**
- **Policies from duplicate IDs: 5,194 (86.6% of all policies)**
- **99.2%** of customers with multiple policies have inconsistent demographic date

These values are expected to remain consistent across all policies belonging to the same customer. Inconsistencies may indicate data integration errors or that some

[1] data_cleaning.py

records describe insured individuals rather than the policyholder.
Using these fields without validation could lead to biased churn analysis or misleading insights.

To enable reliable analysis, demographic attributes (gender and country) were standardized at the customer level by selecting the most frequent value observed for each customer_id (mode). Customers with unresolved conflicts were flagged (as “conflict”) to avoid bias in segmentation analyses.

That resulting in:

- Gender: 79.4% with a usable value (27.3% resolved via mode)
- Country: 69.2% with a usable value (24.3% resolved via mode)

These results suggest that demographic attributes are only partially reliable and should be used with caution when deriving customer level insights.

2. **Age**

The age column represents the customer’s age, but the reference point in time is unclear. It could reflect the age at the joining date or the snapshot date. It may not consistently present the relevant information.

This requires further evaluation, specifically to understand how the table is updated and how frequently it is refreshed.

3. **Income band missing values**

53 records (9.22%) are missing an income_band value. This could bias any income based analysis.

This data can significantly affect the analysis of customer characteristics and behavior. While other customer attributes could be used to estimate missing income bands, doing so may introduce unintended bias. Therefore, values were not inferred and were instead categorized as “Unknown.”

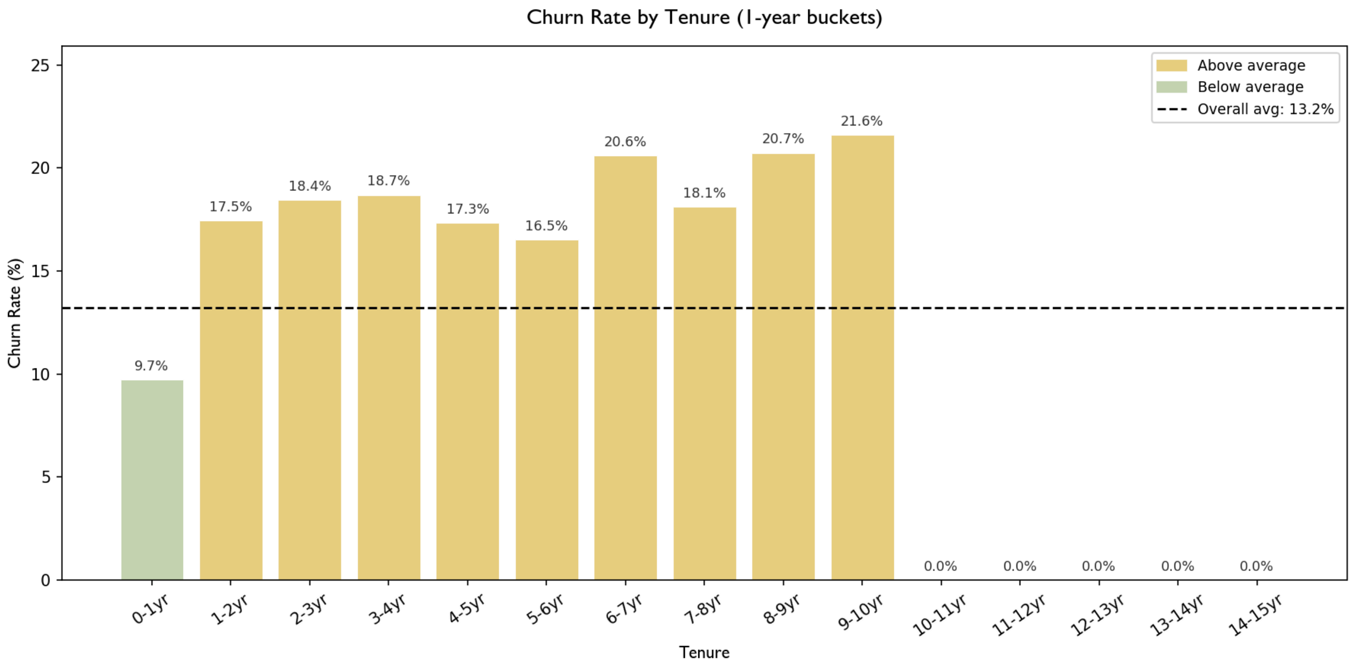
Exploratory Analysis

1. [2]Which policy or customer characteristics are most associated with churn?

- **Late payments** are the strongest signal, 4 late payments = 43% churn.
- **Customer service calls** spike at 3 - 21.2% churn. As the number of calls increases, the churn rate also rises. This may be due either to significant customer dissatisfaction or to calls related to the termination of the policy.

On the positive side, **annual payers are much stickier**. 7.7% vs 15.5% for monthly payers. Customers in annual payment plans tend to churn less.

[2] exploratory_analysis.py



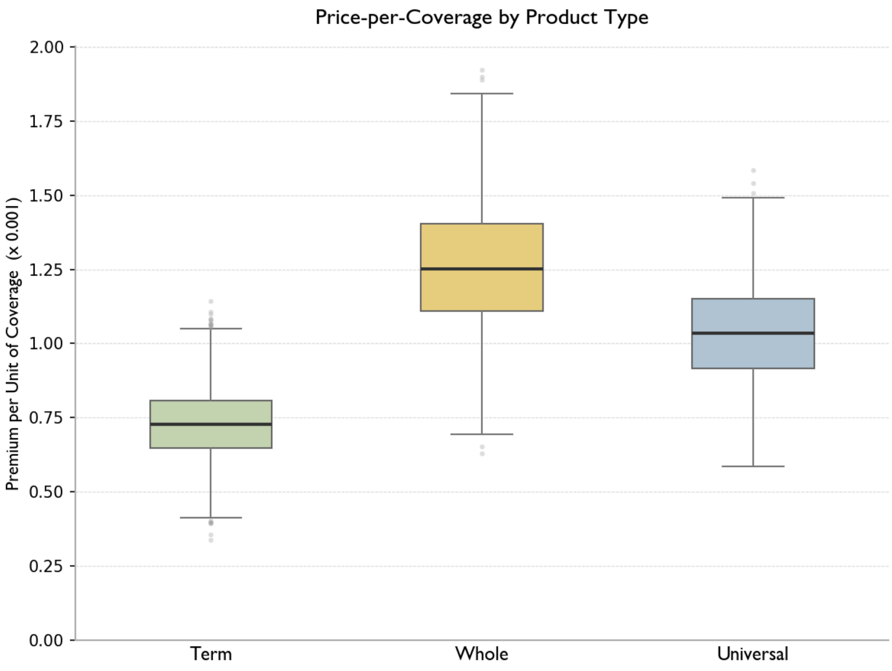
As shown in the graph, the duration for which a policy remains active appears to affect the churn rate. While the rates vary over time, churn drops to zero once the policy duration crosses the 10 year mark.

2. How does premium vs. coverage behave across product types?

An important variable will be the price per coverage, so we can calculate the relative price for each product. Consists of:

price_per_coverage = premium / coverage_amount

With this variable I analyzed (using Claude) the distribution of pricing in the different products:



The premium - coverage relationship differs across product types.
Whole life policies show the highest premium per unit of coverage, indicating more expensive pricing relative to the level of protection provided.
Term policies offer the lowest price per coverage, and therefore appear more efficient.
Universal policies are between these segments but display greater variability, suggesting more diversity pricing structures.

3. Are there channels or payment methods that look risky or attractive?

A channel or payment method might be risky if it shows:

- High churn percentage rate
- Low tenure average
- High late payment average

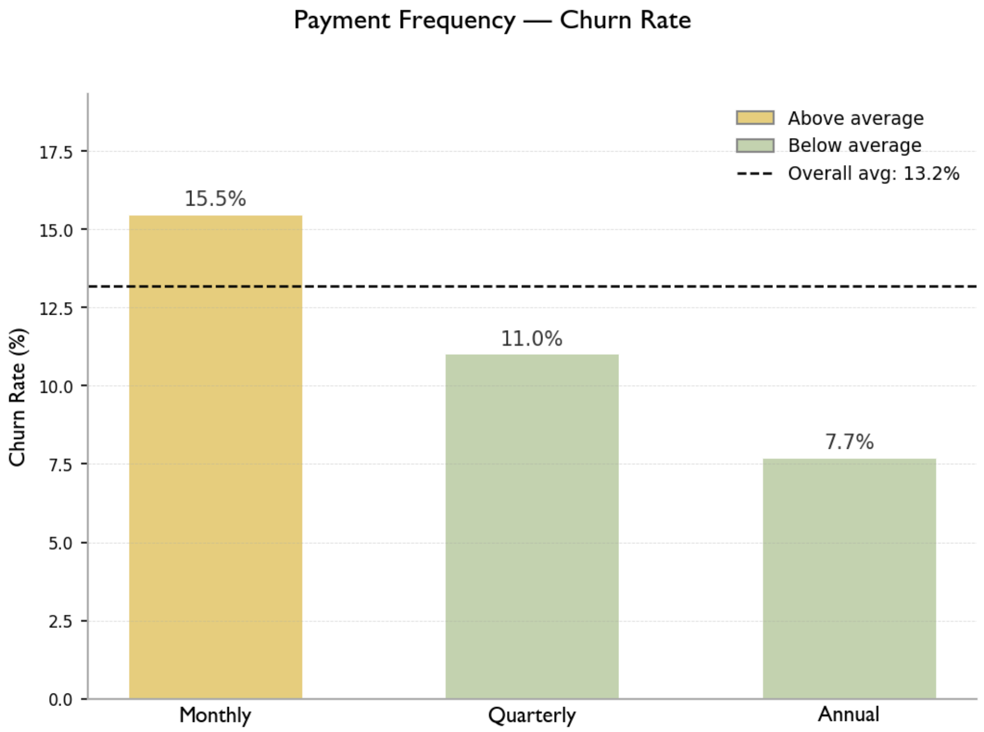
The acquisition channels include online, employer, bank, and agent. Below is an analysis of how the data behaves across each channel:

Acquisition channel	**Churn rate**	**(vs avg churn)**	**Tenure (avg)**	**Late pay (avg)**
Bank	15.5%	+2.3%	83.8m	0.57
Employer	14.4%	+1.2%	82.6m	0.62
Online	14.0%	+0.8%	84.8m	0.59
Agent	11.6%	-1.6%	83.4m	0.60

Agent acquired customers demonstrate the lowest churn and appear the most stable channel.
Bank acquired policies show the highest churn relative to average (suggests elevated retention risk).

Tenure and late payment behavior are broadly similar across channels, indicating that churn differences are likely driven by customer acquisition quality rather than payment performance.

The optional paying methods are annual, monthly and quarterly. Let's analyze the correlation between them and the variables listed above:



Payment frequency is a strong signal: monthly payers churn at 15.5% (+2.3% above avg) while annual payers churn at just 7.7% (-5.5%). Tenure and late payments are nearly identical across frequencies.
So payment frequency itself is the differentiator, not engagement behavior.

Caveats and limitations

This analysis identifies **correlations rather than causal effects**, and some behavioral signals (such as service calls) may reflect churn already in progress.
Additionally, demographic inconsistencies and limited knowledge of the data lowers the precision of the conclusions.

Business Narrative

Churn risk is concentrated in early policy years and is strongly influenced by payment frequency, with **annual payers showing nearly half the churn rate of monthly customers**. Acquisition channel also plays a role as agent sourced policies demonstrate significantly stronger retention than bank acquired policies, despite similar tenure and payment behavior.

These findings suggest that **retention is driven more by customer commitment and acquisition quality** than operational friction.

I would suggest encouraging more customers to move to annual payments, putting more attention on retention in the first years of a policy. Continue analyzing which channels actually bring higher long term value so we can allocate resources more effectively.

Bonus - feature for churn model

According to my analysis these are the key features for a churn model:

1. Payment frequency - showed a strong relationship with churn, with annual payers being significantly more stable than monthly payers.
   This likely reflects a higher level of customer commitment and may serve as an important predictor of churn.
2. Late payment behavior (number of late payments) - appeared to be one of the strongest signals for churn in the analysis. Increasing late payments may indicate financial friction or disengagement, making this a valuable predictive feature.
3. Customer service calls - churn increases as the number of service calls rises, although some of these calls may be directly related to policy termination. Still, call intensity could serve as a short term churn signal if interpreted carefully.

Thank you for taking the time to read and assess :)