Total Goal

1. Understand the business workflows with which data solutions are to be integrated
2. Understanding Relationships Among Data Entities
3. Identifying the Best ML Model
4. How Data Products (ML Model) Integrate with Business Workflows
5. Measuring the Effectiveness of the Data-Driven Solution

While all five goals were introduced in the exploratory data analysis (EDA) phase, Goals 3, 4, and 5 are more aligned with the preprocessing, modeling, and evaluation stages of the project. Therefore, only Goals 1 and 2 are covered in this EDA section, while Goals 3, 4, and 5 are fully addressed in the **Preprocessing & Modeling** phase

Goal 1

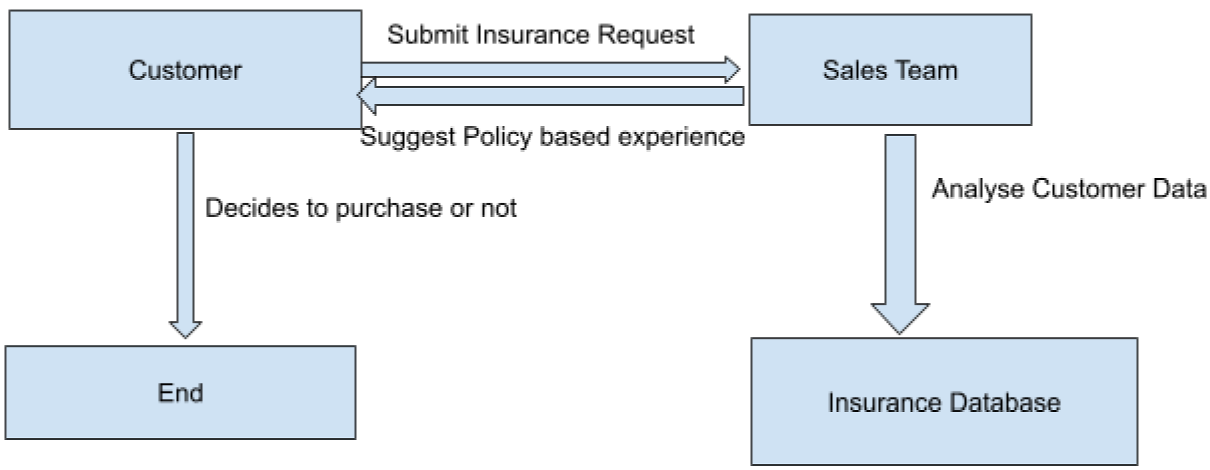
Understand the business workflows with which data solutions are to be integrated

Business Workflow Mapping

Create a **Data Flow Diagram (DFD)** showing how the **insurance company currently sells policies** vs. how the **ML model will enhance this process**.

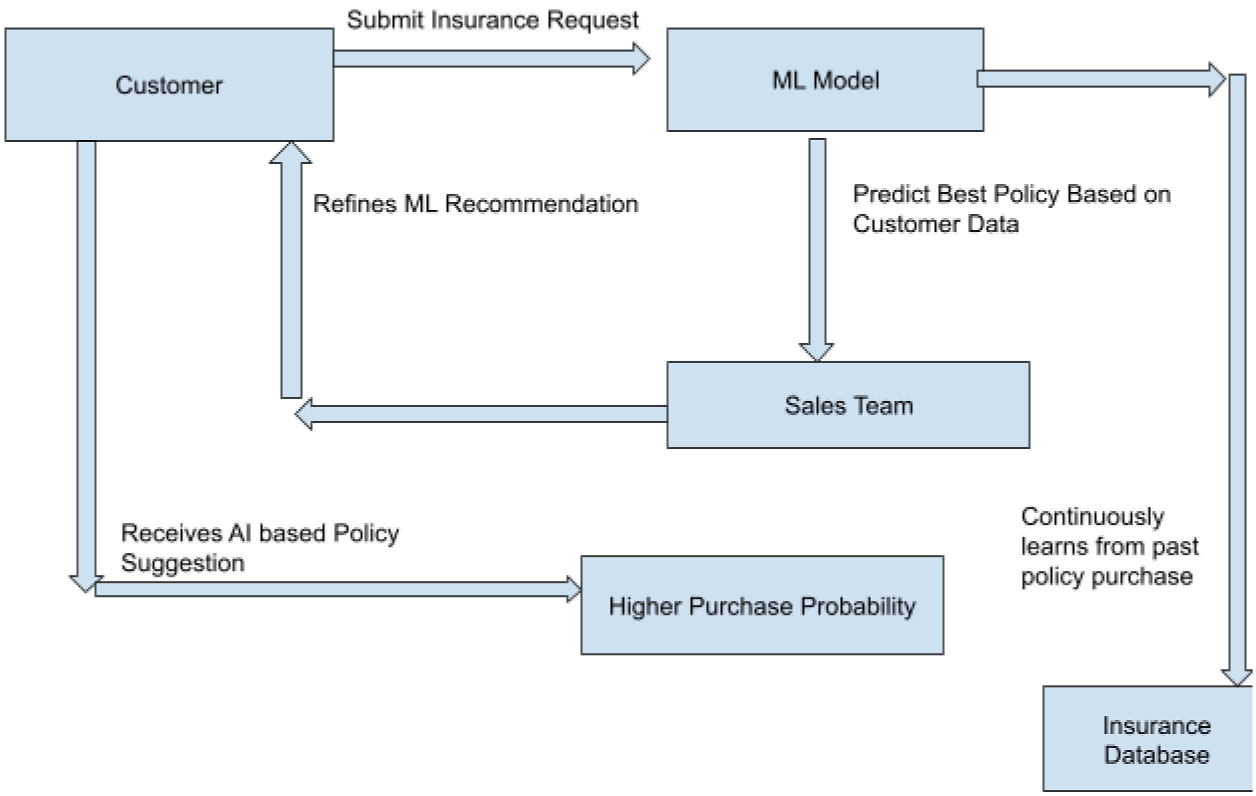
Current Workflow (Without ML)

```
Customer  --->  [Submit Insurance Request]  --->  Sales Team
Sales Team  --->  [Manually Analyzes Customer Data]  --->  Insurance
Database
Sales Team  --->  [Suggests Policy Based on Experience]  --->
Customer
Customer  --->  [Decides to Purchase or Not]  --->  END
```



Enhanced Workflow (With ML Model)

```
Customer  --->  [Submit Insurance Request]  --->  ML Model
ML Model  --->  [Predicts Best Policy Based on Customer Data]  --->
Sales Team
Sales Team  --->  [Refines ML Recommendation]  --->  Customer
Customer  --->  [Receives AI-Based Policy Suggestion]  --->  Higher
Purchase Probability
ML Model  --->  [Continuously Learns from Past Policy Purchases]  --->
Insurance Database
```



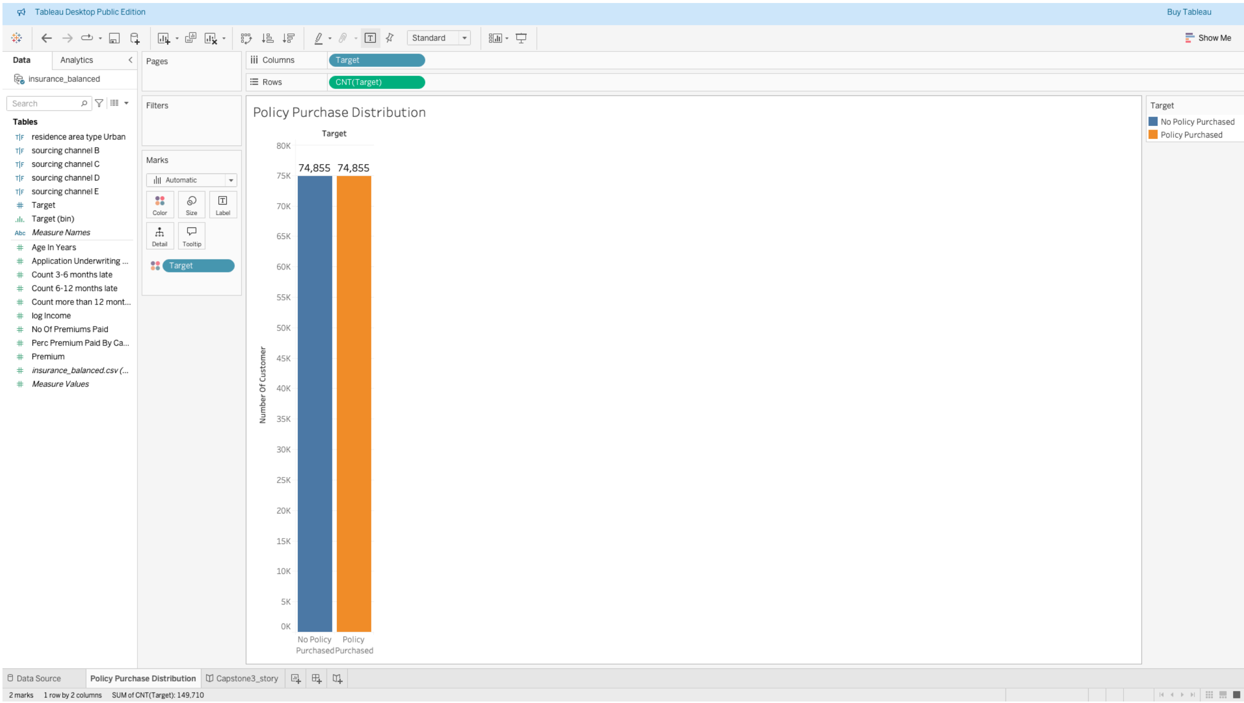
Goal 2

EDA in Python is done in jupyter notebook

EDA in Tableau :

It's done in Tableau. In this document I am sharing the observation of the graph.

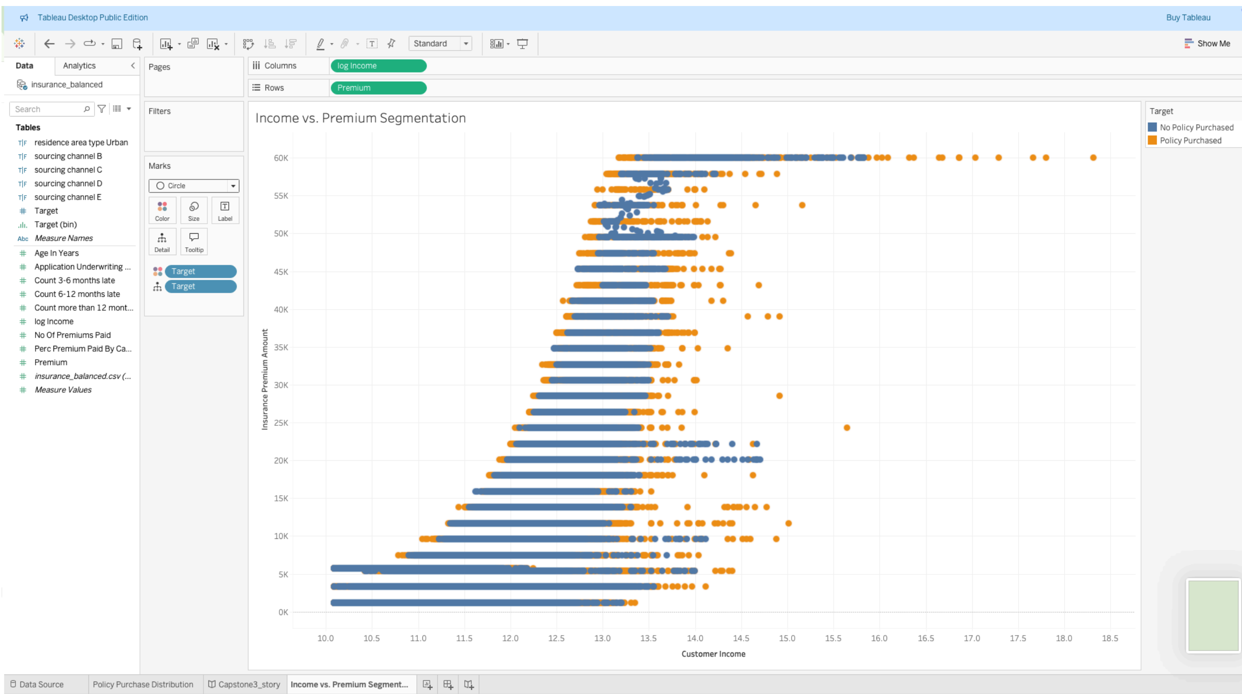
1. Check Policy Purchase Distribution (Class Balance)



Observations:

1. The dataset is now balanced after applying SMOTE (Synthetic Minority Over-sampling Technique).
2. Both classes—"No Policy Purchased" (Blue) and "Policy Purchased" (Orange)—have equal representation (74,855 each).
3. Before SMOTE, there was a class imbalance, where policy purchases (1) were the minority. This could have caused bias in model predictions.
4. After SMOTE, both classes have equal data points, ensuring fairer model training and reducing bias toward the majority class.
5. This balancing allows the machine learning model to learn equally from both customer groups (buyers & non-buyers), improving predictive accuracy.

2. Customer Segmentation by Income & Premium



Observations: "Income vs. Premium Segmentation"

1. **Higher-income customers tend to purchase higher-premium policies**
   - As `log_Income` (X-axis) increases, we see policy buyers (orange) moving towards higher premiums (Y-axis).
   - This suggests a positive correlation between income and premium selectio**n.**

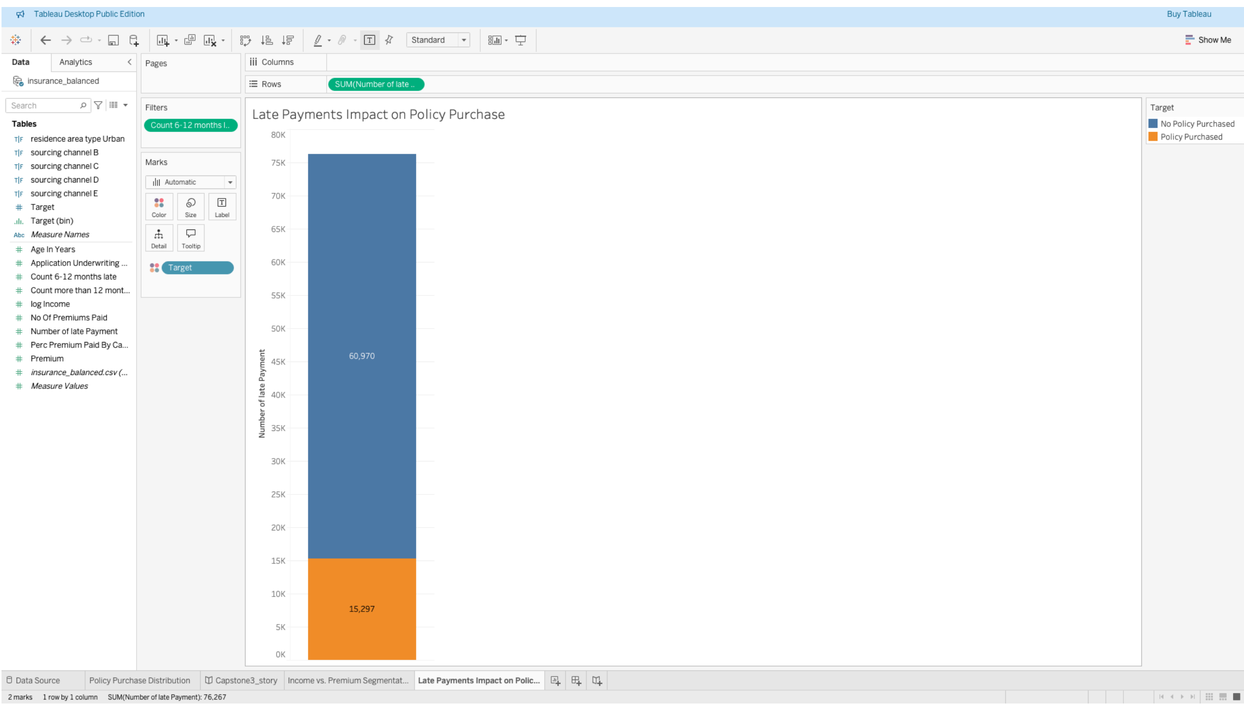
2. **Some lower-income customers still purchase high-premium policies**
   - Not all high-premium purchases come from high-income groups.
   - There are some policy buyers (orange) in the lower-income range, suggesting other factors influence premium choices.

3. **Many non-policyholders (blue) exist across all income levels**
   - Even higher-income customers don't always buy policies (blue dots present at higher `log_Income` values).
   - This could be due to customer awareness, alternative coverage, or lack of need.

Business Insights

1. Higher-income groups are more likely to buy policies with higher premiums.
2. Some low-income customers still opt for high-premium plans—indicating potential financing options or strong policy value perception.
3. Not all high-income customers buy policies—suggesting a potential marketing opportunity to target them.

3. Impact of Late Payments on Policy Purchase



Observations:

1. **Majority of Late Payers Did Not Purchase Policies**

- 60,970 customers (blue) had late payments and did not buy a policy.
- Only 15,297 customers (orange) with late payments still purchased a policy.
  This suggests that late payments might negatively impact policy purchasing decisions.

2. **Late Payments Could Indicate Financial Struggles**

- Customers who delay payments might find it difficult to afford new policies.
  This could be an opportunity for customized policy plans or flexible payment options.

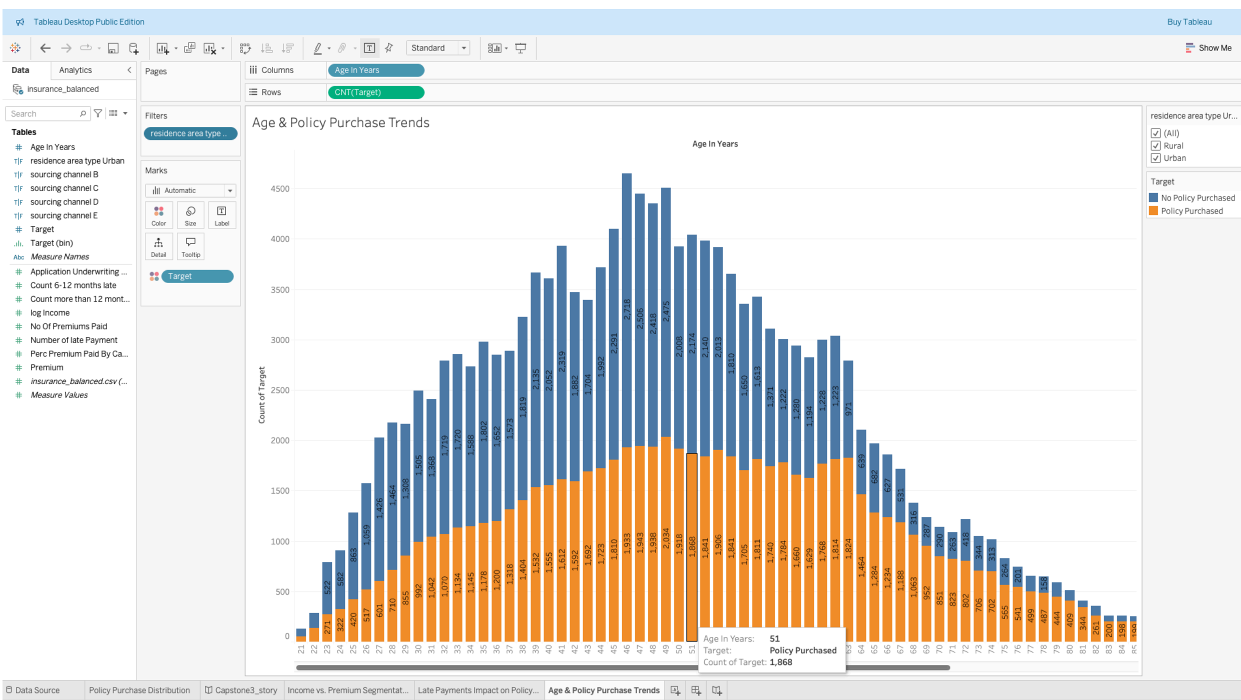
3. **Stacked Distribution Suggests Late Payers Are Riskier Customers**

- Insurers might flag frequent late payers as high-risk customers.
- Offering incentives or penalties for late payments could impact policy purchase behavior.

Business Insights

1. Customers with frequent late payments are significantly less likely to purchase a new policy.
2. This presents an opportunity for insurers to target these customers with special financial plans.
3. Marketing campaigns can focus on educating these customers about timely payments and policy benefits.

4. Customer Demographics & Policy Purchases



Observation:

1. **Peak Policy Purchases:** The highest number of policies purchased (orange bars) is seen among customers aged 30-50 years.
2. **Decline After 50:** Policy purchases gradually **decline after age 50**, with fewer customers buying policies as they age.
3. **Residence Area Impact:** Now that the Urban vs. Rural filter is active, you can compare how policy purchases vary across these areas.

4. **Balanced Representation:** Since SMOTE was applied, both purchased (1) and non-purchased (0) classes are fairly distributed.

Goal 3

Final Model Selection

Model	Accuracy	F1-Score	ROC-AUC
Random Forest	0.94	0.94	N/A
Logistic Regression	0.79	0.79	0.8704
XGBoost	0.90	0.91	0.9673

Final Model Chosen: **Random Forest-** It achieved the highest overall accuracy and F1-score while maintaining balanced precision/recall for both classes. It also showed strong performance in the confusion matrix.

Goal 4

Integration with Business Workflow

Let's answer: How will this model be used in the business process?

Where the Model Fits in the Workflow

- **Before**: Sales team manually identifies prospects → Uses basic rules to target customers
- **After**: Model scores each customer → Predicts likelihood of insurance purchase (1 = likely, 0 = not likely)
- **Action**:
    - High-score customers → prioritized for email campaigns or outbound calls
    - Low-score customers → receive low-cost nurturing sequences

How Will the Model Be Accessed?

- **Option 1:** Embedded into a Tableau dashboard (score by customer)
- **Option 2:** CSV output sent to marketing/sales team weekly
- **Option 3:** Automated batch scoring in production if deployed

Outputs to Show

- Predicted Class (0/1)
- Probability Score (e.g., 0.83 confidence for class 1)
- Top Feature Contributions (optional with SHARP/Feature Importance)

Goal 5

Measuring Model Impact (Define how you'll evaluate business impact)

A/B Testing Plan: Evaluating Model Impact

To evaluate the real-world effectiveness of the insurance policy prediction model, we propose an A/B testing framework. This will allow us to measure how well the model performs in improving customer engagement and policy purchase rates.

Testing Groups

- **Group A (Control Group)**
  Customers receive policy recommendations based on the existing (manual or rule-based) method.
- **Group B (Treatment Group)**
  Customers receive policy recommendations based on the machine learning model's predictions.

Evaluation Metrics (KPIs)

We will track and compare the following KPIs between the two groups:

- **Policy Purchase Rate**: % of customers who purchase a recommended policy.
- **Revenue per Customer**: Total premium value generated per customer.
- **Uplift Score**: Improvement in conversion rate between Group B and Group A.
- **Time to Conversion** *(optional)*: How quickly customers take action after receiving a recommendation.

Success Criteria

The model will be considered successful if:

- Group B shows a **significant uplift** (e.g., >10%) in policy purchase rate compared to Group A.
- The model recommendations lead to **higher revenue or reduced marketing effort**.

Business Integration Notes

- The A/B test can be deployed through an existing CRM or digital marketing platform.
- Model predictions can be surfaced via email campaigns, app notifications, or customer dashboards.