Clustering & Dimensionality Reduction Analysis on Telco Customer Churn

1. Main Objective of the Analysis

The objective of this project is to explore the Telco Customer Churn dataset using **Unsupervised Learning** techniques. Specifically, the focus is on applying **clustering algorithms** and **dimensionality reduction** methods to identify distinct customer segments and uncover hidden patterns in churn behaviour. These insights can help stakeholders design targeted interventions and improve customer retention.

We will train and compare different clustering models and use dimensionality reduction techniques like **PCA** for interpretability and visualisation.

---

2. Data Set Description

The **Telco Customer Churn** dataset contains records for 7043 customers, with 21 variables related to their demographics, service usage, and account information.

Features Include:

- **Demographics**: `gender`, `SeniorCitizen`, `Partner`, `Dependents`
- **Services**: `PhoneService`, `MultipleLines`, `InternetService`, `StreamingTV`, `TechSupport`, etc.
- **Account Info**: `tenure`, `MonthlyCharges`, `TotalCharges`, `Contract`, `PaymentMethod`
- **Churn Indicator**: `Churn` (Yes/No)

Our goal is to find meaningful clusters of customers **without using the target variable (`Churn`)** in the clustering process.

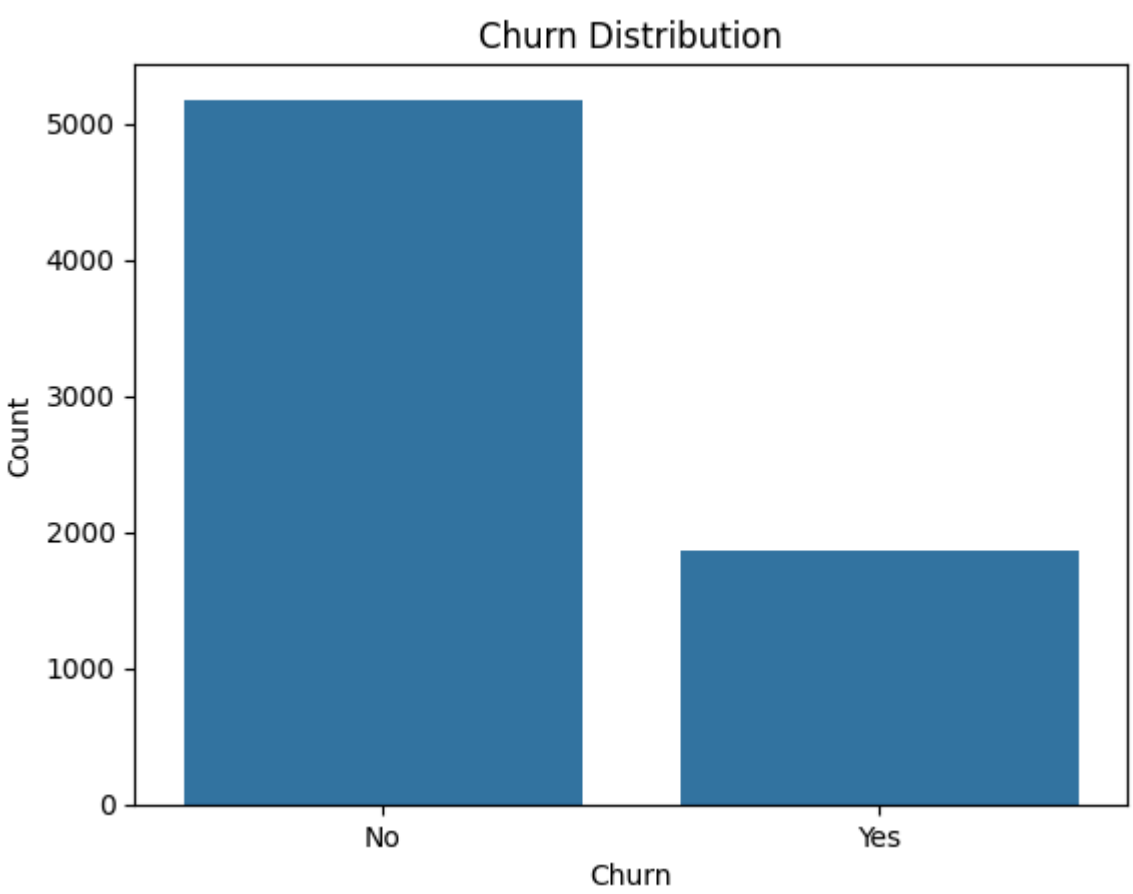
---

3. Data Exploration & Preprocessing

Key Preprocessing Steps:

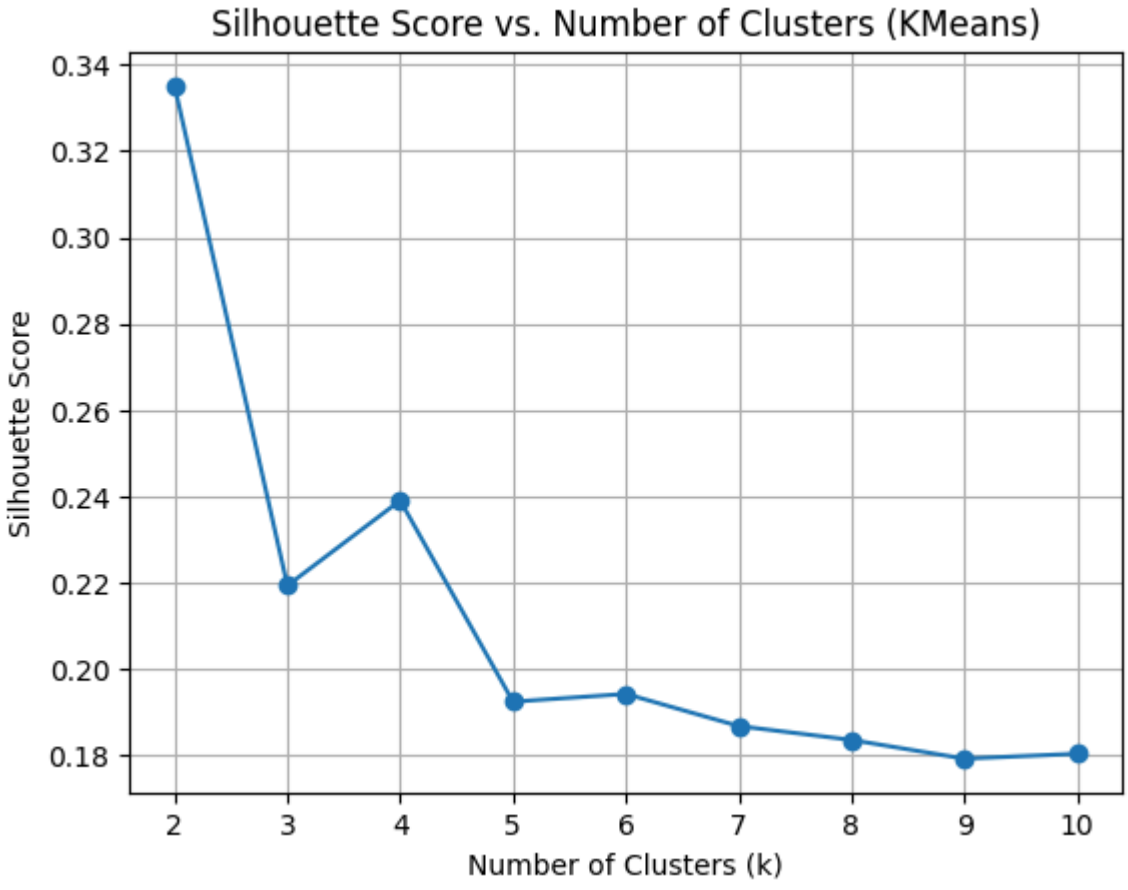
- Converted `TotalCharges` to numeric, handled ~11 missing rows.
- Encoded categorical variables using One-Hot Encoding.
- Scaled numerical features using Min-Max Scaling.
- Created a new feature `ServicesCount` = number of services used per customer.

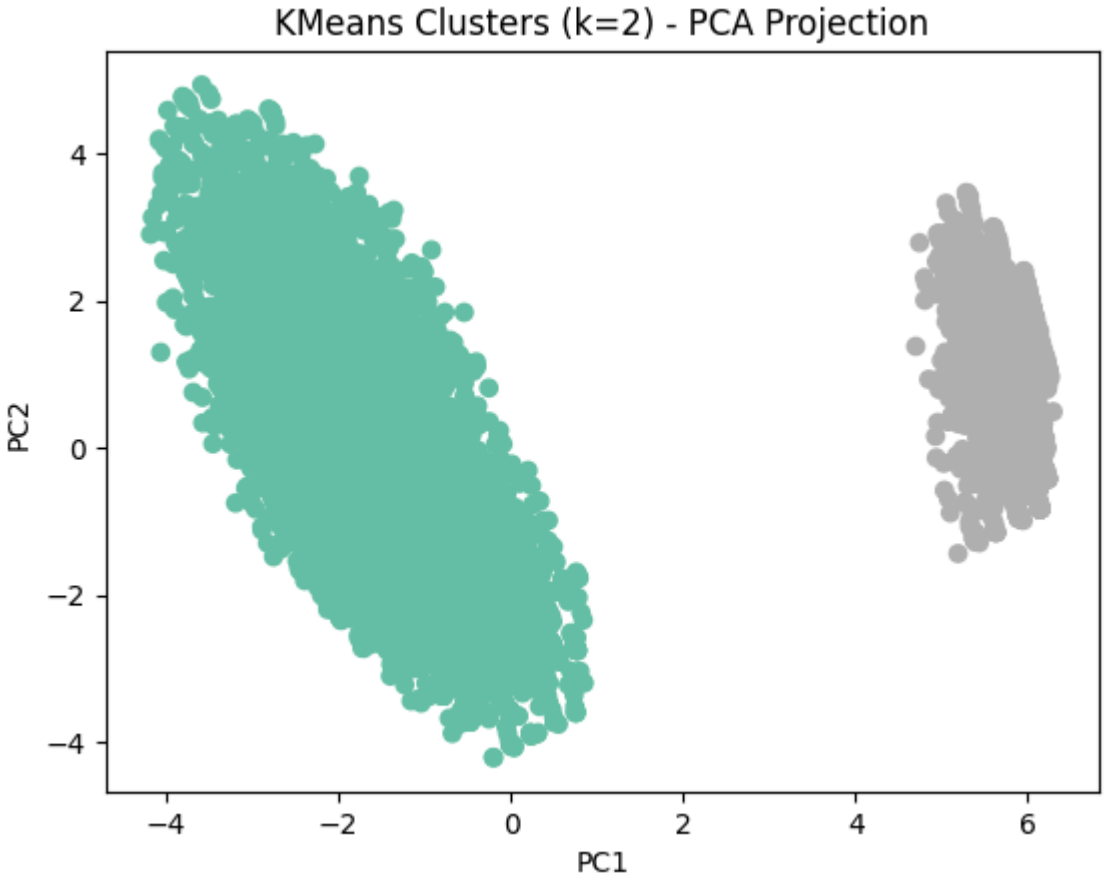
Churn Distribution Visualisation



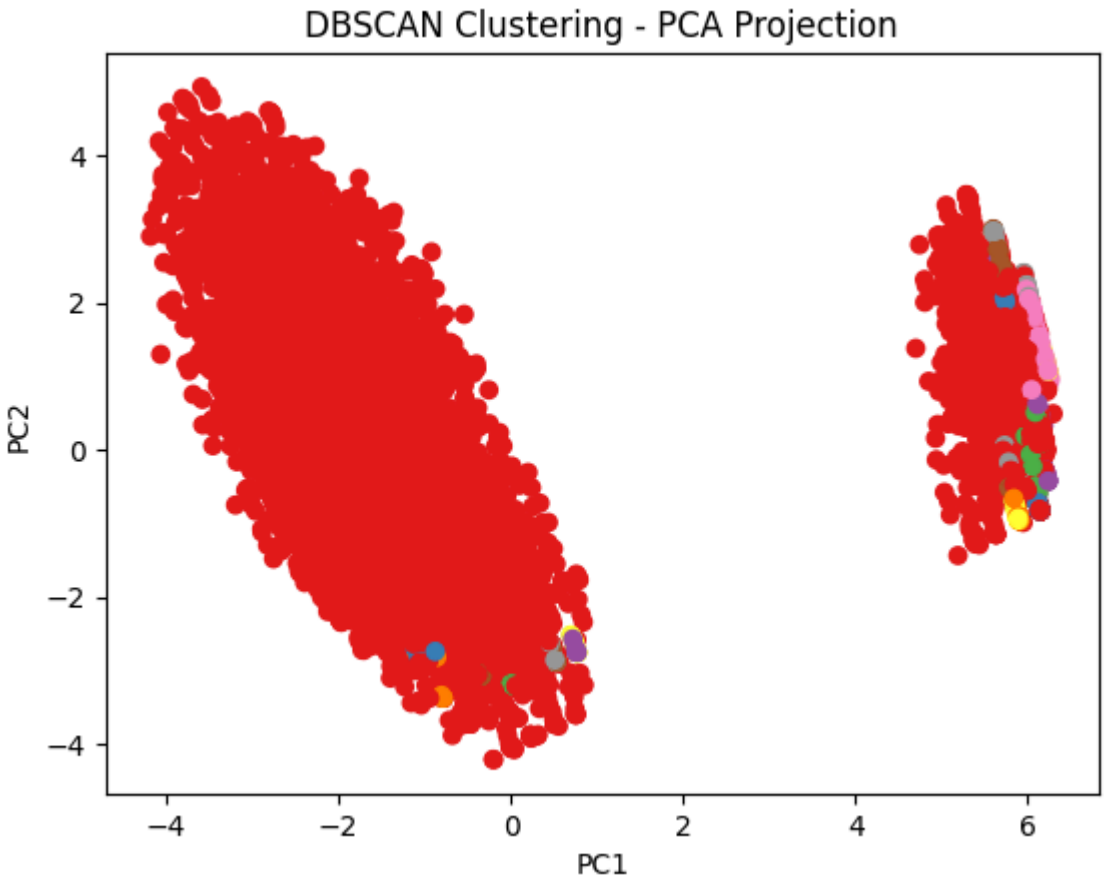
4. Model Training: Clustering Variants

K-Means Clustering

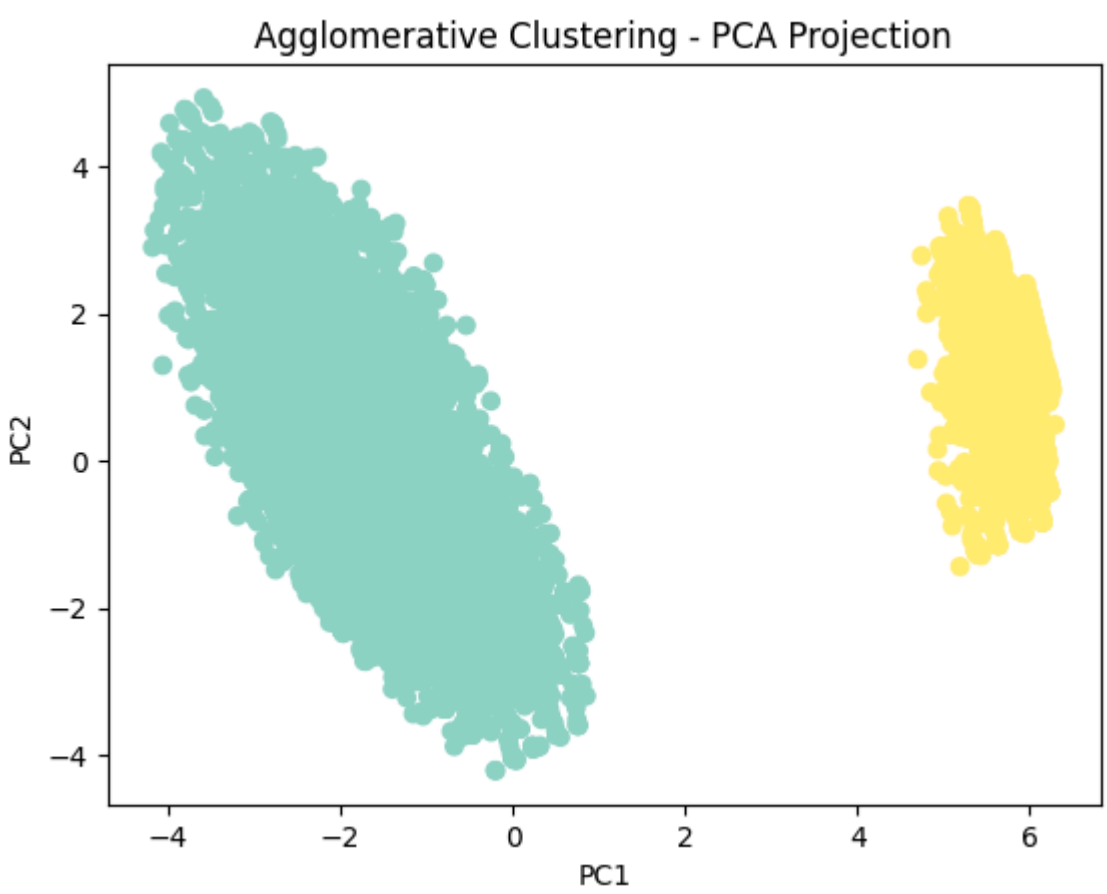




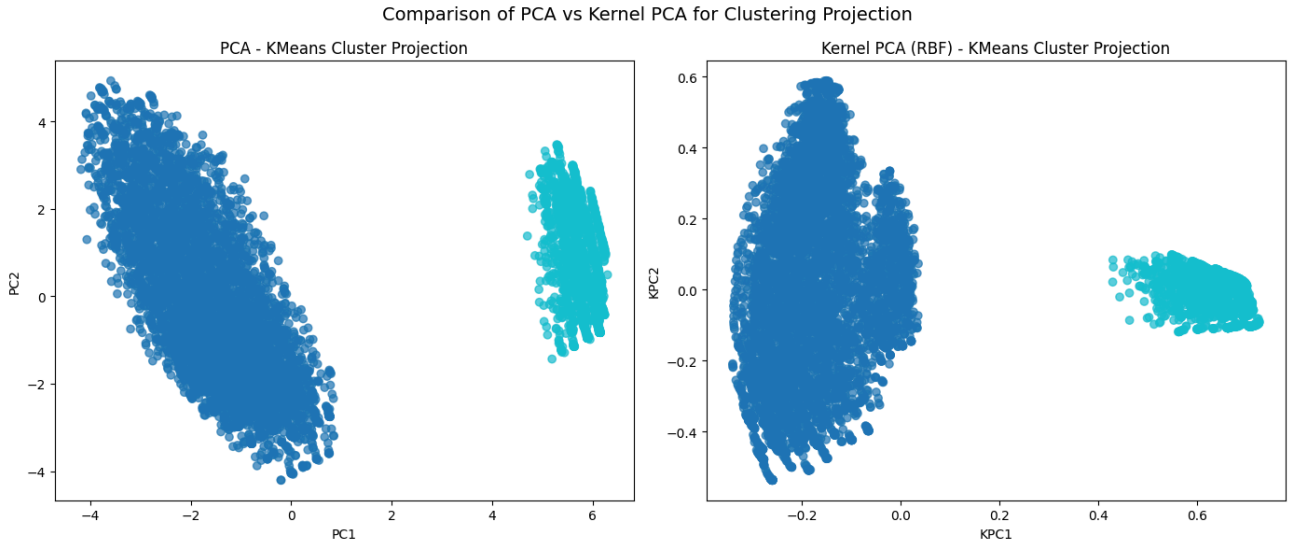
DBSCAN (Density-Based Spatial Clustering)



Hierarchical Clustering (Agglomerative)



5. Dimensionality Reduction



6. Recommended Final Model

After evaluating multiple unsupervised learning models, the **K-Means algorithm with** `k = 2` clusters is recommended as the final model. This decision is based on both the **highest silhouette score** among all tested models and the **simplicity and interpretability** it offers when dividing customers into two distinct segments.

Why K-Means with 2 Clusters?

- Achieved the **highest silhouette score**, suggesting optimal separation between the two groups.
- Results in a **clear, business-actionable split** of customers into two behavioural or risk-based categories.

- Easy to interpret and communicate to stakeholders for decision-making.
- Offers a strong foundation to support more targeted churn intervention strategies.

Interpretation of Clusters

Cluster	Description	Churn Propensity
0	Loyal customers with longer tenure, annual/biannual contracts, auto-pay	Low
1	At-risk customers with short tenure, monthly contracts, manual payment	High

This binary segmentation serves well as a first-tier churn flag system. Customers in cluster 1 can be prioritised for personalised outreach or retention incentives.

---

7. Key Findings & Insights

The clustering analysis revealed the following important insights:

- **Contract Type Strongly Differentiates Customers**: Those on **month-to-month** plans are more likely to be placed in the high-risk cluster, while customers with **1- or 2-year contracts** are more loyal.
- **Tenure Remains a Powerful Indicator**: Short-tenure customers (< 6 months) are disproportionately found in the high-risk cluster. Early engagement plays a crucial role in retention.
- **Payment Method Drives Loyalty**: Customers using **automatic payments** (bank transfer, credit card) are more likely to be found in the low-risk cluster, whereas those using **paper or mailed cheques** tend to fall into the high-risk category.
- **Service Bundling Plays a Role**: Those with **multiple services** (e.g., streaming, tech support, phone) are more engaged and appear more frequently in the low-risk cluster.
- **Visual separability confirmed via both PCA and Kernel PCA**, with KPCA offering smoother, more curved boundaries in the cluster layout — supporting the idea that customer behaviour patterns are somewhat non-linear.

These findings indicate that just two clusters capture a significant portion of the churn behaviour variability, offering a simplified yet effective view for stakeholders.

---

8. Suggestions for Next Steps

To expand and refine this analysis, the following next steps are suggested:

1. **Enhance Feature Depth**:
    - Introduce more customer engagement metrics (e.g., logins, support tickets, net promoter score).
    - Add financial data such as customer lifetime value (CLV).
2. **Supervised Model Comparison**:
    - Build a **churn classification model** using these clusters as pre-segmentation.
    - Compare model performance trained on full data vs. split by cluster.
3. **Business Strategy Testing**:

- Use the high-risk cluster as the **target group** for churn prevention campaigns.
- Run A/B tests on retention offers tailored to this group.

4. **Refine Clustering Techniques**:
    - Experiment with **Gaussian Mixture Models (GMM)** for probabilistic cluster membership.
    - Use **UMAP** as an alternative to PCA/KPCA for visualisation and distance preservation.
5. **Real-Time Cluster Assignment**:
    - Deploy the model to assign new customers to one of the two clusters in real-time.
    - Integrate with a CRM or marketing automation platform for action.