ADVANCED DATA ANALYTICS
DATA 5200

VALENTINA NGUYEN
ARNOLDO RUIZ
COREY THOMPSON

EPA

VAC TRUCK PARTS

Environmental Products & Accessories, LLC.

DATA DRIVEN SALES OPTIMIZATION

TABLE OF CONTENTS

EPA

VAC TRUCK PARTS

Environmental Products & Accessories, LLC.

EPA Sales is a jetting and vacuum truck parts reseller and manufacturer that offers a wide range of components at competitive prices. Unlike niche suppliers that focus on limited product lines, EPA Sales provides parts across multiple industries.

Current Customer Segments

Industrial and Commercial Contractors

Municipalities and Public Works Departments

Environmental Services and Waste Management Companies

OEMs (Original Equipment Manufacturers)

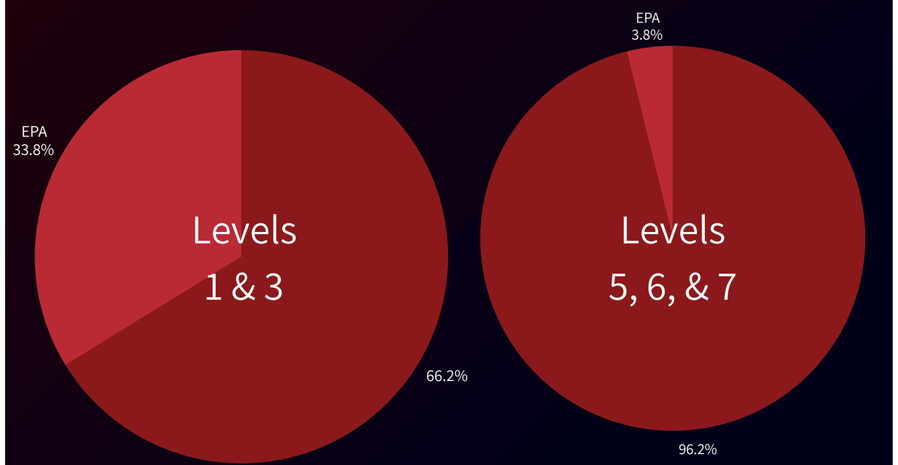
Equipment Dealers

EPA

VAC TRUCK PARTS

Environmental Products & Accessories, LLC.

Current Market Share



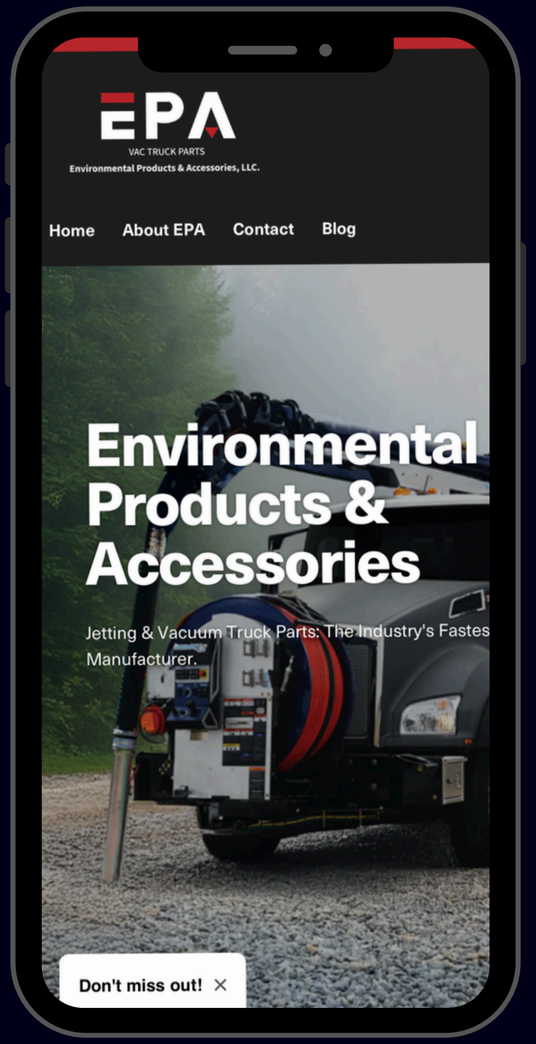
GOALS

- Will a customer reorder within 30 days?
- When will they reorder?

“Reorder” or “Repurchase”: any order placed by the same customer following a prior transaction.

THE DATA

- EPA Sales transaction data directly captures real customer purchasing behavior.
- Includes critical purchase timestamps, order amounts, and product types.
- Ideal for predicting short-term customer loyalty (repurchase behavior) because it tracks actual buying patterns, not just website visits or abandoned carts.
- Data comes from Shopify, a trusted and structured e-commerce platform, minimizing missing or unreliable records.



DATA PREPROCESSING

1. DROPPED UNNECESSARY COLUMNS
2. FILTERED INVALID ROWS
3. RENAMED COLUMNS FOR CLARITY
4. CHANGED DATA TYPES
5. CUSTOMER SEGMENTATION ADJUSTMENT
6. JOINED TABLES FOR ANALYSIS

VARIABLE DICTIONARY

ORDER_NUM

Unique identifier for each order. Primary key.

TAGS

Meta data was originally going to be used as customer segment indicator

EMAIL_DOMAIN

Domain extracted from emails

DAYS_BETWEEN_ORDERS

Created as a derived attribute from order date and last order date.

DOMAIN _TYPE

0 = personal email
1 = corporate email

ADDITIONAL DEFINITIONS & CONSIDERATIONS

- Predict when a customer will reorder a previously purchased product.
- Focus: Model the likelihood of a reorder within 30 days
- A “reorder” = any repeat purchase of a product a customer bought before
- Analyze purchase behavior, including:
  Order frequency
  Spending patterns
  Product selection

EXPLORATORY DATA ANALYSIS

Analysis of the dataset is based on the joined customers and transaction_details table as well as the products table. The analysis is divided into customer purchase behavior, product analysis, and time-based trends to gain insight into purchasing patterns and sales performance.

- CUSTOMER PURCHASE BEHAVIOR
- PRODUCT ANALYSIS
- TIME-BASED ANALYSIS
- CUSTOMER REPURCHASE DISTRIBUTION
- CORRELATION

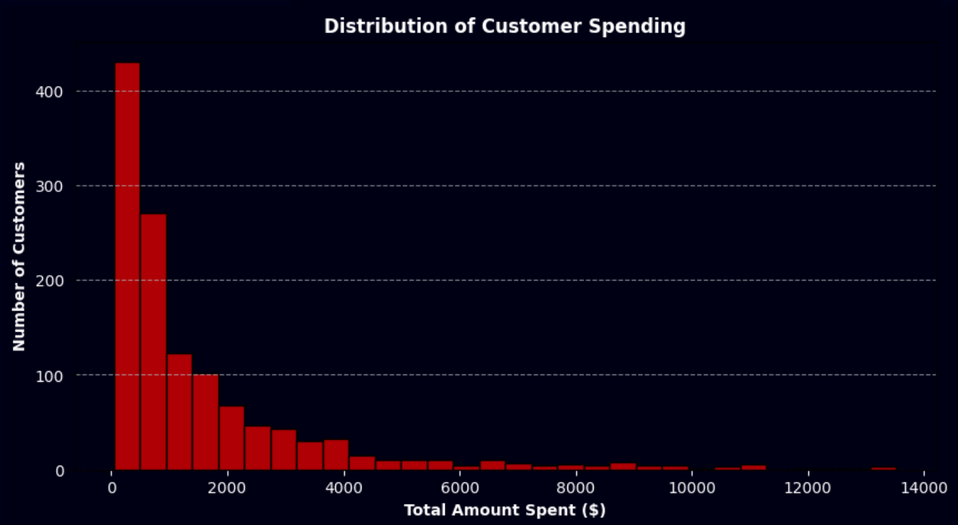
CUSTOMER PURCHASE BEHAVIOR

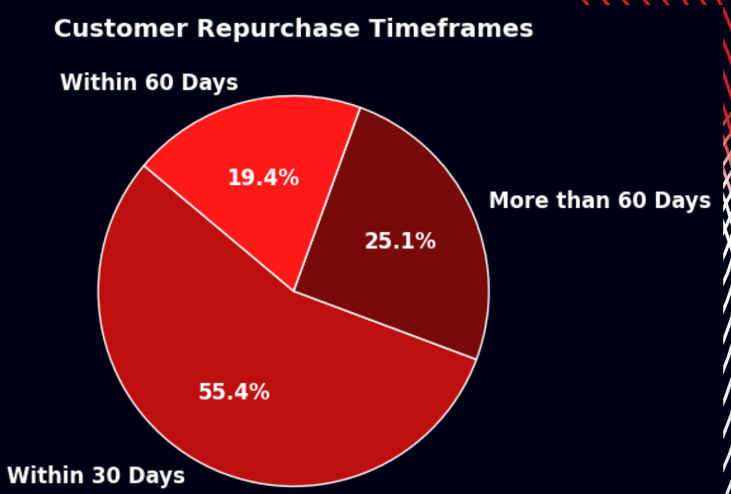
Key Insights

- Spending is heavily right-skewed — most customers spend moderately; a small few spend much more.
- 76% of customers are non-priority (lower spenders, less frequent).
- Priority Access customers represent only 24% but contribute significantly to revenue.
- Most customers only purchased once; frequent buyers are rare but extremely valuable.

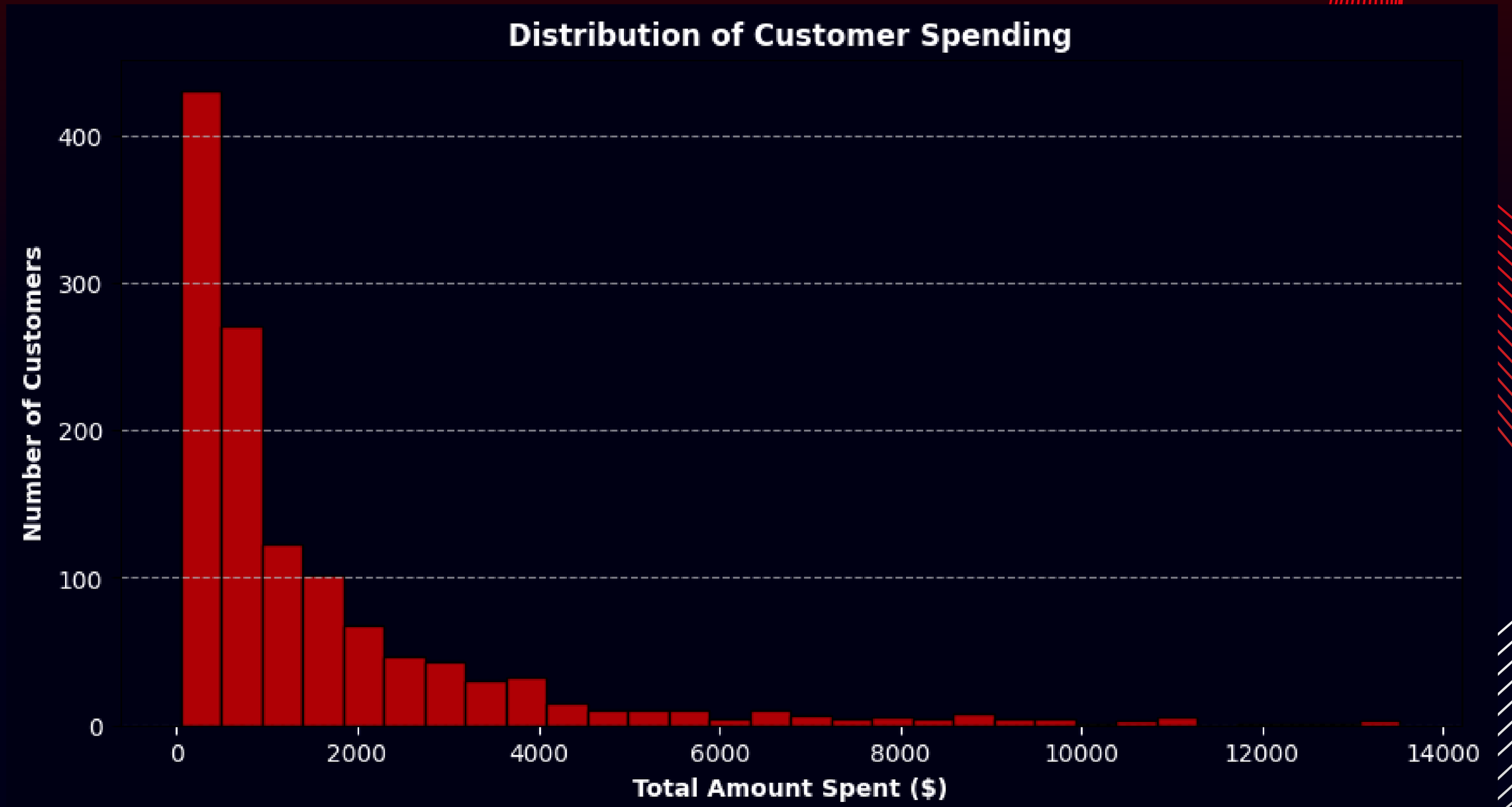
Strategic Implications

- Focus on retention of high spenders and frequent buyers.
- Upsell and nurture the mid-tier ($1K–$10K) customers.
- Identify early loyalty signals (like purchase frequency) to intervene faster.





CUSTOMER PURCHASE BEHAVIOR



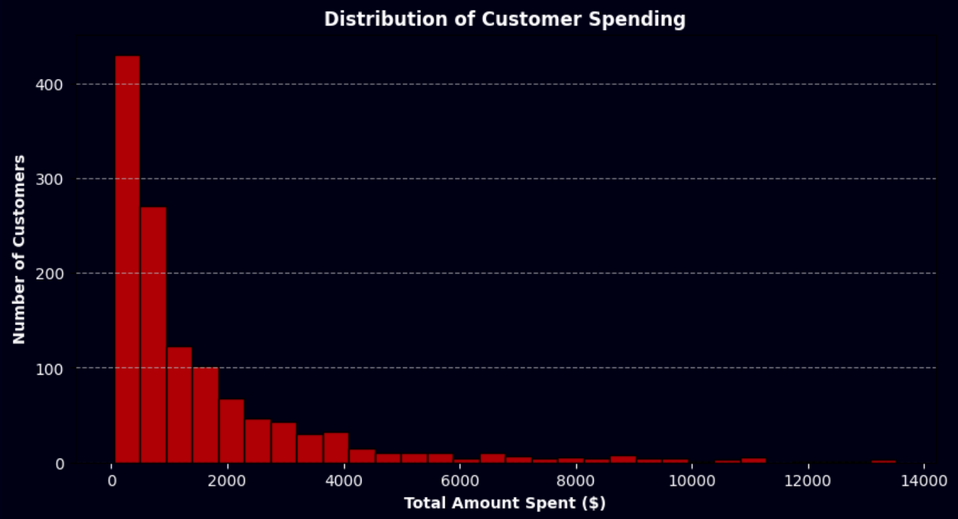
CUSTOMER PURCHASE BEHAVIOR

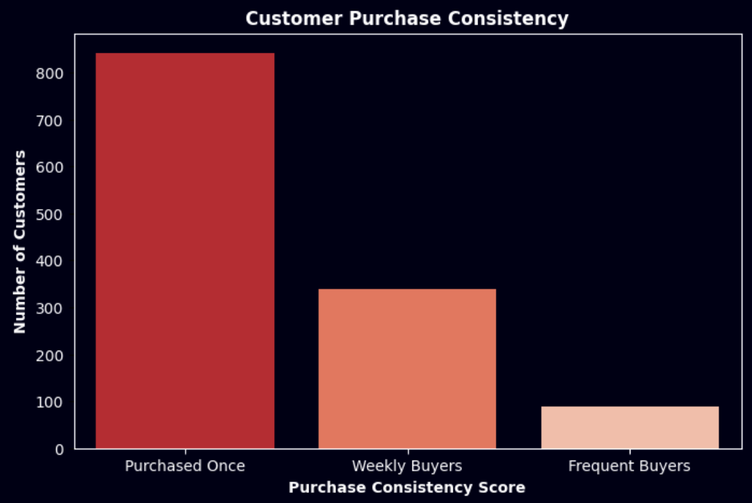
Key Insights

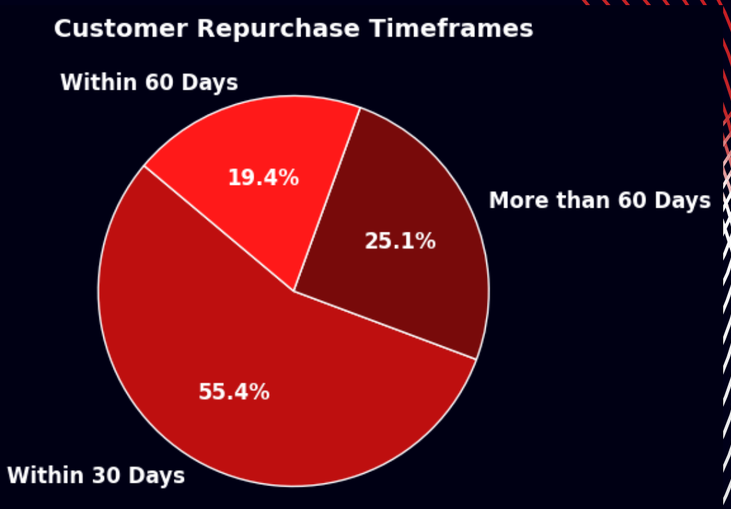
- Spending is heavily right-skewed — most customers spend moderately; a small few spend much more.
- 76% of customers are non-priority (lower spenders, less frequent).
- Priority Access customers represent only 24% but contribute significantly to revenue.
- Most customers only purchased once; frequent buyers are rare but extremely valuable.

Strategic Implications

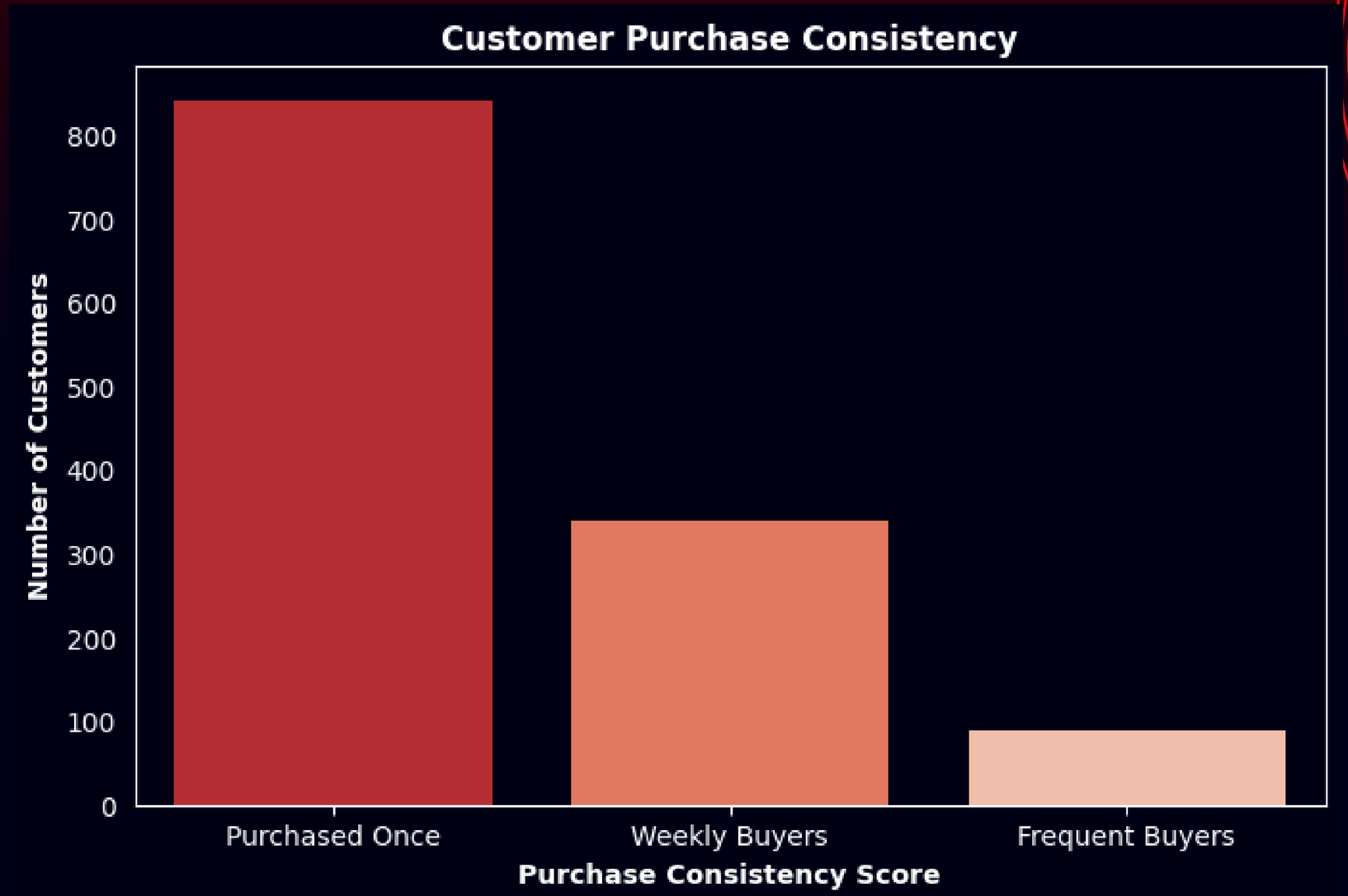
- Focus on retention of high spenders and frequent buyers.
- Upsell and nurture the mid-tier ($1K–$10K) customers.
- Identify early loyalty signals (like purchase frequency) to intervene faster.







CUSTOMER PURCHASE BEHAVIOR



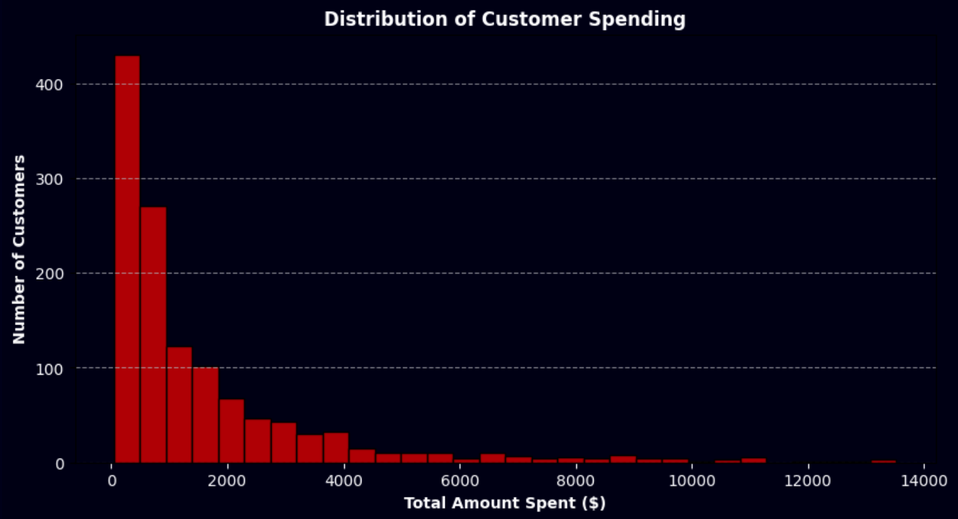
CUSTOMER PURCHASE BEHAVIOR

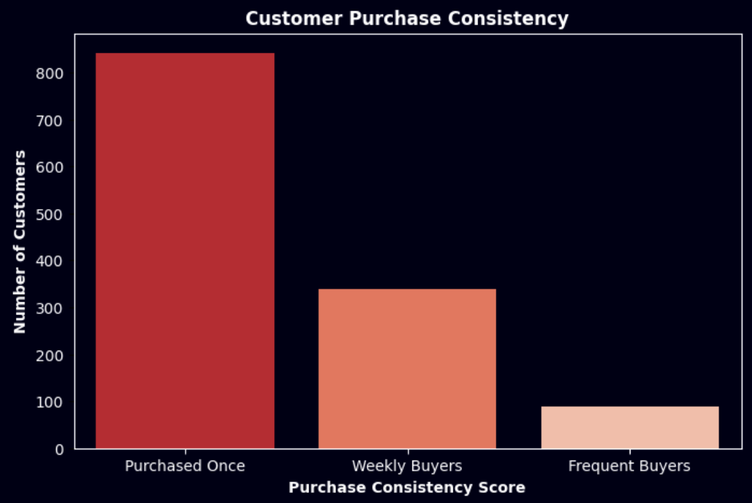
Key Insights

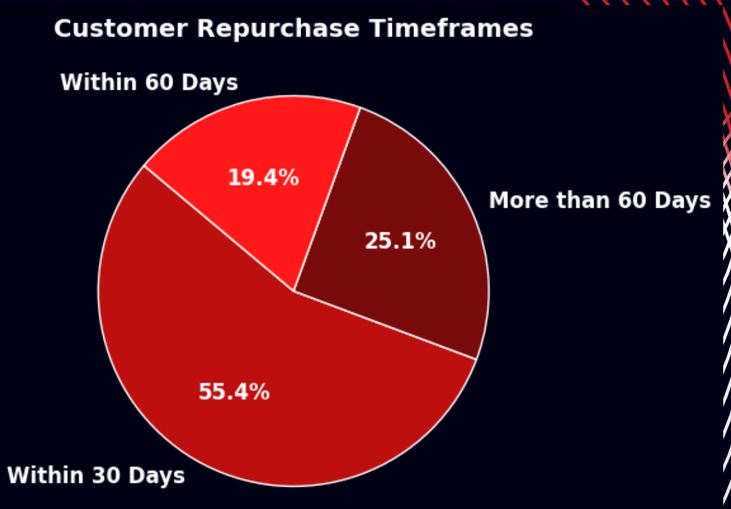
- Spending is heavily right-skewed — most customers spend moderately; a small few spend much more.
- 76% of customers are non-priority (lower spenders, less frequent).
- Priority Access customers represent only 24% but contribute significantly to revenue.
- Most customers only purchased once; frequent buyers are rare but extremely valuable.

Strategic Implications

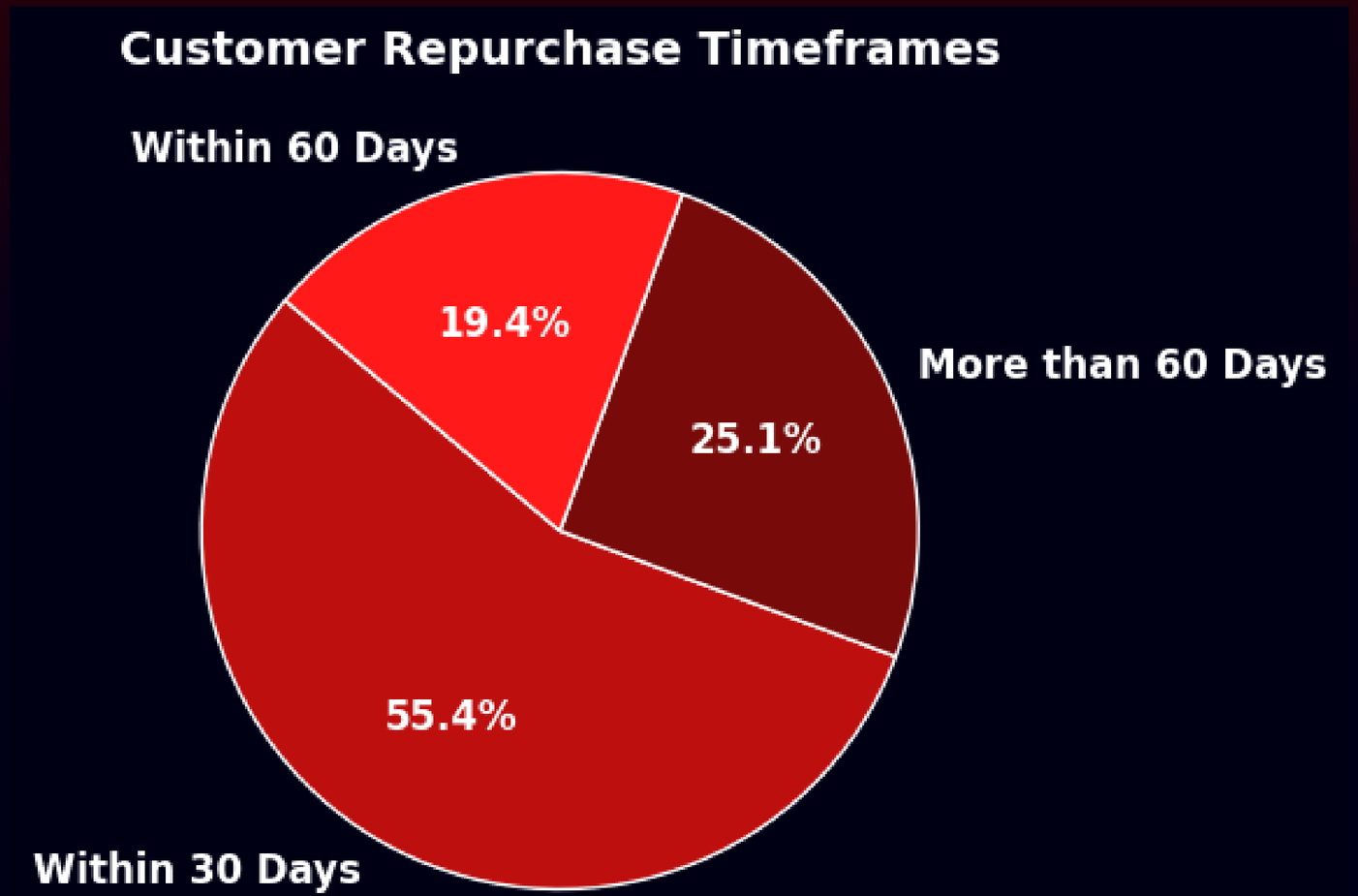
- Focus on retention of high spenders and frequent buyers.
- Upsell and nurture the mid-tier ($1K–$10K) customers.
- Identify early loyalty signals (like purchase frequency) to intervene faster.





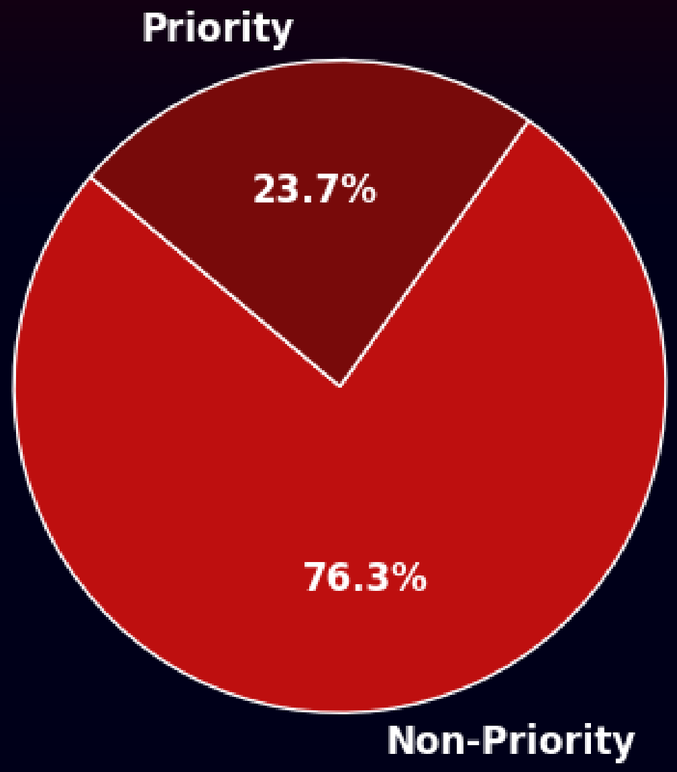


CUSTOMER PURCHASE BEHAVIOR

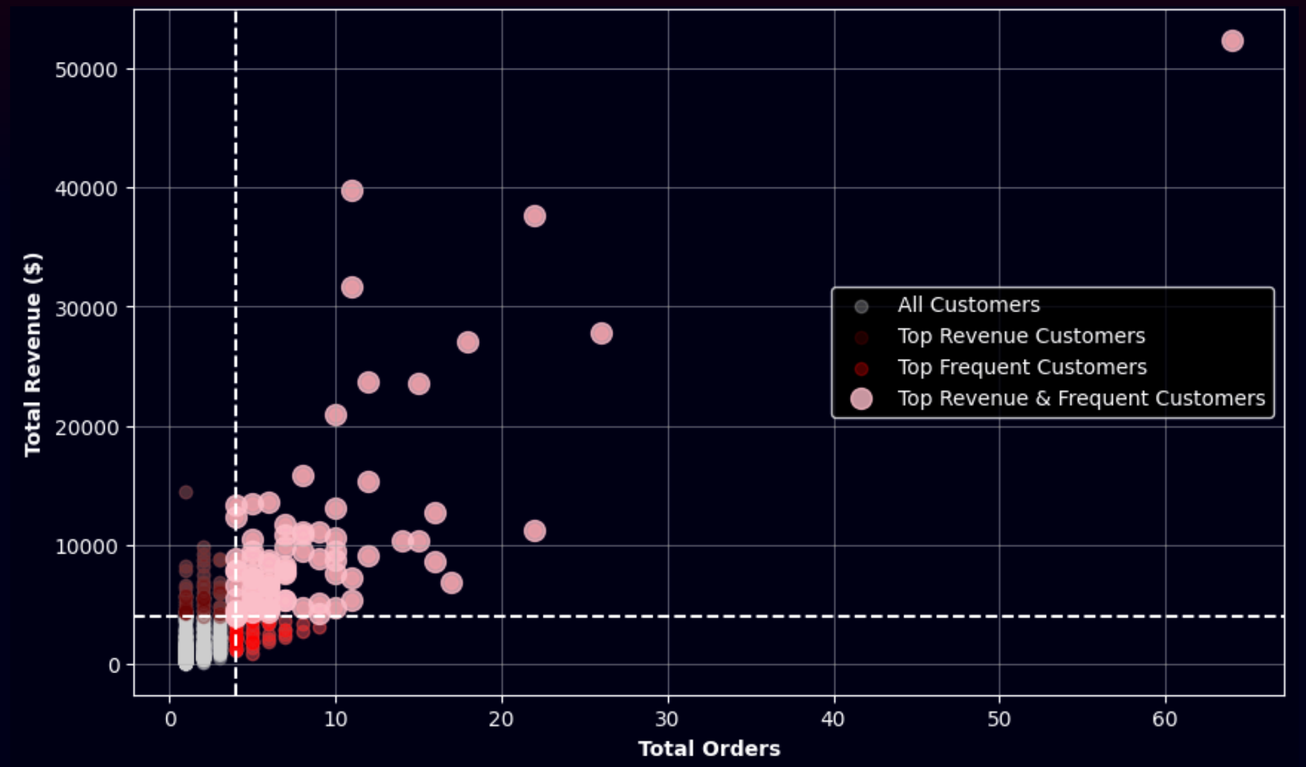


CUSTOMER ACTIVITY SEGMENTATION

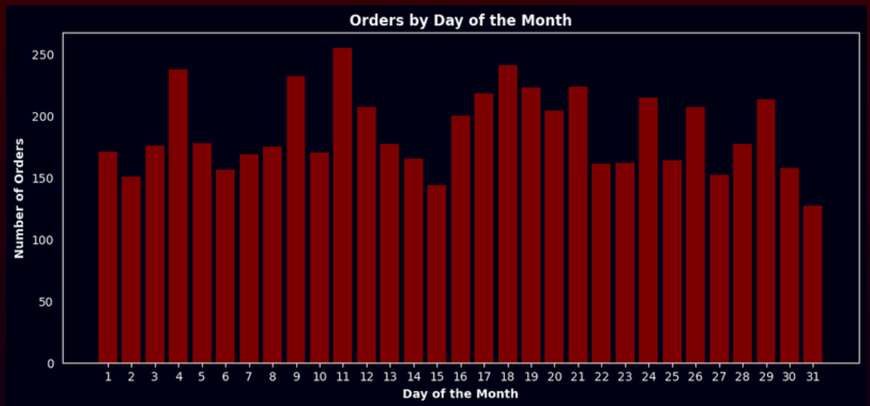
Priority vs. Non-Priority Segments

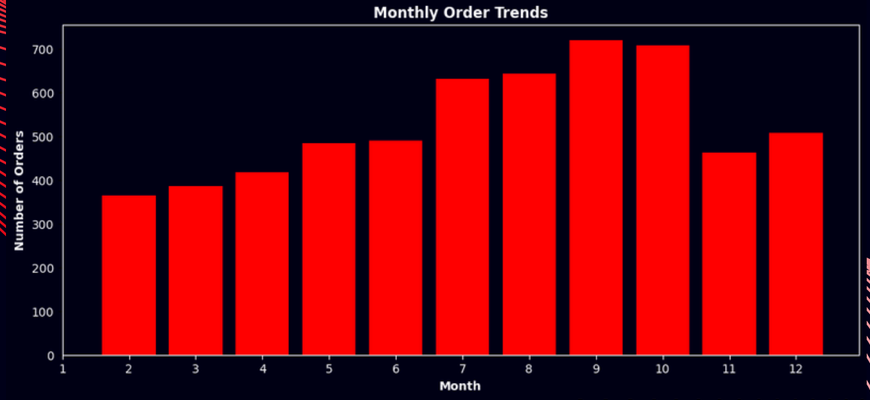


Customer Revenue vs. Purchase Frequency

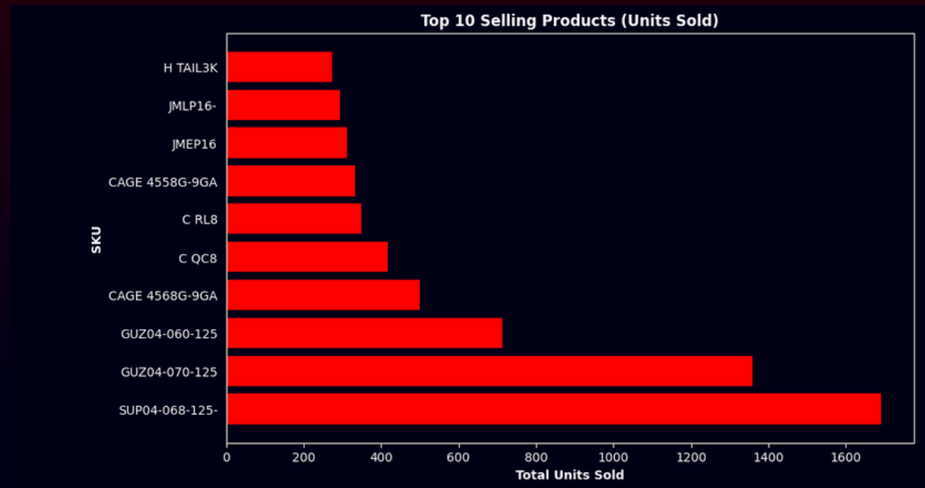


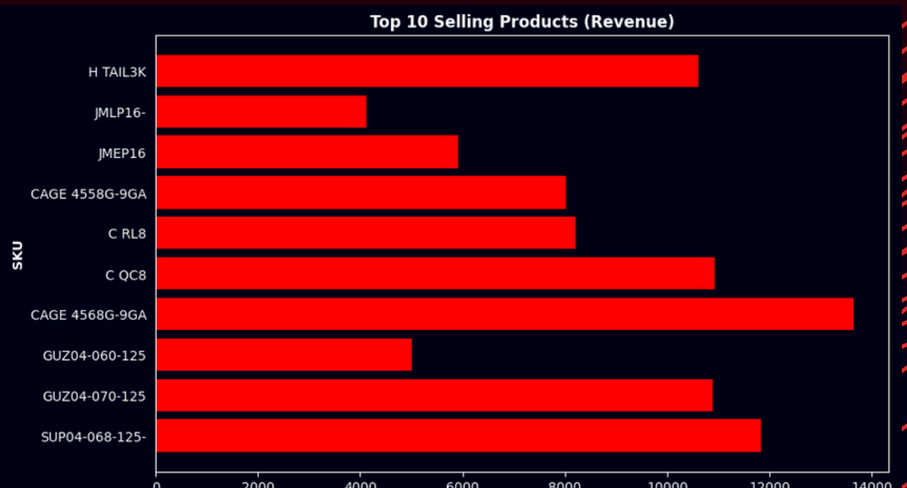
TIME-BASED ANALYSIS



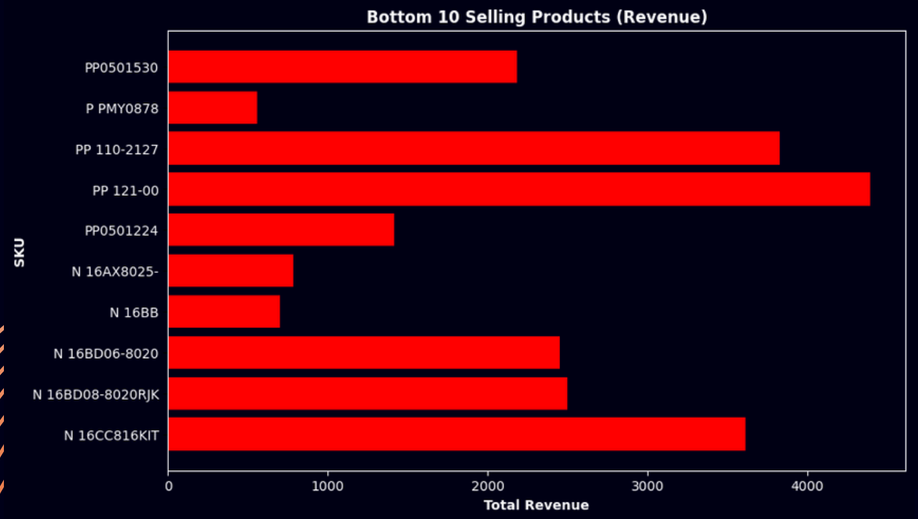


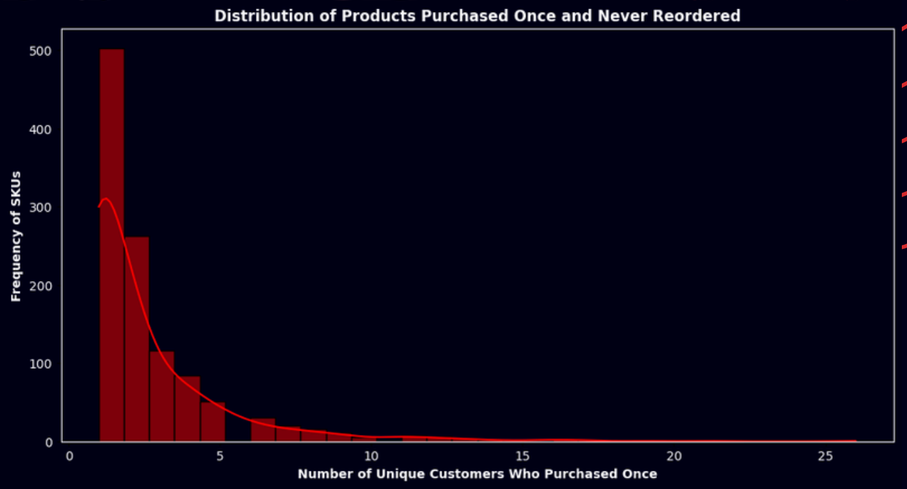
PRODUCT ANALYSIS





Note: All bottom 10 products were only sold once





Super Products
Vacuum Truck Bags

Piranha Hose Jetter Swage Tool
Male End

Piranha Hose Jetter Swage Tool
Mender

1/2" Black Linear Hydro Excavation
Gun

ASSOCIATION RULES

antecedents	consequents	support	confidence	lift
(vacuum truck parts, vacuum truck tubes, hydro...	(vacuum truck hose fittings)	0.005686	0.666667	2.378744
(vacuum truck hose, vacuum truck tubes)	(vacuum truck hose fittings)	0.005686	0.666667	2.378744
(vacuum truck tubes, hydro excavation)	(vacuum truck hose fittings)	0.015435	0.633333	2.259807
(vacuum truck hose, vacuum truck tubes)	(vacuum truck parts)	0.005280	0.619048	2.801646
(vacuum truck parts, hydro excavation)	(vacuum truck hose fittings)	0.018684	0.589744	2.104274
(vacuum truck valves, vacuum truck parts)	(vacuum truck hose fittings)	0.007311	0.580645	2.071809
(vacuum truck tubes, manhole tools)	(vacuum truck parts)	0.007717	0.575758	2.605726
(hydroexcavation trigger)	(hydro excavation)	0.011779	0.547170	3.111160
(manhole tools, hydro excavation)	(vacuum truck hose fittings)	0.007717	0.527778	1.883172
(vacuum truck tubes, jetter hose)	(vacuum truck hose fittings)	0.007311	0.514286	1.835031

K-MEANS CLUSTERING

Customers were segmented into three distinct clusters using K-Means. Segmentation was based on purchase frequency, spending, and customer type. This allows to tailor marketing and retention strategies based on customer value and engagement.

Cluster 0:

- Average of 12.6 days between orders (more spread out).
- Mostly Level 1 customers.
- Fewer items per order and lowest total spend.
- Large concentration of customers from Texas.
- Represents newer or infrequent buyers.

Cluster 1:

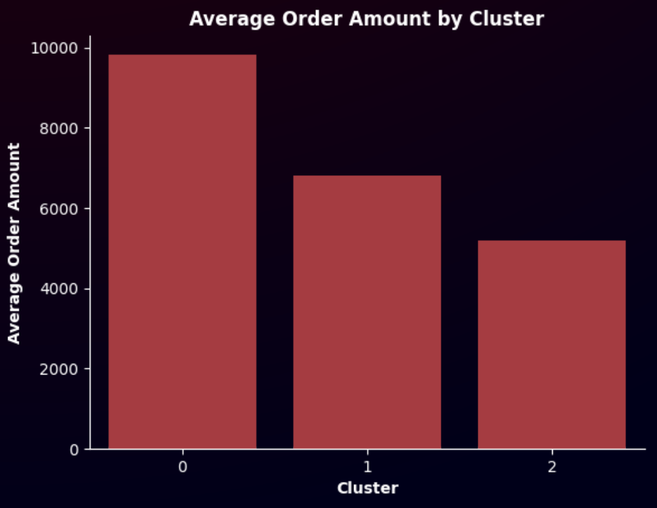
- Average of 11 days between orders.
- Primarily Level 5/6/7 customers.
- Moderate number of items per order; higher spending than Cluster 0.
- Many customers located in Virginia.
- Represents moderately engaged repeat buyers.

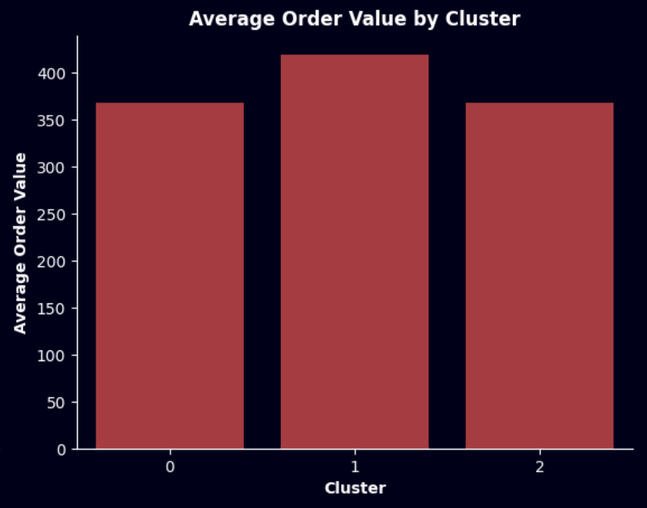
Cluster 2:

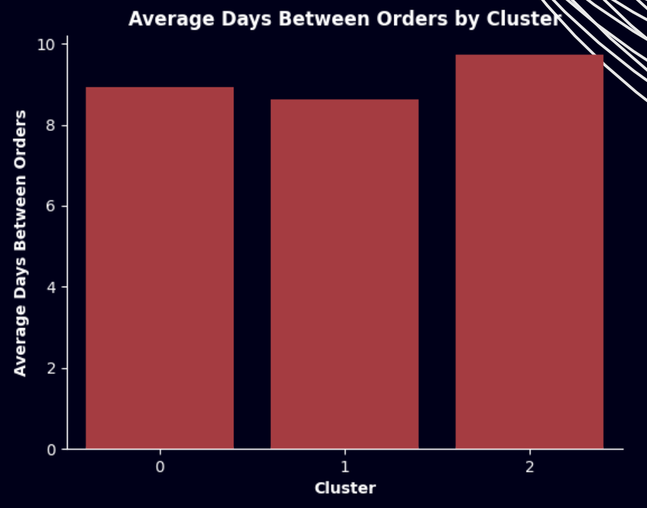
- Average of 10 days between orders (most frequent).
- Also Level 5/6/7 customers.
- Highest number of items per order (~4 items).
- Highest overall spending among clusters.
- Strong presence in Florida.
- Represents frequent, high-value customers.

K-MEANS CLUSTERING PT2

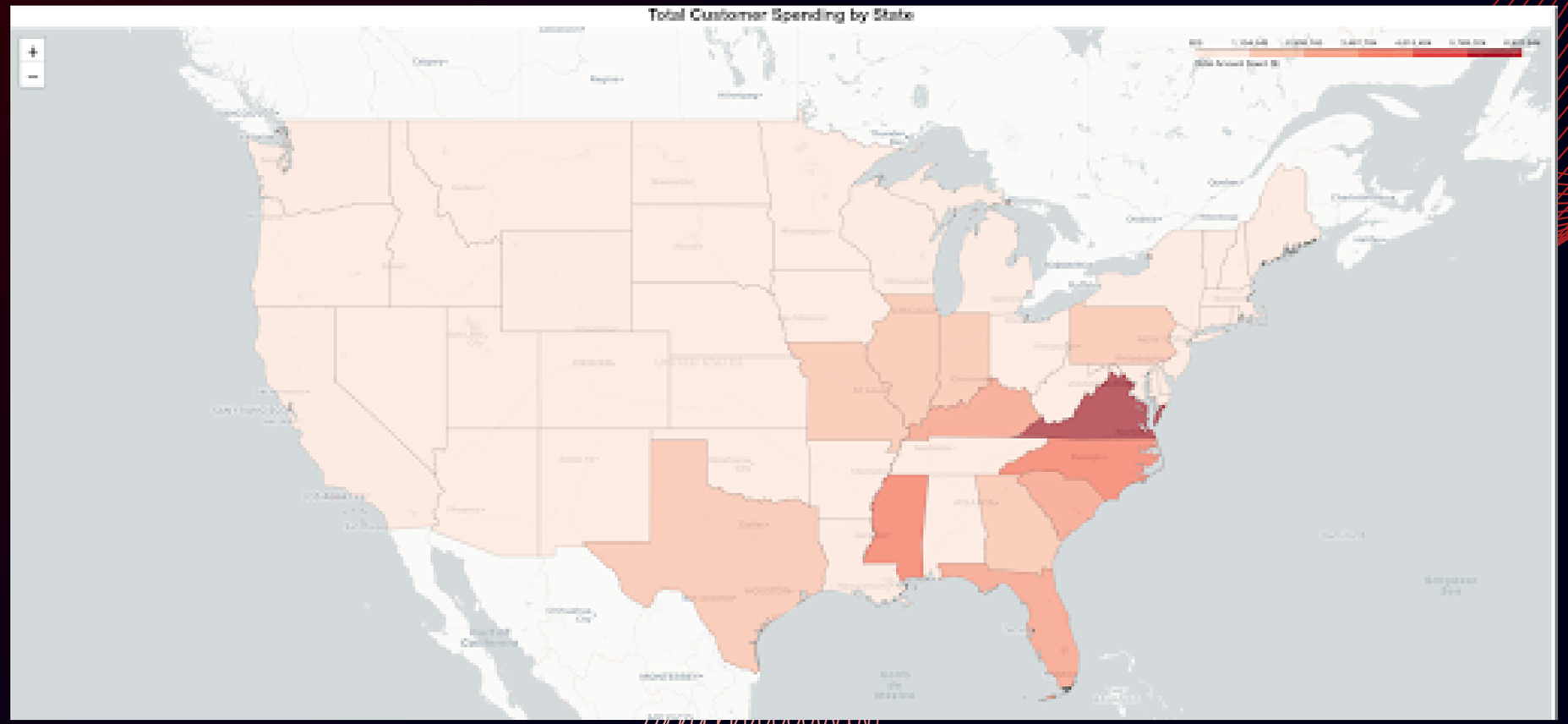
- Customers were segmented into three distinct clusters using K-Means.
- Segmentation was based on purchase frequency, spending, and customer type.
- Helps tailor marketing and retention strategies based on customer value and engagement.



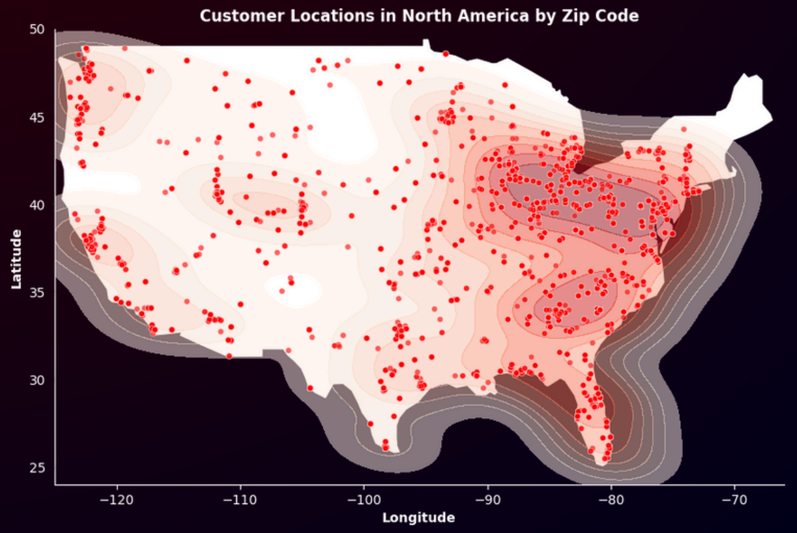


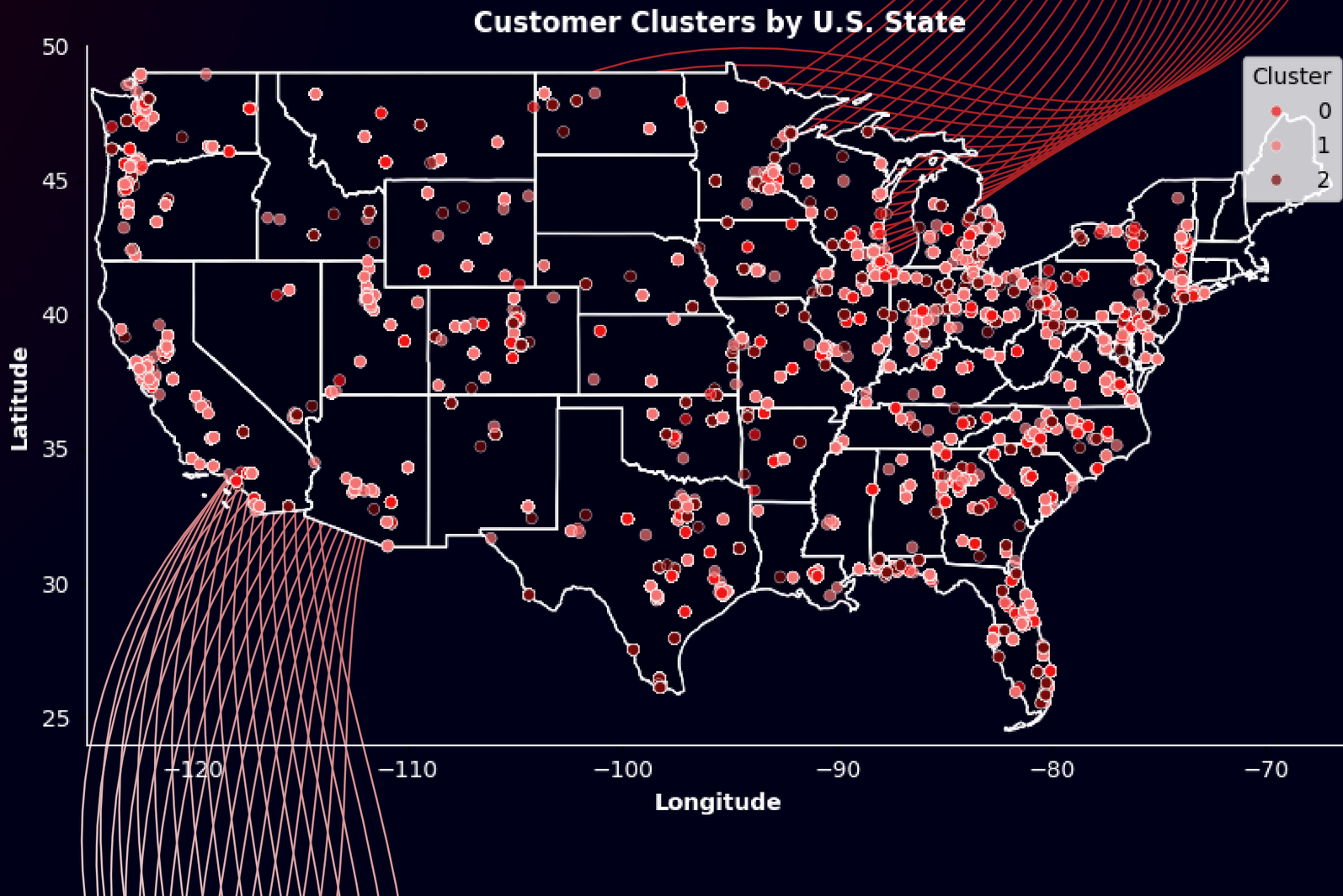


K-MEANS CLUSTERING PT3



K-MEANS CLUSTERING PT3





COX PROPORTIONAL HAZARDS MODEL

MODEL 1

model	lifelines.CoxPHFitter
duration col	'Time_To_Next_Days'
event col	'Repurchased'
baseline estimation	breslow
number of observations	1244
number of events observed	1244
partial log-likelihood	-7506.16
time fit was run	2025-04-28 02:44:24 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Cumulative_Spent_Total	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	4.04	<0.005	14.22
Prior_Order_Count	0.03	1.03	0.01	0.02	0.05	1.02	1.05	0.00	4.10	<0.005	14.55
Customer_Type	0.18	1.19	0.07	0.04	0.31	1.04	1.37	0.00	2.48	0.01	6.25

Concordance	0.62
Partial AIC	15018.32

MODEL 2

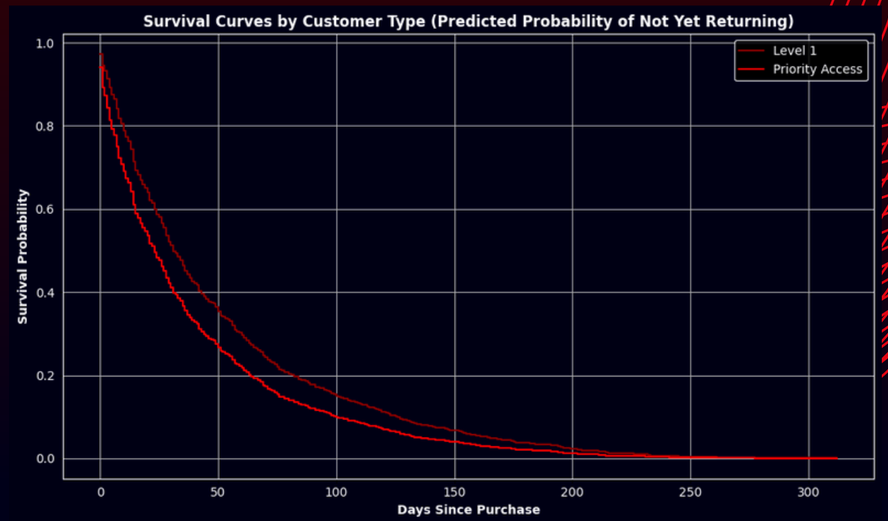
model	lifelines.CoxPHFitter
duration col	'Time_To_Next_Days'
event col	'Repurchased'
baseline estimation	breslow
number of observations	1244
number of events observed	1244
partial log-likelihood	-7428.39
time fit was run	2025-04-28 02:47:38 UTC

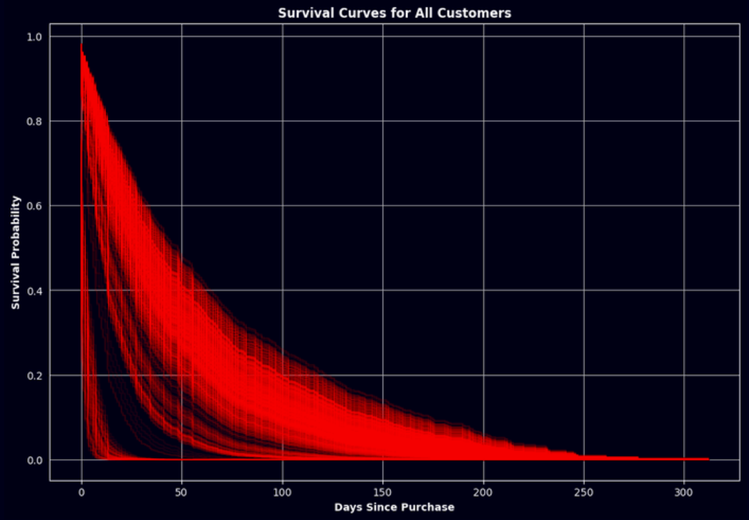
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Cumulative_Spent_Total	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	2.89	<0.005	8.0
Prior_Order_Count	-0.00	1.00	0.01	-0.02	0.01	0.98	1.02	0.00	-0.23	0.82	0.2
Customer_Type	0.11	1.12	0.07	-0.03	0.25	0.97	1.29	0.00	1.58	0.11	3.1
Num_of_Orders	0.03	1.03	0.00	0.03	0.04	1.03	1.04	0.00	13.83	<0.005	142.1
Customer_Tenure_Days	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	0.00	-5.77	<0.005	26.8

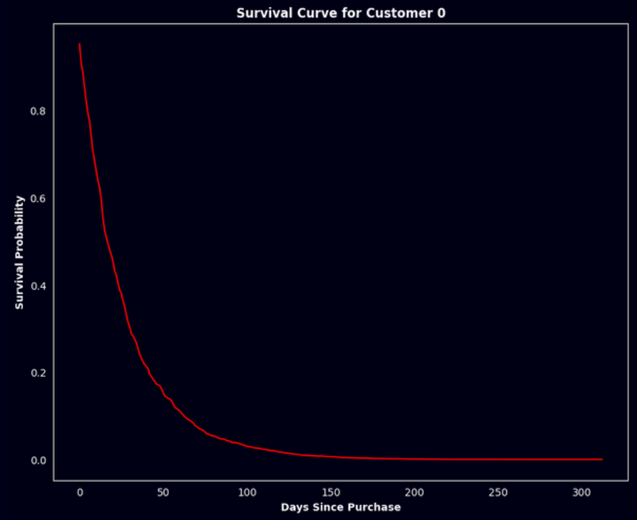
Concordance	0.69
Partial AIC	14866.79
log-likelihood ratio test	393.88 on 5 df
-log2(p) of ll-ratio test	273.09

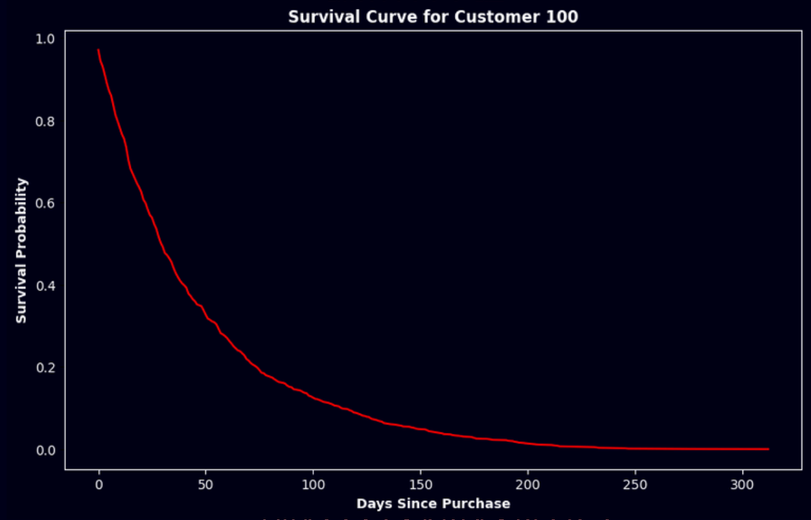
COX PROPORTIONAL HAZARDS MODEL

- Concordance score: 0.62 — model reliably ranks repurchase likelihood.
- Key insights:
- More prior orders → faster repurchase (~3% faster per order).
- Customer type impacts repurchase speed (~19% faster for some types).
- Cumulative spend has minimal practical effect.
- Business takeaway: Most repurchases happen within 50–100 days - early engagement is critical.



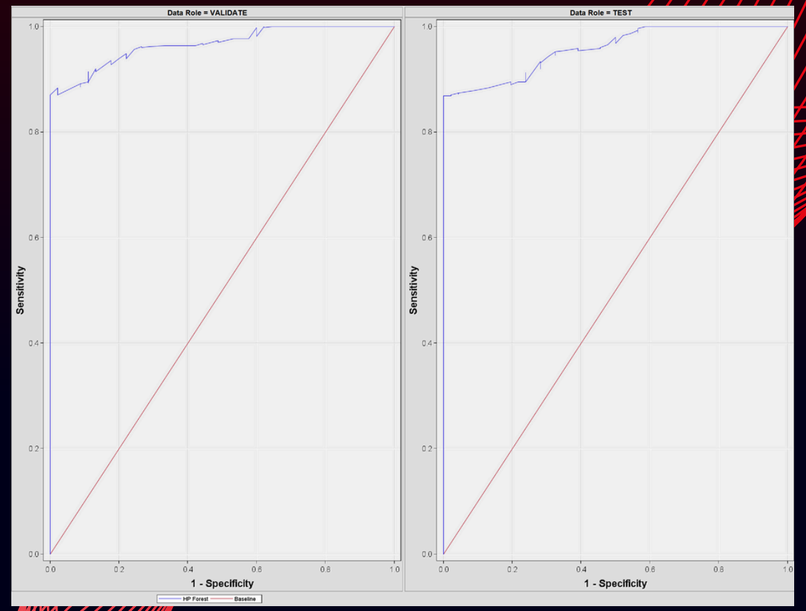


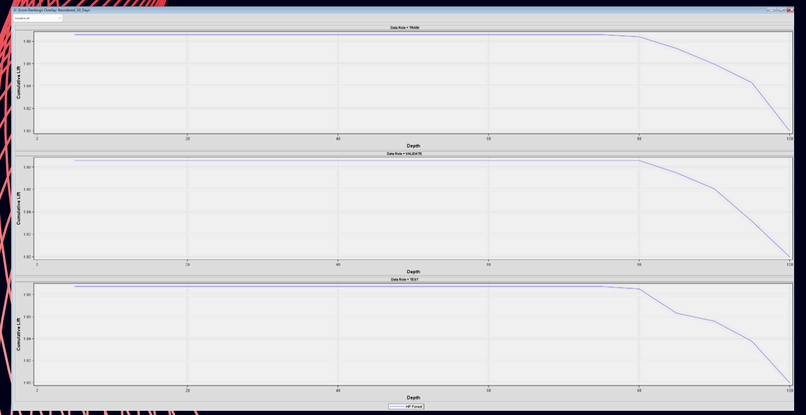




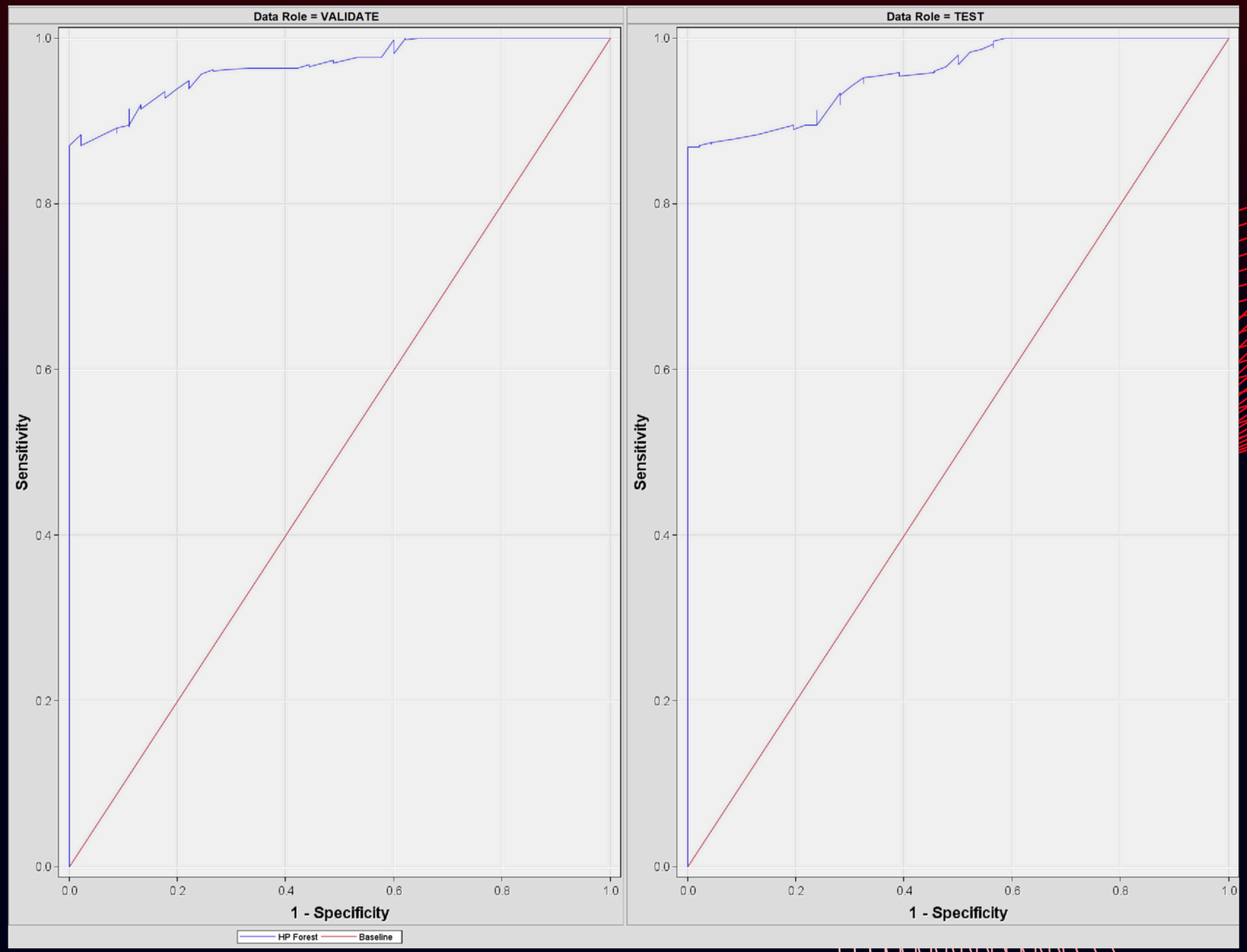
RANDOM FOREST MODEL

- Goal: Predict reorders within 30 days
- Model passed all validation checks — no overfitting or unrealistic metrics.
- Strong separation between reorders and non-reorders
  - 96.6% AUC across Train, Validation, and Test
  - ~8% misclassification rate
  - RASE is low and similar across all sets (~0.22–0.23)
- Consistency across Train, Validation, and Test tells us the model is generalizing well, not just memorizing.
- Top predictors
  - Days between orders
  - Total Price
  - Customer Type.
- This model can help us proactively reach out to customers who are most likely to reorder, improving retention and reducing churn.



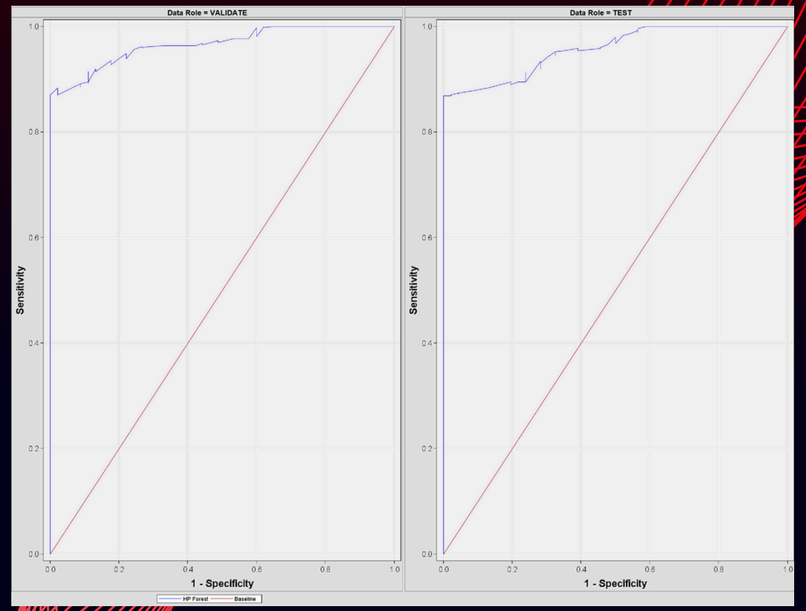


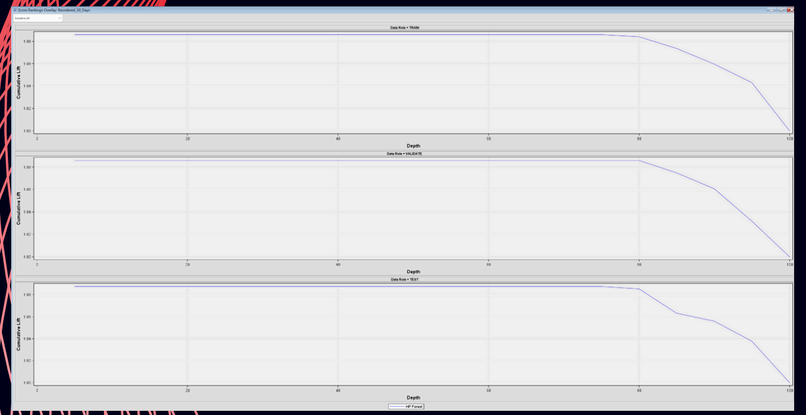
RANDOM FOREST MODEL



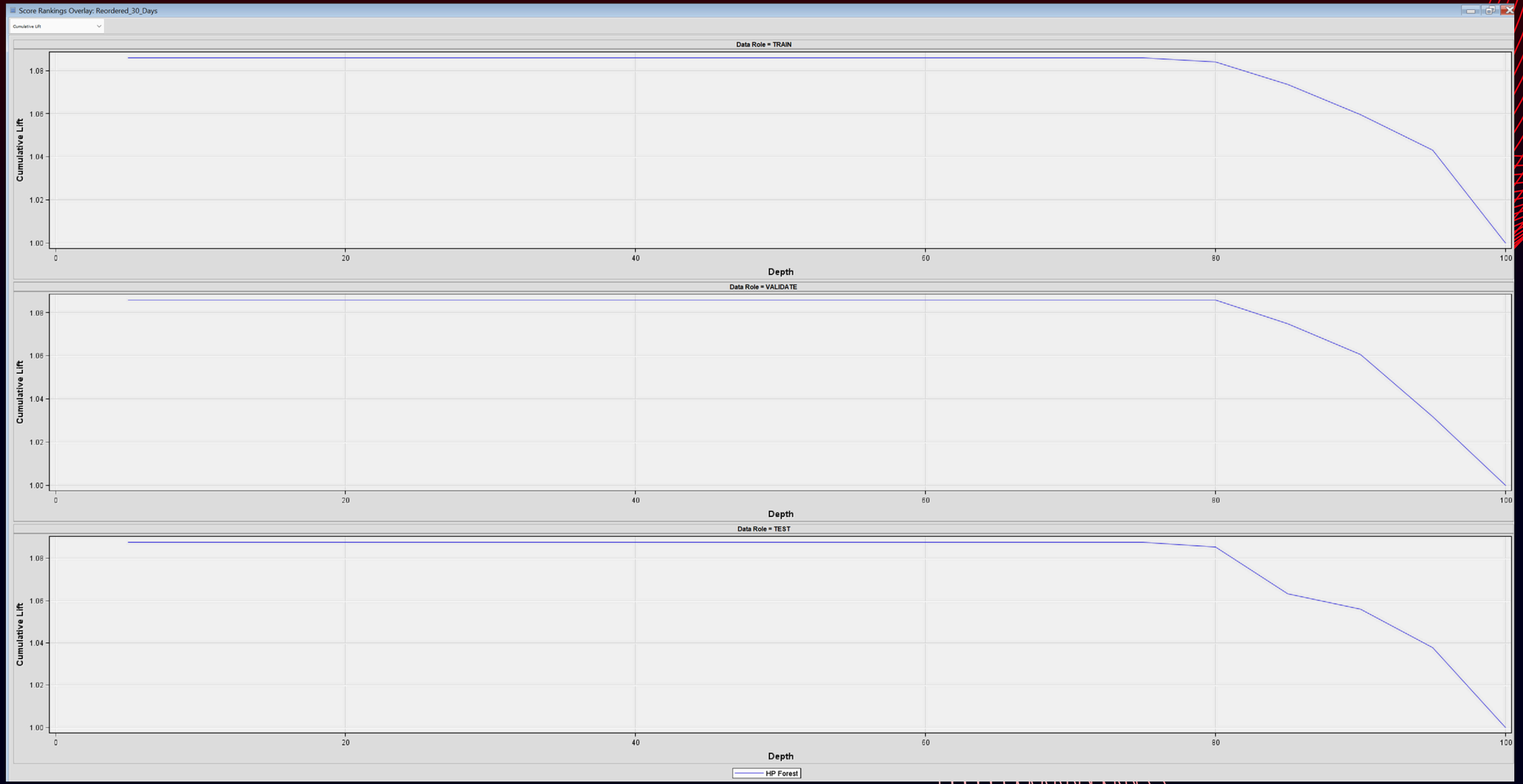
RANDOM FOREST MODEL

- Goal: Predict reorders within 30 days
- Model passed all validation checks — no overfitting or unrealistic metrics.
- Strong separation between reorders and non-reorders
  - 96.6% AUC across Train, Validation, and Test
  - ~8% misclassification rate
  - RASE is low and similar across all sets (~0.22–0.23)
- Consistency across Train, Validation, and Test tells us the model is generalizing well, not just memorizing.
- Top predictors
  - Days between orders
  - Total Price
  - Customer Type.
- This model can help us proactively reach out to customers who are most likely to reorder, improving retention and reducing churn.



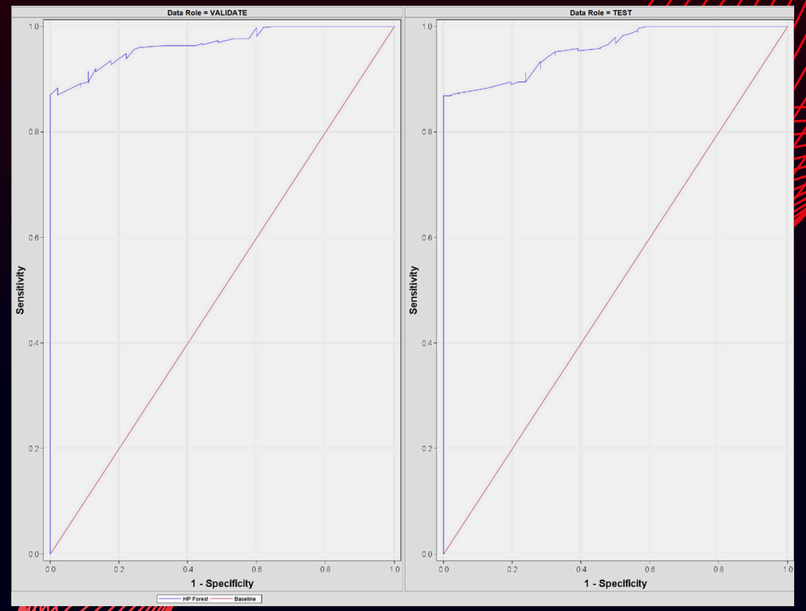


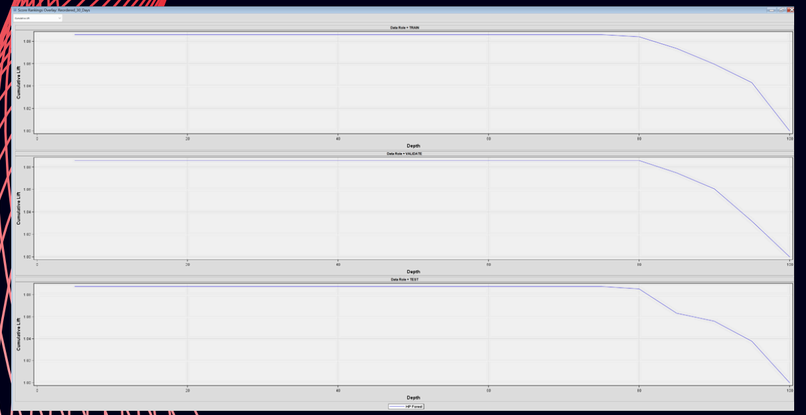
RANDOM FOREST MODEL



RANDOM FOREST MODEL

- Goal: Predict reorders within 30 days
- Model passed all validation checks — no overfitting or unrealistic metrics.
- Strong separation between reorders and non-reorders
  - 96.6% AUC across Train, Validation, and Test
  - ~8% misclassification rate
  - RASE is low and similar across all sets (~0.22–0.23)
- Consistency across Train, Validation, and Test tells us the model is generalizing well, not just memorizing.
- Top predictors
  - Days between orders
  - Total Price
  - Customer Type.
- This model can help us proactively reach out to customers who are most likely to reorder, improving retention and reducing churn.





RECCOMENDATIONS

- Focus on nurturing mid-tier customers ($1K–$10K) through upselling and loyalty campaigns.
- Set up automated touchpoints at 30, 60, and 90 days post-purchase to boost retention.
- Improve product SEO, add "frequently purchased with" recommendations, and create product tutorials to increase repeat orders.
- Prioritize promotion of high-volume, high-margin products while avoiding overuse of discounts.
- Integrate model scores into CRM to flag high-churn-risk customers and identify VIP prospects for targeted outreach.

THANK YOU!