D. Saul Jameson
https://www.linkedin.com/in/dsauljameson/

Clustering Project Documentation

1. Project Overview

Objective

The goal of this project was to segment customer behavior using **K-Means clustering** to identify meaningful groups based on user interactions. The results help businesses understand audience segments, personalize marketing strategies, and optimize resource allocation.

Use Case

The project aims to answer the question: **What ad platforms bring in the best customers?** By segmenting website visitors, we can determine which platforms attract high-value users based on engagement, purchase intent, and other key metrics. The project focuses on segmenting website visitors based on features like **time spent on the site, pages viewed, purchase intent, bounce rate, and referral source**.

---

2. Data Generation & Preprocessing

Dataset

Since real-world data was not available, we generated synthetic data to reflect realistic user behavior.

Features included:

- **Time Spent (minutes)** – Duration of user sessions.
- **Pages Viewed** – Number of pages a user visits.
- **Bounce Rate (%)** – Percentage of sessions where users leave without engagement.
- **Purchase Intent Score** – Derived variable indicating likelihood of conversion.
- **Referral Source** – Categorical variable indicating how users found the site (organic search, paid ads, direct, referral).

D. Saul Jameson
https://www.linkedin.com/in/dsauljameson/

Preprocessing Steps

- **Normalization**: Used **MinMaxScaler** to scale numeric features for better clustering.
- **Encoding**: One-hot encoded categorical variables like referral source.
- **Handling Outliers**: Used **IQR (Interquartile Range)** to cap extreme values.

---

3. Clustering Methodology

K-Means Algorithm

- **K-Means++ Initialization**: Used to select optimal cluster centers.
- **Elbow Method**: Determined the best **K (number of clusters)** by plotting within-cluster sum of squares (WCSS).
- **Silhouette Score**: Validated cluster quality by measuring how well-separated clusters are.

Cluster Analysis

After running K-Means with **K=4**, we identified four distinct segments:

1. **Engaged Buyers** – Users with high time spent, low bounce rates, and high purchase intent.
2. **Casual Browsers** – Users who view multiple pages but have moderate engagement.
3. **One-Time Visitors** – Users with high bounce rates and low engagement.
4. **Referral-Driven Users** – Users coming from referrals with varying engagement levels.

---

4. Visualization & Insights

Looker Dashboard Overview

The clustering results are visualized in Looker through interactive charts and tables that allow filtering and detailed analysis.

Key Visuals in Looker:

1. **Cluster Distribution by Ad Platform** – Displays the percentage of customers in each cluster per advertising platform, helping identify where high-value users originate.

2. **Filter Controls for Custom Segmentation** – Users can apply filters (e.g., time spent, referral source, or pages viewed) to create **dynamic segments**, adjusting averages based on selected criteria.

3. **Validation Table** – A detailed breakdown of the clustered data, allowing users to validate results and cross-check key insights.

Business Implications

- **Personalized Marketing**: Create **custom campaigns** targeting each user segment.

- **Optimized Ad Spend**: Focus retargeting ads on **Engaged Buyers** instead of **One-Time Visitors**.

- **Website Experience Improvement**: Reduce bounce rates by improving content flow for **Casual Browsers**.

---

5. Deployment & Scalability

Storage & Integration

- **BigQuery**: Storing segmented user data for analysis.

- **Looker Studio**: Used for **interactive dashboards** to track cluster behavior over time.

- **Real-Time Processing**: The model can be integrated with **Google Analytics API** for live segmentation.

Automation

- **Scheduled re-training**: Using **Google Cloud Functions** to re-run clustering periodically.
- **Dynamic threshold tuning**: Adjusting cluster boundaries based on seasonal trends.

---

6. Future Enhancements

- **More Features**: Incorporate **session device type, geographic location, and past purchase history**.
- **Hierarchical Clustering**: Explore advanced techniques for more **fine-grained segmentation**.
- **Predictive Modeling**: Use cluster insights to **train ML models** for **conversion prediction**.

---

7. Summary

This clustering project successfully **segmented user behavior**, providing actionable insights for marketing and UX optimization. The methodology ensures **scalability** and **real-world applicability**, making it valuable for businesses aiming to personalize user interactions and increase conversions.

Next Steps

- Deploy the model for **live segmentation**.
- Use insights to **optimize digital marketing campaigns**.
- Continuously refine clusters based on **real-time user behavior**.